

UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE CIENCIAS MATEMÁTICAS
Departamento de Estadística e Investigación Operativa



TEORÍA DE CÓPULAS APLICADA A LA PREDICCIÓN

MEMORIA PARA OPTAR AL GRADO DE DOCTOR
PRESENTADA POR

Daniel Vélez Serrano

Bajo la dirección del doctor:
Vicente Quesada Paloma

Madrid, 2007

ISBN: 978-84-669-3134-2

UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE CIENCIAS MATEMÁTICAS

DEPARTAMENTO DE ESTADÍSTICA E INVESTIGACIÓN OPERATIVA



TEORÍA DE CÓPULAS APLICADA A LA PREDICCIÓN

Daniel Vélez Serrano

Noviembre 2006

TESIS DOCTORAL

D. VICENTE QUESADA PALOMA, CATEDRÁTICO DEL DEPARTAMENTO DE ESTADÍSTICA E INVESTIGACIÓN OPERATIVA DE LA UNIVERSIDAD COMPLUTENSE DE MADRID

CERTIFICA:

Que la presente memoria de título:

TEORÍA DE CÓPULAS APLICADA A LA PREDICCIÓN

Ha sido realizada bajo mi dirección por D. Daniel Vélez Serrano, licenciado en Ciencias Matemáticas, y constituye su tesis para optar al grado de Doctor en Ciencias Matemáticas.

Y para que conste, en cumplimiento de la legislación vigente y a los efectos oportunos, firmo la presente en Madrid a 27 de Noviembre de 2006

- *I saw...I saw a city in the clouds.*
- *Mmm. Friends you have there.*
- *They were in pain.*
- *It is the future you see.*
- *Future? Will they die?*

Yoda closes his eyes and lowers his head.

- *Difficult to see. **Always in motion is the future...***

*(THE EMPIRE STRIKES BACK,
Script adaptation by Lawrence Kasdan and Leigh Brackett from a story by George Lucas)*

Agradecimientos

Creo que una tesis doctoral no es sólo el resultado del esfuerzo compartido entre el director de tesis y su alumno. Al menos en mi caso, debo confesar que esta tesis nunca hubiera visto la luz de no haber sido por una relación de personas a las que considero tengo mucho que agradecer.

En primer lugar, me gustaría citar a un compañero de mi primer trabajo llamado Ignacio Arbués, posiblemente la persona más inteligente con la que he trabajado. Fue su amplio dominio de la materia, su pasión por las matemáticas y la facilidad con la que era capaz de transmitir sus conocimientos, las que me motivaron a empezar los cursos de postgrado para mejorar mi grado de formación. Perdí la pista a este asturiano cuando decidió dejar la empresa en la que trabajaba conmigo, para empezar a prepararse las oposiciones del INE y lo último que sé de él es que las aprobó, algo de lo que nunca tuve duda alguna. En cualquier caso, si alguna vez lees estas líneas, ¡gracias Nacho!.

En segundo lugar quería hacer mención a mi director de tesis, Vicente Quesada, dado que fue él quien, una vez que hube conseguido la suficiencia investigadora, me animó a continuar el doctorado y a hacer esta tesis. Para mi, supuso todo un reto emprender este camino estando ya bien situado laboralmente con casi cinco años de experiencia, y a sabiendas de las dificultades y el sacrificio que con seguridad implicaría un compromiso de este tipo. Entre las cosas que quiero agradecerle me gustaría destacar el hecho de que haya sabido siempre proponerme temas de investigación tan relacionados con los proyectos de análisis que desarrollo y a la vez tan innovadores y desconocidos en el mundillo en el que me muevo. También le agradezco no haberme presionado y haberme dado flexibilidad total a la hora de llevar esta tesis adelante, entendiendo las dificultades derivadas de compatibilizar dicha labor con mi trabajo.

A los doctores Javier Martín y José María García Santesmases, me gustaría agradecer en primer lugar el hecho de que me pusieran en contacto con Vicente para que me tutelase esta tesis. Fueron ellos además, quienes me ofrecieron mi primer trabajo como consultor matemático y les quiero agradecer la confianza que en mí depositaron desde el primer momento para participar y dirigir importantes proyectos de investigación. Entre ellos, guardo un especial recuerdo de los realizados para el **Departamento de Análisis y Simulación** de la compañía **Enagás**. Agradezco a la responsable del departamento Laura Marín y a todo su equipo la ayuda prestada, de forma especial a Ana Belén García que me ha echado una mano con la provisión de datos utilizados para el desarrollo de las aplicaciones prácticas que aquí se exponen y que además, me ha transmitido muchos ánimos a lo largo de estos últimos años.

También tengo mucho que agradecer a mi círculo de amistades por los buenos ratos que hemos compartido y que espero seguir compartiendo. Gracias por sus ánimos constantes a Luis Javier Ferrero, mi primer compañero de fatigas en la carrera, que siempre estuvo ahí. Gracias también a Sandra Martín, a Amanda Sanz, a José Luis Navarro, a José María Picado, a Miguel Redondo (de manera especial por su ayuda y por sus consejos), a David Bueno, a Francisco Fernández, a David Gutiérrez, a Carlos Ruíz, a Prudencio Ureña, a Óscar Villanueva y a muchos otros que seguro deo de nombrar, quienes, a través de su compañía y afecto, han aportado su granito de arena en esta tesis.

Gracias a los padres de mi novia, Tomás y Charo, por sus consejos, su atención, y por esas agradables meriendas y cenas en las que nos hemos reunido para ver alguna película o partido. Ratitos de conversación totalmente necesarios.

Finalmente, quiero hacer mención especial a mi familia. De forma general gracias a todos por sus ánimos y por haberme entendido cuando ni yo mismo era capaz de hacerlo.

Debo agradecerle a mi hermano José todas sus enseñanzas de programación, sus charlas matemáticas y científicas, y todos sus consejos a la hora de dar formato para presentar esta tesis. Además considero que, el hecho de que comenzáramos al mismo tiempo la andadura en una tesis doctoral y de que nos hayamos motivado mutuamente durante todos estos años a seguir adelante, ha sido un factor determinante para el logro de nuestros objetivos.

Quiero agradecer a mi abuelo Paco las “*partiditas de tute*” esporádicas que han salpicado muchas de mis tardes y que me han ayudado a evadirme en momentos de bastante agobio y tensión. Gracias abuelo por tu interés declaradamente sincero por todas y cada una de mis labores profesionales.

También quería recordar a mi padre que sólo pudo verme empezar este largo camino aunque muy probablemente se marchó con la certeza de que intentaría recorrerlo hasta el final. Creo que debo agradecerle a él (o a sus genes) una parte importante de ese carácter constante y de esa cabezonería que considero imprescindibles para haber afrontado este desafío. Cualquiera que sea el rinconcito de cielo en el que estés, ¡gracias papá!

Finalmente quiero dar las gracias a las tres mujeres más importantes en mi vida y a las que quiero dedicar esta tesis. A mi madre que ha sido la que creo que ha aguantado más que nadie mis cambios de humor, ha sabido convivir con montañas de papeles ininteligibles sobre mi escritorio y aún así, me ha seguido animando hasta el último momento. A mi chica, Leticia, por sus ánimos, por su apoyo, por haberme sabido escuchar y sobre todo, por haber tenido la paciencia suficiente para seguir conmigo después de un último año (o dos, o quizá tres) especialmente duro en el que tal vez haya estado un “poquito” más volcado en mis intereses profesionales que personales.

Para terminar quiero dar las gracias a mi abuela Consolación que en paz descanse, quien siempre mostró de forma incondicional una confianza ciega en que terminaría consiguiendo cuanto me propusiera. Tal vez llevaras razón, abuela.

Prólogo

Dicen que el saber no ocupa lugar. Si el saber es sinónimo de experiencia tal vez no debiera haber duda alguna al respecto. Si en cambio ese saber es reflejo de la formación cultural de una persona, los límites podrían venir marcados por su afán de conocimiento pero también por su memoria. Estos mismos límites serían aplicables a cualquier individuo para el que “saber” no fuera otra cosa que la capacidad para retener en su cabeza un volumen determinado de información sin importar que ésta estuviese referida a los números de la guía telefónica o a las palabras del diccionario enciclopédico. En cualquier caso, en el mundo empresarial, la frase se nos antoja aprovechable en uno u otro sentido.

Según van pasando los días, el ordenador personal permite almacenar más información, más saber, en menos espacio. Existen grandes bases de datos que recogen históricos de una compañía al nivel de detalle que se desee. Si bien existen restricciones impuestas en función de la memoria de los equipos, la cantidad de datos recogidos suele ser suficiente para plantear estudios de mercado y tomar decisiones que puedan mejorar la evolución del negocio.

Puede ser que este saber no exista. Sería el caso, por ejemplo, de una compañía nueva en cualquiera de los sectores (telecomunicaciones, industria, banca, seguros, etc.) que no tuviese ningún dato almacenado sobre el que basar algún análisis específico. También podría ser ilustrativo el ejemplo de una compañía que quisiera lanzar un producto nuevo y no tuviera por tanto historia del mismo para poder conocer la respuesta que pudiera esperar tras una campaña comercial. Un tercer caso podría venir inducido por un cambio significativo de comportamiento que experimentara el mercado haciendo inservible cualquier información almacenada. Es entonces cuando el saber del experto de negocio, basado en la experiencia de la que hablábamos, juega un papel determinante siendo su opinión la que terminará conduciendo a la toma de una decisión final.

La situación ideal se plantea cuando se combina la experiencia del analista en un negocio determinado con unas circunstancias del mercado que hacen que los históricos de datos de la compañía resulten de un valor incuestionable. El saber del experto se fusiona con el saber almacenado en sus equipos informáticos y esto le permite realizar una planificación de las acciones a emprender a corto, medio y largo plazo.

Obviamente, “saber” no va siempre asociado a algo que ya ha pasado como pudiera entenderse si lo relacionamos con la experiencia o el grado de formación cultural de la persona. No sorprende a nadie la atracción del ser humano por saber lo que pasará en el futuro, inquietud que se remonta a nuestros antepasados. Ya en la Edad Antigua, los oráculos, estaban vinculados a dos clases de instituciones: la adivinación y el profetismo. En la Grecia clásica el oráculo se establecía en un lugar determinado (el ordenador personal), presidido e inspirado por un dios (la matemática) y secundado por sacerdotes que servían de intérpretes (los analistas que diseñan los modelos y que, junto con los expertos de negocio, interpretan los resultados proporcionados por éstos para tomar las últimas decisiones).

Índice

Prólogo.....	9
0. Introducción.....	15
0.1 Estado del arte.....	16
0.2 Objetivos.....	18
0.3 La tesis.....	21
1. Funciones cópula.....	25
1.1 Introducción.....	25
1.2 Definición.....	28
1.3 Teorema de Sklar.....	29
1.4 Capacidad de las funciones cópula para reflejar relaciones de dependencia.....	32
1.5 Tipos de cópulas.....	33
1.5.1 Tipos de cópulas en función de su soporte.....	34
1.5.2 Tipos de cópulas en función del conocimiento explícito de su forma.....	35
1.5.3 Tipos de cópulas en función de la relación de dependencia que son capaces de reflejar.....	36
2. Elección de la cópula que mejor refleja una relación de dependencia.....	39
2.1 Introducción al problema.....	39
2.2 Determinación de las distribuciones marginales.....	40
2.3 Propuesta de un conjunto inicial de familias de cópulas candidatas.....	41
2.4 Determinación de la cópula óptima dentro de una familia.....	41
2.5 Selección de la mejor familia a partir de las cópulas representantes.....	42
2.5.1 Método 1: Empleo de la cópula empírica.....	42
2.5.2 Método 2: Métodos gráficos.....	43
2.5.3 Método 3: Aproximación analítica de los métodos gráficos.....	44
2.5.4 Método 4: Criterio de información de Akaike (AIC).....	44
2.5.5 Método 5: Contrastes de bondad de ajuste de una cópula.....	45
2.5.6 Método 6: Calidad de las predicciones que proporciona una cópula.....	46
3. Metodologías para la construcción de funciones cópula.....	49
3.1 Método de inversión.....	49
3.2 Métodos geométricos.....	50
3.2.1 Cópulas singulares con soporte conocido.....	50
3.2.2 Cópulas construidas como Suma Ordinal de cópulas.....	50
3.2.3 Cópulas construidas a partir de “Cambios” de M (cota superior de Fréchet-Hoeffding).....	50
3.2.4 Cópulas construidas como “Suma Convexa” de cópulas.....	52
3.2.5 Métodos basados en el conocimiento de las secciones de una cópula.....	52
3.3 Métodos algebraicos.....	53
3.4 Método de construcción de cópulas arquimedianas.....	54
3.5 Método de construcción de cópulas con soporte fractal.....	54
4. Construcción de cópulas con valor óptimo del estadístico de Pearson.....	57
4.1 Etapa I: Construcción de la subcópula que minimiza el estadístico de Pearson.....	59
4.2 Etapa II: Construcción de una cópula por interpolación de una subcópula.....	64
4.2.1 Método de interpolación bilineal.....	64
4.2.2 Método de interpolación mediante polinomios de Bernstein.....	67

4.2.3 Método de interpolación mediante polinomios cúbicos de Hermite.....	69
4.2.3.1 Condiciones a imponer para que la función construida por interpolación cúbica verifique las propiedades frontera de las cópulas.....	72
4.2.3.2 Condiciones a imponer para que la función construida por interpolación cúbica verifique la propiedad 2-creciente de las cópulas.....	73
4.2.3.2.1 Condición a imponer para que una superficie interpoladora de Ferguson sea 2-creciente.....	74
4.2.3.2.2 Condición a imponer para que una superficie interpoladora de Hermite sea 2-creciente.....	81
4.2.3.3 Condiciones para construir cópulas por interpolación cúbica.....	95
4.2.4 Algunas observaciones sobre las cópulas interpoladoras.....	100
5. Simulación de valores para las cópulas construidas por interpolación.....	105
5.1 Simulación de valores para algunas familias de cópulas conocidas.....	106
5.2 Simulación de valores para la cópula construida por interpolación bilineal.....	107
5.3 Simulación de valores para la cópula construida por interpolación cúbica.....	107
5.4 Simulación de valores para la cópula construida por interpolación mediante polinomios de Bernstein.....	109
6. Empleo de funciones cópula para predecir.....	111
6.1 Empleo de funciones cópula como modelos de función de transferencia.....	111
6.2 Empleo de funciones cópula para explicar dependencia dinámica.....	116
7. Aplicación práctica: Predicción de la demanda de gas natural.....	119
7.1 Introducción.....	119
7.2 Predicción de la demanda de gas natural a medio plazo.....	122
7.2.1 Introducción al problema.....	122
7.2.2 Planteamiento de la solución.....	123
7.2.3 FASE I: Construcción de un histórico de demanda en condiciones normales de temperatura.....	127
7.2.4 FASE II: Construcción de un modelo autorregresivo para hacer una predicción a medio plazo.....	144
7.2.4.1 Identificación del patrón estacional semanal.....	145
7.2.4.2 Identificación de la variabilidad.....	149
7.2.4.3 Identificación del ciclo anual de la serie.....	153
7.2.4.3.1 Generación de la variable Ciclo a partir de wavelets.....	154
7.2.4.3.2 Generación de la variable Ciclo a partir de splines de regresión.....	159
7.2.4.3.3 Comparativa entre los métodos planteados para la generación de la variable Ciclo.....	164
7.2.5 FASE III: Simulación del valor esperado de la demanda en situaciones climatológicas extremas.....	168
7.2.5.1 Generación de distribuciones marginales.....	170
7.2.5.2 Determinación de una cópula a través del criterio de selección basado en el estadístico de Pearson.....	171
7.2.5.3 Construcción de cópulas que optimizan el estadístico de Pearson.....	178
7.2.5.3.1 Construcción de una cópula por interpolación bilineal.....	179
7.2.5.3.2 Construcción de una cópula por interpolación cúbica.....	181
7.2.5.3.3 Construcción de una cópula por interpolación mediante polinomios de Bernstein.....	184
7.2.5.4 Comparativa de resultados entre las diferentes cópulas candidatas.....	185

7.3 Predicción de la demanda de gas natural a corto plazo.....	198
7.3.1 Introducción al problema.....	198
7.3.2 Planteamiento de la solución.....	199
7.3.2.1 Ajuste de una serie mediante un ARIMA complementado con modelos de función de transferencia.....	201
7.3.2.2 Ajuste de una serie mediante un ARIMA complementado con un algoritmo basado en funciones cópula.....	217
8. Conclusiones.....	239
9. Futuros trabajos.....	243
ANEXOS.....	245
ANEXO I: Algunas familias de cópulas.....	245
AI.1 Familias no paramétricas.....	245
AI.1.1 Familias de dependencia extrema.....	245
AI.1.2 Otras familias.....	245
AI.2 Familias uniparamétricas.....	246
AI.2.1 Familias elípticas.....	246
AI.2.2 Familias de valor extremo.....	246
AI.2.3 Familias arquimedianas.....	247
AI.2.4 Otras familias uniparamétricas.....	250
AI.3 Familias biparamétricas.....	251
AI.3.1 Familias elípticas.....	251
AI.3.2 Familias de valor extremo.....	251
AI.3.3 Familias arquimedianas.....	251
AI.3.4 Otras familias biparamétricas.....	252
ANEXO II: Demostración del lema 4.2.....	255
ANEXO III: Demostración del lema 4.3.....	257
ANEXO IV: Demostración del lema 4.4.....	261
ANEXO V: Demostración del lema 4.5.....	279
ANEXO VI: Demostración del lema 4.6.....	283
ANEXO VII: Demostración del teorema 4.6.....	299
ANEXO VIII: Contrastes de bondad de ajuste a una distribución dada.....	307
ANEXO IX: Contraste de Kruskal-Wallis.....	309
ANEXO X: Análisis mediante wavelets.....	311
ANEXO XI: Análisis mediante splines de regresión.....	315
Bibliografía.....	321

0. Introducción

Una de las principales funciones de una empresa transportista de gas es diseñar un plan de operaciones a través del cual pueda garantizar la continuidad y seguridad del suministro de gas natural y la correcta coordinación entre los puntos de acceso, los almacenamientos, el transporte y la distribución teniendo en cuenta las necesidades y características de cada uno de sus tipos de clientes tanto dentro del ámbito industrial como del doméstico. En nuestro país, **Enagás** es la principal empresa transportista de gas y entre sus activos, cuenta con 7.538 km de gasoductos de alta presión, tres plantas de regasificación (Barcelona, Cartagena y Huelva) y la propiedad o gestión de los almacenamientos de gas natural. La gestión del plan de operaciones es una tarea muy compleja que se realiza no sólo en función de los recursos de gas disponibles sino también de la demanda prevista a abastecer y del cumplimiento de los compromisos contractuales tanto con las empresas suministradoras como con las distribuidoras y comercializadoras.

La aparición de las comercializadoras de gas ha tenido lugar a raíz de la reciente liberalización del mercado del gas en España (desde Enero de 2003). Según la nueva normativa, un cliente no tiene porqué mantenerse en el mercado regulado que se rige de acuerdo a las tarifas fijadas por el gobierno, sino que puede comprar directamente en dicho mercado, elegir una comercializadora a la que pague una tarifa de menor coste o establecer contratos bilaterales. El crecimiento del número de comercializadoras, el incremento esperado del consumo y la capacidad limitada de la red de gasoductos, son factores a tener en cuenta a la hora de elaborar el plan de operaciones. Resulta por ello imprescindible para el departamento encargado de su gestión, disponer de unas buenas herramientas de predicción de la demanda a distintos horizontes para poder adaptarse de manera adecuada a las necesidades de suministro a corto (10 días con detalle horario) y medio (1-3 años) plazo. Cuanto mayor sea la bondad de las predicciones que proporcionen estos sistemas, más fielmente se ceñirán a la realidad los programas y planes de gestión citados y por tanto más beneficiosos resultarán los resultados de la operación.

Existe un amplio abanico de metodologías matemáticas para el tratamiento de problemas de predicción de demanda de gas a corto y medio plazo. Queremos observar que por lo general éstas técnicas son igualmente aplicables al sector eléctrico y por ello no debe de extrañar que la literatura que nos ha servido de consulta haga referencia indistintamente a uno u otro sector. La diferencia fundamental radica en el periodo de influencia del factor meteorológico sobre la demanda, que abarca al invierno (calefacción eléctrica) y al verano (aparatos de aire acondicionado) en el caso eléctrico y sólo al invierno en el caso gasista. Esta circunstancia no afecta desde un punto de vista metodológico sino únicamente a la hora de definir las variables climáticas que participan en los modelos y es por ello que, aun cuando la exposición que hacemos en esta tesis esta centrada en el mercado gasista, ya hemos contrastado la validez de los métodos que proponemos en uno y otro ámbito.

0.1 Estado del arte

Si bien como hemos dicho los métodos de predicción que se manejan en la actualidad son aplicables a datos de demanda referidos tanto al gas como a la electricidad, la utilización de ellos está condicionada por el horizonte de predicción para el que se plantea el estudio.

En el tratamiento a corto plazo se pretende obtener predicciones con detalle diario para un horizonte que suele depender del máximo periodo para el que se disponga de una previsión fiable de los agentes climatológicos y que habitualmente oscila entre 10 y 15 días. Después de haber trabajado durante más de tres años en problemas de predicción de demanda a corto plazo y haber contrastado la eficacia de diversas alternativas creemos que, por las características de este tipo de series, es la metodología **Box-Jenkins** fundamentada en el empleo de modelos **ARIMA**, la que proporciona unos mejores resultados (véase por ejemplo el estudio para la *British Gas North Western* que se expone en [FIREAST]). Estos modelos plantean a través de una ecuación lineal (aditiva en las variables) la relación entre la demanda en un instante de tiempo “ t ”, el valor de la propia variable en instantes anteriores de tiempo (parte ARI del modelo) y el error que el modelo ha podido cometer en dichos instantes (parte MA). Es esta capacidad para corregirse en función de los errores de predicción más recientes la que los hace especialmente orientados para la predicción a corto plazo. A través de una ecuación autorregresiva el modelo refleja la tendencia, variabilidad y estacionalidad de los datos, captura mediante un “*análisis de intervención*” el efecto del calendario laboral presente en series cuyo comportamiento es marcadamente industrial e identifica el efecto de las variables climáticas, característico de la demanda doméstica, a través de “*modelos de función de transferencia*” convenientemente ajustados. Hay que tener en cuenta que esta metodología asume ciertas condiciones de estabilidad del histórico y una hipótesis fuerte de linealidad en la relación existente entre la variable a predecir, su pasado y los agentes climatológicos que participan.

Otra técnica muy utilizada para hacer predicciones a corto plazo es la **red neuronal** que pierde la capacidad de interpretar fácilmente las asociaciones que plantea entre los elementos que en ella intervienen pero que, a cambio, permite detectar comportamientos no lineales complejos y aprender de ellos para aumentar su potencia predictiva. Conceptualmente estos modelos enlazan los valores de las variables explicativas (almacenados en los nodos de una capa denominada “*capa de entrada*”) con los valores correspondientes de la variable a predecir (almacenado en un único nodo de la denominada “*capa de salida*”) asignando un valor (peso) a cada par de nodos enlazados. El objetivo que se persigue es la estimación de los pesos de la red, es decir, el cálculo de los valores que permiten reflejar de forma óptima la relación de dependencia existente. En su estructura más básica (una capa de entrada y otra de salida) la red neuronal establece una relación lineal entre las variables que participan en el modelo. Sin embargo, también es posible intercalar capas intermedias denominadas “*capas ocultas*” que permiten plantear relaciones no lineales (*ocultas*) entre las variables. Con el fin de esclarecer las posibles dependencias no lineales que llevan a algunos autores a catalogar de “*caja negra*” a los modelos de red neuronal, surgen más recientemente los **sistemas difusos de inferencia** basados en la implementación de una relación de reglas que utilizan la información en términos lingüísticos y por tanto más transparentes (“si hace *bastante* frío en invierno, la demanda de gas crece *mucho*”). Entre los inconvenientes principales de esta metodología destaca el crecimiento desmesurado de reglas conforme crece el número de variables de entrada y la necesidad de traducir en valores numéricos las predicciones también lingüísticas (difusas) que estos métodos ofrecen.

Con el fin de mejorar los altos costes computacionales que conlleva la implementación de los sistemas difusos, algunos autores proponen un modelo híbrido denominado **red neuronal difusa** que básicamente consiste en alimentar con una base de reglas difusas a una red neuronal. En este contexto de las redes de neuronas y los sistemas difusos aplicados a la predicción de la demanda energética podemos citar por ejemplo los artículos de [LYXUE] , [MUHAJI] o [VACHUNTA].

Cuando el número de variables explicativas crece de forma desmesurada, las posibilidades de sobreajuste de los modelos neuronales aumentan a la vez que disminuye su potencia predictiva. Para resolver este problema existen técnicas de **Inteligencia Artificial** como son los **algoritmos genéticos** o los **sistemas expertos**, de cuya aplicación al problema de predicción de la demanda a corto plazo podemos encontrar una breve descripción en [PERWHI].

Respecto del tratamiento a medio plazo, encontramos también diferentes propuestas de modelización para la predicción con detalle mensual o diario en función del objetivo que se persiga. Así por ejemplo según se cita en [MEHRA], para el tratamiento mensual acostumbran a utilizarse **modelos econométricos** planteados a través de una relación de ecuaciones en las que la demanda se expresa como una función de factores económicos tales como la población, el ingreso per cápita, el precio de los distintos medios de combustible alternativos a aquél que se predice, etc. Otra alternativa que se comenta vuelve a ser el empleo de modelos ARIMA que suele concretarse en **modelos X11** los cuales están especialmente orientados al tratamiento de series mensuales, incorporando variables que hacen referencia al número de días laborables y festivos que tiene un mes, el momento del año en el que cae la Semana Santa, si el mes de Febrero es bisiesto, etc. En estos modelos puede utilizarse como variable explicativa un dato de temperatura media mensual si bien dicho efecto suele venir recogido en gran medida por la propia periodicidad anual de la serie.

También es frecuente hacer la predicción a medio plazo en términos diarios. Aunque no tiene sentido plantearse el conocimiento del valor esperado para la demanda en un día concreto de alguno de los dos años a los que se plantea el estudio, sí resulta de interés el consumo máximo diario que pueden llegar a alcanzarse en un periodo determinado (denominado “*pico*” en la terminología energética) así como su probabilidad de ocurrencia. No sólo se trata de realizar la mejor predicción posible del pico de consumo sino la probabilidad de tener que cubrir una demanda que supere en un determinado porcentaje el valor pronosticado. A este nivel, la influencia meteorológica es determinante pero obviamente desconocida pues no son realistas las previsiones que puedan proporcionarse más allá de la primera semana o quincena. Por ello los modelos que normalmente se utilizan se mueven en el ámbito de la simulación, es decir, la propuesta de una predicción “*base*” que se considera en condiciones normales de temperatura y las posibles variaciones que cabe esperar ante diferentes escenarios térmicos configurados por el analista. La tesis de [GUTIER] que aborda la predicción de la demanda eléctrica residencial en Andalucía, cita algunos de los enfoques más utilizados para la previsión del pico de demanda como el **modelo econométrico simple de Spann y Beauvais (1977)** en el que la punta de demanda se considera como el resultado de una componente base y otra que refleja el efecto de la temperatura. También menciona la posibilidad de utilizar modelos ARIMA de mayor o menor grado de complejidad como por ejemplo el modelo de **Abraham, B. (1983)** para la previsión de la demanda eléctrica en el estado americano de Iowa.

Otra forma alternativa de predecir la demanda a medio plazo con detalle diario es realizando una predicción mensual y desagregando el resultado en valores diarios. El procedimiento propuesto en el artículo de [GIDEDU] consiste en realizar una predicción de la demanda de gas a partir de la facturación mensual y de las temperaturas medias mensuales esperadas y posteriormente descomponer dicha predicción en valores diarios a través de una ecuación que permite

obtener las distribuciones de consumos diarios a partir de las distribuciones de consumo mensuales.

En esta tesis proponemos nuevos métodos para predecir a corto y medio plazo la demanda de gas, y por extensión la demanda eléctrica, conjugando algunas de las técnicas citadas con el empleo de una clase de funciones denominadas **cóputas**. Las posibilidades de este tipo de funciones en el ámbito de la simulación las hacen muy atractivas para abordar problemas de predicción a medio plazo y hemos considerado que es ésta una razón de peso para considerar la metodología que proponemos como una alternativa más de predicción a dicho horizonte, sin intención de establecer una comparativa rigurosa con los resultados a los que puedan conducir algunas de las técnicas que acabamos de enumerar.

En cuanto al corto plazo se refiere, las funciones cóputa suelen ser utilizadas para predecir series de carácter financiero sujetos a una fuerte volatilidad que en el contexto de la demanda energética acostumbra a traducirse en la predicción de los precios del gas y/o la electricidad (véase por ejemplo [KETTLER]). Por ello, dado que su empleo no parece estar tan orientado a la precisión que exige la predicción puntual de la demanda energética para un día determinado y teniendo en cuenta la amplia variedad de métodos que de antemano existe para tratar el problema, hemos estimado oportuno contrastar con alguno de ellos, los resultados a los que conduce un algoritmo que proponemos basado en su utilización. De esta forma pretendemos hacernos una idea de las posibilidades que pueden llegar a ofrecer estas funciones en un estudio predictivo a corto plazo y sentar las bases de futuras líneas de investigación. En concreto, hemos decidido comparar los resultados conseguidos con dicho algoritmo con aquellos proporcionados por una solución de eficacia contrastada en el tratamiento de este tipo de series. La solución a la que hacemos referencia y que también hemos mencionado en este capítulo consiste en aplicar la metodología Box-Jenkins basada en el empleo de modelos ARIMA complementado con modelos de función de transferencia para reflejar el efecto de las variables de temperatura.

0.2 Objetivos

El objetivo fundamental que persigue esta tesis es definir una metodología que automatice el proceso que conlleva realizar una predicción diaria a medio plazo (1-2 años) para una serie de demanda energética y, a través de ella, conocer el “pico” que se puede presentar durante dicho horizonte. A tal fin hemos utilizado un histórico de datos que refleja la demanda de gas natural en Madrid así como las temperaturas máximas y mínimas registradas en algunos observatorios de la provincia. Estos datos han sido proporcionados por el **Departamento de Análisis y Simulación** de la compañía Enagás.

El método que proponemos sugiere como punto de partida realizar una predicción diaria de la demanda para los dos próximos años en condiciones normales de temperatura siendo así posible conocer, bajo esta hipótesis de normalidad, el valor máximo que se puede esperar durante un periodo concreto (un invierno, un año, etc.). Este planteamiento inicial sigue en cierta medida las directrices marcadas por autores como **Engle, Granger, Mitchem y Ramanathan** (1979) o **Stanton y Gupta** (1980) quienes consideran que una forma adecuada de abordar esta problemática es estimar la demanda L_t de un día genérico “t”, y posteriormente considerar el pico esperado para ella como el $\max_t \{L_t\}$.

Para hacer esta predicción inicial, establecemos qué se entiende por condiciones climáticas normales y construimos a

partir de un histórico diario de demanda de gas, aquél que hipotéticamente se hubiera presentado en dichas condiciones. Así, a partir de un nuevo histórico cuyo comportamiento está aislado de la influencia meteorológica, realizamos en una segunda fase, una predicción a dos años vista a través de un modelo autorregresivo que incorpora entre sus variables explicativas el patrón cíclico de la serie, la extrapolación del cual permite plantear dicha predicción. Para la detección de este patrón cíclico hemos utilizado técnicas de suavizado de curvas como son el análisis mediante **wavelets** y los **splines de regresión**.

En una tercera fase, aplicamos la **Teoría de Cópulas** para analizar la relación de dependencia existente entre las desviaciones de los datos reales de demanda con respecto a los datos construidos en condiciones normales y las desviaciones entre los valores reales de temperatura y los supuestamente normales y, a partir de dicho análisis, simulamos el comportamiento que cabría esperar para la demanda en determinados escenarios térmicos. Esta es una de las características más innovadora que aporta esta metodología, el empleo de funciones cópula para analizar a fondo la relación demanda/temperatura, cuantificar la respuesta del consumo ante situaciones climatológicas especialmente adversas y dar un resultado basado, no en un único valor esperado alrededor del cual se construye un intervalo de confianza simétrico, sino en toda una distribución de valores sujeta a posibles asimetrías y apuntamientos que permita valorar la probabilidad de que la demanda alcance ciertas cotas y evaluar el riesgo asociado a la toma de determinadas decisiones. Es precisamente esta propuesta de utilizar esta teoría en el ámbito de la predicción de la demanda energética la que nos ha llevado a profundizar en las posibilidades que puede ofrecer y la que, en cierto modo, nos ha motivado a definir a esta tesis como una tesis sobre cópulas.

Queremos hacer hincapié en que, en este contexto de la predicción del pico de demanda, no es tan importante conocer con exactitud el valor máximo esperado para la demanda como la distribución de valores esperados que permita evaluar la probabilidad de que éste se presente y el riesgo que puede conllevar. Obviamente es de esperar que el pico de consumo diario se presente, casi con toda seguridad, en una condición climatológica especialmente adversa, o lo que es lo mismo, que valores extremos de la demanda de gas vayan a ir asociados a valores extremos de las variables de temperatura. A este respecto, la denominada **Teoría del Valor Extremo**, introducida en los años cincuenta por el matemático alemán **Emil Julius Gumbel**, es una rama de la estadística que estudia los extremos de distribuciones de probabilidad, analiza la relación existente entre las colas de dichas distribuciones y evalúa el riesgo asociado de que se presenten situaciones poco habituales. La teoría de cópulas presenta algunos vínculos con esta teoría puesto que se basa en encontrar la función de distribución conjunta que mejor representa a un conjunto de variables univariantes, la cual, en el caso que nos ocupa, debe enfatizar la relación de dependencia característica que se establece entre las colas de las variables *demanda* y *temperatura*. Veremos que cobran especial interés las denominadas **cópulas de valor extremo** a las cuales el propio Gumbel hizo su particular aportación (**cópula de Gumbel**) y que va a participar frecuentemente en nuestros análisis.

La idea que hemos perseguido es construir una herramienta informática que, en función de unos históricos de demanda energética (gas o electricidad) y temperaturas (máxima y/o mínima) de la zona en la que han sido registrados, genere un patrón de consumo en condiciones climáticas normales, proponga la función cópula que mejor identifique la relación entre ambas variables en un mes concreto y, a partir de ella, simule el comportamiento de la demanda bajo condiciones climatológicas desfavorables para, en particular, poder conocer el valor máximo que puede llegar a alcanzar y su probabilidad de ocurrencia. En este proceso, uno de los principales problemas que hemos encontrado y que surge en

general cuando se trabaja con funciones cópula es determinar aquella familia que mejor se ajuste a la muestra de datos dada e identifique de forma adecuada la relación de dependencia en ella implícita. Hemos utilizado un criterio ya existente, basado en someter a un contraste de bondad de ajuste a una relación de cópulas candidatas (véase [DOSCH]) y proponer como mejor alternativa aquella para la cual el estadístico de **Pearson**, evaluador del contraste, presente un menor valor respecto de una partición dada del cuadrado unidad, el cual, dicho sea de paso, contiene al soporte de estas funciones.

Ante la posibilidad de que ninguna de las familias candidatas proporcione un buen valor para dicho estadístico que permita garantizar que la cópula se ajuste en cierta medida a la muestra, hemos propuesto un método de construcción de funciones cópula empíricas no paramétricas que constituye el pilar teórico de esta tesis. Las funciones cuya construcción detallamos se obtienen por interpolación polinómica del dominio de definición de una clase de funciones denominadas **subcópulas** de la cual seleccionamos aquella que presenta un valor mínimo (y por tanto óptimo) para el estadístico de Pearson. La intención es disponer de una función cópula alternativa a la mejor familia candidata de expresión conocida en caso de que, por alguna razón, no consigamos a través de esta última unos resultados suficientemente satisfactorios. La naturaleza empírica de las cópulas que vamos a construir permite reflejar dependencias locales en función de la mayor o menor concentración de los pares muestrales en distintas regiones del cuadrado unidad, pudiendo desempeñar el papel de cópulas pertenecientes a clases de comportamiento tan dispar como la “gaussiana” o la de “valor extremo”, aunque creemos que su principal virtud será la de capturar relaciones multimodales incapaces de recoger en ocasiones por las familias más convencionales.

Si bien las posibilidades de las funciones cópula dentro del ámbito de la simulación las hacen especialmente atractivas para abordar problemas de predicción a medio plazo, también proponemos una manera de tratar la predicción de demanda energética a corto plazo (y en general cualquier tipo de predicción de series temporales), a través de un algoritmo innovador basado en el empleo de este tipo de funciones. El método que proponemos parte de las predicciones obtenidas con un modelo ARIMA ajustado únicamente a partir del histórico de la variable a predecir y suple el empleo de funciones de transferencia por el de funciones cópula con las que trata de ir reduciendo en sucesivas iteraciones el error medio porcentual que se obtiene conforme se tiene en cuenta la información aportada por distintas variables explicativas. La idea básica que subyace bajo el algoritmo es encontrar la cópula que mejor explique la relación de dependencia entre los errores de predicción derivados del ajuste del modelo y los valores que toma una variable explicativa. Conocida esta relación es posible predecir el error que se espera cometer de acuerdo a un valor concreto del regresor para, en función de él, llevar a cabo la corrección correspondiente y proponer una nueva predicción a la que irá asociada otro error de predicción que podrá ser relacionado con una nueva variable.

0.3 La tesis

Esta tesis se encuadra dentro de los objetivos que persigue el programa de doctorado del **Departamento de Estadística e Investigación Operativa de la Facultad de Matemáticas de la Universidad Complutense de Madrid**, cuya finalidad es potenciar la investigación de metodologías y resultados teóricos orientados a dar solución a un problema práctico.

La exposición que realizamos está estructurada en dos grandes bloques, uno puramente teórico en el que exponemos algunas nociones de la teoría de cópulas que consideramos básicas para el mejor entendimiento del método de construcción de cópulas que planteamos y otro práctico en el que abordamos la predicción de la demanda de gas a corto y medio plazo mediante el empleo de este tipo de funciones.

El bloque teórico se compone de un total de 6 capítulos de los cuales damos a continuación una breve descripción:

- En el primero de los capítulos, se hace una presentación de las funciones cópula, dándose su definición, haciéndose mención de los diferentes tipos y familias que existen y citándose las diferentes versiones del teorema de Sklar, resultado que supone uno de los pilares sobre el cual se construye esta teoría.
- El segundo capítulo enumera diversas técnicas existentes que permiten seleccionar la familia de cópulas que mejor se adapta a la relación de las variables que intervienen en un análisis y dentro de la familia seleccionada aquél miembro que se considera más adecuado para ello.
- El tercer capítulo hace un repaso de los métodos existentes para la construcción de funciones cópula tanto desde la perspectiva geométrica como desde la puramente analítica y, en conjunción con el capítulo segundo, sirve de introducción al siguiente que constituye la aportación teórica fundamental de esta tesis.
- En el capítulo cuarto se presenta una metodología para la generación de funciones cópula alternativa a las expuestas en el anterior. Se basa en la interpolación polinómica del dominio de definición de aquella función subcópula que optimiza el valor del estadístico de Pearson. Según se explica en el capítulo segundo, este estadístico es utilizado por algunos autores como criterio de evaluación de la bondad del ajuste de las funciones cópula a los datos.

Para mostrar cómo la construcción de cópulas a partir de subcópulas se complica conforme crece el orden de los polinomios interpoladores comparamos, a modo de ejemplo, la interpolación bilineal (realizada mediante polinomios de orden 1) con la cúbica, llevada a cabo a través de **polinomios de Hermite**. En este último caso, se presenta un amplio desarrollo, a raíz del cual se obtienen restricciones a imponer sobre dichos polinomios para que el resultado de la interpolación sea una función cópula. Además, se presenta otro método de

aproximación mediante **polinomios de Bernstein** el cual conduce a superficies que, si bien no interpolan exactamente el dominio de definición de la subcópula de partida, dan lugar a cópulas de aspecto más suave y apropiado para ser utilizadas con fines predictivos.

- El quinto capítulo explica la utilidad de las funciones cópula para realizar predicciones a medio plazo. Cuando el horizonte de predicción se prolonga, es difícil disponer de un valor fiable para todos y cada uno de los agentes explicativos que intervienen en el sistema. En tales circunstancias, el analista acostumbra a plantear diferentes escenarios configurados por hipotéticos valores de estos factores y, de acuerdo a éstos, realiza una simulación de la evolución del sistema. Explicaremos cómo la determinación de una función cópula que se ajuste bien a los datos permite obtener una buena representación de la función de densidad de la variable a predecir condicionada a las explicativas (densidad cópula condicionada). La simulación de la evolución de nuestro sistema se puede establecer a partir de la simulación de valores de la cópula condicionada.
- Finalmente el sexto capítulo detalla una posible forma de aplicar este tipo de funciones para realizar predicciones a corto plazo dada una serie temporal. Se argumentan los posibles problemas que pueden surgir con el empleo de este tipo de funciones cuando las variables sobre las que se definen presentan algún tipo de dependencia temporal y la posibilidad de utilizarlas como método alternativo al empleo de modelos de función de transferencia para reflejar la influencia de una variable explicativa en aquella que se desea predecir.

Por otra parte, el bloque práctico se compone de un único capítulo (el séptimo) que consta de dos apartados orientados respectivamente a la utilización de las funciones cópula para la predicción a corto y a medio plazo:

- El primero de los apartados trata la predicción a medio plazo (entre 1 o 2 años) con detalle diario. Para una compañía del sector gasista el conocimiento del “pico” diario que puede esperarse para el invierno del año siguiente le permite evaluar si la capacidad del gasoducto es suficiente para soportar la demanda de una determinada población y, en función de ello, plantear una posible expansión de su infraestructura.

Si bien en esta tesis analizamos como hemos dicho, una serie de demanda de gas, la metodología sería igualmente aplicable a una de consumo eléctrico. Para una compañía de este sector, el conocimiento del “pico” diario de demanda le permite realizar planes para que sus sistemas estén preparados ante posibles sobrecargas de la red. En las provincias del norte peninsular, este valor punta suele darse en invierno como resultado de una ola de frío que provoca un efecto de saturación de los sistemas de calefacción eléctrica. En la zona sur, el problema está más asociado a la estación del verano derivado de los aires cálidos procedentes del Sáhara traducidos en un preocupante incremento de la venta de aparatos de aire acondicionado.

El método que se detalla permite, en una primera fase, la obtención de un patrón de comportamiento diario de la demanda a dos años vista en condiciones normales de temperatura (patrón climático). En una fase posterior, se describe cómo aislar el histórico de datos de demanda del efecto climático para llevar a cabo la predicción a

tan largo horizonte (de 365 a 2*365 días) a partir de un modelo autorregresivo y de las componentes de baja frecuencia resultantes de la descomposición de la serie de demanda mediante funciones wavelets o splines de regresión. En una fase final, las funciones cópula van a permitir simular el crecimiento o decrecimiento esperado de la demanda supuesto un incremento del valor de los agentes climatológicos respecto del patrón estándar establecido. La relación entre la demanda y la temperatura no es lineal y además presenta un comportamiento difícil de capturar en situaciones extremas. Existen familias de cópulas que hacen énfasis en la relación entre sucesos “cola” y que, como veremos, se ajustan bastante bien a este tipo de situaciones. También veremos que las cópulas obtenidas por la metodología de construcción que se describe en el capítulo cuarto del bloque teórico, pueden ser una buena aproximación a estas familias.

- El segundo de los apartados trata la predicción a corto plazo de demanda de gas.

Los programas de una empresa responsable de la gestión de las redes de transporte de gas se establecen en función de los recursos disponibles (yacimientos), la demanda prevista que debe ser abastecida, el cumplimiento de compromisos contractuales con empresas suministradoras y comercializadoras, etc. La dificultad para mantener la calidad del gas (presión) dentro del gasoducto justifica la necesidad de que la predicción no exceda en demasía el dato real de demanda: el gas calienta menos y, en consecuencia se necesita más, con la consiguiente pérdida económica para la compañía. Si por el contrario el dato predicho se sitúa significativamente por debajo del real, los clientes industriales podrían ver reducidas sus tasas de producción y los domésticos plantearse el empleo de otro combustible como forma de calefacción para su hogar.

Las principales dificultades de modelización surgen a la hora de realizar una predicción condicionada a unas circunstancias climáticas especialmente desfavorables, más aún cuando éstas se presentan de manera brusca e inesperada (olas de frío en Marzo). Como ya hemos comentado, entre las diferentes técnicas que pueden ser contrastadas, los modelos ARIMA con funciones de transferencia para reflejar el efecto de las temperaturas constituyen tal vez la alternativa que proporciona unos mejores resultados. La idea que proponemos en su lugar, sugerida como veremos por la manera de proceder en el ajuste a medio plazo, consiste en reemplazar las funciones de transferencia por funciones cópula que cuantifiquen de manera diferente la relación de dependencia en condiciones extremas (pico de demanda – ola de frío).

Los resultados de este bloque práctico han sido obtenidos a partir de una relación de programas implementados fundamentalmente con el software *SAS* dada las buenas y contrastadas referencias que de él encontramos para el desarrollo de proyectos de *Data Mining*. En concreto, los módulos utilizados de esta herramienta han sido:

- *SAS/BASE*.- que soporta el lenguaje de programación de la herramienta.
- *SAS/GRAPH*.- para todo lo referido al aspecto visual y de representación gráfica de resultados.

- *SAS/STAT*.- para la obtención de estadísticas básicas, realización de contrastes de hipótesis y planteamiento de modelos de regresión.
- *SAS/ETS*.- para la implementación de modelos ARIMA.
- *SAS/OR*.- para la obtención de soluciones a problemas de programación matemática (lineal y no lineal, continua y entera).

También se ha utilizado el paquete estadístico *SPLUS/WAVELETS* para la construcción de funciones wavelets. Si bien éstas pueden también ser empleadas disponiendo del correspondiente módulo IML de *SAS*, hemos considerado que es el software de *SPLUS* el que ofrece un mayor abanico de opciones asociadas a este tipo de funciones ampliando de manera notable sus posibilidades de uso. De todas formas, hemos programado con *SAS/BASE* otra técnica de suavizado de curvas, alternativa a las funciones wavelets, basada en splines, con la idea de que la herramienta para el tratamiento de series de demanda energética que hemos desarrollado estuviera totalmente integrada en *SAS*.

En lo que a las cópulas concierne, no hemos encontrado ningún procedimiento de *SAS* que permita gran versatilidad a la hora de realizar análisis basados en esta teoría. Tan sólo el módulo *RISK DIMENSIONS* ofrece algunas posibilidades. Por ello, y aun cuando *SPLUS* tiene alguna solución orientada al tratamiento con cópulas (*S+FINMETRICS*), hemos considerado de gran utilidad, no sólo de cara a la integración con la herramienta mencionada, sino además para futuros trabajos, implementar en *SAS/BASE* una relación de macros específicas que permitan trabajar con ellas.

En el capítulo octavo de la tesis presentamos las principales conclusiones teóricas y prácticas a las que hemos llegado y en el noveno, las próximas líneas de investigación que pretendemos seguir en el futuro.

Finalmente se incluye una relación de Anexos en los cuales se pueden consultar las demostraciones rigurosas de los lemas y teoremas que proponemos en el bloque teórico y que constituyen la aportación teórica que hacemos al campo de la teoría de cópulas.

También se pueden consultar en este capítulo de anexos las expresiones correspondientes a las familias de cópulas más populares, así como una introducción a algunos de los métodos matemáticos utilizados en el desarrollo práctico como son las wavelets o los splines de regresión. Pretendemos con ello dar una idea que facilite el mejor entendimiento de esta exposición.

1. Funciones cópula

1.1 Introducción

Un modelo matemático no es más que una forma de intentar explicar la realidad, una forma de medir la relación que existe entre las variables que definen un sistema. Cuando el objetivo es claramente predictivo, acostumbra a plasmarse en una relación de ecuaciones paramétricas que plantean una hipótesis sobre el funcionamiento de dicha realidad. Cada parámetro de cada ecuación tiene por finalidad cuantificar esa supuesta relación. Lógicamente es imposible predecir el futuro con exactitud y es por ello que una parte del modelo esté reservada al error que asume que cometerá, un término que se considera desconocido.

Existen muchos tipos de modelos que pueden ser utilizados para predecir, algunos lineales y paramétricos como los **modelos de regresión** o los ARIMA, otros no paramétricos como los **árboles de decisión**, algunos no lineales como las redes neuronales, etc. Todos ellos tratan de detectar la relación de dependencia que liga a las variables del sistema y además comparten un aspecto común que es la necesidad de disponer de un histórico de datos sobre el que estudiar dicha relación. Cuanto mayor sea su volumen y más fiable su contenido, de mayor calidad serán los resultados que el modelo proporcione.

En el ámbito matemático, predecir consiste en conocer el comportamiento futuro de un sistema que puede ser más o menos complejo. En su versión más simplificada, predecir es conocer el valor futuro que se espera que tome una variable en función de su relación con otra u otras variables. Si el comportamiento de la variable a predecir está condicionado por su propia historia como ocurre cuando se plantean modelos diseñados para el tratamiento de series temporales, la relación se establecerá igualmente entre dos variables: el presente y el pasado de la misma.

Hemos empleado con cierta familiaridad dos términos que tienen un equivalente matemático directo:

- La “*esperanza*” matemática de una variable identifica el valor que se “**espera**” que tome la variable.
- El término “**condicionado**” presupone que existe un conocimiento del valor de las variables que van a intervenir como explicativas en el modelo.

La conjunción de ambos conceptos da lugar al de “*esperanza condicionada*” de una variable que viene a ser el valor que se espera que tome la variable condicionado a que se tiene un conocimiento del valor de otras que actúan como explicativas. El cálculo de la esperanza condicionada de una variable Y al valor “ x ” de otra (u otras) X se realiza a partir de la distribución condicionada $Y|X=x$ que es una función que permite conocer como se distribuyen los valores de la variable Y cuando se sabe que la variable X toma el valor “ x ”. El conocimiento de esta distribución permite

localizar cuál es el valor de Y que se presenta más habitualmente cuando X vale “ x ”, viendo cuál es el valor (o valores) de Y alrededor del cual la distribución es más densa (tiene mayor masa o densidad). Éste puede ser por ejemplo la media, la mediana o la moda de la distribución y puede ser propuesto como valor esperado o predicción. Además, la distribución informa también del nivel de concentración alrededor del valor que damos como predicción, lo cual permite proporcionar una idea de su volatilidad (o variabilidad). Esta medida de la dispersión esperada para la predicción resulta de gran utilidad para analizar el riesgo derivado de la toma de una posible decisión fundamentada en ella. Nuevamente se repite el término matemático “*esperado*” y aparece uno nuevo, “*dispersión*”, palabras que conducen al concepto de varianza, que viene a ser lo que se espera que valga la dispersión (la media de las desviaciones con respecto a la predicción).

Es importante destacar la necesidad cada vez mayor de proporcionar estas dos medidas. En ocasiones, no es tan importante la predicción puntual como el grado de confianza en ella que nos puede dar el conocimiento de su dispersión esperada. Y aún más, puede interesarnos conocer la distribución de valores esperada dado que en ocasiones pueden existir varios valores candidatos para la predicción (distribuciones multimodales) y, asociados a ellos, diferentes medidas de su variación. Los métodos de predicción más convencionales suelen proporcionar una predicción y un intervalo de confianza asociado a la misma que por lo general suele ser simétrico. Esto es así porque el método presupone ciertas hipótesis de normalidad de los agentes que intervienen en el sistema, hipótesis que en muchas circunstancias no son realistas.

En resumen, estamos interesados en conocer la distribución esperada de una variable del sistema, condicionado al valor de otras con las que mantiene algún tipo de relación desconocida que no tiene porqué ser cuantificable por un parámetro. No hay que buscar necesariamente una o varias fórmulas que plasmen la relación mediante sumas y/o productos de pares de parámetros y/o variables. Nuestro objetivo es determinar la forma de esa distribución condicionada y, a partir de ella, proponer una predicción.

La función de densidad (equivalentemente, de masa) asociada a la distribución condicionada $Y|X$ se puede calcular a partir de la densidad (o masa) de la distribución conjunta de (X,Y) y de la densidad (o masa) de la distribución marginal de X de la siguiente manera:

- En el caso en el que las variables tomen valores continuos, $f(y|x) = \frac{f_{XY}(x,y)}{f_X(x)}$ siendo f_{XY} la función de densidad de la conjunta (X,Y) y f_X la función de densidad de X .
- En el caso en el que las variables tomen valores discretos, $P(y|x) = \frac{P_{XY}(x,y)}{P_X(x)}$ siendo P_{XY} la función de masa de la conjunta (X,Y) y P_X la función de masa de X .

La distribución condicionada que buscamos podrá expresarse en función de la distribución conjunta de las variables y de alguna de sus marginales. Determinar las marginales que mejor se ajustan a cada una de las variables puede ser más

o menos sencillo dado que la propia representación de los valores de la variable nos puede sugerir el empleo de alguna de las muchas distribuciones conocidas que existen (Poisson, Uniforme, Normal, Exponencial, etc.) y, aun en el peor de los casos, siempre podemos optar por utilizar una distribución empírica.

Mucho más complicado resultaría proponer una expresión para la distribución conjunta, entre otras cosas porque en ella debe venir representada implícitamente la relación de dependencia existente entre las variables X e Y . La solución a este problema no es inmediata como veremos a continuación.

Supongamos un caso muy sencillo en el que tenemos dos variables X e Y con sus respectivas funciones de distribución marginales $F_X(x)$ y $G_Y(y)$. Entonces, existen infinitas funciones de distribución bivariantes conjuntas $H_{XY}(x,y)$ cuyas marginales son $F_X(x)$ y $G_Y(y)$. Más aun, podemos encontrarnos con pares de variables (X_1, Y_1) y (X_2, Y_2) que responden a estructuras de dependencia diferentes aun cuando las X_i y las Y_i presentan la misma distribución y siendo además el coeficiente de correlación lineal entre X_1 e Y_1 el mismo que entre X_2 e Y_2 . Es decir, pueden existir dos funciones de distribución conjuntas y distintas H_1 y H_2 asociadas a (X,Y) , que explicarían respectivamente y de forma diferente la relación de dependencia entre ellas.

Nuestra intención es encontrar la función de distribución conjunta H que mejor refleje la relación entre X e Y ; aquella que nos garantice que la distribución condicionada que construyamos represente fielmente esa relación y pueda ser utilizada por el analista con fines predictivos. Es en este contexto en el que aparecen en 1959 unas funciones bautizadas por el matemático **Abe Sklar** como **cópulas**, funciones que, como el autor define, unen (o copulan) funciones de distribución multivariantes a sus marginales unidimensionales de igual forma que el término gramatical “cópula” sirve de enlace entre el sujeto y el predicado de una oración. Es el concepto lingüístico el que sirve de inspiración a Sklar para dar nombre a esta clase de funciones. Citaremos algunos referentes históricos que han sido extraídos del capítulo 1 de [NELSEN], al cual invitamos al lector a consultar para conocer más en detalle algunos de los hitos más significativos en la historia de esta teoría.

Quizá los precedentes más remotos de la teoría los encontramos en 1940-1941, cuando **Wassily Hoeffding** define una clase de distribuciones bivariantes estandarizadas cuyo soporte está contenido en el cuadrado $[-1/2, 1/2]^2$ y cuyas marginales son uniformes sobre $[-1/2, 1/2]$. Según **Schweizer** (1991), si Hoeffding hubiese utilizado como dominio de definición el intervalo $[0,1]$ en vez de $[-1/2, 1/2]$, hubiese sido él quien hubiese descubierto las cópulas.

Sin embargo, el antecedente más claro al desarrollo de Sklar lo encontramos en el trabajo de Ferón en 1956. Ferón realizó un estudio sobre distribuciones tridimensionales en el que definía unas funciones auxiliares de dominio el cuadrado unidad que le permitieron enlazar dichas distribuciones con sus marginales univariantes. Sklar observa, según nos comenta en uno de sus artículos más recientes que data de 1996, que funciones con características similares a las de Ferón podían ser definidas para dimensiones mayores o iguales que 2, y que de igual manera servirían de enlace entre las distribuciones conjuntas y sus marginales univariantes. A partir de esta apreciación, Sklar establece el teorema que lleva su nombre y que constituye el pilar fundamental de una teoría que ha despertado a partir de la década de los noventa y que, dada su “corta” edad, mantiene hoy día muchas líneas de investigación abiertas.

1.2 Definición

La mayor parte de los artículos que podemos encontrar sobre cópulas comienzan dando su definición sin hacer alusión a otra clase de funciones a partir de la cual podría definirse en una única frase: las cópulas son subcópulas cuyo dominio es el cuadrado unidad. Esta forma de presentarlas que hace **Nelsen** (véase capítulo 2 de [NELSEN]) nos parece la más apropiada dado que la metodología que constituye la base teórica de esta tesis se fundamenta en la posibilidad de extender cualquier subcópula a una cópula (véase el lema 2.3.5 de [NELSEN]). Sin embargo, de igual manera que desconoceríamos las tareas de un director si lo único que supiéramos de él es que en su ausencia es el subdirector quien las desempeña, y a pesar de las ventajas de espacio que ofrece la definición propuesta, parece que estamos obligados a explicar las propiedades que caracterizan a una función subcópula.

Definición 1.1.- Una **2-subcópula** es una función C' con las siguientes propiedades:

1. $Dom(C')=S_1 \times S_2$ siendo $S_1, S_2 \subset [0,1]; 0,1 \in S_1, S_2$ (1.1)

2. $C'(u,0)=0; C'(0,v)=0 \quad \forall u \in S_1, v \in S_2$ (1.2)

3. $C'(u,1)=u; C'(1,v)=v \quad \forall u \in S_1, v \in S_2$ (1.3)

4. C' es 2-creciente, es decir, el volumen a través de C' de cualquier rectángulo B contenido en su dominio de definición, también llamado **C' -volumen**, es mayor o igual que cero, o lo que es lo mismo,

$$V_c([u_1, u_2] \times [v_1, v_2]) = C'(u_2, v_2) - C'(u_2, v_1) - C'(u_1, v_2) + C'(u_1, v_1) \geq 0 \quad \forall B = [u_1, u_2] \times [v_1, v_2] \subset Dom(C')$$

(1.4)

El “2” que precede al término subcópula identifica que es una función bivalente y que implícitamente es utilizada para reflejar la relación entre pares de variables. Si bien esta teoría es extrapolable al caso n-variante y tiene sentido por tanto hablar de n-subcópulas (y n-cópulas), nos centraremos en el caso bivalente dado que resulta más fácil de manejar en cuanto a notación, se puede visualizar gráficamente (la relación entre X e Y se puede representar en el plano y la función cópula que depende de ellas, en el espacio) y es la que normalmente encontraremos en la mayor parte de la literatura existente al respecto. Por ello, a partir de ahora los términos subcópula y cópula irán referidos al caso bivalente.

Definición 1.2.- Una **cópula** es una función $C: [0,1] \times [0,1] \rightarrow [0,1]$ con las siguientes propiedades:

1. $C(u,0)=0; C(0,v)=0 \quad \forall u, v \in [0,1]$ (1.5)

2. $C(u,1)=u; C(1,v)=v \quad \forall u, v \in [0,1]$ (las marginales de C son uniformes) (1.6)

3. C es 2-creciente, es decir, $V_c([u_1, u_2] \times [v_1, v_2]) = C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \geq 0$
 $\forall u_1, u_2, v_1, v_2 \in I^2$ t.q. $u_1 \leq u_2, v_1 \leq v_2$ (1.7)

1.3 Teorema de Sklar

Muchas de las propiedades de las cópulas y las subcópulas son comunes. Sin embargo, el hecho de que se diferencien en el dominio de definición, resulta determinante como pone de manifiesto el **Teorema de Sklar** (1959), que constituye el pilar fundamental y una referencia obligada en la teoría de cópulas pues establece la relación que existe entre las distribuciones multivariantes y sus marginales univariantes a través de un cópula.

Teorema 1.1.- Teorema de Sklar

“Sean X e Y variables aleatorias con funciones de distribución marginales F_X y G_Y respectivamente y función de distribución conjunta H_{XY} . Entonces existe una cópula C tal que

$$H_{XY}(x, y) = C(F_X(x), G_Y(y)) \quad \forall x, y \in \bar{\mathbb{R}} = [-\infty, \infty] \quad (1.8)$$

Si F_X y G_Y son continuas, entonces C es única; si no, C está unívocamente determinada sobre $\text{Ran}(F_X) \times \text{Ran}(G_Y)$. Recíprocamente, si C es una cópula y F_X y G_Y son funciones de distribución, entonces la función H_{XY} definida por $H_{XY}(x, y) = C(F_X(x), G_Y(y))$ es una función de distribución conjunta con marginales F_X y G_Y ”.

No entraremos en la demostración de este teorema que puede ser encontrada en [NELSEN] (véase teorema 2.3.3). Sí nos gustaría hacer una observación que ayuda a comprender mejor el empleo de la notación (u, v) como par sobre el que actúa una función cópula. Verdaderamente, se puede ver que C actúa sobre puntos de I^2 puesto que el par $(F_X(x), F_Y(y))$ se mueve en el cuadrado unidad al ser valores probabilísticos. Este par se corresponde con dos valores asociados a sendas distribuciones uniformes estándar (u, v) según establece el **Teorema de Transformación Integral de Probabilidad de Fisher** (1932) (véase apartado 4.2 de [HEIREN]).

Teorema 1.2.- Teorema integral de probabilidad

“Si X es una variable aleatoria continua con función de distribución acumulada F_X , la variable aleatoria $Z = F_X(X)$ se distribuye según una variable aleatoria uniforme $(0, 1)$ ”.

Dicho teorema viene a ser, en cierto modo, un caso particular de un resultado aún más genérico, que permite determinar la distribución de la transformación de una variable aleatoria continua. Este resultado (véase el teorema 7.6 de [QUEGAR]) establece lo siguiente:

Teorema 1.3.- Transformada de variables aleatorias continuas

“Sea X una variable aleatoria absolutamente continua con soporte C_X y función de densidad f_X . Sea $\phi: \mathbb{R} \rightarrow \mathbb{R}$ estrictamente creciente o decreciente en C_X y en consecuencia biyectiva y por tanto inyectiva, lo cual garantiza la existencia de ϕ^{-1} . Si ϕ^{-1} admite derivada continua sobre $C_Z = \phi(C_X)$ entonces $Z = \phi(X)$ es una variable aleatoria absolutamente continua con soporte C_Z y función de densidad $k_Z(z) = f_X(\phi^{-1}(z)) * |(\phi^{-1})'(z)| * I_{C_Z}(z)$ ”.

En el caso que nos ocupa, $\phi=F_X$ es monótona creciente por ser función de distribución.

$$z=\phi(x)=F_X(x)\Rightarrow\phi^{-1}(z)=F_X^{-1}(z) \quad y \quad C_Z=\phi(C_X)=F_X(C_X)=(0,1)$$

El soporte transformado es el intervalo abierto (0,1) dado que obviamente

$$F_X(\min\{x|x\in C_X\})=F_X(-\infty)=0 \quad y \quad F_X(\max\{x|x\in C_X\})=F_X(+\infty)=1$$

En consecuencia, podemos aplicar el resultado y establecer que:

$$k_Z(z)=f_X(F_X^{-1}(z))*\left|\frac{\partial(F_X^{-1}(z))}{\partial z}\right|*I_{(0,1)}(z)$$

Dado que $F'_X(x)=f_X(x)\Rightarrow F'_X(F_X^{-1}(z))=f_X(F_X^{-1}(z))$, la expresión anterior es igual a:

$$k_Z(z)=F'_X(F_X^{-1}(z))*\left|\frac{\partial(F_X^{-1}(z))}{\partial z}\right|*I_{(0,1)}(z)=\frac{d(F_X(F_X^{-1}(z)))}{dz}*I_{(0,1)}(z)=\frac{dz}{dz}*I_{(0,1)}(z)=I_{(0,1)}(z) \quad \text{que es la}$$

función de densidad de una uniforme en el intervalo (0,1)

En consecuencia podemos concluir que $F_X(X)=U=^d U(0,1)$ y $G_Y(Y)=V=^d U(0,1)$ y que la relación que establece el teorema de Sklar entre las marginales y la conjunta a través de la cópula C es

$$H_{XY}(x,y)=C(F_X(x),G_Y(y))=C(u,v) \quad \forall x,y\in\bar{\mathbb{R}}=[-\infty,\infty] \quad .$$

Estos resultados ayudan a entender mejor la definición de las cópulas que se presenta en [MATTEIS] (definición 2.2).

Definición 1.3.- Una **cópula** es una función de distribución multivariante F de variables aleatorias X_1, X_2, \dots, X_n cuyas distribuciones marginales son uniformes estándar, es decir, $X_i \sim F_i \quad \forall i=1, 2, \dots, n$

Como se puede leer en [PATTON2], existen otras versiones del teorema de Sklar enunciadas a partir de dos funciones asociadas a una cópula y cuya existencia está garantizada por ser C una función monótona.

Definición 1.4.- Las derivadas parciales respecto de U y V , que en virtud de la monotonía de C existen para casi todo “ u ” y “ v ” respectivamente (salvo conjunto de medidas **Lebesgue** nula), definen las denominadas **cópulas condicionadas** asociadas a C y son:

$$\text{Dado “}u\text{” fijo, la cópula condicionada a “}u\text{” es la función de }V, \quad v \rightarrow C_1(u,v)=C(v|u)=\frac{\partial C}{\partial u}(u,v) \quad (1.9)$$

$$\text{Dado “}v\text{” fijo, la cópula condicionada a “}v\text{” es la función de }U, \quad u \rightarrow C_2(u,v)=C(u|v)=\frac{\partial C}{\partial v}(u,v) \quad (1.10)$$

Se puede demostrar que estas funciones así definidas no sólo existen, sino que además son no decrecientes para casi todo punto de $[0,1]$ (véase el teorema 2.7.7 de [NELSEN]).

La consecuencia directa de la adaptación del teorema de Sklar para distribuciones condicionadas continuas es que las distribuciones condicionadas de $Y|X=x$ y $X|Y=y$ vienen dadas, respectivamente, por

$$H_{Y|X}(y)=C_1(F(x),G(y)) \quad (1.11)$$

$$H_{X|Y}(x)=C_2(F(x),G(y)) \quad (1.12)$$

pues por ejemplo:

$$\frac{\partial C}{\partial u}(u, v) = \lim_{h \rightarrow 0} \frac{C(u, v) - C(u-h, v)}{C(u, \infty) - C(u-h, \infty)} = \underset{\text{marg. uniformes}}{\lim_{h \rightarrow 0}} \frac{C(F_X(x), G_Y(y)) - C(F_X(x-h), G_Y(y))}{F_U(u) - F_U(u-h)} \underset{\text{Sklar}}{=} \\ \lim_{h \rightarrow 0} \frac{H(x, y) - H(x-h, y)}{h} = H(y|x) \quad \text{siendo } U \sim^d U(0,1) \text{ y } F_U \text{ su F.D}$$

Esta relación es para nosotros de gran importancia dado que la predicción de una variable Y a partir del conocimiento de otra variable X la plantearemos a partir de la simulación de valores de la distribución condicionada $Y|X=x$ mediante el **método de la transformada inversa** (véase el apartado 2.3 de [PARVAL]), siendo para ello necesario disponer de su función de distribución $H(y|x)$ que, como vemos, puede ser aproximada mediante una función cópula. Finalmente, cabe hablar de función de densidad conjunta asociada a las variables U y V que viene a denominarse como densidad cópula y cuya existencia está igualmente garantizada en función del citado teorema 2.7.7 que encontramos en [NELSEN].

Definición 1.5.- Se define la **densidad cópula** condicionada asociada a una cópula C como:

$$C_{12}(u, v) = \frac{\partial C}{\partial u \partial v}(u, v) \quad (1.13)$$

Podemos encontrar en [PATTON2] (véase apartado 2.1) una versión del teorema de Sklar asociada a las funciones de densidad.

Teorema 1.4.- Teorema de Sklar para Densidades Cópulas

“Sea h_{XY} función de densidad bivalente conjunta de las variables X e Y con densidades marginales f_X y g_Y . Entonces h_{XY} se puede descomponer en el producto de las marginales f_X y g_Y y la densidad cópula c de la siguiente manera:

$$h_{XY}(x, y) = c(F_X(x), G_Y(y)) * f_X(x) * g_Y(y) \quad (1.14)”$$

Estas son las dos principales versiones que encontraremos del Teorema de Sklar, una asociada a funciones de distribución y otra a funciones de densidad. Sin embargo, existen algunas otras como por ejemplo la que se cita en [PATTON2] (véase teorema 2 en el apartado 2.2) que está más orientada al tratamiento de series temporales pues maneja distribuciones condicionadas que dependen de un subíndice “ t ” el cual hace referencia a un instante de tiempo. A este respecto propone la siguiente versión del teorema:

Teorema 1.5.- Teorema de Sklar para Distribuciones Condicionadas Continuas

“Sea H_t función de distribución bivalente condicionada con marginales continuas F_t y G_t y sea F_{t-1} un conjunto al que se condicionará. Existe entonces una única cópula condicional $C_t: [0,1] \times [0,1] \rightarrow [0,1]$ tal que

$$H_t(x, y|F_{t-1}) = C_t(F_t(x|F_{t-1}), G_t(y|F_{t-1})|F_{t-1}) \quad \forall x, y \in \bar{\mathbb{R}} = [-\infty, \infty] \quad (1.15)$$

Recíprocamente, si C_t es una cópula condicionada y F_t y G_t son las funciones de distribución condicionadas de dos variables aleatorias X e Y , entonces la función H_t definida por (1.15) es una función de distribución condicionada bivalente con marginales F_t y G_t ”.

El teorema emplea una función denominada *cópula condicional*, cuya definición también se presenta en [PATTON2] (definición 2 del apartado 2.2):

Definición 1.6.- Se define una **2-cópula condicional** como una función $C_t: [0,1] \times [0,1] \rightarrow [0,1]$ que cumple las siguientes propiedades:

$$1. \quad C_t(u, 0|F_{t-1})=0; C_t(0, v|F_{t-1})=0 \quad \forall u, v \in [0,1] \quad (1.16)$$

$$2. \quad C_t(u, 1|F_{t-1})=u; C_t(1, v|F_{t-1})=v \quad \forall u, v \in [0,1] \quad (1.17)$$

$$3. \quad V_{C_t}([u_1, u_2] \times [v_1, v_2]|F_{t-1}) = C_t(u_2, v_2|F_{t-1}) - C_t(u_2, v_1|F_{t-1}) - C_t(u_1, v_2|F_{t-1}) + C_t(u_1, v_1|F_{t-1}) \geq 0 \\ \forall u_1, u_2, v_1, v_2 \text{ t.q. } u_1 \leq u_2, v_1 \leq v_2 \quad (1.18)$$

En cualquiera de sus versiones, el resultado pone de manifiesto que dada una *cópula* C , y dos funciones de distribución F_X y G_Y asociadas respectivamente a dos variables X e Y , es posible construir una función de distribución conjunta H_{XY} para la bidimensional (X, Y) , que tiene por funciones de distribución marginales a F_X y G_Y .

Una propiedad muy atractiva de las funciones *cópula* es su independencia con respecto a las unidades en las que son medidas las variables X e Y . Así, si por ejemplo ϕ y ψ son dos transformaciones crecientes y no lineales, entonces la *cópula* asociada al par $(\phi(X), \psi(Y))$ es la misma que la asociada al par (X, Y) , invarianza que no se mantiene para el coeficiente de correlación.

Han sido muchos los autores que han propuesto funciones *cópula* orientadas a reflejar uno o varios aspectos característicos de la relación existente entre las variables a analizar. Cada una de ellas proporciona una función H distinta cuando se aplica al par $(F_X(x), G_Y(y))$. El problema que inevitablemente surge es encontrar la *cópula* C que permita la obtención de la distribución conjunta H_{XY} , que mejor explique el grado de vinculación existente entre X e Y .

1.4 Capacidad de las funciones *cópula* para reflejar relaciones de dependencia

La selección de una u otra función *cópula* C suele estar condicionada por la forma en que ésta establece la relación de dependencia entre las variables U y V , relación que es cuantificable de muchas formas.

Así por ejemplo, el **coeficiente de correlación** $\rho(X, Y)$ (que se mueve entre -1 y 1) nos proporciona un indicador con el que valorar la dependencia lineal que existe entre X e Y . Cuanto más próximo a 1 esté en valor absoluto, mayor es la relación lineal que vincula a las variables. Además, el signo de este coeficiente nos informa del “sentido” de la relación: si es positivo, X crece conforme crece Y , mientras que si es negativo, una y otra variable se mueven, de forma lineal, en sentido opuesto. El hecho de que $\rho(X, Y)$ valga 0 es representativo de ausencia de relación lineal entre X e Y , lo cual no quiere decir que no pueda existir algún otro tipo de relación (no lineal).

Existen también las denominadas “**medidas de asociación**”, algunas tan populares como la **Tau de Kendall** y el **coeficiente de correlación de Spearman**, que cuantifican relaciones no necesariamente lineales, sino que se utilizan directamente como funciones de evaluación del contraste de independencia:

$$H_0: F_{XY}(x, y) = F_X(x)F_Y(y) \quad (X \text{ e } Y \text{ independientes}) \quad \text{vs} \quad H_1: F_{XY}(x, y) \neq F_X(x)F_Y(y) \quad .$$

Estas medidas, se mueven también entre -1 y 1. Cuando toman alguno de estos valores extremos, reflejan respectivamente una relación de dependencia negativa o positiva “perfecta”. Conforme se aleja de ellos, la medida es sinónimo de falta de dependencia entre las variables. En términos coloquiales, vienen a determinar cómo se relacionan los valores “grandes” y “pequeños” de la variable aleatoria X con los de la variable Y .

El capítulo 5 de [NELSEN] analiza en profundidad las diferentes vías en las que las cópulas pueden ser utilizadas para estudiar la dependencia entre variables. Entre los resultados más importantes, destacamos aquéllos que ponen de manifiesto la relación entre una función cópula ajustada a las muestras asociadas a un par de variables U y V y el valor muestral de sus medidas de asociación (teoremas 5.1.3 y 5.1.6).

Teorema 1.6.- Relación entre una cópula y la Tau de Kendall

“Sean X e Y variables aleatorias continuas cuya cópula es C . Entonces, la popular versión de la Tau de Kendall para X e Y viene dada por

$$\tau_{XY} = 4 \iint_{I^2} C(u, v) dC(u, v) - 1 = 4 * E[C(U, V)] - 1 \quad (1.19)''.$$

Teorema 1.7.- Relación entre una cópula y el coeficiente de correlación de Spearman

“Sean X e Y variables aleatorias continuas cuya cópula es C . Entonces, la popular versión del coeficiente de correlación de Spearman para X e Y viene dada por

$$\rho_{XY} = 12 \iint_{I^2} C(u, v) dudv - 3 \quad (1.20)''.$$

Gran parte de la importancia de estas relaciones es su utilidad para concretar la cópula más adecuada de entre todas las pertenecientes a una misma familia paramétrica puesto que, por lo general, es fácil calcular el valor del parámetro a partir del estimador muestral de estas medidas de asociación mediante las expresiones (1.19) y (1.20).

1.5 Tipos de cópulas

Existen muchos tipos de funciones cópula y es difícil encontrar en la literatura una clasificación clara de todas ellas dado que existen muy diversos criterios para hacerlo: en función de la dependencia o no de parámetros, de su soporte (continuo o discreto), del tipo de relación que reflejan (cópulas elípticas, cópulas de valor extremo, etc.) Por ello, en vez de presentar un esquema general que permita ubicar cada cópula de acuerdo a una jerarquía concreta, enumeraremos algunos de estos criterios y citaremos algunos ejemplos asociados a las clases que resultan de su aplicación.

1.5.1 Tipos de cópulas en función de su soporte

Según se presenta en [NELSEN] (expresión 2.4.1) , toda cópula C se puede descomponer como

$$C(u, v) = A_C(u, v) + S_C(u, v) \quad (1.21)$$

siendo $A_C(u, v) = \int_0^u \int_0^v \frac{\partial^2}{\partial s \partial t} C(s, t) ds dt$ la parte absolutamente continua de la cópula y $S_C(u, v)$ la parte singular. Así, podemos hacer una primera clasificación en:

- **Cópulas singulares.-** aquéllas que no tienen parte absolutamente continua,

$$\int_0^u \int_0^v \frac{\partial^2}{\partial s \partial t} C(s, t) ds dt = 0 \Rightarrow \frac{\partial^2 C(s, t)}{\partial s \partial t} = 0 \text{ para casi todo } (s, t) \in I^2 .$$

Se trata de cópulas cuyo soporte son puntos o rectas.

Son ejemplos de este tipo de cópulas las populares cotas inferior y superior de **Fréchet-Hoeffding** dadas respectivamente por

$$W(u, v) = \max(0, u + v - 1) \quad (1.22)$$

$$M(u, v) = \min(u, v) \quad (1.23).$$

La importancia de estas cópulas es que proporcionan una acotación para cualquier otra función cópula verificándose la siguiente relación $W(u, v) \leq C(u, v) \leq M(u, v) \quad \forall \text{ cópula } C$ (1.24)

La primera de ellas, W , es reflejo de dependencia negativa perfecta entre las variables X e Y mientras que la segunda, M , es reflejo de dependencia positiva perfecta.

Los soportes de estas cópulas son las diagonales secundaria (para W) y principal (para M) del cuadrado unidad, cuyas pendientes -1 y 1 son representativas del tipo de dependencia que reflejan. El siguiente gráfico ilustra al soporte de estas funciones.

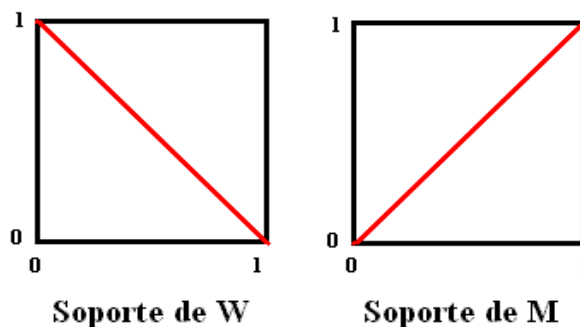


Ilustración 1.1: Soporte de las cópulas de Fréchet-Hoeffding

- **Cópulas absolutamente continuas.**- aquéllas que no tienen parte singular, es decir, tales que $S_C(u, v) = 0$.

Un ejemplo de este tipo de cópulas es la cópula producto

$$\Pi(u, v) = u * v \quad (1.25)$$

A través de ella queda caracterizada la independencia entre las variables X e Y . Basta observar que

$C_1(u, v) = C(v|u) = v$, es decir, la cópula condicionada a un valor “ u ” es independiente de dicho valor (y de igual manera $C_2(u, v) = C(u|v) = u$).

El hecho de que X e Y sean independientes implica que el valor de cualquiera de las medidas de asociación citadas ((1.19) o (1.20)) sea 0. Así por ejemplo, para la correlación rango de Spearman se puede comprobar

$$\text{que: } \rho_{XY} = 12 \iint_{I^2} C(u, v) dudv - 3 = 12 \int_0^1 \int_0^1 uv dudv - 3 = 12 \int_0^1 v * \left[\frac{u^2}{2} \right]_0^1 dv - 3 = 12 \int_0^1 \frac{v}{2} = 12 \left[\frac{v^2}{4} \right]_0^1 - 3 = 0$$

- **Cópulas mixtas.**- aquéllas con parte singular $S_C(u, v) \neq 0$ y parte absolutamente continua $A_C(u, v) \neq 0$, no constituyendo ninguna de estas partes una cópula por sí misma por no tener marginales uniformes. La presencia de una componente singular hace que estas cópulas sean de mayor utilidad cuando las marginales univariantes se suponen discretas.

Algún ejemplo es la familia de **Cuadras-Augé** (véase (A1.33)) que responde a la expresión:

$$C_\theta(u, v) = [\min\{u, v\}]^\theta * [u * v]^{1-\theta} \quad \text{con } 0 \leq \theta \leq 1 \quad (1.26)$$

Obsérvese que cuando $\theta = 0$ se tiene la cópula absolutamente continua Π mientras que cuando $\theta = 1$ se tiene la cópula singular M .

En relación a este primer criterio de clasificación en función del soporte, mencionaremos la existencia de cópulas cuyo soporte son fractales (véase [FRENERO]). Éstos, de acuerdo a la definición que da **Mandelbrot**, son conjuntos cuya dimensión topológica es inferior a su dimensión de Hausdorff.

1.5.2 Tipos de cópulas en función del conocimiento explícito de su forma

Las cópulas se pueden clasificar también en función de que su expresión responda o no a una ecuación paramétrica, pudiendo distinguir entre:

- **Cópulas paramétricas.**- Todas las cópulas que responden a una misma ecuación paramétrica definen una familia de cópulas. En ella, el parámetro (uniparamétricas) o parámetros (multiparamétricas) cuantifican de algún modo la relación de dependencia entre las variables que asocian.
- **Cópulas no paramétricas.**- De igual manera existen familias de cópulas no paramétricas que son aquellas en cuya definición no participa ningún parámetro sino que, por su estructura empírica, se ajustan de forma local a los datos. El método de construcción de cópulas que se presenta en esta tesis (véase capítulo 4) está orientado a la generación de este tipo de cópulas.

Dentro de uno y otro grupo, gozan de popularidad la clase de las **cóputas arquimedianas** caracterizada por la facilidad con que pueden ser construidas y por la gran variedad de estructuras de dependencia que permiten reproducir. Este grupo de cóputas será comentado en el siguiente apartado.

1.5.3 Tipos de cóputas en función de la relación de dependencia que son capaces de reflejar

Las cóputas también pueden ser clasificadas en función del tipo de relación que permiten reflejar. A este respecto, citamos algunas de las clases más características y frecuentemente utilizadas:

- **Cóputas de dependencia extrema.**- se caracterizan, como su propio nombre indica, por capturar una relación de dependencia extrema entre las variables: dependencia positiva perfecta (cóputa M), dependencia negativa perfecta (cóputa W) o independencia (cóputa I).
- **Cóputas elípticas.**- se definen como las cóputas asociadas a las distribuciones elípticas. Su rasgo más característico es que representan relaciones de dependencia simétricas sin importar que se analice la cola izquierda o derecha de las distribuciones implicadas. Normalmente no se utilizan para análisis financieros y de compañías aseguradoras en las que surgen asimetrías derivadas del hecho de que grandes pérdidas suelen ir acompañadas de grandes ganancias.

Los ejemplos más populares son la cóputa **gaussiana** (véase (A1.5) en el Anexo I), la **t-cóputa** o cóputa de la **t** de **Student** (véase (A1.40)) y, como caso particular de esta última, la cóputa de **Cauchy** (véase (A1.6)). La principal diferencia entre ellas radica en las posibilidades que presentan de cara a buscar asociaciones entre fenómenos extremos. A este respecto, la t-cóputa permite tratar con colas más pesadas que la gaussiana.

- **Cóputas de valor extremo.**- estas cóputas serán de gran utilidad para representar relaciones que ponen mayor énfasis entre los sucesos “cola” (extremos) de las distribuciones marginales. Como se define por ejemplo en [SEGBERS], las cóputas de valor extremo son los posibles límites (en caso de que existan) de cóputas asociadas a los máximos de muestras independientes e idénticamente distribuidas. Entendamos mejor esta definición.

Sea una muestra de variables aleatorias bidimensionales $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$ independientes e idénticamente distribuidas de acuerdo a unas mismas marginales F_X y G_Y y a una misma distribución conjunta H_{XY} que, en virtud del teorema de Sklar llevará asociada una cóputa C : $H_{XY}(x, y) = C(F_X(x), G_Y(y))$.

Sean las variables $M_n = \max\{X_1, X_2, \dots, X_n\}$ y $N_n = \max\{Y_1, Y_2, \dots, Y_n\}$ cuyas funciones de distribución vienen dadas por $F^n(x) = P[M_n \leq x]$ y $G^n(y) = P[N_n \leq y]$ y con distribución conjunta $H^n(x, y) = P[M_n \leq x, N_n \leq y]$.

Si C es también la cóputa asociada al par (M_n, N_n) y a su posible límite cuando “ n ” tiende a infinito se dice entonces que C es una cóputa de valor extremo (CVE). Como se explica en [SEGBERS], de acuerdo al teorema de **Dehuelves** una cóputa C de valor extremo queda caracterizada por la condición

$$C^t(u^{\frac{1}{t}}, v^{\frac{1}{t}}) = C(u, v) \quad \forall t > 0 \quad (1.27)$$

siendo un corolario de ésta el que las cópulas de valor extremo sólo modelizan dependencia positiva.

Además, existe un teorema propuesto por **Pickands** (1981) que permite asociar una representación asociada a este tipo de cópulas.

Teorema 1.8.- Representación de cópulas de valor extremo

“Una cópula C es una cópula de valor extremo si y sólo si existe una función real valorada A , definida sobre el intervalo $[0,1]$, que verifica la siguiente relación:

$$C(u, v) = \exp \left\{ \log(u*v) * A\left(\frac{\log(v)}{\log(u*v)}\right) \right\} \quad (1.28)$$

o equivalentemente

$$C(e^{-u}, e^{-v}) = \exp \left\{ -(u+v) * A\left(\frac{v}{u+v}\right) \right\} \quad (1.29) ”.$$

La función A recibe el nombre de **función de dependencia de Pickands** y verifica las siguientes propiedades:

1. Es convexa en $[0,1]$
2. $\max(t, 1-t) \leq A(t) \leq 1 \quad \forall t \in [0,1]$

- **Cópulas arquimedianas.-** existe una gran diversidad de familias que pertenecen a la clase arquimediana y gracias a esta variedad permiten, a diferencia de las elípticas (simétricas) y de las de valor extremo (muy orientadas a dependencias en las colas), recoger muchos tipos de estructuras de dependencia adicionales. Otra ventaja de este tipo de cópulas es la facilidad con la que pueden ser construidas, lo cual queda claramente puesto de manifiesto a partir de su definición.

Definición 1.7.- Sea Φ el conjunto de funciones $\varphi: [0,1] \rightarrow [0,\infty]$ que son continuas, estrictamente decrecientes, convexas y para los cuales $\varphi(0)=\infty$ y $\varphi(1)=0$. Schweizer y Sklar demuestran que cada miembro de Φ , genera una cópula C a través de la expresión:

$$C(u, v) = \varphi^{-1}(\varphi(u) + \varphi(v)) \quad \text{con } 0 \leq u, v \leq 1 \quad (1.30)$$

La función φ recibe el nombre de **“generador de la cópula”**.

Las cópulas arquimedianas más conocidas son de tipo paramétrico y pueden ser consultadas en [NELSEN] (uniparamétricas) y en [JOE] (biparamétricas). Sin embargo, existen también algunas aportaciones dentro del ámbito no paramétrico (véase por ejemplo [VANLAM]).

Las cópulas cuya construcción se describe en el apartado cuarto de esta tesis persiguen evitar el compromiso de decantarse por alguno de los diferentes tipos de cópulas en función de la estructura de dependencia ligada a la muestra. Se trata de emplear una clase de cópulas que, por su naturaleza empírica y no paramétrica, se ajusten bien a los datos sin necesidad de preguntarse si serán capaces de capturar ciertas asimetrías o de resaltar la relación entre los valores extremos observados en las colas de las distribuciones marginales.

De acuerdo a lo expuesto, queda patente la dificultad de establecer una clasificación única de las cópulas conforme a un criterio general dado que, incluso de acuerdo a un mismo criterio, pueden existir solapamiento entre las clases establecidas. Así, la cópula de valor extremo de **Gumbel** (véase (AI.8) o (AI.15)) pertenecen también a la clase arquimediana, si bien ninguna de las clases es subconjunto de la otra pues por ejemplo la cópula de **Husler y Reiss** (véase (AI.9)) es de valor extremo pero no arquimediana mientras que la cópula de **Frank** (véase (AI.13)) es arquimediana pero no de valor extremo).

En el Anexo I se presenta un esquema que trata de ubicar las cópulas más populares de acuerdo a una jerarquía concreta que trata de responder a los dos últimos criterios que hemos definido. El criterio de clasificación de acuerdo al soporte (singular, absolutamente continuo o mixto) se ha obviado dado que en su mayor parte (salvo las cotas de Fréchet-Hoeffding) se trata de cópulas no singulares. Así:

- El primer nivel de esta jerarquía se establece de acuerdo al número de parámetros del que depende la familia.

- El segundo criterio clasifica conforme al tipo de relación de dependencia: cópulas de dependencia extrema, cópulas elípticas, cópulas de valor extremo y la clase arquimediana que pudiera considerarse comodín en el sentido de que permite reflejar distintos tipos de relaciones.

2. Elección de la cópula que mejor refleja una relación de dependencia

2.1 Introducción al problema

Uno de los problemas clásicos en la estadística es conocer la distribución a la que responde una muestra dada de forma que ésta quede bien caracterizada y puedan extraerse conclusiones con fines descriptivos o predictivos. Dentro del contexto de la teoría de cópulas, este problema presenta una doble vertiente: una univariante asociada a la especificación de las funciones de distribución F_X y G_Y (en adelante F y G) correspondientes a las marginales de X e Y , y otra bivariante (en general multivariante) asociada a la determinación de aquella conjunta H_{XY} (en adelante H), de las infinitas que comparten dichas marginales, que mejor captura la relación entre ellas. La vertiente bivariante desemboca en la búsqueda de una función cópula C cuyas características puedan esperarse para la verdadera distribución conjunta H , siendo el teorema de Sklar el que establece la transformación final de C en H .

En ocasiones, por las características del problema que se está estudiando, se puede tener una idea preconcebida de la familia de cópulas que puede ser más apropiada para explicar la relación entre las variables que se manejan. Así por ejemplo, si el estudio está orientado a medir el grado de asociación para valores extremos de dos variables, que se intuye presenta un comportamiento especial respecto del grado de asociación que pudieran tener para otros valores no extremos, suelen ser aconsejable utilizar cópulas que enfatizan la relación entre las colas de las distribuciones marginales (cópulas del valor extremo), como por ejemplo las pertenecientes a la familia de Gumbel.

Obsérvese que venimos empleado con cierta regularidad el concepto de “familia”. Generalmente cuando se habla de los diferentes tipos de cópulas que existen, solemos referirnos intrínsecamente a diferentes tipos de familias. Todas las cópulas que pertenecen a una misma familia, presentan una misma estructura (o ecuación) que puede depender de uno o varios parámetros (o también de ninguno, si hablamos de cópulas no paramétricas), de forma que, para cada uno de los valores del espacio paramétrico de definición, se obtendrá un miembro de esa familia.

De esta forma, en un primer paso, el analista determina varias familias de cópulas que a su juicio considera candidatas a reflejar un “tipo” de relación entre las variables de estudio. Dentro de cada una de ellas, selecciona aquel miembro (normalmente dado por el valor de uno o varios parámetros) que mejor refleja una relación “concreta” (la observada en los datos). Finalmente debe decidirse por aquél representante que, en función de ciertos criterios, mejores resultados le proporcione. Podemos resumir que las etapas que encontraremos en el proceso de selección de cópulas y que a continuación pasaremos a detallar son las siguientes:

1. Determinación de las distribuciones marginales asociadas a cada una de las variables en función de las muestras de datos disponibles.
2. Propuesta de un conjunto inicial de familias de cópulas candidatas que, por sus características, se perfilan como adecuadas para reflejar la relación existente entre las variables. Esta propuesta se hará de acuerdo al conocimiento o en su defecto intuición, que se tenga sobre la forma de dicha relación.

3. Selección de una cópula por familia. En el caso paramétrico se trata de determinar los valores asociados a los parámetros correspondientes a cada familia para lo cual, se suelen utilizar expresiones que permitan el cálculo de dichos parámetros a partir de la estimación muestral de alguna medida de asociación como el coeficiente de correlación de Spearman o la Tau de Kendall.
4. Elección de la cópula de entre todas las que representan a cada una de las familias candidatas. A este respecto, existen muy diversos criterios que exponemos en el apartado 2.5.

2.2 Determinación de las distribuciones marginales

La vertiente univariante conlleva la especificación de funciones de distribución asociadas a cada una de las variables. Si bien existen los clásicos contrastes de bondad de ajuste que permiten evaluar el grado de parentesco con alguna distribución conocida, una buena aproximación podría venir dada por la versión continua de la función de distribución empírica de cada variable que, como explica **De Matteis** (véase capítulo 4 de [MATTEIS]), se calcula de la siguiente manera:

Dada la muestra $\{x_1, x_2, \dots, x_n\}$ extraída de la variable X , la función de distribución empírica (discreta) viene dada por

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n 1_{[x_i, \leq x]} \quad (2.1).$$

Consideremos entonces “ a ” y “ b ” dos números reales tales que $a \leq x_1, x_2, \dots, x_n$ y $b \geq x_1, x_2, \dots, x_n$. Ordenamos las x_i de menor a mayor y denotamos por “ z ” a las variables “ x ” ordenadas, $\{z_1, z_2, \dots, z_n\}$. Además, se definen dos puntos auxiliares $z_0 = a$ y $z_{n+1} = b$. A partir de esta nueva muestra de $n+2$ elementos se define la función de distribución empírica continua mediante rectas que unen los puntos medios de los segmentos que conforman la función de distribución empírica (discreta).

Pongamos como ejemplo la muestra $\{x_1, x_2, x_3\} = \{11, 2, 25\}$ a partir de la cual se puede calcular la función de distribución empírica discreta (en azul). Sean $a = 1$ y $b = 26$ (una unidad menos y más respectivamente que el mínimo y el máximo del conjunto de valores). Así, tenemos z_0 y z_{n+1} . Sea $\{z_1, z_2, z_3\} = \{2, 11, 25\}$ la muestra ordenada. Hallando los puntos medios entre estos valores, obtenemos $\{(2+11)/2, (11+25)/2\} = \{6.5, 18\}$ que determinarán las inflexiones de la función de distribución empírica continua (en rojo). Como inflexiones asociadas a los extremos se utilizan los puntos “ a ” y “ b ”. El siguiente gráfico ilustra el ejemplo descrito:

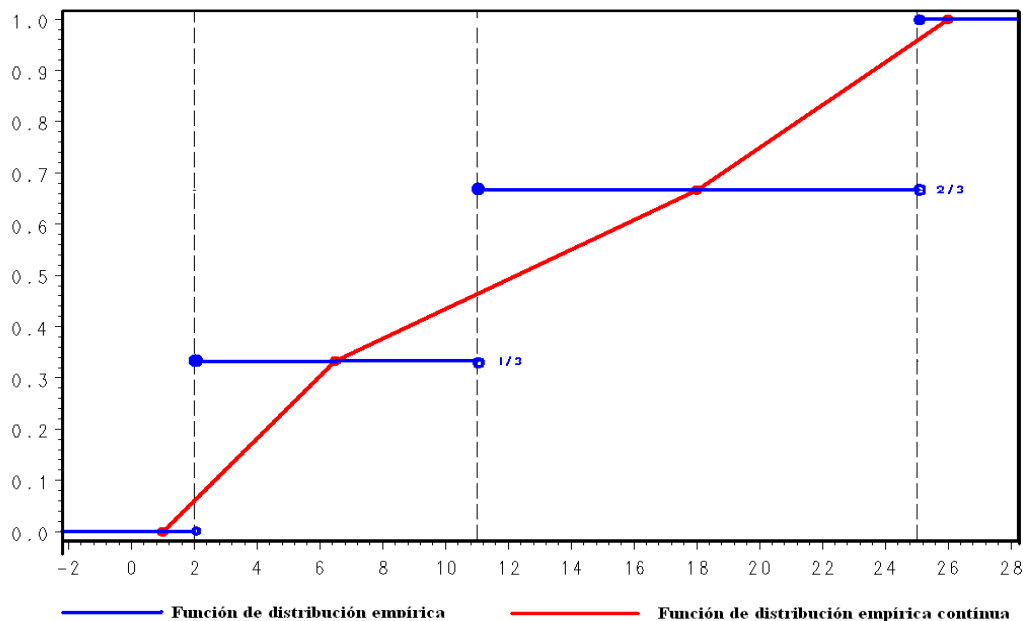


Ilustración 2.1: Funciones de distribución empírica discreta y continua asociada a una variable

2.3 Propuesta de un conjunto inicial de familias de cópulas candidatas

El experto en cópulas conoce las propiedades que caracterizan a las diferentes familias existentes y que las pueden hacer más o menos apropiadas para reflejar algún tipo de relación que, a priori, puede presuponer que exista entre las variables. Así por ejemplo, según citábamos en el apartado 1.5.3, las familias elípticas resultan más convenientes para reflejar relaciones simétricas mientras que las definidas como cópulas de valor extremo enfatizan asimetrías que ganan fuerza entre los sucesos “cola” de las distribuciones.

Debe ser el conocimiento del analista sobre la relación subyacente a los datos y el que tiene sobre las características de las familias de cópulas a su alcance, los factores principales que le lleven a descartar de antemano alguna de estas familias y seleccionar algunas otras como candidatas de partida.

2.4 Determinación de la cópula óptima dentro de una familia

Si bien como decíamos, el conocimiento del experto puede llevarle a hacerse una idea de cuál es la familia más conveniente para su problema en función del tipo de relación que sospeche existe entre sus variables, también admitimos que esta intuición puede no existir si el analista desconoce la relación que cabe esperar a priori. También puede darse el caso de que el analista encuentre varias familias que desempeñan papeles parecidos y no termine de ver claro cuál de ellas es la más apropiada. En estas ocasiones, resulta beneficioso disponer de una amplia gama de familias candidatas para poder seleccionar de cada una de ellas un representante (aquella cópula que mejor se ajustase a la muestra de datos disponible) entre los cuales acabará eligiendo uno.

En este contexto, las cópulas arquimedianas, que presentamos en el capítulo anterior, suelen resultar especialmente

útiles pues gracias a su gran diversidad permite recoger relaciones de muy distintos tipos. Como también se explica en el capítulo 4 de [MATTEIS], el procedimiento para determinar la cópula que mejor se ajusta a una muestra aleatoria bivalente de “ n ” observaciones $\{(x_1, y_1), \dots, (x_n, y_n)\}$ comienza asumiendo que ésta ha sido generada por una distribución bivalente desconocida $H(x, y)$ con marginales continuas $F(x)$ y $G(y)$ y cópula arquimediana $C(u, v)$. Hecha esta consideración, se trata de determinar a qué familia pertenece C , o lo que es lo mismo, la forma del generador φ de la cópula que, recordemos, caracteriza a la cópula arquimediana. Se podrá disponer de varias familias candidatas $\{C_\theta\}_{\theta \in \Theta}$ (con varios tipos de generadores φ_θ) y, para cada una de ellas, elegir un representante $C_{\theta_{opt}}$. Como cita De Matteis, para la estimación de θ , existen diferentes alternativas:

- Estimar en un primer paso las funciones de distribución marginales mediante métodos paramétricos o no paramétricos y, posteriormente, a partir de ellas, estimar θ mediante el principio de máxima verosimilitud.
- También es posible hacer la estimación de las marginales y del parámetro θ en un sólo paso. En este caso, la estimación de θ puede hacerse de dos formas:
 - Empleando un método paramétrico como es el procedimiento de estimación de máxima verosimilitud, siendo la función de verosimilitud $L(\alpha, \theta, X, Y)$ donde α identifica a los parámetros de las marginales.
 - Empleando un método no paramétrico, recomendado por **Genest y Rivest** (1993) donde θ es estimado en un sólo paso, con independencia de las funciones de distribución marginales. La estimación se hace empleando la correlación rango de Kendall.

Procediendo de la forma que describe [MATTEIS] para cada una de las familias candidatas, el resultado es un conjunto de cópulas (una por familia) entre las que será necesario hacer la selección final.

2.5 Selección de la mejor familia a partir de las cópulas representantes

En este apartado se proponen algunos de los criterios más utilizados para decantarse por la selección de una determinada cópula dentro de un conjunto de candidatas. Dicho conjunto puede haber sido el resultado de alguno de los procesos descritos en el apartado anterior. La decisión obviamente lleva implícita el optar por una determinada familia.

2.5.1 Método 1: Empleo de la cópula empírica

Las **cópulas empíricas** fueron estudiadas originalmente por Deheuvels (1979). La idea consiste en construir una función cópula a partir de valores muestrales $\{(x_1, y_1), \dots, (x_n, y_n)\}$ recogidos para las variables univariantes sin establecer dependencia de ningún parámetro. De esta forma, la cópula es no paramétrica y queda definida únicamente a partir de la muestra de datos disponible.

La definición de la cópula empírica puede encontrarse en [NELSEN] (definición 5.5.1) y responde a la expresión:

$$C_n\left(\frac{i}{n}, \frac{j}{n}\right) = \frac{(n^\circ \text{ de pares } (x, y) \text{ en la muestra tales que } x \leq x_{(i)} \text{ e } y \leq y_{(j)})}{n} \quad (2.2)$$

para $\{(x_i, y_i), \dots, (x_n, y_n)\}$ muestra de una distribución bivalente conjunta, y siendo $x_{(i)}$ e $y_{(j)}$ con $1 \leq i, j \leq n$ los estadísticos de orden definidos a partir de dicha muestra.

Este primer método, citado por **Claudio Romano** (véase el apartado 4.2 de [ROMANO]), plantea la selección de una cópula dentro de una familia de candidatas $\{C_k\}_{1 \leq k \leq K}$ eligiéndose aquélla que minimiza la distancia a la empírica. La medida de distancia que se propone está basada en la norma discreta L^n , siendo “ n ”, el número de variables analizadas (en el caso bivalente, $n = 2$):

$$\bar{d}_n(\hat{C}, C_k) = \left(\sum_{t_1=1}^T \dots \sum_{t_n=1}^T \left(\hat{C}\left(\frac{t_1}{T}, \dots, \frac{t_n}{T}\right) - C_k\left(\frac{t_1}{T}, \dots, \frac{t_n}{T}\right) \right)^2 \right)^{1/2} \quad (2.3)$$

Si bien estamos partiendo de que ya disponemos de una cópula concreta dentro de cada una de las familias y esta distancia nos va a ayudar a seleccionar una de ellas, también es posible aplicar esta medida a todas las cópulas de una misma familia para determinar el valor del parámetro más conveniente. Es decir, podríamos prescindir del paso previo de estimación que hemos comentado (la solución paramétrica o la no paramétrica de Genest y Rivest) y hacer la selección de la cópula dentro de cada una de las familias valiéndonos de \bar{d}_n . Como se explica en [ROMANO], el vector de parámetros $\theta \in \Theta$ (o el parámetro si hablamos de cópulas uniparamétricas) se puede estimar de la siguiente forma:

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}} \left(\sum_{u \in I} [\hat{C}(u, v) - C(u, v; \theta)]^2 \right)^{1/2} \quad (2.4)$$

2.5.2 Método 2: Métodos gráficos

Existen diferentes métodos gráficos que permiten un contraste visual del ajuste de una cópula a los datos. Aquí se proponen dos basados en el empleo de *QQ-plots* que han sido extraídos del apartado 4.3 de [MATTEIS].

- Gráfico basado en la función de distribución condicional $Y|X$. - Para ello, basta observar que la cópula condicionada $C_1(F(x), G(y)) = H_{Y|X}(x, y)$ se debería distribuir teóricamente según una $U(0,1)$. Así, mediante un gráfico *QQ-plot* se puede establecer el contraste observando si el resultado se aproxima a la recta $y = x$. Se trataría de ver para cuál de todas las cópulas candidatas se obtiene una mejor aproximación.
- Gráfico basado en la función de distribución de la cópula.- Se define $K_c(t)$ como la **C-medida** del conjunto $\{(u, v) \in [0,1] \times [0,1] : C(u, v) \leq t\}$. Se puede demostrar que si C es una cópula arquimediana generada por

$$\text{una función } \varphi, \text{ entonces } K_c(t) \text{ se puede escribir como } K_c(t) = t - \frac{\varphi(t)}{\varphi(t^{\text{derecha}})} \quad (2.5)$$

siendo dicha función la función de distribución de la variable aleatoria $C(u, v)$.

Por tanto, de igual modo que si X tiene función de distribución F entonces $F(X) \sim U(0,1)$, podemos concluir que si C está suficientemente bien aproximada a los datos cabe esperar que

$K_c(C(F(X), G(Y))) \sim U(0,1)$. Valdría nuevamente un gráfico *QQ-plot* asociado a las funciones de distribución “*K*” correspondientes a cada una de las cópulas candidatas para terminar decantándose por aquella que más se aproxime a la recta $y = x$.

2.5.3 Método 3: Aproximación analítica de los métodos gráficos

Si bien los métodos gráficos anteriores pueden proporcionar una idea bastante buena de cuál es la cópula más apropiada, se puede eliminar la subjetividad asociada a la agudeza visual del analista y plantear un test de hipótesis para contrastar si las distribuciones de $C_l(F(x), G(y))$ o $K_c(t)$ se aproximan a una $U(0,1)$.

A este respecto, en [MATTEIS] (apartados 4.4.1 y 4.4.2) se citan los dos contrastes clásicos de bondad de ajuste a una distribución dada, el de la **Chi-cuadrado** y el de **Kolmogorov-Smirnov**. Para llevar a cabo estos contrastes se trocea el rango de variación de la distribución a contrastar en una serie de intervalos y se comprueba si el número de valores muestrales observados en cada una de ellos (O_i) se parece al número de ellos que cabría esperar (E_i) bajo el supuesto de que siguieran una distribución $U(0,1)$. Para una descripción más detallada de estos contrastes se puede consultar el Anexo VIII.

– Contraste de la Chi-cuadrado (basado en el estadístico de Pearson)

La muestra de partida es el conjunto de valores $C_l(F(x_i), G(y_i))$ (o $K_c(t_i)$) que nos proporcionará las frecuencias observadas (O_i) dadas por el número de pares $(F(x_i), G(y_i))$ que caen en cada uno de los intervalos. El objetivo es contrastar si éstas frecuencias se aproximan a las esperadas (E_i) para una distribución uniforme estándar. Aquella cópula C cuya condicionada C_l dé un mayor grado de proximidad entre estas frecuencias (menor valor del estadístico de Pearson) será la propuesta para representar la relación entre X e Y .

– Contraste de Kolmogorov-Smirnov (basado en el estadístico D_n)

En este caso, $C_l(F(x_i), G(y_i))$ (o $K_c(t_i)$) proporcionará la muestra de valores a partir de la cual se construirá la función de distribución empírica F_n . El objetivo es ver si ésta se parece a la función de distribución de una $U(0,1)$. Así, la cópula C cuya condicionada C_l nos proporcione un valor del estadístico D_n más pequeño será aquella para la que la muestra de valores se aproxime más a una distribución uniforme estándar y por tanto, la más apropiada para representar la relación entre X e Y .

2.5.4 Método 4: Criterio de información de Akaike (AIC)

Otra posibilidad que se plantea es evaluar la bondad de la estimación de los parámetros θ realizada en la etapa previa que, recordemos, podía hacerse en dos pasos mediante el método de la máxima verosimilitud o en un único paso empleando la correlación rango tal y como explican Genest y Rivest. La medida que se propone en [MATTEIS] (apartado 4.4.3) es el criterio de información de **Akaike** que viene dado por

$$AIC = 2 * (-\log L) + 2 * \rho \quad (2.6)$$

donde L es la función de verosimilitud de la muestra que se utiliza para la estimación del parámetro θ y ρ es el número

de parámetros que se estima.

Utilizar el criterio de Akaike supone decantarse por aquel “modelo” que presenta un menor valor del AIC lo cual es representativo de un modelo más parsimonioso (menos parámetros) y que mejor se ajusta a los datos (menor error de predicción). Dado que L representa la probabilidad de que la muestra quede bien representada por los parámetros, interesa que L sea próxima a 1 lo cual se traducirá en que el logaritmo de L se aproxime a 0. Es por ello que un AIC pequeño es indicativo de que el estimador representa fielmente a la muestra a la vez que castiga el posible sobreajuste derivado del empleo de un excesivo número de parámetros. Con más parámetros es más fácil conseguir un buen ajuste a la muestra de referencia pero los resultados del mismo no suelen adaptarse bien a otras muestras distintas (el “modelo” no es bueno, sus conclusiones no son extrapolables).

En nuestro caso, el método consiste en seleccionar aquella cópula C que proporcione un menor valor del AIC que probablemente será independiente de ρ si todas las familias de cópulas candidatas son uniparamétricas (pues en dicho caso ρ vale siempre 1).

2.5.5 Método 5: Contrastes de bondad de ajuste de una cópula

Según se explica en [DOSCH] (véase su capítulo introductorio), existen varios autores como **Breymann, Malevergne y Sornette, Mashal, Zeevi y Fermanian o Savu y Trede** que han propuesto diferentes contrastes de bondad de ajuste de una cópula. El método que a continuación citamos es el que se propone en dicho artículo. Este método consiste en considerar $C(u,v)$ como una cópula desconocida asociada a la variable aleatoria bidimensional (X,Y) y contrastar si dicha cópula pertenece a una familia paramétrica conocida $C(u,v; \theta)$ con $\theta \in \Theta \subset \mathbb{R}^d$.

Es decir, la hipótesis nula asociada al contraste sería $H_0: C(u, v) = C(u, v; \theta)$ para algún $\theta \in \Theta$

Se puede utilizar como valor para θ aquél que se estimara para dicha familia mediante el método de la máxima verosimilitud o a través de la correlación rango de Kendall, θ_{OPT} .

Se trata de un contraste de bondad de ajuste pero, a diferencia del que se planteara en la aproximación analítica del método descrito en el apartado 2.5.3, se trata de un test cuyo estadístico irá asociado a una muestra bidimensional y en consecuencia, las clases de las que depende no serán intervalos del eje real sino rectángulos del plano real (concretamente del rectángulo unidad). Por ello, nuevamente instamos a la consulta del Anexo VIII para un mejor entendimiento del desarrollo que a continuación exponemos.

En primer lugar, dividiremos el intervalo $[0,1]$ del eje X en “ r ” subintervalos de igual longitud y se procede de manera idéntica con el intervalo $[0,1]$ del eje Y , si bien el número de clases puede ser distinto, “ s ”. Como resultado de esta partición, el rectángulo $[0,1]^2$ quedaría dividido en un total de $r \times s$ rectángulos del mismo área B_{ij} .

Para construir el estadístico del contraste será necesario disponer del número de puntos muestrales que caen en cada uno de los rectángulos y de la frecuencia teórica esperada para cada uno de ellos a través de la cópula que se contrasta.

Así, dada una cópula C perteneciente a la familia $C(u,v; \theta)$ para la que ya ha sido estimado su parámetro θ_{OPT} la frecuencia teórica asociada al rectángulo B_{ij} es $n \cdot p_{ij}(\theta_{OPT})$, donde

$$p_{ij}(\theta_{OPT}) = P((U, V) \in B_{ij} | \theta_{OPT}) = \iint_{B_{ij}} dC(u, v; \theta_{OPT}) \quad (2.7)$$

Por otro lado, para calcular el número de puntos muestrales que caen en cada uno de los rectángulos, N_{ij} (este número es

independiente de la familia de cópulas considerada), se deben distinguir dos casos en función del conocimiento que se tenga de las distribuciones marginales asociadas a X e Y :

- Si se supone que F y G son conocidas, basta con hallar la imagen a través de F y G de la muestra bidimensional, es decir, $(u_i, v_i) = (F(x_i), G(y_i))$ y calcular el número de puntos transformados que caen en cada rectángulo.
- Por el contrario si las distribuciones de F y G son desconocidas, la forma de calcular (u_i, v_i) es a partir de las funciones de distribución empíricas asociadas a X e Y , es decir, haciendo

$$\hat{u}_i = \hat{F}_n(x_i) = \frac{1}{n} \sum_{i=1}^n 1_{[x_i \leq x]} \quad \text{y} \quad \hat{v}_i = \hat{G}_n(y_i) = \frac{1}{n} \sum_{i=1}^n 1_{[y_i \leq y]} \quad .$$

$$\text{Lógicamente,} \quad \hat{u}_i = \frac{\text{rango de } x_i \text{ en } x_1, x_2, \dots, x_n}{n} \quad \text{y} \quad \hat{v}_i = \frac{\text{rango de } y_i \text{ en } y_1, y_2, \dots, y_n}{n} \quad .$$

A partir de todos estos datos, se puede plantear nuevamente como estadístico del contraste, el estadístico de Pearson

$$\sum_{i=1}^r \sum_{j=1}^s \frac{(N_{ij} - n * p_{ij}(\theta_{OPT}))^2}{n * p_{ij}(\theta_{OPT})} \quad (2.8)$$

o bien el del cociente de verosimilitudes

$$\sum_{i=1}^r \sum_{j=1}^s (N_{ij} * \ln(p_{ij}(\theta_{OPT}))) \quad (2.9)$$

En el primer caso (F y G conocidas), estos estadísticos se distribuyen según un Chi-cuadrado con $r * s - 1 - d$ grados de libertad (donde d normalmente valdrá 1 pues consideramos cópulas uniparamétricas); en el segundo caso (F y G desconocidas), el estadístico se distribuye según una Chi-cuadrado con $(r - 1) * (s - 1) - d$ grados de libertad.

En cualquiera de ellos, se calculará uno de los dos estadísticos para cada una de las cópulas candidatas y se seleccionará de entre todas ellas, aquella C para la que se obtenga un menor valor del estadístico del contraste dado que será la que refleje mayor proximidad entre el volumen empírico de cada uno de los rectángulos y el volumen esperado para ellos a través de la cópula.

2.5.6 Método 6: Calidad de las predicciones que proporciona una cópula

Aun cuando todos los métodos descritos puedan ser válidos diremos que, por lo general, la selección de una u otra cópula estará sujeta a la finalidad que persiga el estudio que se está llevando a cabo. En nuestro caso, en el que la idea consiste en utilizar estas funciones para realizar predicciones, parece conveniente decidirse por aquélla que mejores resultados proporcione, resultados que se medirán en términos de error. Supongamos que estamos interesados en predecir el comportamiento de una variable Y en función de los valores conocidos de una variable X . Para ello, disponemos de un histórico dado por una muestra bidimensional que relaciona ambas variables $\{(x_1, y_1), \dots, (x_n, y_n)\}$. Supongamos que $H_{Y|X}$ es la función de distribución de la variable condicionada (desconocida). Sabemos que ésta se encuentra relacionada con las marginales de X e Y a través de la derivada de una cópula respecto de la primera de las variables (U), C_1 , mediante la expresión (1.11). La obtención de predicciones de Y a partir de X se realizará mediante la

simulación de valores de la función $H_{Y|X}$. Así, nuestro estudio predictivo se puede plantear en una relación de etapas:

- I. **Determinación de un conjunto de entrenamiento y otro de validación.-** Reservamos parte del histórico para validar la calidad de las predicciones que vamos a realizar. Dicha parte recibirá el nombre de conjunto de validación y no participará en el ajuste de las distribuciones marginales ni en el de la selección de la función cópula que mejor representa a una de las familias candidatas sino que será utilizado con posterioridad para la evaluación de las mismas. Por lo general, se reserva una cuarta o quinta parte de los datos disponibles para validar los resultados. El resto del histórico se utilizará para realizar los ajustes y recibe el nombre de conjunto de entrenamiento.

- II. **Determinación de marginales.-** Tomaremos como marginales las funciones de distribución empíricas continuas asociadas a X e Y y las aplicaremos a los valores muestrales (de entrenamiento) para obtener la muestra transformada $\{(F_X(x_1), F_Y(y_1)), \dots, (F_X(x_n), F_Y(y_n))\}$. Ésta será una muestra bidimensional de variables uniformes estándar $\{(u_1, v_1), \dots, (u_n, v_n)\}$, siempre y cuando las aproximaciones dadas por estas marginales sean suficientemente buenas. Por supuesto, también pueden proponerse como marginales algunas de las distribuciones univariantes conocidas (Normal, Exponencial, Pareto, etc.), si bien conviene previamente plantear el correspondiente contraste que confirme la hipótesis lanzada. En cualquier caso, la cópula C y de igual manera su condicionada respecto de U , C_U , actuarán sobre esta muestra de uniformes, lo cual permite intuir la transformación del problema de predicción original de Y en función de X , en un problema de predicción de V en función de U . Efectivamente, el planteamiento original busca saber cuál es el valor “ y ” asociado a un valor “ x ” conocido. El hecho de que para este último podamos calcular la imagen $F_X(x) = “u”$ y de que además dispongamos de una distribución conjunta C que presupone cierta relación de dependencia sobre pares del tipo (u, v) , nos permite proponer como alternativa la predicción del valor de “ v ” = $G_Y(y)$ (en lugar de “ y ”). Una vez realizada la predicción de la transformada $V=G(Y)$ se puede deshacer la transformación (mediante G^{-1}) para obtener la predicción de la variable Y .

- III. **Selección preliminar de cópulas.-** Se consideran una relación de familias de cópulas candidatas y dentro de cada una de ellas se escoge un representante. En el caso de las cópulas uniparamétricas esto se traduce en determinar el valor del parámetro θ a partir del valor muestral de la Tau de Kendall o del coeficiente de Spearman desde el cual se define. Este valor muestral se debe estimar también a partir del conjunto de entrenamiento.

- IV. **Simular valores para la cópula condicionada asociada a cada uno de los representantes.-** Hemos dicho que las predicciones se obtendrán por simulaciones de $H_{Y|X}$. Dado que estamos trabajando con la muestra transformada, parece lógico que las simulaciones las realicemos para la cópula condicionada $C_U = C_{V|U}$. La predicción de V condicionada a $U = u$ se realizará mediante simulaciones de la variable $V|U=u$. Para realizar dichas simulaciones se puede utilizar el método de la transformada inversa, siendo necesario para ello disponer de la función de distribución de dicha variable, C_U . Así, nuestra relación de cópulas C candidatas a

explicar la relación entre X e Y , proporciona a través de la ecuación (1.9) otra relación de cópulas condicionadas $C_1=C_{V|U}$ candidatas a ser la mejor representación de la función de distribución $H_{Y|X}$. En el apartado 5 se da un mayor nivel de detalle de este método de la transformada inversa y se presentan además expresiones aportadas por distintos autores que permiten llevar a cabo la simulación para diferentes familias de cópulas. Cualquiera de estas expresiones permite, para un “ u ” conocido (“ x ” conocido), simular valores de la variable V a través de la expresión correspondiente. Para cada uno de los valores “ v ” simulados se puede calcular su contraimagen a través de la propia marginal de Y (por ejemplo su función de distribución empírica continua) y obtener de esta forma valores simulados de la variable de interés Y supuesto un valor de X . Es decir, se utiliza la cópula para simular valores de la distribución condicionada de $H_{Y|X=x}$ donde “ x ” es conocido. El resultado final de esta etapa es una relación de valores simulados para la variable Y para cada uno de los valores propuestos para la variable X y cada una de las cópulas que representan a las familias candidatas.

- V. **Propuesta de un valor predicho.-** El paso anterior proporciona una distribución de valores simulados de Y a partir de un “ x ” y una cópula C concreta. El valor que se puede esperar para Y podrá venir dado por la media de las simulaciones o, en su defecto, por alguna otro parámetro de tendencia central más robusto como por ejemplo la mediana. Cualquiera de ellas se puede utilizar como valor predicho. Además, el conocimiento de la distribución permite también proporcionar alguna medida de dispersión asociada a la predicción como la varianza, la desviación media absoluta o el rango intercuartílico que podría traducirse en una evaluación del riesgo de la misma. De hecho, el alcance va aún más allá puesto que proporciona conocimiento sobre posibles asimetrías, densidad en las colas, apuntamiento, presencia de varias “modas”, etc. La simulación de valores, y por tanto la obtención de las predicciones, se pueden realizar a partir de cualquier “ x ” tanto de la muestra de entrenamiento como de la de validación. Comparando el valor predicho con el real es posible hablar de un error de predicción para cada familia de cópulas que se considerará asociado a la muestra en el primer caso (*in-sampling*) y fuera de muestra en el segundo (*out-of-sampling*) siendo este último por lo general de mayor magnitud dado que los datos de dicho conjunto no participan directamente en el ajuste. Aquella familia de cópulas que a través de su representante proporcione en media (o en mediana) menores errores (principalmente a futuro) será la que se seleccione finalmente para los fines predictivos que se persiguen.

3. Metodologías para la construcción de funciones cópula

A lo largo del tiempo se han propuesto muy diversas técnicas para construir funciones cópula. El hecho de que éstas puedan proporcionar distribuciones multivariantes con unas marginales concretas (según demuestra el teorema de Sklar) ha llevado a los distintos autores a la búsqueda de métodos que permitan generar funciones de este tipo intentando que además posean ciertas características orientadas a identificar algún tipo de relación de dependencia. Aquí presentamos un esquema paralelo, aunque más resumido, al que expone Nelsen (véase capítulo 3 de [NELSEN]) y se citan algunos otros propuestos por otros autores. Además del libro citado, referenciamos otros artículos través de las cuales el lector podrá entrar en un mayor nivel de detalle.

3.1 Método de inversión

Como corolario directo del teorema de Sklar, encontramos una forma directa de construir cópulas a partir de funciones de distribución conocidas. El resultado, que podría considerarse como una versión recíproca de dicho teorema, establece lo siguiente:

Corolario 3.1

“ Si H es función de distribución bivalente con marginales continuas y estrictamente crecientes F y G entonces $C(u, v) = H(F^{(-1)}(u), G^{(-1)}(v))$ define una función cópula”.

Nelsen presenta un resultado aún más general (véase el apartado 3.1 de [NELSEN]) en función de una subcópula C' y las cuasi-inversas de F y G según el cual no es necesario imponer que las distribuciones marginales sean estrictamente crecientes (de serlo, tienen una única cuasi-inversa que es precisamente la inversa ordinaria) ni continuas (condición impuesta para que el resultado se dé no sólo para subcópulas sino también para cópulas). La técnica recibe el nombre de **“método de inversión”** pues permite obtener funciones cópula a partir de las inversas de funciones de distribución. A modo de curiosidad, se puede observar que la función cópula construida C puede servir para, dadas otras dos funciones de distribución univariantes F' y G' (distintas de F y G), obtener otra conjunta H' (distinta de H) con marginales F' y G' . Efectivamente, en virtud del teorema de Sklar, $H'(x, y) = C(F'(x), G'(y))$ es función de distribución bivalente con marginales F' y G' . Esto justifica el hecho de que haya infinitas distribuciones conjuntas con marginales comunes.

3.2 Métodos geométricos

Existe otro tipo de métodos que aprovechan rasgos característicos de la geometría de las cópulas tales como su soporte o sus secciones y vuelcan sus esfuerzos en determinar las restricciones a imponer sobre ciertas funciones para que cumplan las tres propiedades que identifican a las cópulas ((1.5)-(1.7)). En [NELSEN] (véase apartado 3.2) se citan los siguientes:

3.2.1 Cópulas singulares con soporte conocido

Este método se basa en que se tiene algún conocimiento del soporte de una cópula C cuya definición suele depender de un parámetro θ . Se utiliza para construir cópulas singulares cuyo soporte (de medida Lebesgue nula) yace en un conjunto dado. Este conocimiento permite plantear un número finito de expresiones a las que responderán los C -volúmenes de los distintos rectángulos contenidos en el cuadrado unidad en función de la intersección de éstos con el soporte conocido. El principal objetivo es buscar el rango de valores de θ que permiten garantizar que todos esos volúmenes son mayores o iguales que 0 (véase expresión (1.7)).

3.2.2 Cópulas construidas como Suma Ordinal de cópulas

Este método permite obtener una cópula C como combinación de otras C_i . El soporte de C es el resultado de escalar el soporte de cada C_i a un cuadrado $J_i^2 = [a_i, b_i] \times [a_i, b_i] \subseteq I^2$ es decir, el soporte de C es una combinación de los soportes de una sucesión de cópulas. Dado que cada uno de éstos se define sobre el cuadrado unidad es necesario hacer la correspondiente conversión (escalado) al cuadrado J_i^2 al que se quiere ajustar. Formalmente, la definición es la siguiente:

Definición 3.1.- Sea $\{J_i\}_{i \in I}$ una partición contable de I , es decir, un conjunto de intervalos cerrados $[a_i, b_i]$ y disjuntos cuya unión es el cuadrado unidad. Asociada a ella, sea $\{C_i\}_{i \in I}$ un conjunto contable de cópulas. Se define la **suma ordinal** de $\{C_i\}_{i \in I}$ respecto de $\{J_i\}_{i \in I}$ como la cópula C dada por

$$C(u, v) = \begin{cases} a_i + (b_i - a_i) * C_i\left(\frac{u - a_i}{b_i - a_i}, \frac{v - a_i}{b_i - a_i}\right) & \text{si } (u, v) \in J_i^2 \\ M(u, v) = \min(u, v) & \text{si } (u, v) \notin J_i^2 \end{cases} \quad (3.1)$$

3.2.3 Cópulas construidas a partir de “Cambios” de M (cota superior de Fréchet-Hoeffding)

Este método comparte cierta similitud con el anterior. Consiste en generar una cópula C cuyo soporte se construye escalando el soporte de la cota superior de Fréchet-Hoeffding M o alguna permutación del mismo (soporte de la cota

inferior, W) a una relación de rectángulos definidos sobre el cuadrado unidad. A grandes rasgos la técnica queda definida en cuatro pasos:

1. Considerar el soporte de la cópula M o cota superior de Fréchet-Hoeffding (1.23) (véase ilustración (1.1)).
2. Realizar una partición del cuadrado unidad en una relación de cuadrados. La partición se denomina regular si todos los cuadrados son del mismo área. Los cuadrados que componen la diagonal principal de I^2 contienen el soporte de M .
3. El trozo de soporte contenido en cada cuadrado de la diagonal principal y que tiene pendiente positiva (como M) puede cambiar de orientación (a pendiente negativa). Hacer cuantos cambios se desee. El resultado es una cópula nueva (salvo que no “gire” ningún trozo) cuyo soporte sigue contenido en los cuadrados de la diagonal principal.
4. Finalmente, cada trozo de soporte contenido en los cuadrados de la diagonal principal (con pendiente positiva o negativa) puede “trasladarse” a otro cuadrado de la misma fila (o columna) respecto de la partición siempre y cuando no existan dos trozos del soporte original de M en una misma fila o columna

Los siguientes gráficos ilustran bastante bien este método.

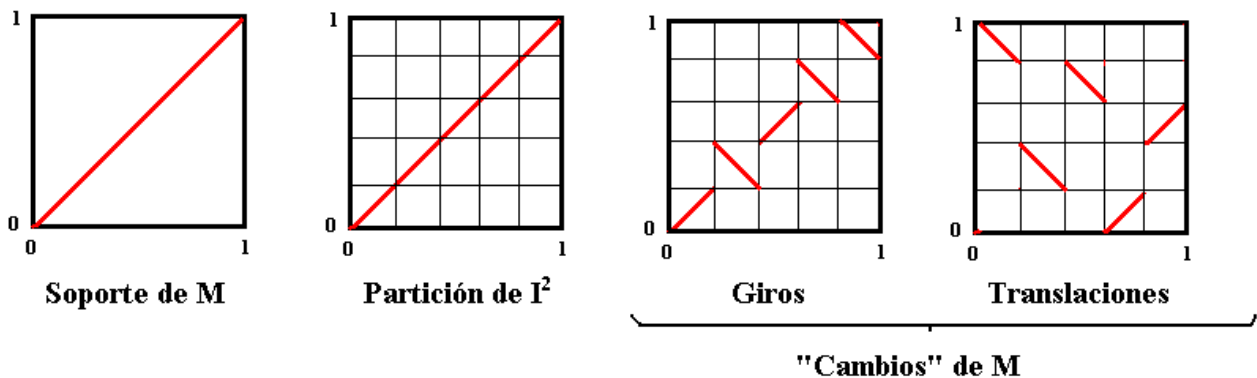


Ilustración 3.1: Método de construcción de cópulas mediante “cambios” de M

Definición 3.2.- Un “cambio” de M viene determinado por una 4-upla $M(n, \{J_i\}, \pi, \omega)$ donde “ n ” es el número de trozos en que se divide cada intervalo $[0,1]$, $\{J_i\}$ define los rectángulos que resultan de dicha partición, ω es un vector cuyos elementos son 1 ó -1 en función de la pendiente de cada trozo, es decir, dependiendo de si gira o no cada trozo de M y π es una permutación de $S_n = \{1, 2, \dots, n\}$.

En los gráficos anteriores, $n = 5$, $\{J_i\} = [0, \frac{1}{5}], [\frac{1}{5}, \frac{2}{5}], \dots, [\frac{4}{5}, 1]$, $\omega = (+1, -1, +1, -1, -1)$ (los trozos 1 y 3 no “giran”) y $\pi = (4, 2, 5, 3, 1)$ (el trozo uno se “traslada” de la columna primera a la cuarta y el dos no permuta, no se “traslada”).

Como resultado importante, Nelsen observa (teorema 3.2.1 de [NELSEN]) que toda cópula C se puede aproximar mediante ciertos “cambios” de M .

3.2.4 Cópulas construidas como “Suma Convexa” de cópulas

Este método de construcción se basa en un resultado que establece que cualquier combinación lineal convexa de un conjunto finito de cópulas es también una cópula (ejercicio 2.3 propuesto en [NELSEN]). Es decir, dadas las cópulas C_0 y C_1 y dado $\theta \in I = [0,1]$, $C(u, v) = (1-\theta) * C_0(u, v) + \theta * C_1(u, v)$ es también una cópula. De igual manera se puede demostrar (ejercicio (3.17), también propuesto en [NELSEN]) que se pueden construir cópulas como combinación de un conjunto infinito de ellas o lo que es lo mismo, como una suma convexa de cópulas $\{C_\theta\}_{\theta \in \Theta}$ donde θ es una observación de una variable aleatoria continua Θ con función de distribución Λ .

Definición 3.3.- Se define la **suma convexa** de una sucesión de cópulas $\{C_\theta\}_{\theta \in \Theta}$ con respecto a Λ , función que recibe el nombre de “**función de combinación**”, como
$$\int_{\mathbb{R}} C_\theta(u, v) d\Lambda(\theta) \quad (3.2).$$

Como hemos señalado, se puede demostrar que la expresión anterior define una cópula. Veremos más adelante que las cópulas arquimedianas se definen a partir de la transformada de **Laplace** de la función de combinación.

3.2.5 Métodos basados en el conocimiento de las secciones de una cópula

Estos métodos permiten construir cópulas cuyas “secciones” horizontal y vertical responden a una expresión conocida. Estas secciones, como su propio nombre indica, se obtienen al hacer una sección (o corte) respecto de un valor “ u ” (o “ v ”) y permitir que sea la otra variable V (o U) la que varíe; son por tanto funciones en una única variable.

Definición 3.4.- Sea C una cópula y $a \in I$ un número cualquiera.

– Se define la **sección horizontal** de C en “ a ” como la función $S_{ah}(t) = C(t, a) \quad \forall t \in I \quad (3.3)$

– Se define la **sección vertical** de C en “ a ” como la función $S_{av}(t) = C(a, t) \quad \forall t \in I \quad (3.4)$

– Se define la **sección diagonal** de C como la función $\delta_C(t) = C(t, t) \quad \forall t \in I \quad (3.5)$

Una cópula cuya expresión responde a un polinomio de grado “ n ” en “ u ” presentará la forma $u \rightarrow C(u, v) = a_n(v) * u^n + a_{n-1}(v) * u^{n-1} + \dots + a_1(v) * u + a_0(v)$ y, equivalentemente en “ v ”, será del tipo

$v \rightarrow C(u, v) = b_n(u) * v^n + b_{n-1}(u) * v^{n-1} + \dots + b_1(u) * v + b_0(u)$ que vienen a ser, respectivamente, las secciones horizontal y vertical de una cópula.

El método consiste en suponer un conocimiento sobre las secciones de una cópula y a partir de él, preguntarse por las condiciones que deben satisfacer los polinomios $a_i(v) \forall i \in \{0, 1, \dots, n\}$ y $b_i(u) \forall i \in \{0, 1, \dots, n\}$ para que la función $C(u, v)$ sea efectivamente una cópula.

En el capítulo siguiente veremos cómo obtener funciones cópula por interpolación polinómica (lineal y cúbica) de una subcópula para la que el valor del estadístico de Pearson es óptimo. No resulta por eso casual que algunas de las restricciones a imponer sobre estas funciones interpoladoras para que sean cópulas (por ejemplo en el caso de los polinomios cúbicos de Hermite) guarden un claro paralelismo con las citadas en [NELSEN] (véase apartado 3.2.5) para cópulas con secciones lineal, cuadrática o cúbica (a este último respecto, véase [NELQUERO]).

Respecto de la sección diagonal de una cópula, comentaremos únicamente un resultado presentado en [NELSEN] (teorema 3.2.11), según el cual se pueden construir también cópulas a partir del conocimiento de su sección diagonal δ .

Dicho resultado establece que, si δ es una función diagonal, esto es, una función

$$\delta: I \rightarrow I \text{ t.q. } \delta(1)=1; 0 \leq \delta(t_2) - \delta(t_1) \leq 2(t_2 - t_1) \forall t_1, t_2 \in I, t_1 \leq t_2; \delta(t) \leq t \forall t \in I$$

entonces $C(u, v) = \min(u, v, \frac{1}{2} * [\delta(u) + \delta(v)])$ define una cópula con sección diagonal δ .

3.3 Métodos algebraicos

Los métodos algebraicos basan la generación de funciones cópula en una relación algebraica planteada entre la función de distribución conjunta y sus marginales univariantes.

A modo ilustrativo encontramos en [NELSEN] (expresión (3.3.1) del apartado 3.3.1) la relación algebraica

$$\theta = \frac{H(x, y) * (1 - F(x) - G(y) + H(x, y))}{(F(x) - H(x, y)) * (G(y) - H(x, y))} \quad (3.6).$$

Dicha ecuación es el cociente entre el producto de los elementos de la diagonal principal y el producto de los situados en la diagonal secundaria de la siguiente tabla de contingencia.

	$X \leq x$	$X > x$	
$Y \leq y$	Nº pares (x_i, y_i) t.q $x_i \leq x, y_i \leq y$ $H(x, y)$	Nº pares (x_i, y_i) t.q $x_i > x, y_i \leq y$ $G(y) - H(x, y)$	$G(y)$
$Y > y$	Nº pares (x_i, y_i) t.q $x_i \leq x, y_i > y$ $F(x) - H(x, y)$	Nº pares (x_i, y_i) t.q $x_i > x, y_i > y$ $1 - F(x) - G(y) + H(x, y)$	$1 - G(y)$
	$F(x)$	$1 - F(x)$	

Realmente proporciona una medida de la asociación entre las variables X e Y pues por ejemplo, si $\theta = 1$, entonces $H(x, y) = F(x) * G(y)$ y en consecuencia las variables son independientes.

Además, haciendo las transformaciones $u = F(x)$ y $v = G(y)$ y teniendo en cuenta el teorema de Sklar es posible expresar

$$\theta = \frac{C(u, v) * (1 - u - v + C(u, v))}{(u - C(u, v)) * (v - C(u, v))} \quad (3.7).$$

La búsqueda de una solución en C permite obtener una familia de cópulas uniparamétrica que es la familia de **Plackett** (véase (AI.37)).

3.4 Método de construcción de cópulas arquimedianas

Ya comentamos que una de las principales ventajas de las cópulas arquimedianas es la facilidad con la que pueden ser construidas. Atendiendo a la definición (1.7), bastaría con encontrar una función $\varphi: [0, 1] \rightarrow [0, \infty]$ (denominada generador) continua, estrictamente decreciente, convexa y tal que $\varphi(0) = \infty$ y $\varphi(1) = 0$. A partir de ella, según demuestran Schweizer y Sklar es posible generar una cópula a partir de la relación $C(u, v) = \varphi^{-1}(\varphi(u) + \varphi(v))$. La obtención de generadores de cópulas (funciones φ) se puede conseguir calculando la transformada de Laplace de una función de distribución.

Ya adelantábamos en el apartado 3.2.4 la posibilidad de construir cópulas arquimedianas a partir de la transformada de Laplace de la función de distribución Λ de una variable aleatoria Θ . Dicha transformada viene dada por la expresión

$$\psi(t) = \int_0^{\infty} e^{-\theta t} d\Lambda(\theta) \quad \text{y permite generar una cópula a partir de la ecuación} \quad C(u, v) = \psi(\psi^{-1}(u) + \psi^{-1}(v)).$$

Si bien las familias de cópulas arquimedianas más conocidas son paramétricas (y sobre todo uniparamétricas), algunos autores han propuesto versiones que no lo son. Así por ejemplo, en [VANLAM] se propone un método de construcción de cópulas arquimedianas no paramétricas a partir de una combinación log-lineal continua de generadores arquimedianos y se demuestra que los coeficientes de estas funciones ajustadas a trozos pueden ser interpretadas como medidas de dependencia local.

3.5 Método de construcción de cópulas con soporte fractal

Finalmente vamos a hacer referencia al trabajo reciente que encontramos en [FRENERO] como un ejemplo más de la gran cantidad de aplicaciones e investigaciones que hoy día siguen surgiendo dentro del ámbito de la teoría de cópulas. En este artículo se propone un sistema que permite la construcción de funciones cópula cuyo soporte es un fractal, entendiéndose por tal, un conjunto cuya dimensión topológica es inferior a su dimensión de Hausdorff.

El algoritmo parte de una cópula C y una matriz T denominada **matriz de transformación** la cual se define como una matriz con las siguientes propiedades:

- Ninguno de sus elementos es negativo.

- La suma de todos sus elementos es 1.
- Ninguna fila o columna tiene todos sus elementos iguales a 0.

En sucesivas iteraciones, la matriz T actúa sobre la cópula C dando lugar a una secuencia de cópulas $\{C, T(C), T^2(C)=T(T(C)), T^3(C)=T(T^2(C)), \dots, T^m(C)=T(T^{m-1}(C)), \dots\}$ con límite C_T .

Bajo ciertas condiciones impuestas sobre T y C_T , los autores demuestran que dicho límite es una cópula cuyo soporte σ es un fractal de dimensión Hausdorff “ s ” comprendida entre 1 y 2.

4. Construcción de cópulas con valor óptimo del estadístico de Pearson

Según hemos comentado, uno de los principales problemas que encontrará el analista que trabaje con funciones cópula es seleccionar de entre una gran diversidad de familias, aquella que mejor represente la distribución conjunta entre las variables de estudio, disponiendo según hemos visto, de diferentes criterios de selección que le servirán de ayuda en la toma de esta complicada decisión.

En esta tesis proponemos un método de construcción de cópulas alternativo especialmente dirigido a obtener la mejor candidata respecto de uno de los criterios enumerados en el capítulo 2. En concreto, la metodología consiste en obtener la subcópula que proporciona el menor valor posible respecto del estadístico de Pearson el cual, como hemos visto en el apartado 2.5.5, sirve de evaluación del contraste de bondad de ajuste. En un segundo paso, esta subcópula será extendida mediante técnicas de interpolación polinómica para definir una cópula que conserve el valor de dicho estadístico. El resultado permitirá al analista disponer de una cópula no paramétrica que presenta características a priori atractivas:

- Presenta el menor valor posible respecto del estadístico de Pearson para una partición dada del cuadrado unidad, lo cual la convierte en la cópula óptima respecto de dicho criterio de selección.
- Su naturaleza no paramétrica evita el paso de buscar al mejor representante (el mejor valor del parámetro) dentro de una familia de cópulas.
- Para el analista es indiferente conocer el tipo de dependencia que existe entre las variables puesto que la cópula se adapta localmente a los cambios que se producen respecto de dicha relación. Desaparece el problema de tener un conocimiento de antemano de la relación entre ellas y tener que plantearse si la cópula a utilizar debe poner mayor énfasis entre las colas de las distribuciones marginales, resaltar algún tipo de asimetría o reflejar efectos “multimodales”.

Para explicar la manera en que vamos a construir esta cópula, comencemos considerando una muestra de K pares distribuidos sobre el cuadrado unidad sobre el que se realiza una partición en $n \times m$ rectángulos del mismo área. Fusionando las expresiones (2.7) y (2.8) que dan forma al estadístico de Pearson se obtiene:

$$\sum_{i=1}^n \sum_{j=1}^m \frac{(N_{ij} - K * p_{ij}(\theta))^2}{K * p_{ij}(\theta)} = \sum_{i=1}^n \sum_{j=1}^m \frac{(N_{ij} - K * \iint_{B_{ij}} dC(u, v; \theta))^2}{K * \iint_{B_{ij}} dC(u, v; \theta)}$$

Nuestra intención es construir la cópula que proporciona un mejor valor respecto de este estadístico. Teniendo en cuenta que la función que obtendremos va a ser no paramétrica, podemos, en primer lugar, omitir en la expresión anterior la dependencia del parámetro.

Observemos además que las cantidades p_{ij} representan el volumen del rectángulo $B_{ij} = [u_i, u_j] \times [v_i, v_j]$ a través

de la cópula C . Otra forma de escribir este volumen, como se muestra en la tercera condición que debe satisfacer una función cópula (1.7), sería $p_{ij} = V_c(B_{ij}) = C(u_{i+1}, v_{j+1}) - C(u_i, v_{j+1}) - C(u_{i+1}, v_j) + C(u_i, v_j)$, con lo cual podemos concluir que la contribución de la cópula al estadístico viene dada por el valor de dicha cópula en una relación de puntos que son los vértices de cada uno de los rectángulos en que ha sido dividido el cuadrado unidad.

De esta forma, se trata de construir una función C con las siguientes características:

- Que sea una función cópula, es decir, que verifique las propiedades (1.5) - (1.7).
- Que sea no paramétrica, es decir, que no dependa de ningún parámetro.
- Que su evaluación sobre los nodos (u_i, v_j) , los cuales sirven de intersección entre las líneas horizontales y verticales que generan la partición del cuadrado unidad, garanticen un valor mínimo respecto del estadístico de

Pearson dado por
$$\varphi = \sum_{i=1}^n \sum_{j=1}^m \frac{(N_{ij} - n*(C(u_i, v_j) - C(u_i, v_{j-1}) - C(u_{i-1}, v_j) + C(u_{i-1}, v_{j-1})))^2}{n*(C(u_i, v_j) - C(u_i, v_{j-1}) - C(u_{i-1}, v_j) + C(u_{i-1}, v_{j-1}))} \quad (4.1)$$

Este objetivo lo alcanzaremos en dos etapas:

- Obtención de la **subcópula** óptima.- Consideraremos el espacio de todas las posibles subcópulas con soporte los nodos resultantes de la partición del cuadrado unidad y buscaremos aquella cuyo valor sobre los puntos de su soporte proporcionen el menor valor del estadístico de Pearson.
- Obtención de la **cópula** óptima.- Extensión de la subcópula óptima a una cópula mediante interpolación con funciones polinómicas de forma que la cópula resultante tome el mismo valor que ella en cada uno de los nodos de la rejilla que trocea el cuadrado unidad. En consecuencia tendremos una cópula cuyo estadístico de Pearson superará en bondad al de otras candidatas pues, como veremos, estará conformado por la combinación óptima de valores que una cópula puede tomar sobre los puntos de evaluación del estadístico.

Podría pensarse que esta forma de proceder es sinónimo de sobreajuste en el sentido de que estamos forzando a la cópula a pasar por una relación de puntos para que el estadístico correspondiente proporcione el mejor p-valor posible. Es decir, algo parecido a construir un modelo que minimice el error cuadrático medio definiendo para ello la ecuación que interpola exactamente todos los valores muestrales ($MSE = 0$) pero descuidando con ello la posibilidad de extrapolar dicho modelo a otras muestras o lo que es lo mismo, su poder predictivo.

Sin embargo, lo que realmente estamos tratando de hacer es utilizar el estadístico en cuestión para hacernos una idea de lo que debería valer la cópula que estamos buscando en una relación de puntos de su dominio, concretamente en aquellos nodos que comprenden la rejilla que hacen una partición del cuadrado unidad. La aproximación al estadístico va a permitir que los valores que se estiman para esos puntos se muevan en consonancia con la muestra y puedan constituir un soporte fiable. No estamos construyendo con esto una función cópula propiamente dicha, sino una función subcópula que sirva de punto de partida. La verdadera función cópula se construye a partir de ella mediante alguna

técnica de interpolación pero, por supuesto, dicha función no está ajustada (sobreajustada) a todos los puntos de su dominio. Sí habrá que tener precaución a la hora de considerar la dimensión de la rejilla que provoca la partición. No conviene que la cantidad de rectángulos que surjan ($n*m$) sea desmesurada ya que como consecuencia, el número de puntos que podrían caer en cada uno de ellos podría ser demasiado pequeño (uno o ninguno por rectángulo) y es ahí donde sí podría existir el problema del sobreajuste (la función subcópula que sirve de punto de partida se ajusta a todos los valores de la muestra). En principio, podría servir de punto de partida el convenio empírico que citamos en el Anexo VIII según el cual el número de clases debe ser mayor o igual que 5. De hecho en los ejemplos que vamos a plantear vamos a considerar este número de divisiones en cada uno de los intervalos $[0,1]$, dando lugar así a una partición del cuadrado unidad en 25 rectángulos. Particiones de mayor dimensión reducirían en exceso el número de pares muestrales por rectángulo.

A pesar de que el método que se va a desarrollar busca la optimización del criterio descrito en el apartado 2.5.5, nuestro verdadero argumento de selección estará basado en la calidad de las predicciones que proporcione cada cópula (apartado 2.5.6). Es decir, en ningún caso vamos a decantarnos de forma incondicional por el empleo de la cópula que resulte de aplicar esta metodología sino que recomendamos establecer siempre el contraste con los resultados que puedan proporcionar los representantes de otras familias de cópulas conocidas.

4.1 Etapa I: Construcción de la subcópula que minimiza el estadístico de Pearson

Consideremos un conjunto de K pares de puntos (x_i, y_i) , que han sido debidamente transformados a través de sus respectivas funciones de distribución en $(F(x_j), G(y_j)) = (u_i, v_i)$. De esta forma, los posibles rangos de variación de las variables X e Y se han adaptado al rango de variación del cuadrado unidad dado que las variables transformadas se distribuyen según uniformes estándares. Se trata de encontrar la función cópula C que mejor describa la relación de dependencia entre las variables U y V , y aplicarla sobre los pares (u_i, v_i) para, en virtud del teorema de Sklar, obtener la función de distribución conjunta que mejor se ajuste a la relación que mantienen las variables originales.

Como hemos dicho en el apartado anterior, la selección de esa cópula óptima la vamos a basar en la búsqueda de la subcópula que proporciona el menor valor del estadístico de Pearson respecto de una partición del cuadrado unidad en $n \times m$ rectángulos del mismo área.

Para el ejemplo que servirá de referencia a este capítulo vamos a utilizar una parte de los datos diarios de demanda doméstica de gas natural en Madrid que son empleados en el capítulo 7. En dicho capítulo se explica cómo estimar a partir de una serie temporal de demanda de gas, el valor de los datos que supuestamente se hubieran presentado en condiciones normales de temperatura, es decir, se detalla la generación de una serie temporal de demanda de gas aislando el efecto de la climatología. A partir del resultado obtenido, se puede disponer de una muestra bidimensional donde la componente “ x ” de cada par mide la diferencia entre el dato de temperatura máxima del día “ t ” y otro que se considera como temperatura normal del mes al que pertenece dicho día, es decir,

$$X(t) = \text{Temperatura_Maxima_Normal}_{MES} - \text{Temperatura_Maxima_Real}(t)$$

La componente “ y ” mide la variación porcentual entre el dato real de demanda y el que se estima que debió presentarse en condiciones normales de temperatura, es decir,

$$Y(t) = \frac{100 * (Demanda_Real(t) - Demanda_en_condiciones_normales_de_temperatura(t))}{Demanda_en_condiciones_normales_de_temperatura(t)}$$

Lógicamente, como el modelo no recoge el efecto de la temperatura es de esperar que, si un día “t” la temperatura cae por debajo de la media esperada para el mes ($X(t)$ positivo), la estimación del dato normal de demanda se quede por debajo del dato real (pues al hacer frío, subirá la demanda de gas) y, en consecuencia, $Y(t)$ será también positivo.

El siguiente gráfico muestra pares de datos asociados al mes de Noviembre cuando ambas componentes son positivas. Se trata de analizar la relación existente entre ellas utilizando funciones cópula.

Para ello, empezamos generando la muestra transformada a través de las distribuciones empíricas continuas de X e Y (véase el apartado 2.2), y continuamos haciendo una partición de I^2 en 5×5 rectángulos.

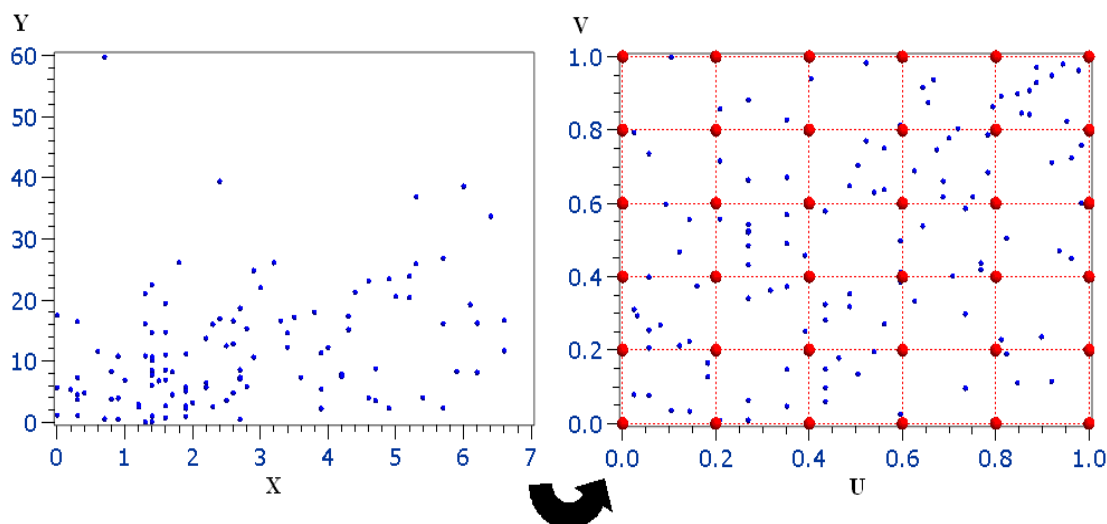


Ilustración 4.1: Conversión de la muestra (x,y) en la muestra de uniformes (u,v) que sirve de soporte a una cópula

El estadístico de Pearson se construye a partir del número de puntos que caen en cada uno de los rectángulos y el número esperado de ellos a través de la subcópula cuya bondad se esté contrastando. Esta última cantidad viene dada por el volumen a través de la subcópula de cada uno de los rectángulos, que viene a ser la probabilidad a través de ella de que un par (u,v) se encuentre en el mismo. Para el cálculo de estos volúmenes basta con disponer del valor de la subcópula en cada uno de los nodos que surgen al hacer la partición (puntos rojos del gráfico anterior).

Así, vamos a considerar el espacio constituido por todas las subcópulas cuyo dominio son precisamente esos nodos:

$$\Omega = \{ C' \mid C' \text{ subcópula con } Dom(C') = S_1 \times S_2 \}, \quad S_1 = \{ 0, \frac{1}{n}, \frac{2}{n}, \dots, \frac{n-1}{n}, 1 \} \quad S_2 = \{ 0, \frac{1}{m}, \frac{2}{m}, \dots, \frac{m-1}{m}, 1 \} \quad (4.2)$$

Las condiciones a imponer para que una subcópula C' pertenezca a Ω son:

1. $Dom(C') = S_1 \times S_2$ (4.3)

2. $C'(\frac{i}{n}, 0) = 0 \quad \forall i \in \{0, 1, 2, \dots, n\}$ y $C'(0, \frac{j}{m}) = 0 \quad \forall j \in \{0, 1, 2, \dots, m\}$ (4.4)

3. $C'(\frac{i}{n}, 1) = \frac{i}{n} \quad \forall i \in \{0, 1, 2, \dots, n\}$ y $C'(1, \frac{j}{m}) = \frac{j}{m} \quad \forall j \in \{0, 1, 2, \dots, m\}$ (4.5)

4. C' debe ser 2-creciente, es decir, $V_{C'}([\frac{i-1}{n}, \frac{i}{n}] \times [\frac{j-1}{m}, \frac{j}{m}]) \geq 0 \quad \forall i \in \{1, 2, \dots, n\} \quad \forall j \in \{1, 2, \dots, m\}$ (4.6)

Obsérvese que basta asegurar que cualquiera de los $n \times m$ rectángulos disjuntos tenga volumen mayor o igual que 0 a través de la subcópula, dado que, de ser así, el volumen de cualquier otro rectángulo cuyos vértices pertenezcan al dominio de $C'(S_1 \times S_2)$ también será mayor o igual que 0 puesto que su volumen se podrá expresar como suma de los volúmenes de algunos de los rectángulos disjuntos.

Es decir, si en el ejemplo gráfico que hemos planteado, aseguramos que el volumen de $B_{22} = [1/5, 2/5] \times [1/5, 2/5]$ y el volumen de $B_{23} = [1/5, 2/5] \times [2/5, 3/5]$ son mayores o iguales que cero, entonces, el volumen del rectángulo $B = [1/5, 2/5] \times [1/5, 3/5]$ (también contenido en el dominio de C') será mayor o igual que 0 por ser suma de los otros.

El objetivo es encontrar, entre todas las subcópulas de Ω , aquella para la que la evaluación del estadístico de Pearson (4.1) proporcione el menor valor posible. Este problema puede ser abordado mediante un modelo de programación matemática con las siguientes características:

- Existirá una variable c'_{ij} por cada nodo de la rejilla que resulta de hacer la partición. Dichas variables van a representar el valor que debe tomar la subcópula en cada uno de los nodos.
- La función objetivo a minimizar será el propio estadístico de Pearson que, por su estructura, da carácter no lineal al problema. Se trata de encontrar los c'_{ij} que minimizan dicho estadístico.
- Las restricciones que determinan la región factible van a ser lineales y deben asegurar que la función cuyo valor en cada uno de los puntos de su dominio va a calcularse, sea una subcópula. Para ello:
 - Variables asociadas a nodos del tipo $(i/n, 0)$ y $(0, j/m)$ deben valer 0 (se debe cumplir la condición (4.4)).
 - Variables asociadas a nodos del tipo $(i/n, 1)$ y $(1, j/m)$ deben valer respectivamente i/n y j/m (se debe cumplir la condición (4.5)).
 - Para cada rectángulo $[\frac{i-1}{n}, \frac{i}{n}] \times [\frac{j-1}{m}, \frac{j}{m}]$ debe satisfacerse la restricción lineal

$$C'(\frac{i}{n}, \frac{j}{m}) - C'(\frac{i-1}{n}, \frac{j}{m}) - C'(\frac{i}{n}, \frac{j-1}{m}) + C'(\frac{i-1}{n}, \frac{j-1}{m}) \geq 0 \quad (\text{se debe cumplir la condición (4.6)}).$$

Formalmente, el modelo matemático de programación no lineal quedaría planteado de la siguiente manera:

$$\begin{aligned} & \text{Min} \sum_{i \in S_1} \sum_{j \in S_2} \frac{(N_{ij} - n * (c'_{i,j} - c'_{i-1,j} - c'_{i,j-1} + c'_{i-1,j-1}))^2}{n * (c'_{i,j} - c'_{i-1,j} - c'_{i,j-1} + c'_{i-1,j-1})} \\ \text{s.a} \quad & c'_{i0} = 0 \quad \forall i \in S_1 \quad ; \quad c'_{0j} = 0 \quad \forall j \in S_2 \quad ; \quad c'_{im} = \frac{i}{n} \quad \forall i \in S_1 \quad ; \quad c'_{nj} = \frac{j}{m} \quad \forall j \in S_2 \\ & p_{ij} = c'_{i,j} - c'_{i-1,j} - c'_{i,j-1} + c'_{i-1,j-1} \geq 0 \quad \forall i \in S_1, j \in S_2 \\ & S_1 = \{1, 2, \dots, n-1, n\} \quad S_2 = \{1, 2, \dots, m-1, m\} \end{aligned} \quad (4.7)$$

Conviene observar que realmente, las únicas variables de decisión del problema van a ser aquellas asociadas a nodos que se encuentran en el interior del cuadrado unidad dado que el valor de las fronteras ya viene prefijado de antemano por las condiciones (4.4) y (4.5).

Se trata de un problema de programación fraccional no lineal (véase capítulo 13.3 de [HILIE]) que presenta la dificultad añadida de que, por las características de la región factible, los valores que van a tomar las variables se pueden mover en entornos muy cercanos a 0 y ello puede llevar a que el denominador de la función objetivo se anule. Ante la posible falta de convergencia del algoritmo que se aplique para resolver el problema, se debe poner especial cuidado en la selección de una solución inicial factible que evite acercarse a valores que puedan anular este denominador.

Se propone como punto de partida, la subcópula producto (véase (1.25) o (AI.3) en el Anexo I) que se define en cada

$$\text{vértice } \left(\frac{i}{n}, \frac{j}{m}\right) \text{ como } C'\left(\frac{i}{n}, \frac{j}{m}\right) = \frac{i*j}{n*m}.$$

Teorema 4.1: *Existencia de solución factible al problema de programación no lineal que permite obtener una subcópula que optimiza el valor del estadístico de Pearson*

“Sea una partición del cuadrado unidad en $n \times m$ rectángulos del mismo área. Entonces el problema de programación fraccional (4.7) tiene solución factible. La solución óptima a dicho problema va a proporcionar la relación de valores en cada uno de los puntos del dominio de definición de la subcópula que optimiza el estadístico de Pearson”.

demostración

Como adelantábamos una solución factible viene dada por la propia subcópula producto $C'\left(\frac{i}{n}, \frac{j}{m}\right) = \frac{i*j}{n*m}$ (aunque lógicamente no es necesariamente la óptima).

Se puede demostrar que dichos valores satisfacen el problema de programación matemática y que por tanto, la función definida a través de ellos es una subcópula pues, efectivamente,

- $c'_{i0} = \frac{i*0}{n*m} = 0 \quad \forall i \in S_1$ y $c'_{0j} = \frac{0*j}{n*m} = 0 \quad \forall j \in S_2$ (se cumple la condición (4.4))
- $c'_{im} = \frac{i*m}{n*m} = \frac{i}{n} \quad \forall i \in S_1$ y $c'_{nj} = \frac{n*j}{n*m} = \frac{j}{m} \quad \forall j \in S_2$ (se cumple la condición (4.5))
- $c'_{ij} - c'_{i-1j} - c'_{ij-1} + c'_{i-1j-1} = \frac{i*j}{n*m} - \frac{(i-1)*j}{n*m} - \frac{i*(j-1)}{n*m} + \frac{(i-1)*(j-1)}{n*m} = \frac{1}{n*m} > 0 \quad \forall (i, j) \in S_1 \times S_2$
(se cumple la condición (4.6)).

c.q.d

De acuerdo a la verificación de (4.6), esta relación de valores p_{ij} no anula los denominadores correspondientes a cada uno de los sumandos de la función objetivo dado que “ i ” y “ j ” son estrictamente mayores que 0 (por supuesto “ n ” y “ m ” también) por ir asociados a puntos interiores del cuadrado unidad. Por tanto, pueden constituir una buena solución de partida al problema de optimización que se resolverá finalmente aplicando alguno de los algoritmos destinados a ello.

El algoritmo que se recomienda, está condicionado al hecho de que las restricciones del problema sean de tipo lineal y también a la dimensión del mismo en lo que al número de variables se refiere, o lo que es lo mismo, al número de nodos que resulten al realizar la partición del cuadrado unidad.

Los métodos que ofrecen mayor fiabilidad y velocidad son aquéllos que utilizan la matriz Hessiana. Sin embargo, cuando la dimensión del problema es muy grande, no se aconseja el empleo de métodos que impliquen el cálculo de las derivadas de segundo orden de la función objetivo que componen dicha matriz puesto que suelen requerir demasiado tiempo y memoria computacional. En el enlace [SASDOC] (ayuda al usuario de la herramienta *SAS*) se proporciona una relación de técnicas a utilizar en función de las características del modelo a resolver. Dado que *SAS* ha sido la herramienta utilizada para la implementación de los modelos matemáticos que se presentan en esta tesis, hemos considerado que es el asistente de dicha herramienta el que mejor puede conocer la eficiencia y complejidad de los algoritmos que tiene programados para tratar este tipo de problemas. Este asistente puede ser consultado para un mayor conocimiento de cada una de las técnicas que a continuación se citan:

- Cuando el número de variables es inferior a 40, se propone alguno de los algoritmos de **Newton-Raphson** o el método de la región de confianza (Trust Region Method). Estas técnicas utilizan el **Hessiano**, el cual, para problemas de dimensión pequeña, no resulta demasiado costoso de calcular y cuyo uso ofrece siempre mayor fiabilidad.
- Cuando el número de variables se encuentra entre 40 y 200, se propone el algoritmo **Cuasi-Newton** que no utiliza la matriz hessiana, sino la propia función objetivo y su gradiente. Si bien esto implica la ejecución de un mayor número de iteraciones que los métodos anteriores, éstas van a ser más rápidas y el resultado desde un punto de vista computacional se ve más beneficiado.
- Si el número de variables es superior a 200, se propone el algoritmo del **Gradiente Conjugado**. En este caso, el cálculo del Hessiano resulta aún menos recomendado dada la gran cantidad de memoria computacional que es necesario reservar para ello. Si bien el número de iteraciones que precisa es mayor que el de los algoritmos Cuasi-Newton, la velocidad con la que éstas son llevadas a cabo es mucho mayor.

Una vez resuelto el problema, dispondremos de una subcópula C^* cuyo dominio son los puntos de evaluación del estadístico de Pearson. La forma en que dicha subcópula ha sido construida asegura que los valores que toma en cada uno de los puntos de su dominio proporcionan el menor valor posible para el estadístico de Pearson y por tanto aquella que mejor se ajusta a los datos de acuerdo al contraste de bondad de ajuste, la que más se aproxima a la cópula teórica desconocida.

4.2 Etapa II: Construcción de una cópula por interpolación de una subcópula

Realmente estamos interesados en la cópula que mejor se ajusta a los datos y no en la subcópula que mejor lo hace porque como veremos, necesitamos conocer cuánto vale la función en todos los puntos del cuadrado unidad para poder simular valores de la variable Y a partir de cualquier valor de la variable X y no sólo en aquellos “ x ” cuyos transformados son los vértices de los rectángulos de la partición. Sin embargo, vamos a aprovechar el resultado del apartado anterior para construir una cópula que tome los mismos valores que los que toma en su dominio de definición la subcópula que nos ha proporcionado el modelo de optimización y, por tanto, que conserve el valor del estadístico de Pearson respecto de la partición.

Parece obvio que la forma de fabricar una cópula que tome el mismo valor que una subcópula en una relación de puntos (concretamente en aquellos que constituyen su soporte), es mediante algún método de interpolación. Así, se propone

definir una función C que pase por los puntos $(\frac{i}{n}, \frac{j}{m}, C'(\frac{i}{n}, \frac{j}{m})) \quad \forall i \in \{1, 2, \dots, n\}, j \in \{1, 2, \dots, m\}$, o lo que es lo mismo, construir una superficie que interpole los valores que toma la subcópula óptima que ha resultado del modelo matemático, pero sin descuidar el aspecto más importante: la función C debe ser una cópula.

4.2.1 Método de interpolación bilineal

Nelsen enuncia un resultado (véase lema 2.3.5 de [NELSEN]) que comenta la posibilidad de construir una cópula a partir de la extensión de una subcópula. El lema al que hacemos referencia dice lo siguiente:

Lema 4.1

“Sea C' subcópula. Entonces existe una cópula C tal que $C'(u,v)=C(u,v) \quad \forall u, v \in \text{Dom}(C')$; es decir, cualquier subcópula puede extenderse a una cópula. La extensión, generalmente, no es única”.

Nelsen plantea la construcción de una cópula por interpolación lineal de los valores que toma una subcópula en su dominio de definición. Dado que estamos particularmente interesados en las cópulas que se construyen a partir de subcópulas, detallamos a continuación el método que sirve además de demostración al lema enunciado.

demostración

Sea $\text{Dom}(C') = S_1 \times S_2$. Se verifica entonces que C' es uniformemente continua en $S_1 \times S_2$ (teorema 2.2.4 de [NELSEN]). Existe un resultado topológico que nos garantiza que en estas condiciones, C' puede extenderse de forma continua a su clausura y así, obtenemos $C'' : \bar{S}_1 \times \bar{S}_2 \rightarrow I \subset \mathbb{R}$ que seguirá siendo subcópula porque C' lo es.

A continuación, se extenderá C'' a una función C con dominio I^2 , $C: I^2 \rightarrow I \subset \mathbb{R}$. Para ello, sea $(a, b) \in I^2$ arbitrario y sean a_1 y a_2 el mayor y menor elemento de \bar{S}_1 que satisfacen $a_1 \leq a \leq a_2$ y, análogamente sean b_1 y b_2 el mayor y menor elemento de \bar{S}_2 que satisfacen $b_1 \leq b \leq b_2$.

Se observa que, si $a \in \bar{S}_1 \Rightarrow a_1 = a = a_2$ e, igualmente, si $b \in \bar{S}_2 \Rightarrow b_1 = b = b_2$.

Sean ahora

$$\lambda_1 = \begin{cases} \frac{a-a_1}{a_2-a_1} = \frac{1}{a_2-a_1} * a + \frac{-a_1}{a_2-a_1} & \text{si } a_1 < a_2 \\ 1 & \text{si } a_1 = a_2 \end{cases} \quad \text{y} \quad \mu_1 = \begin{cases} \frac{b-b_1}{b_2-b_1} = \frac{1}{b_2-b_1} * b + \frac{-b_1}{b_2-b_1} & \text{si } b_1 < b_2 \\ 1 & \text{si } b_1 = b_2 \end{cases}$$

Se define entonces,

$$C(a, b) = (1-\lambda_1) * (1-\mu_1) * C''(a_1, b_1) + (1-\lambda_1) * \mu_1 * C''(a_1, b_2) + \lambda_1 * (1-\mu_1) * C''(a_2, b_1) + \lambda_1 * \mu_1 * C''(a_2, b_2) \quad (4.8)$$

El hecho de que los parámetros λ_1 y μ_1 sean funciones lineales en “a” y “b” garantiza que la interpolación así definida sea lineal en cada componente (interpolación bilineal). Además, la demostración de que la función C así construida es una cópula puede consultarse en [NELSEN] (lema 2.3.5). Dicha prueba completa la demostración del lema.

c.q.d

Adjuntamos a continuación los gráficos asociados a la distribución y densidad cópula construida de esta forma a partir de la relación de puntos presentados en la ilustración (4.1). Se puede observar la forma escalonada que presenta la densidad cópula. Este aspecto característico es el que cabe esperar si tenemos en cuenta que la densidad cópula toma un valor constante sobre cada uno de los rectángulos en los que queda dividido el cuadrado unidad una vez realizada la partición del mismo. Este valor constante se obtiene derivando con respecto a las dos variables (según (1.13)) la expresión (4.8), siendo el resultado de esta operación

$$\frac{V_{C''}(B)}{(a_2-a_1) * (b_2-b_1)} \quad (4.9)$$

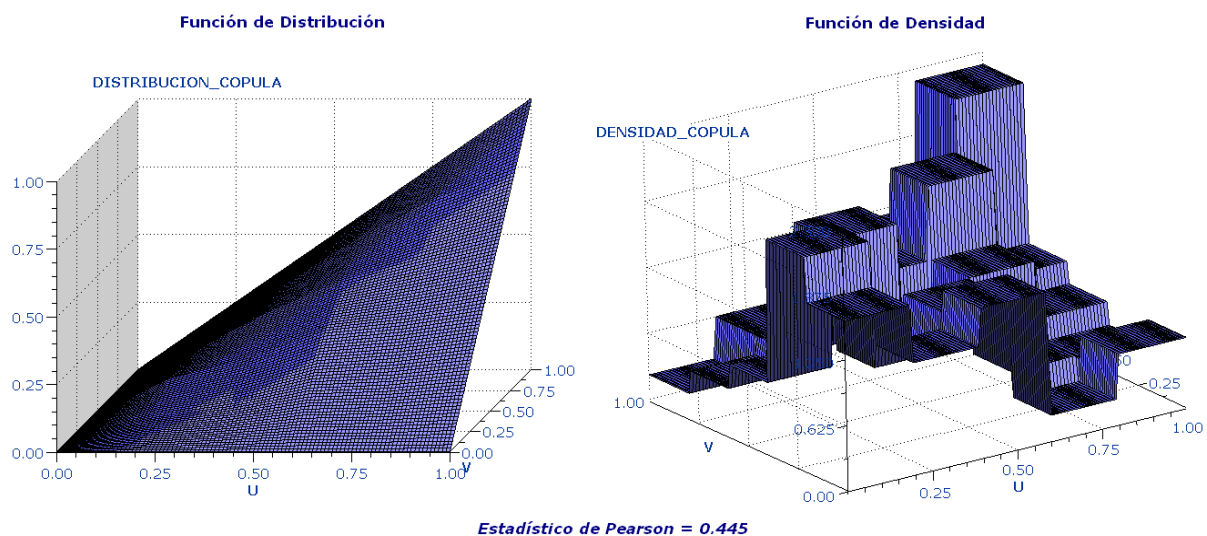


Ilustración 4.2: Cópula construida por interpolación bilineal de la subcópula que optimiza el estadístico de Pearson

Al pié de dicha función cópula se proporciona el valor obtenido al evaluar la expresión de Pearson respecto de la partición realizada. Dicha expresión mide, a fin de cuentas, la diferencia entre el número de pares muestrales observados en cada rectángulo de la partición y el número de ellos que cabría esperar a través de la cópula generada. Esta diferencia es ínfima en comparación con los valores obtenidos respecto del contraste $H_0: C(u, v) = C(u, v; \theta)$ para algún $\theta \in \Theta$ para una relación de familias de cópulas: la de **Cola Derecha Pesada** (Heavy Right Tail o **HRT**, véase (AI.12) en Anexo I), la de Gumbel (véase (AI.8) o (AI.15)), la de Frank (AI.13), la de Plackett (AI.37), la **Normal** (AI.5), la de **Ali-Mikhail-Haq** (AI.10), la **Cópula con Secciones Cúbicas** (véase ecuación (4.3) en [NELQUERO]), la de **Clayton** (AI.11) y finalmente la cópula Producto (*II*) que caracteriza independencia. En la siguiente tabla, presentamos el valor de los estadísticos así como sus correspondientes p-valores. Hemos obtenido dos regiones críticas al 95%: una de ellas asociada a las familias uniparamétricas ($d = 1$) que se calcula a partir de una Chi-cuadrado con 15 grados de libertad ($(n-1)*(m-1)-d=15$) y otra asociada a la cópula con secciones cúbicas que es la única biparamétrica ($d = 2$) y para la que el número de grados de libertad es 14 (véase apartado 2.5.5).

Contraste de bondad de ajuste a una Cópula

Mes = 11

COPULA	PARAMETRO_COPULA_CHAR	PEARSON	P_VALOR
GUM	1.3689	14.8734	0.46058
HRT	1.3550	15.0270	0.44948
PLA	3.3650	16.1800	0.37019
NOR	0.4108	17.6479	0.28163
CSC	1.1568/0.1800	18.0633	0.25935
FRA	1.0527	23.9931	0.06521
CLA	0.7379	25.2079	0.04722
ALI	0.4313	25.4144	0.04465
PRO	.	37.3915	0.00111

La región crítica viene dada por $\{X|X > 24.99579014\}$

La región crítica para la biparamétrica viene dada por $\{X|X > 23.684791305\}$

Nivel de significación del contraste, $\alpha = 0.05$

Tabla 4.1: Evaluación del estadístico de Pearson para una serie de cópulas

Las cópulas presentadas han sido ordenadas de mayor a menor p-valor, es decir, de mejor a peor ajuste. Así, se puede observar que aun cuando la mayoría de los valores obtenidos para el estadístico de Pearson llevaría a no poder rechazar la hipótesis para un nivel de significación de $\alpha = 0.05$ (únicamente la de Clayton y la de Ali-Mikhail-Haq la rechazarían por muy poco), las cópulas que mejor se ajustan son la HRT y la de Gumbel, caracterizadas por capturar relaciones en las que la dependencia es mayor entre las colas derechas de las distribuciones (entre los máximos decrementos de temperatura del mes de Noviembre y las máximas desviaciones del patrón de demanda construido en condiciones normales de temperatura). Por el contrario la que peor se ajusta es la cópula producto lo cual, de acuerdo al teorema 2.4.2 que encontramos en [NELSEN], es indicativo de que no puede considerarse independencia entre ambas variables. Los siguientes gráficos, asociados a las densidades de las cópulas HRT y Gumbel, hacen explícita esta relación enfatizada entre los extremos. Obsérvese también que en la ilustración (4.2) esta característica queda también puesta de manifiesto pues el “escalón” de mayor altura se presenta entre los valores más altos de U y V (los más próximos al punto (1,1)).

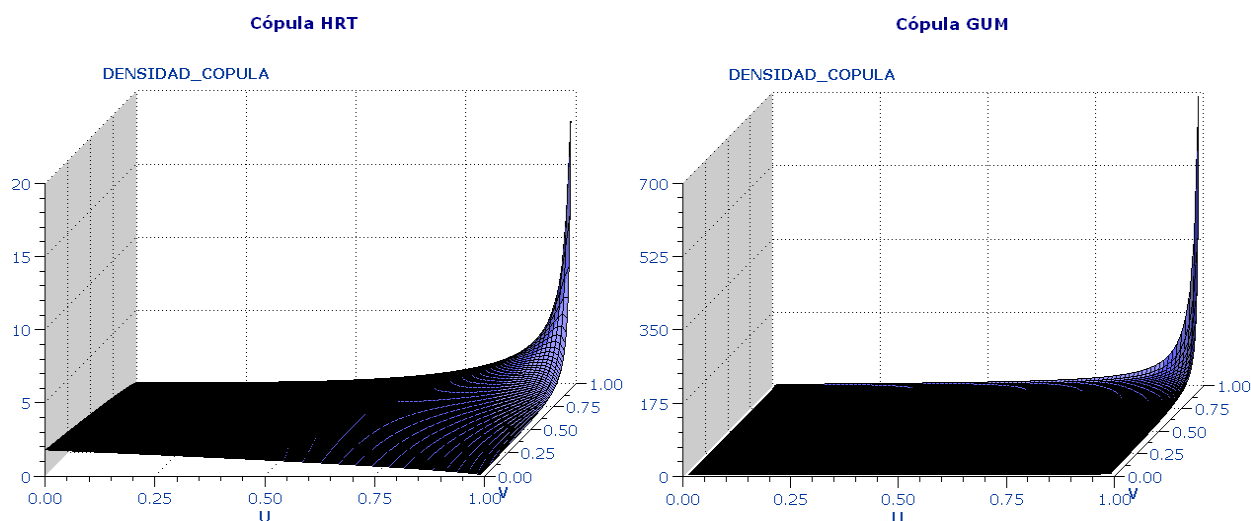


Ilustración 4.3: Densidades asociadas a la cópula de Cola Derecha Pesada (HRT) y a la cópula de Gumbel

Vamos a ampliar el repertorio de funciones cópula que se generan por interpolación de subcópulas para así disponer de un mayor abanico de posibilidades a la hora de plantearnos cuál es la cópula que, proporcionando un buen valor para el valor del estadístico de Pearson, mejor se ajusta a los datos. Vamos a proponer como ejemplo la construcción de una función cópula por interpolación cúbica mediante polinomios de Hermite. Veremos que la construcción de dicha cópula no está siempre garantizada sino que es necesario imponer algunas condiciones adicionales al problema de programación fraccional (4.7).

Sin embargo, antes de entrar en este extenso desarrollo vamos a presentar otra cópula que también se obtiene por interpolación mediante unas funciones polinómicas denominadas polinomios de Bernstein. Como principal ventaja de esta cópula, veremos que su densidad no presenta los bruscos cambios de comportamiento que se aprecian en la cópula resultante de la interpolación bilineal y que también se observarán en la obtenida por interpolación cúbica. De hecho una de las razones por las que anteponeamos el método de interpolación de Bernstein al de Hermite es extraer algunas ideas de dicha técnica para poder suavizar la superficie cúbica interpoladora que construiremos.

4.2.2 Método de interpolación mediante polinomios de Bernstein

Una forma de llevar a cabo la interpolación del dominio de la subcópula es mediante una expansión de Bernstein que se basa en una combinación lineal de una familia de polinomios denominados polinomios de Bernstein. Esta expansión fue propuesta por **Li** (1998) y la cópula que se obtiene como consecuencia ha sido discutida por **Sancetta** (2003) y Sancetta y **Satchell** (2004). Esta cópula posee una propiedad importante y es su capacidad de “aproximarse”, este matiz es importante, a cualquier función que satisfaga las propiedades de las cópulas ((1.5)-(1.7)).

En el apéndice B de [HUSASC] encontramos la siguiente definición:

Definición 4.1.- Sea $\alpha(\omega)$ una función definida sobre una rejilla de $(m+1)^2$ puntos, siendo

$$\omega = \frac{k}{m} + \frac{l}{m} \quad k, l \in \{0, 1, 2, \dots, m\} \quad . \text{ Los "m" intervalos en que han sido divididos los ejes de ordenadas y abscisas que dan lugar a ella se suponen de la misma longitud.}$$

Se define además el polinomio de Bernstein de orden "m" como

$$P_{j,m}(x) = \binom{m}{j} x^j (1-x)^{m-j} \quad \forall j \in \{0, 1, 2, \dots, m\} \quad (4.10)$$

Entonces, se define la cópula de Bernstein como $C^B(u, v) = \sum_{k=0}^m \sum_{l=0}^m \alpha\left(\frac{k}{m}, \frac{l}{m}\right) P_{k,m}(u) P_{l,m}(v) \quad (4.11)$

Sancetta y Satchell (2004) demuestran que la función así definida será una cópula siempre que $\alpha(\omega)$ satisfaga las tres propiedades básicas de las cópulas.

Como explican Sancetta y Satchell (véase el capítulo 1 de [SANSAT]), la cópula de Bernstein puede ser utilizada como una aproximación de una cópula conocida o desconocida. En el caso en que se utilice como aproximación de una cópula conocida recibe el nombre de **Cópula de Bernstein Aproximada (ABC)**. Esta representación es particularmente útil cuando el manejo de la cópula paramétrica a la que se aproxima resulta incómodo en cuanto a su operativa, por presentar una forma muy compleja. La otra posibilidad es utilizar la cópula de Bernstein para aproximar cópulas desconocidas en cuyo caso, recibe el nombre de **Cópula de Bernstein Empírica (EBC)**.

En lo que a nosotros concierne, vamos a utilizar la cópula de Bernstein para interpolar una subcópula, lo cual no supone un problema de cara a su construcción dado que este tipo de funciones, por su propia definición, satisface las tres propiedades básicas de las cópulas, lo cual permite aplicar (4.11). Así pues, no se trata de una aproximación a una cópula conocida sino a una subcópula para la cual conocemos el valor que toma en cada uno de los puntos de su dominio de definición. En particular, si $\alpha(\omega)$ es la subcópula que optimiza el valor del estadístico de Pearson y que se obtiene como solución del modelo de programación fraccional, la función $C^B(u, v)$ construida a partir de ella será una cópula. Sin embargo, conviene observar que la interpolación por polinomios de Bernstein no es exacta sino que, como hemos ido comentando, se trata de una aproximación. En consecuencia, los valores que la cópula de Bernstein así construida tomará en cada uno de los nodos que resultan de la partición no constituirán exactamente una solución óptima del modelo. De hecho, puede comprobarse que el valor que se obtiene al evaluar la expresión de Pearson respecto de la partición presentada en la ilustración (4.1), ha aumentado sensiblemente, pasando de ser 0.445 (ilustración (4.2)) a ser 17.012 (ilustración (4.4)). Sin embargo, la ganancia obtenida a costa de esta pérdida de optimalidad, es el aspecto suave que estos polinomios interpoladores proporcionan a la densidad cópula, característica que, por lo general, es de esperar en una relación de dependencia entre variables

A continuación se adjuntan los gráficos asociados a la cópula construida de esta forma y a la densidad de la que hablamos. Podemos apreciar la mayor concentración probabilística alrededor del punto $(u, v) = (1, 1)$, siendo en este aspecto parecida a las cópulas que reflejan una relación de dependencia más fuerte entre sucesos extremos, las cuales, de acuerdo a la tabla (4.1), constituían unas buenas candidatas para el ajuste (HRT y Gumbel).

De igual forma que hiciéramos anteriormente con la cópula que se obtenía por interpolación bilineal, adjuntamos la ecuación de la densidad asociada a $C^B(u, v)$. Esta expresión ha sido extraída del capítulo 2.2 de [SANSAT]:

$$c_B(u_1, u_2) = \sum_{v_1=0}^m \sum_{v_2=0}^m \beta\left(\frac{v_1}{m}, \frac{v_2}{m}\right) * \prod_{j=1}^2 \binom{m}{v_j} * u_j^{v_j} * (1-u_j)^{m-v_j}$$

siendo

$$\beta\left(\frac{v_1}{m}, \frac{v_2}{m}\right) = (m+1)^2 * \Delta_{1,2} \alpha\left(\frac{v_1}{m+1}, \frac{v_2}{m+1}\right) =$$

$$(m+1)^2 * \left[\alpha\left(\frac{v_1+1}{m+1}, \frac{v_2+1}{m+1}\right) - \alpha\left(\frac{v_1+1}{m+1}, \frac{v_2}{m+1}\right) - \alpha\left(\frac{v_1}{m+1}, \frac{v_2+1}{m+1}\right) + \alpha\left(\frac{v_1}{m+1}, \frac{v_2}{m+1}\right) \right]$$
(4.12)

Al ser α una función 2-creciente (por poseer las condiciones que la definen como subcópula) la expresión definida por β es no negativa y de esta forma, también lo será la densidad en ella basada.

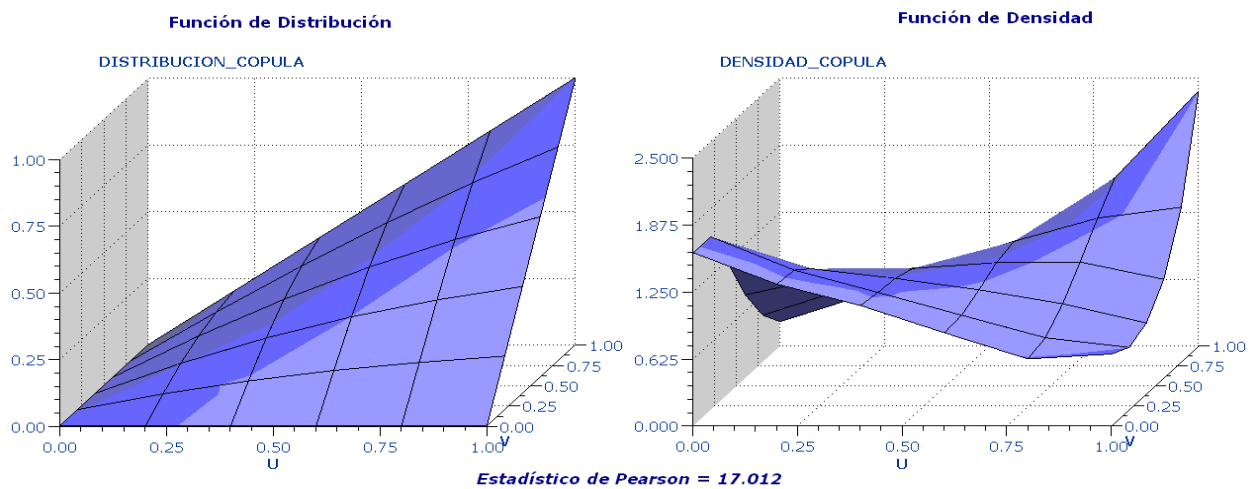


Ilustración 4.4: *Cópula construida por interpolación mediante polinomios de Bernstein, de la subcópula que optimiza el valor del estadístico de Pearson*

Es cierto que desde la óptica de nuestro problema, la cópula de Bernstein es una buena aproximación a la subcópula que optimiza el estadístico de Pearson pero podríamos estar más interesados en funciones cópula que supongan una interpolación “exacta” de la misma pues de esta forma presentará el mismo valor para dicho estadístico.

4.2.3 Método de interpolación mediante polinomios cúbicos de Hermite

En este apartado comentaremos la forma en que se propone construir una cópula por interpolación cúbica de la subcópula cuyos valores en los puntos de su dominio proporciona el problema de optimización. Vamos a construir una función a trozos que vaya interpolando los valores que toma la subcópula en cada uno de los 4 vértices de cada uno de los $n \times m$ rectángulos en los que ha sido dividido el cuadrado unidad. Así, dado un rectángulo $[u_i, u_{i+1}] \times [v_j, v_{j+1}]$ (lógicamente los vértices van a ser puntos consecutivos dentro de la partición realizada en cada eje), consideremos los valores que ha determinado óptimos el problema de programación matemática ($C^*(u_i, v_j)$, $C^*(u_{i+1}, v_j)$, $C^*(u_i, v_{j+1})$ y $C^*(u_{i+1}, v_{j+1})$) para que la subcópula por ellos definida minimice el valor del estadístico de Pearson. Nuestro objetivo es definir una superficie cúbica que pase por todos ellos.

En primer lugar es necesario seleccionar el tipo de funciones que se utilizará para llevar a cabo dicha interpolación. Tal y como se describe en [FOVAFEHU] una posible alternativa son los polinomios de Hermite. De acuerdo a éstos, la ecuación que define una superficie interpoladora entre los valores que toma una subcópula cualquiera C' (en nuestro caso $C' = C^*$) en los vértices (u_i, v_j) , (u_{i+1}, v_j) , (u_i, v_{j+1}) y (u_{i+1}, v_{j+1}) es la siguiente:

$$x(u, v) = U * M_{H_u} * G_{H_x} * M_{H_v}^T * V^T \quad (4.13)$$

siendo

$$U = \left[\left(\frac{u-u_i}{u_{i+1}-u_i} \right)^3 \left(\frac{u-u_i}{u_{i+1}-u_i} \right)^2 \left(\frac{u-u_i}{u_{i+1}-u_i} \right) 1 \right] \quad (4.14)$$

$$V = \left[\left(\frac{v-v_j}{v_{j+1}-v_j} \right)^3 \left(\frac{v-v_j}{v_{j+1}-v_j} \right)^2 \left(\frac{v-v_j}{v_{j+1}-v_j} \right) 1 \right] \quad (4.15)$$

$$M_{H_u} = \begin{bmatrix} 2 & -2 & u_{i+1}-u_i & u_{i+1}-u_i \\ -3 & 3 & -2*(u_{i+1}-u_i) & -(u_{i+1}-u_i) \\ 0 & 0 & u_{i+1}-u_i & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \quad (4.16)$$

$$M_{H_v}^T = \begin{bmatrix} 2 & -3 & 0 & 1 \\ -2 & 3 & 0 & 0 \\ v_{j+1}-v_j & -2*(v_{j+1}-v_j) & v_{j+1}-v_j & 0 \\ v_{j+1}-v_j & -(v_{j+1}-v_j) & 0 & 0 \end{bmatrix} \quad (4.17)$$

$$G_{H_x} = \begin{bmatrix} C'(u_i, v_j) & C'(u_i, v_{j+1}) & \frac{\partial C'(u_i, v_j)}{\partial v} & \frac{\partial C'(u_i, v_{j+1})}{\partial v} \\ C'(u_{i+1}, v_j) & C'(u_{i+1}, v_{j+1}) & \frac{\partial C'(u_{i+1}, v_j)}{\partial v} & \frac{\partial C'(u_{i+1}, v_{j+1})}{\partial v} \\ \frac{\partial C'(u_i, v_j)}{\partial u} & \frac{\partial C'(u_i, v_{j+1})}{\partial u} & \frac{\partial^2 C'(u_i, v_j)}{\partial u \partial v} & \frac{\partial^2 C'(u_i, v_{j+1})}{\partial u \partial v} \\ \frac{\partial C'(u_{i+1}, v_j)}{\partial u} & \frac{\partial C'(u_{i+1}, v_{j+1})}{\partial u} & \frac{\partial^2 C'(u_{i+1}, v_j)}{\partial u \partial v} & \frac{\partial^2 C'(u_{i+1}, v_{j+1})}{\partial u \partial v} \end{bmatrix} \quad (4.18)$$

Los vectores U y V definen las incógnitas asociadas a los polinomios de Hermite, los cuales responden a una estructura del tipo $x(t) = a_x * t^3 + b_x * t^2 + c_x * t + d_x$. Los coeficientes de estos polinomios se estiman a través de la **matriz base de Hermite** (M_H) y la **matriz de geometría de Hermite** (G_H). Esta última consta de cuatro bloques (matrices 2 x 2) asociados a cada una de las esquinas:

- El primero de los bloques (esquina superior izquierda) va asociado a los cuatro valores que van a ser interpolados; en nuestro caso, se refiere al valor que toma la subcópula C' en cada vértice del rectángulo $[u_i, u_{i+1}] \times [v_j, v_{j+1}]$.

- Los bloques segundo y tercero (esquinas superior derecha e inferior izquierda) van asociados a las derivadas respecto de la segunda y primera de las variables respectivamente en cada uno de los vértices de la superficie interpoladora; en nuestro caso, van referidos a las coordenadas de los vectores tangentes a la superficie interpoladora en cada uno de los vértices del rectángulo $[u_i, u_{i+1}] \times [v_j, v_{j+1}]$.
- El bloque cuarto (esquina inferior derecha) está constituido por las derivadas segundas respecto de una y otra variable en cada uno de los vértices de la superficie interpoladora. Cuando estos cuatro valores valen 0, la superficie interpoladora se denomina **superficie de Ferguson**.

Como resultado de desarrollar la ecuación (4.13) se obtiene la ecuación de la función que interpola los valores de la subcópula C' en cada uno de los vértices del rectángulo $[u_i, u_{i+1}] \times [v_j, v_{j+1}]$:

$$\begin{aligned}
C_{HERMITE}(u, v) = & \left[\left(\frac{2*(u-u_i)^3}{(u_{i+1}-u_i)^3} - \frac{3*(u-u_i)^2}{(u_{i+1}-u_i)^2} + 1 \right) * C'(u_i, v_j) + \left(\frac{-2*(u-u_i)^3}{(u_{i+1}-u_i)^3} + \frac{3*(u-u_i)^2}{(u_{i+1}-u_i)^2} \right) * C'(u_{i+1}, v_j) + \right. \\
& \left. \left(\frac{(u-u_i)^3}{(u_{i+1}-u_i)^2} - \frac{2*(u-u_i)^2}{(u_{i+1}-u_i)} + (u-u_i) \right) * \frac{\partial C'(u_i, v_j)}{\partial u} + \left(\frac{(u-u_i)^3}{(u_{i+1}-u_i)^2} - \frac{(u-u_i)^2}{(u_{i+1}-u_i)} \right) * \frac{\partial C'(u_{i+1}, v_j)}{\partial u} \right] * \\
& \left(\frac{2*(v-v_j)^3}{(v_{j+1}-v_j)^3} - \frac{3*(v-v_j)^2}{(v_{j+1}-v_j)^2} + 1 \right) + \\
& \left[\left(\frac{2*(u-u_i)^3}{(u_{i+1}-u_i)^3} - \frac{3*(u-u_i)^2}{(u_{i+1}-u_i)^2} + 1 \right) * C'(u_i, v_{j+1}) + \left(\frac{-2*(u-u_i)^3}{(u_{i+1}-u_i)^3} + \frac{3*(u-u_i)^2}{(u_{i+1}-u_i)^2} \right) * C'(u_{i+1}, v_{j+1}) + \right. \\
& \left. \left(\frac{(u-u_i)^3}{(u_{i+1}-u_i)^2} - \frac{2*(u-u_i)^2}{(u_{i+1}-u_i)} + (u-u_i) \right) * \frac{\partial C'(u_i, v_{j+1})}{\partial u} + \left(\frac{(u-u_i)^3}{(u_{i+1}-u_i)^2} - \frac{(u-u_i)^2}{(u_{i+1}-u_i)} \right) * \frac{\partial C'(u_{i+1}, v_{j+1})}{\partial u} \right] * \\
& \left(\frac{-2*(v-v_j)^3}{(v_{j+1}-v_j)^3} + \frac{3*(v-v_j)^2}{(v_{j+1}-v_j)^2} \right) + \\
& \left[\left(\frac{2*(u-u_i)^3}{(u_{i+1}-u_i)^3} - \frac{3*(u-u_i)^2}{(u_{i+1}-u_i)^2} + 1 \right) * \frac{\partial C'(u_i, v_j)}{\partial v} + \left(\frac{-2*(u-u_i)^3}{(u_{i+1}-u_i)^3} + \frac{3*(u-u_i)^2}{(u_{i+1}-u_i)^2} \right) * \frac{\partial C'(u_{i+1}, v_j)}{\partial v} + \right. \\
& \left. \left(\frac{(u-u_i)^3}{(u_{i+1}-u_i)^2} - \frac{2*(u-u_i)^2}{(u_{i+1}-u_i)} + (u-u_i) \right) * \frac{\partial^2 C'(u_i, v_j)}{\partial u \partial v} + \left(\frac{(u-u_i)^3}{(u_{i+1}-u_i)^2} - \frac{(u-u_i)^2}{(u_{i+1}-u_i)} \right) * \frac{\partial^2 C'(u_{i+1}, v_j)}{\partial u \partial v} \right] * \\
& \left(\frac{(v-v_j)^3}{(v_{j+1}-v_j)^2} - \frac{2*(v-v_j)^2}{(v_{j+1}-v_j)} + (v-v_j) \right) + \\
& \left[\left(\frac{2*(u-u_i)^3}{(u_{i+1}-u_i)^3} - \frac{3*(u-u_i)^2}{(u_{i+1}-u_i)^2} + 1 \right) * \frac{\partial C'(u_i, v_{j+1})}{\partial v} + \left(\frac{-2*(u-u_i)^3}{(u_{i+1}-u_i)^3} + \frac{3*(u-u_i)^2}{(u_{i+1}-u_i)^2} \right) * \frac{\partial C'(u_{i+1}, v_{j+1})}{\partial v} + \right. \\
& \left. \left(\frac{(u-u_i)^3}{(u_{i+1}-u_i)^2} - \frac{2*(u-u_i)^2}{(u_{i+1}-u_i)} + (u-u_i) \right) * \frac{\partial^2 C'(u_i, v_{j+1})}{\partial u \partial v} + \left(\frac{(u-u_i)^3}{(u_{i+1}-u_i)^2} - \frac{(u-u_i)^2}{(u_{i+1}-u_i)} \right) * \frac{\partial^2 C'(u_{i+1}, v_{j+1})}{\partial u \partial v} \right] * \\
& \left(\frac{(v-v_j)^3}{(v_{j+1}-v_j)^2} - \frac{(v-v_j)^2}{(v_{j+1}-v_j)} \right)
\end{aligned}$$

$$\forall (u, v) \in [u_i, u_{i+1}] \times [v_j, v_{j+1}]$$

(4.19)

Se trata ahora de saber cuáles son las condiciones que debe cumplir esta función interpoladora para ser una cópula.

4.2.3.1 Condiciones a imponer para que la función construida por interpolación cúbica verifique las propiedades frontera de las cópulas

Veremos que las restricciones a imponer para que se cumplan las dos primeras condiciones que caracterizan a las funciones cópula, (1.5) y (1.6) (condiciones frontera), van a ser independientes del hecho de que la superficie interpoladora sea o no de Ferguson, es decir de que los términos asociados a las derivadas de segundo orden sean o no distintos de cero. Cuando hayamos de determinar las restricciones a imponer para que se cumpla la propiedad (1.7) (función 2-creciente), sí nos va a resultar más cómodo empezar considerando este tipo de superficies para que los desarrollos numéricos sean más sencillos y estableceremos, en un paso posterior, las condiciones a satisfacer por una superficie genérica cualquiera construida por interpolación mediante polinomios de Hermite para que el resultado sea una función cópula. Los dos lemas que se exponen a continuación vienen a determinar unas condiciones suficientes para garantizar que la función cúbica (4.19) verifica las propiedades frontera de las cópulas.

Lema 4.2: Condición suficiente para que se cumpla la primera de las propiedades de las funciones cópula

“Sea una partición del cuadrado unidad en $n \times m$ rectángulos del mismo área y sea la solución óptima al problema de programación fraccional (4.7) dada por $c_{ij}^* = C^*(\frac{i}{n}, \frac{j}{m}) \forall i \in \{0, 1, 2, \dots, n\} \quad \forall j \in \{0, 1, 2, \dots, m\}$ que define una subcópula C^* de dominio los puntos de intersección entre los rectángulos que originan la partición.

Sea $C_{HERMITE}(u, v)$ la función que se obtiene por interpolación cúbica de C^* mediante polinomios de Hermite.

Si $\frac{\partial C_{HERMITE}(u_k, 0)}{\partial u} = 0 \forall k \in \{0, 1, 2, \dots, n\}$ y $\frac{\partial C_{HERMITE}(0, v_l)}{\partial v} = 0 \forall l \in \{0, 1, 2, \dots, m\}$, entonces la función $C_{HERMITE}(u, v)$ cumple la primera de las condiciones para ser cópula: $C_{HERMITE}(u, 0) = C_{HERMITE}(0, v) = 0$ para cualesquiera u y v ”.

La demostración de este resultado puede ser consultada en el Anexo II.

Observaremos que las condiciones han sido impuestas para $\frac{\partial C_{HERMITE}(u_k, 0)}{\partial u}$ y $\frac{\partial C_{HERMITE}(0, v_l)}{\partial v}$ pero no para

$\frac{\partial C_{HERMITE}(u_k, 0)}{\partial v}$ y $\frac{\partial C_{HERMITE}(0, v_l)}{\partial u}$ que no tienen por qué valer 0. También conviene observar que no ha sido necesario imponer que C sea una superficie de Ferguson.

Lema 4.3: Condición suficiente para que se cumpla la segunda de las propiedades de las funciones cópula

“Sea una partición del cuadrado unidad en $n \times m$ rectángulos del mismo área y sea la solución óptima al problema de programación fraccional (4.7) dada por $c_{ij}^* = C^*(\frac{i}{n}, \frac{j}{m}) \forall i \in \{0,1,2,\dots,n\} \quad \forall j \in \{0,1,2,\dots,m\}$ que define una subcópula C^* de dominio los puntos de intersección entre los rectángulos que originan la partición.

Sea $C_{HERMITE}(u,v)$ la función que se obtiene por interpolación cúbica de C^* mediante polinomios de Hermite.

Si $\frac{\partial C_{HERMITE}(u_k,1)}{\partial u} = 1 \forall k \in \{0,1,2,\dots,n\}$ y $\frac{\partial C_{HERMITE}(1,v_l)}{\partial v} = 1 \forall l \in \{0,1,2,\dots,m\}$, entonces la función $C_{HERMITE}(u,v)$ cumple la segunda de las condiciones para ser cópula: $C_{HERMITE}(u,1) = u$ y $C_{HERMITE}(1,v) = v$ para cualesquiera u y v , es decir, C tiene marginales uniformes”.

La demostración de este resultado puede ser consultada en el Anexo III.

Observemos nuevamente que las condiciones se imponen sobre $\frac{\partial C_{HERMITE}(u_k,1)}{\partial u}$ y $\frac{\partial C_{HERMITE}(1,v_l)}{\partial v}$, no siendo

necesario que $\frac{\partial C_{HERMITE}(u_k,1)}{\partial v}$ y $\frac{\partial C_{HERMITE}(1,v_l)}{\partial u}$ tomen el valor 1. De igual manera, tampoco ha sido necesario imponer que C sea una superficie de Ferguson.

4.2.3.2 Condiciones a imponer para que la función construida por interpolación cúbica verifique la propiedad 2-creciente de las cópulas

A partir de estos dos lemas que hacen referencia a las condiciones frontera de las cópulas, bastaría saber qué condiciones adicionales son suficientes añadir para que la función C , obtenida por interpolación cúbica, sea 2-creciente (véase expresión (1.7)) y, en consecuencia, una función cópula. Para ello, basta aplicar la expresión (4.19) sobre un rectángulo B cualquiera y estudiar las restricciones que es preciso imponer sobre los elementos variables de dicha expresión para que el resultado sea mayor o igual que cero. Analizando (4.19) podemos distinguir tantos bloques como los que fueron identificados en la matriz (4.18). Los valores interpolados $C(u_i, v_j)$ que aparecen en el bloque 1 son los únicos que vienen dados de antemano como solución del problema (4.7). Por tanto, los elementos variables de (4.19) a los que nos estamos refiriendo son las derivadas de primer y segundo orden respecto de U y V en los vértices (u_i, v_j) de la partición.

4.2.3.2.1 Condición a imponer para que una superficie interpoladora de Ferguson sea 2-creciente

Llegados a este punto, vamos a empezar considerando una simplificación que nos permita eliminar variables en los primeros desarrollos. Así, partiremos de una superficie interpoladora de Ferguson, es decir, anularemos las derivadas segundas de la función respecto de U y V , si bien, como decíamos, extenderemos después el resultado a una superficie de Hermite cualquiera. Sea $C_{FERGUSON}$ la superficie que resulta de hacer 0 en (4.19) las derivadas de segundo orden, es decir,

$$\begin{aligned}
C_{FERGUSON}(u, v) = & \left[\left(\frac{2*(u-u_i)^3}{(u_{i+1}-u_i)^3} - \frac{3*(u-u_i)^2}{(u_{i+1}-u_i)^2} + 1 \right) * C'(u_i, v_j) + \left(\frac{-2*(u-u_i)^3}{(u_{i+1}-u_i)^3} + \frac{3*(u-u_i)^2}{(u_{i+1}-u_i)^2} \right) * C'(u_{i+1}, v_j) + \right. \\
& \left. \left(\frac{(u-u_i)^3}{(u_{i+1}-u_i)^2} - \frac{2*(u-u_i)^2}{(u_{i+1}-u_i)} + (u-u_i) \right) * \frac{\partial C'(u_i, v_j)}{\partial u} + \left(\frac{(u-u_i)^3}{(u_{i+1}-u_i)^2} - \frac{(u-u_i)^2}{(u_{i+1}-u_i)} \right) * \frac{\partial C'(u_{i+1}, v_j)}{\partial u} \right] * \\
& \left(\frac{2*(v-v_j)^3}{(v_{j+1}-v_j)^3} - \frac{3*(v-v_j)^2}{(v_{j+1}-v_j)^2} + 1 \right) + \\
& \left[\left(\frac{2*(u-u_i)^3}{(u_{i+1}-u_i)^3} - \frac{3*(u-u_i)^2}{(u_{i+1}-u_i)^2} + 1 \right) * C'(u_i, v_{j+1}) + \left(\frac{-2*(u-u_i)^3}{(u_{i+1}-u_i)^3} + \frac{3*(u-u_i)^2}{(u_{i+1}-u_i)^2} \right) * C'(u_{i+1}, v_{j+1}) + \right. \\
& \left. \left(\frac{(u-u_i)^3}{(u_{i+1}-u_i)^2} - \frac{2*(u-u_i)^2}{(u_{i+1}-u_i)} + (u-u_i) \right) * \frac{\partial C'(u_i, v_{j+1})}{\partial u} + \left(\frac{(u-u_i)^3}{(u_{i+1}-u_i)^2} - \frac{(u-u_i)^2}{(u_{i+1}-u_i)} \right) * \frac{\partial C'(u_{i+1}, v_{j+1})}{\partial u} \right] * \\
& \left(\frac{-2*(v-v_j)^3}{(v_{j+1}-v_j)^3} + \frac{3*(v-v_j)^2}{(v_{j+1}-v_j)^2} \right) + \\
& \left[\left(\frac{2*(u-u_i)^3}{(u_{i+1}-u_i)^3} - \frac{3*(u-u_i)^2}{(u_{i+1}-u_i)^2} + 1 \right) * \frac{\partial C'(u_i, v_j)}{\partial v} + \left(\frac{-2*(u-u_i)^3}{(u_{i+1}-u_i)^3} + \frac{3*(u-u_i)^2}{(u_{i+1}-u_i)^2} \right) * \frac{\partial C'(u_{i+1}, v_j)}{\partial v} \right] * \\
& \left(\frac{(v-v_j)^3}{(v_{j+1}-v_j)^2} - \frac{2*(v-v_j)^2}{(v_{j+1}-v_j)} + (v-v_j) \right) + \\
& \left[\left(\frac{2*(u-u_i)^3}{(u_{i+1}-u_i)^3} - \frac{3*(u-u_i)^2}{(u_{i+1}-u_i)^2} + 1 \right) * \frac{\partial C'(u_i, v_{j+1})}{\partial v} + \left(\frac{-2*(u-u_i)^3}{(u_{i+1}-u_i)^3} + \frac{3*(u-u_i)^2}{(u_{i+1}-u_i)^2} \right) * \frac{\partial C'(u_{i+1}, v_{j+1})}{\partial v} \right] * \\
& \left(\frac{(v-v_j)^3}{(v_{j+1}-v_j)^2} - \frac{(v-v_j)^2}{(v_{j+1}-v_j)} \right)
\end{aligned}$$

$$\forall (u, v) \in [u_i, u_{i+1}] \times [v_j, v_{j+1}]$$

(4.20)

Podemos comprobar que el resultado de aplicar esta expresión a cada uno de los cuatro vértices de un rectángulo $B \subset I^2$ va a dar lugar a otra en la que los únicos factores desconocidos van a ser las derivadas respecto de U y V . Efectivamente, como ya indicamos, los valores de C' en cada par (u_i, v_j) serán la solución dada por el problema de optimización de la subcópula (4.7). Por otra parte, los valores u_i y v_j serán también conocidos dado que son los vértices

$\left(\frac{i}{n}, \frac{j}{m} \right)$ de la partición. Finalmente los valores “ u ” y “ v ” irán referidos a cada uno de los vértices del rectángulo B .

El resultado que permite establecer las restricciones que garantizan la “2-crecencia” de C proporciona unos rangos de variación para las derivadas primeras. Estos rangos vienen determinados por la región factible asociada al siguiente problema de programación matemática no lineal y entera:

$$\text{Max/Min} \sum_{i=0}^{n-1} \sum_{j=0}^{m-1} \left[\frac{\partial C'(u_i, v_j)}{\partial u} + \frac{\partial C'(u_i, v_j)}{\partial v} \right]$$

Restricciones asociadas a las derivadas respecto de U

$$\begin{aligned} & \frac{2*(u_{i+1}-u_i)}{V_c'(B)} * \left[\frac{\partial C'(u_i, v_j)}{\partial u} - \frac{\partial C'(u_i, v_{j+1})}{\partial u} \right] + \frac{2*2*(u_{i+1}-u_i)}{V_c'(B)} * \left[\frac{\partial C'(u_{i+1}, v_j)}{\partial u} - \frac{\partial C'(u_{i+1}, v_{j+1})}{\partial u} \right] + 3 \leq 3*(1-\delta_1) \\ & -2 * \frac{2*(u_{i+1}-u_i)}{V_c'(B)} * \left[\frac{\partial C'(u_i, v_j)}{\partial u} - \frac{\partial C'(u_i, v_{j+1})}{\partial u} \right] - \frac{2*(u_{i+1}-u_i)}{V_c'(B)} * \left[\frac{\partial C'(u_{i+1}, v_j)}{\partial u} - \frac{\partial C'(u_{i+1}, v_{j+1})}{\partial u} \right] - 3 \leq \left(\frac{24+15*\sqrt{2}}{7} \right) * (1-\delta_1) \\ & 2 * \frac{2*(u_{i+1}-u_i)}{V_c'(B)} * \left[\frac{\partial C'(u_i, v_j)}{\partial u} - \frac{\partial C'(u_i, v_{j+1})}{\partial u} \right] + \frac{2*(u_{i+1}-u_i)}{V_c'(B)} * \left[\frac{\partial C'(u_{i+1}, v_j)}{\partial u} - \frac{\partial C'(u_{i+1}, v_{j+1})}{\partial u} \right] + 3 \leq 3*(1-\delta_2) \\ & - \frac{2*(u_{i+1}-u_i)}{V_c'(B)} * \left[\frac{\partial C'(u_i, v_j)}{\partial u} - \frac{\partial C'(u_i, v_{j+1})}{\partial u} \right] - \frac{2*2*(u_{i+1}-u_i)}{V_c'(B)} * \left[\frac{\partial C'(u_{i+1}, v_j)}{\partial u} - \frac{\partial C'(u_{i+1}, v_{j+1})}{\partial u} \right] - 3 \leq \left(\frac{24+15*\sqrt{2}}{7} \right) * (1-\delta_2) \\ & - \frac{2*(u_{i+1}-u_i)}{V_c'(B)} * \left[\frac{\partial C'(u_i, v_j)}{\partial u} - \frac{\partial C'(u_i, v_{j+1})}{\partial u} \right] - \frac{2*(u_{i+1}-u_i)}{V_c'(B)} * \left[\frac{\partial C'(u_{i+1}, v_j)}{\partial u} - \frac{\partial C'(u_{i+1}, v_{j+1})}{\partial u} \right] - 2 \leq 4*(1-\delta_3) \\ & \left\{ \frac{2*(u_{i+1}-u_i)}{V_c'(B)} * \left[\frac{\partial C'(u_i, v_j)}{\partial u} - \frac{\partial C'(u_i, v_{j+1})}{\partial u} \right] + \frac{2*(u_{i+1}-u_i)}{V_c'(B)} * \left[\frac{\partial C'(u_{i+1}, v_j)}{\partial u} - \frac{\partial C'(u_{i+1}, v_{j+1})}{\partial u} \right] + 3 \right\}^2 - \\ & \frac{2*(u_{i+1}-u_i)}{V_c'(B)} * \left[\frac{\partial C'(u_i, v_j)}{\partial u} - \frac{\partial C'(u_i, v_{j+1})}{\partial u} \right] * \frac{2*(u_{i+1}-u_i)}{V_c'(B)} * \left[\frac{\partial C'(u_{i+1}, v_j)}{\partial u} - \frac{\partial C'(u_{i+1}, v_{j+1})}{\partial u} \right] \leq 9*(1-\delta_4) \\ & \left[\frac{\partial C'(u_i, v_j)}{\partial u} - \frac{\partial C'(u_i, v_{j+1})}{\partial u} \right] \leq 0 \\ & \left[\frac{\partial C'(u_{i+1}, v_j)}{\partial u} - \frac{\partial C'(u_{i+1}, v_{j+1})}{\partial u} \right] \leq 0 \end{aligned}$$

Restricciones asociadas a las derivadas respecto de V

$$\begin{aligned} & \frac{2*(v_{j+1}-v_j)}{V_c'(B)} * \left[\frac{\partial C'(u_i, v_j)}{\partial v} - \frac{\partial C'(u_{i+1}, v_j)}{\partial v} \right] + \frac{2*2*(v_{j+1}-v_j)}{V_c'(B)} * \left[\frac{\partial C'(u_i, v_{j+1})}{\partial v} - \frac{\partial C'(u_{i+1}, v_{j+1})}{\partial v} \right] + 3 \leq 3*(1-\gamma_1) \\ & -2 * \frac{2*(v_{j+1}-v_j)}{V_c'(B)} * \left[\frac{\partial C'(u_i, v_j)}{\partial v} - \frac{\partial C'(u_{i+1}, v_j)}{\partial v} \right] - \frac{2*(v_{j+1}-v_j)}{V_c'(B)} * \left[\frac{\partial C'(u_i, v_{j+1})}{\partial v} - \frac{\partial C'(u_{i+1}, v_{j+1})}{\partial v} \right] - 3 \leq \left(\frac{24+15*\sqrt{2}}{7} \right) * (1-\gamma_1) \\ & 2 * \frac{2*(v_{j+1}-v_j)}{V_c'(B)} * \left[\frac{\partial C'(u_i, v_j)}{\partial v} - \frac{\partial C'(u_{i+1}, v_j)}{\partial v} \right] + \frac{2*(v_{j+1}-v_j)}{V_c'(B)} * \left[\frac{\partial C'(u_i, v_{j+1})}{\partial v} - \frac{\partial C'(u_{i+1}, v_{j+1})}{\partial v} \right] + 3 \leq 3*(1-\gamma_2) \\ & - \frac{2*(v_{j+1}-v_j)}{V_c'(B)} * \left[\frac{\partial C'(u_i, v_j)}{\partial v} - \frac{\partial C'(u_{i+1}, v_j)}{\partial v} \right] - \frac{2*2*(v_{j+1}-v_j)}{V_c'(B)} * \left[\frac{\partial C'(u_i, v_{j+1})}{\partial v} - \frac{\partial C'(u_{i+1}, v_{j+1})}{\partial v} \right] - 3 \leq \left(\frac{24+15*\sqrt{2}}{7} \right) * (1-\gamma_2) \\ & - \frac{2*(v_{j+1}-v_j)}{V_c'(B)} * \left[\frac{\partial C'(u_i, v_j)}{\partial v} - \frac{\partial C'(u_{i+1}, v_j)}{\partial v} \right] - \frac{2*(v_{j+1}-v_j)}{V_c'(B)} * \left[\frac{\partial C'(u_i, v_{j+1})}{\partial v} - \frac{\partial C'(u_{i+1}, v_{j+1})}{\partial v} \right] - 2 \leq 4*(1-\gamma_3) \\ & \left\{ \frac{2*(v_{j+1}-v_j)}{V_c'(B)} * \left[\frac{\partial C'(u_i, v_j)}{\partial v} - \frac{\partial C'(u_{i+1}, v_j)}{\partial v} \right] + \frac{2*(v_{j+1}-v_j)}{V_c'(B)} * \left[\frac{\partial C'(u_i, v_{j+1})}{\partial v} - \frac{\partial C'(u_{i+1}, v_{j+1})}{\partial v} \right] + 3 \right\}^2 - \\ & \frac{2*(v_{j+1}-v_j)}{V_c'(B)} * \left[\frac{\partial C'(u_i, v_j)}{\partial v} - \frac{\partial C'(u_{i+1}, v_j)}{\partial v} \right] * \frac{2*(v_{j+1}-v_j)}{V_c'(B)} * \left[\frac{\partial C'(u_i, v_{j+1})}{\partial v} - \frac{\partial C'(u_{i+1}, v_{j+1})}{\partial v} \right] \leq 9*(1-\gamma_4) \\ & \left[\frac{\partial C'(u_i, v_j)}{\partial v} - \frac{\partial C'(u_{i+1}, v_j)}{\partial v} \right] \leq 0 \\ & \left[\frac{\partial C'(u_i, v_{j+1})}{\partial v} - \frac{\partial C'(u_{i+1}, v_{j+1})}{\partial v} \right] \leq 0 \end{aligned}$$

$$\delta_1 + \delta_2 + \delta_3 + \delta_4 = 1$$

$$\gamma_1 + \gamma_2 + \gamma_3 + \gamma_4 = 1$$

$$\delta_1, \delta_2, \delta_3, \delta_4 \in \{0, 1\}$$

$$\gamma_1, \gamma_2, \gamma_3, \gamma_4 \in \{0, 1\}$$

$$\forall i \in \{0, 1, \dots, n-1\} \quad j \in \{0, 1, \dots, m-1\}$$

(4.21)

Obsérvese que existen 2 bloques de restricciones (asociados a los vectores tangentes en las variables U y V) que a su vez se subdividen en “ n ” y “ m ” bloques respectivamente. De ahí que esté planteado para cualesquiera “ i ” y “ j ” comprendidos entre 0 y $n-1$ y 0 y $m-1$ respectivamente. Dando solución a este problema, se dispondrá de los vectores tangentes a la superficie interpoladora en los puntos del dominio de definición de la subcópula. Los valores de estos

vectores en dichos puntos, valores que constituyen las variables de decisión del problema, son siempre positivos pues el hecho de que la cópula sea 2-creciente implica que sea no decreciente en cada una de sus componentes (véase lema 2.1.4 de [NELSEN]) y en consecuencia las derivadas en todos sus puntos deberán ser mayores o iguales que cero. A raíz de esta última observación hemos propuesto como función objetivo del problema, maximizar o minimizar la suma de estos valores positivos con la idea de aumentar o disminuir las pendientes de crecimiento de la distribución cópula en los puntos citados. Así, conforme por ejemplo crezcan las pendientes (si se maximiza la función objetivo), más grandes serán los saltos en términos de probabilidad de la función a construir y menos suave resultará la correspondiente densidad. En cualquier caso, dada una solución del problema (4.21) y teniendo en cuenta que los valores que toma la cópula ya han sido estimados (aquellos que minimizan el estadístico de Pearson) y que las derivadas de segundo orden valen 0 (por ser una superficie de Ferguson), se dispondrá de todos los elementos que componen la expresión (4.20). Podemos proponer entonces el siguiente resultado:

Lema 4.4: Condición suficiente para que se cumpla la tercera de las propiedades de las funciones cópula en una superficie interpoladora de Ferguson

“Sea una partición del cuadrado unidad en $n \times m$ rectángulos del mismo área y sea la solución óptima al problema de programación fraccional (4.7) dada por $c_{ij}^* = C^*\left(\frac{i}{n}, \frac{j}{m}\right) \forall i \in \{0,1,2,\dots,n\} \quad \forall j \in \{0,1,2,\dots,m\}$ que define una subcópula C^* de dominio los puntos de intersección entre los rectángulos que originan la partición.

Sea $C_{HERMITE}(u,v)$ la función que se obtiene por interpolación cúbica de una subcópula C^* mediante polinomios de Hermite. Supongamos que las derivadas de segundo orden en el dominio de definición de la subcópula valen cero, es

decir, $\frac{\partial^2 C_{HERMITE}(u_i, v_j)}{\partial u \partial v} = 0 \quad \forall i \in \{0,1,2,\dots,n\} \quad \forall j \in \{0,1,2,\dots,m\}$, en cuyo caso, $C_{HERMITE}(u,v)$ define una

superficie interpoladora de Ferguson $C_{FERGUSON}(u,v)$.

Si $\frac{\partial C_{FERGUSON}(u_i, v_j)}{\partial u}$ y $\frac{\partial C_{FERGUSON}(u_i, v_j)}{\partial v} \quad \forall i \in \{0,1,2,\dots,n\} \quad \forall j \in \{0,1,2,\dots,m\}$

proporcionan junto con una relación de δ 's y γ 's una solución del problema de programación no lineal entera (4.21), entonces la función $C_{FERGUSON}(u,v)$ cumple la tercera de las condiciones para ser cópula:

$$V_{C_{FERGUSON}}([u_1, u_2] \times [v_1, v_2]) = C_{FERGUSON}(u_2, v_2) - C_{FERGUSON}(u_2, v_1) - C_{FERGUSON}(u_1, v_2) + C_{FERGUSON}(u_1, v_1) \geq 0 \quad , ,$$

$$\forall u_1, u_2, v_1, v_2 \text{ t. q. } u_1 \leq u_2, v_1 \leq v_2$$

El lema, cuya demostración puede ser consultada en el Anexo IV, establece una relación de regiones a la que deben pertenecer las siguientes variables, para que la función dada por la ecuación (4.20) sea 2-creciente:

$$K_{ij} = \frac{2*(u_{i+1} - u_i)}{V_{C^*}(B)} * \left[\frac{\partial C'(u_i, v_j)}{\partial u} - \frac{\partial C'(u_i, v_{j+1})}{\partial u} \right] \quad (4.22)$$

$$L_{ij} = \frac{2*(u_{i+1} - u_i)}{V_{C^*}(B)} * \left[\frac{\partial C'(u_{i+1}, v_j)}{\partial u} - \frac{\partial C'(u_{i+1}, v_{j+1})}{\partial u} \right] \quad (4.23)$$

$$P_{ij} = \frac{2*(v_{j+1}-v_j)}{V_{C'}(B)} * \left[\frac{\partial C'(u_i, v_j)}{\partial v} - \frac{\partial C'(u_{i+1}, v_j)}{\partial v} \right] \quad (4.24)$$

$$Q_{ij} = \frac{2*(v_{j+1}-v_j)}{V_{C'}(B)} * \left[\frac{\partial C'(u_i, v_{j+1})}{\partial v} - \frac{\partial C'(u_{i+1}, v_{j+1})}{\partial v} \right] \quad (4.25)$$

Obsérvese que la función $C_{HERMITE}$ que hemos definido (y por tanto $C_{FERGUSON}$) ofrece un polinomio cúbico en V si se fija el valor de U y, de igual manera, ofrece un polinomio cúbico en U si se fija el valor de V . Por ello, una vez establecidas las condiciones bajo las cuales dicha función es una cópula, podremos asegurar también que va a tener secciones cúbicas en U y V (véanse las ecuaciones (3.3) y (3.4)). No es por ello de extrañar la simetría existente entre la región factible del problema (4.21) y la región que puede encontrarse en el capítulo 2 de [NELQUERO] (véase ilustración (4.5)). En su artículo, los autores presentan un método de generación de familias de cópulas bivariantes con secciones cúbicas. En concreto, algunos de los resultados más importantes a los que llegan vienen dados por los teoremas 2.4 y 4.1, que a continuación presentamos:

Teorema 4.2

“Sean $\alpha, \beta: [0,1] \rightarrow \mathbb{R}$ dos funciones que satisfacen $\alpha(0)=\alpha(1)=\beta(0)=\beta(1)=0$ y sea $C(x, y)=xy+x(1-x)[\alpha(y)(1-x)+\beta(y)x] \quad \forall (x, y) \in [0,1]^2$. Entonces C es una cópula si y sólo si:

- i) $\alpha(y)$ y $\beta(y)$ son absolutamente continuas y,
- ii) para casi todo $y \in [0,1]$, el punto $(\alpha'(y), \beta'(y))$ se encuentra en S . En otras palabras, para casi todo $y \in [0,1]$,
 - ii.1) o $-1 \leq \alpha'(y) \leq 2$ y $-2 \leq \beta'(y) \leq 1$
 - ii.2) o $[\alpha'(y)]^2 - \alpha'(y)\beta'(y) + [\beta'(y)]^2 - 3\alpha'(y) + 3\beta'(y) \leq 0$

Además, C es absolutamente continua”.

Teorema 4.3

“Supongamos que C tiene secciones cúbicas en X e Y , es decir, para casi todo $x, y \in [0,1]$ sea C , dada por :

$$C(x, y)=xy+x(1-x)[\alpha(y)(1-x)+\beta(y)x] \quad y \quad C(x, y)=xy+y(1-y)[\gamma(x)(1-y)+\delta(x)y] \quad \text{donde}$$

$$\alpha, \beta, \gamma, \delta: [0,1] \rightarrow \mathbb{R} \quad y \quad \text{tales que} \quad \alpha(0)=\alpha(1)=\beta(0)=\beta(1)=\gamma(0)=\gamma(1)=\delta(0)=\delta(1)=0 \quad .$$

Entonces, $C(x, y)=xy+xy(1-x)(1-y)[A_1(y)(1-x)+A_2(1-y)(1-x)+B_1xy+B_2x(1-y)]$ donde A_1, A_2, B_1, B_2 son constantes reales tales que los puntos $(A_2, A_1), (B_1, B_2), (B_1, A_1)$ y (A_2, B_2) están todos en S ”.

Es esta región S , a la que hace referencia el teorema, la que mantiene el parentesco con la región factible del problema (4.21). Basta hacer una traslación del origen (0,0) al punto (-1,1) y una simetría respecto del eje de ordenadas para poder conseguir una a partir de la otra. En esta ocasión, son las derivadas de primer orden de (α, β) las que deben pertenecer a la región S . En nuestro caso, dicho papel lo juegan las expresiones (4.22)-(4.25), que combinan los valores que deben tomar las derivadas primeras de la subcópula óptima en los nodos que resultan de hacer la partición de I^2 . La evaluación de dichas derivadas forma parte de las matrices de geometría de Hermite (véase (4.18)).

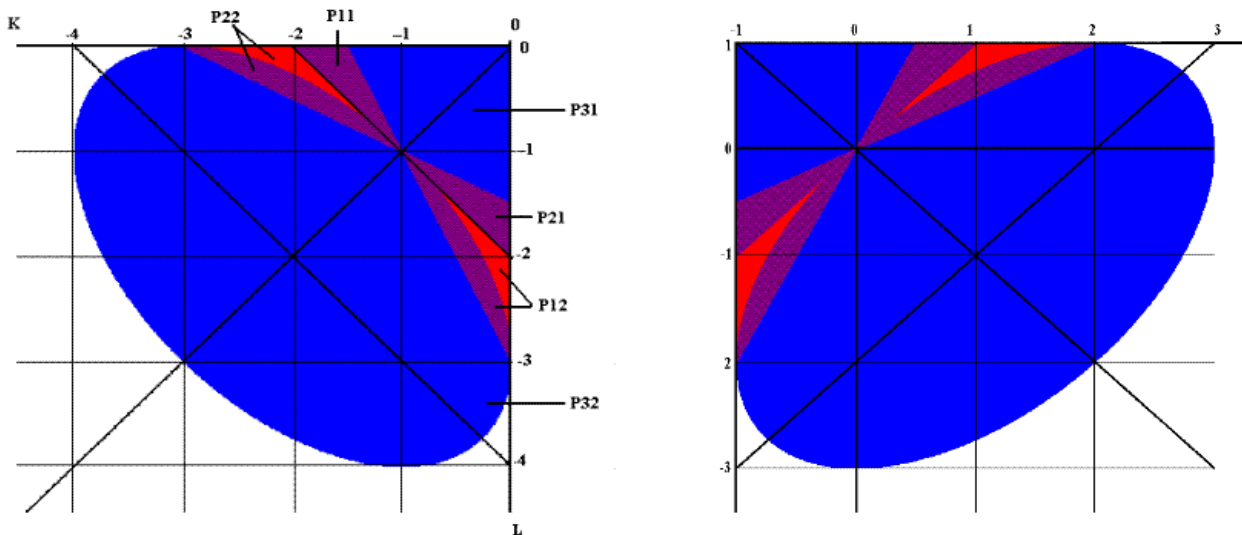


Ilustración 4.5: Región de factibilidad del problema que permite construir una cópula por interpolación cúbica

El lema (4.4) expuesto establece que, encontrando solución factible al problema (4.21), podremos construir una función 2-creciente. Sin embargo nada nos garantiza que esa solución vaya a existir siempre y por ello convendría estudiar bajo qué condiciones es posible determinar una relación de valores para que todas las derivadas primeras que participan en (4.20) (o equivalentemente las variables (4.22)-(4.25)) pertenezcan a una región de tipo S .

Comencemos analizando el modelo planteado por (4.21). Se trata de un problema de programación no lineal entera (MINLP). Ambas características vienen impuestas por la forma de la región factible presentada en la ilustración (4.5). Por un lado la no linealidad viene reflejada por la elipse contenida en la misma, mientras que por otro, el carácter entero surge de la necesidad de imponer restricciones disyuntivas ante la dificultad de expresar la región factible con una única ecuación. Como se explica en el apartado 10.6 de [SIXTO], los problemas de programación entera pueden ser resueltos mediante el método de ramificación y acotación también en el caso no lineal, con el inconveniente añadido de que en su resolución no está garantizada la optimalidad global. Sin embargo el principal problema que encontraremos de cara a la implementación informática de este problema es la escasez de software que permite resolver el MINLP. No habiendo encontrado a tal fin un procedimiento adecuado en la herramienta *SAS*, se propone resolver en su lugar otro problema cuya región factible no precise de ser expresada a trozos. En concreto, se propone como región factible asociada a cada una de las variables (K,L) (o (P,Q)) el rectángulo $[-3,0] \times [-3,0]$. Ésta permite obtener una expresión “relajada” (lineal y continua) del problema (4.21) que es la que a continuación se presenta:

$$\text{Max/Min} \sum_{i=0}^{n-1} \sum_{j=0}^{m-1} \left[\frac{\partial C'(u_i, v_j)}{\partial u} + \frac{\partial C'(u_i, v_j)}{\partial v} \right]$$

Restricciones asociadas a las derivadas respecto de U

$$\begin{aligned} -3 &\leq \frac{2*(u_{i+1}-u_i)}{V_C'(B)} * \left[\frac{\partial C'(u_i, v_j)}{\partial u} - \frac{\partial C'(u_i, v_{j+1})}{\partial u} \right] \leq 0 \\ -3 &\leq \frac{2*(u_{i+1}-u_i)}{V_C'(B)} * \left[\frac{\partial C'(u_{i+1}, v_j)}{\partial u} - \frac{\partial C'(u_{i+1}, v_{j+1})}{\partial u} \right] \leq 0 \end{aligned}$$

Restricciones asociadas a las derivadas respecto de V

$$\begin{aligned} -3 &\leq \frac{2*(v_{j+1}-v_j)}{V_C'(B)} * \left[\frac{\partial C'(u_i, v_j)}{\partial v} - \frac{\partial C'(u_{i+1}, v_j)}{\partial v} \right] \leq 0 \\ -3 &\leq \frac{2*(v_{j+1}-v_j)}{V_C'(B)} * \left[\frac{\partial C'(u_i, v_{j+1})}{\partial v} - \frac{\partial C'(u_{i+1}, v_{j+1})}{\partial v} \right] \leq 0 \end{aligned}$$

$$\forall i \in \{0, 1, \dots, n-1\} \quad j \in \{0, 1, \dots, m-1\}$$

(4.26)

El resultado equivalente al lema (4.4) para esta relajación del problema (4.21) es el siguiente:

Lema 4.5: Condición suficiente “relajada” para que se cumpla la tercera de las propiedades de las funciones cópula en una superficie interpoladora de Ferguson

“Sea una partición del cuadrado unidad en $n \times m$ rectángulos del mismo área y sea la solución óptima al problema de programación fraccional (4.7) dada por $c_{ij}^* = C^*\left(\frac{i}{n}, \frac{j}{m}\right) \forall i \in \{0, 1, 2, \dots, n\} \quad \forall j \in \{0, 1, 2, \dots, m\}$ que define una subcópula C^* de dominio los puntos de intersección entre los rectángulos que originan la partición.

Sea $C_{\text{HERMITE}}(u, v)$ la función que se obtiene por interpolación cúbica de una subcópula C^* mediante polinomios de Hermite. Supongamos que las derivadas de segundo orden en el dominio de definición de la subcópula valen cero, es

decir, $\frac{\partial^2 C_{\text{HERMITE}}(u_i, v_j)}{\partial u \partial v} = 0 \quad \forall i \in \{0, 1, 2, \dots, n\} \quad \forall j \in \{0, 1, 2, \dots, m\}$, en cuyo caso, $C_{\text{HERMITE}}(u, v)$ define una

superficie interpoladora de Ferguson $C_{\text{FERGUSON}}(u, v)$.

Si $\frac{\partial C_{\text{FERGUSON}}(u_i, v_j)}{\partial u}$ y $\frac{\partial C_{\text{FERGUSON}}(u_i, v_j)}{\partial v} \quad \forall i \in \{0, 1, 2, \dots, n\} \quad \forall j \in \{0, 1, 2, \dots, m\}$

proporcionan una solución del problema de programación matemática (4.26), entonces la función $C_{\text{FERGUSON}}(u, v)$ cumple la tercera de las condiciones para ser cópula:

$$\begin{aligned} V_{C_{\text{FERGUSON}}}([u_1, u_2] \times [v_1, v_2]) &= C_{\text{FERGUSON}}(u_2, v_2) - C_{\text{FERGUSON}}(u_2, v_1) - C_{\text{FERGUSON}}(u_1, v_2) + C_{\text{FERGUSON}}(u_1, v_1) \geq 0 \quad ,, \\ \forall u_1, u_2, v_1, v_2.t.q u_1 &\leq u_2, v_1 \leq v_2 \end{aligned}$$

Si bien parece claro a partir de la ilustración (4.5) que la nueva región factible (dada por el rectángulo $[-3, 0] \times [-3, 0]$) está contenida en la región factible del problema (4.21), proponemos consultar su demostración analítica (Anexo V).

Aunque resolver el problema alternativo (4.26) tiene el inconveniente de omitir soluciones factibles (todas aquellas que pertenezcan a la elipse y no estén contenidas en dicho rectángulo), presenta la ventaja de ser fácil de tratar por la mayoría de los paquetes estadísticos, dado que se trata de un problema de programación lineal continua que puede ser abordado aplicando el **algoritmo del Simplex**.

Como resumen de este apartado, podemos establecer el siguiente teorema a partir de los lemas (4.2), (4.3) y (4.4 o 4.5):

Teorema 4.4: Condición suficiente para poder generar una cópula por interpolación de una subcópula mediante una superficie de Ferguson

“Sea una partición del cuadrado unidad en $n \times m$ rectángulos del mismo área y sea la solución óptima al problema de programación fraccional (4.7) dada por $c_{ij}^* = C^*\left(\frac{i}{n}, \frac{j}{m}\right) \forall i \in \{0, 1, 2, \dots, n\} \quad \forall j \in \{0, 1, 2, \dots, m\}$ que define una subcópula C^* de dominio los puntos de intersección entre los rectángulos que originan la partición.

Sea $C_{FERGUSON}(u, v)$ la función que se obtiene por interpolación cúbica de una subcópula C^* mediante superficies de Ferguson (dada por (4.20)).

Si $\frac{\partial C_{FERGUSON}(u_i, v_j)}{\partial u}$ y $\frac{\partial C_{FERGUSON}(u_i, v_j)}{\partial v} \quad \forall i \in \{0, 1, 2, \dots, n\} \quad \forall j \in \{0, 1, 2, \dots, m\}$ cumplen las siguientes condiciones:

$$1. \quad \frac{\partial C_{FERGUSON}(u_k, 0)}{\partial u} = 0 \quad \forall k \in \{0, 1, 2, \dots, n\} \quad \text{y} \quad \frac{\partial C_{FERGUSON}(0, v_l)}{\partial v} = 0 \quad \forall l \in \{0, 1, 2, \dots, m\}$$

$$2. \quad \frac{\partial C_{FERGUSON}(u_k, 1)}{\partial u} = 1 \quad \forall k \in \{0, 1, 2, \dots, n\} \quad \text{y} \quad \frac{\partial C_{FERGUSON}(1, v_l)}{\partial v} = 1 \quad \forall l \in \{0, 1, 2, \dots, m\}$$

3. Proporcionan una solución del problema de programación no lineal entera (4.21) (o bien su relajación lineal (4.26))

entonces la función $C_{FERGUSON}$ es una cópula”.

A modo ilustrativo, adjuntamos el gráfico de una cópula construida siguiendo estas pautas para la relación de puntos presentados en la ilustración (4.1). La densidad cópula presenta un aspecto peculiar conformado por una sucesión de montículos asociados a cada uno de los rectángulos y que llegan a valer 0 en los vértices de los mismos. Ésta es una consecuencia directa de considerar una superficie interpoladora de Ferguson. Al imponer que las derivadas de segundo orden valgan 0 en dichos vértices, como la densidad cópula es a fin de cuenta una interpolación de las derivadas segundas respecto de U y V en dichos vértices (donde vale 0), la superficie resultante tiene drásticas caídas a 0. Al pie de dicha función cópula se proporciona el valor del estadístico de Pearson (5.007) que se observa no coincide con el mínimo valor que se obtuvo para la cópula que fue construida por interpolación bilineal (0.445). Esto es así porque, como veremos, la construcción de esta cópula por interpolación cúbica no está siempre garantizada, es decir, no existe siempre una solución factible asociada al problema de programación matemática (4.21). La solución al problema, pasa

por construir subcópulas que cumplen una propiedad más restrictiva que “ser 2-creciente”. De hecho veremos, que los rectángulos cuyos vértices son puntos del dominio de definición de la subcópula, deberán tener un volumen mínimo (estrictamente mayor que 0) cuyo valor dependerá de la dimensión de la partición. Es por ello lógico que el valor del estadístico de Pearson para estas cópulas más restrictivas sea mayor.

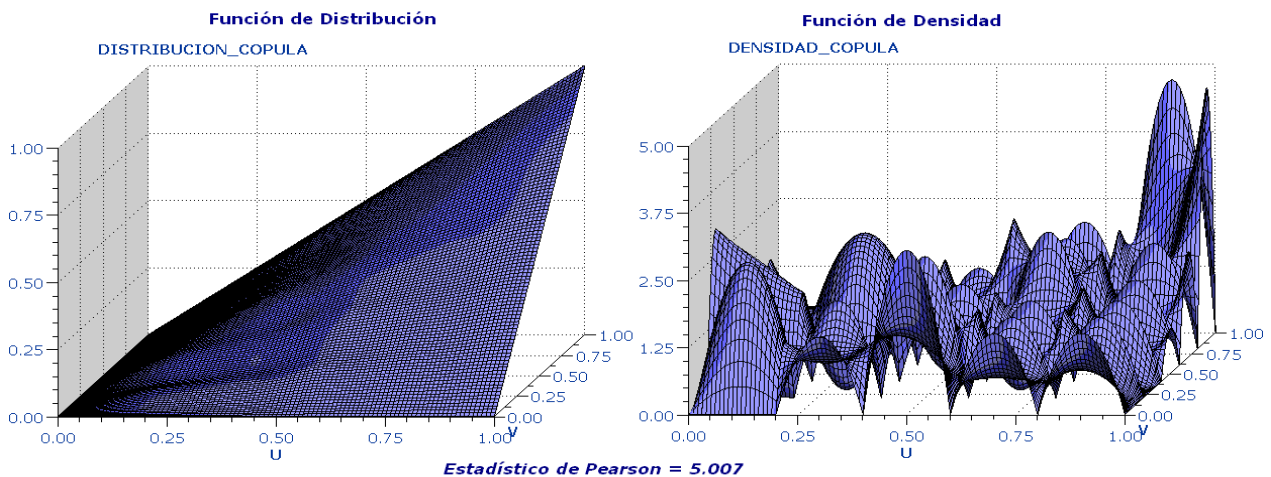


Ilustración 4.6: *Cópula construida por interpolación cúbica mediante superficies de Ferguson, de la subcópula que optimiza el valor del estadístico de Pearson*

Antes de determinar esa condición suficiente que nos asegure que el problema dado por (4.21) tiene solución factible, vamos a estudiar qué restricciones adicionales es preciso añadir, para contemplar la posibilidad de asignar valores distintos de cero a las derivadas segundas, es decir, para que la superficie interpoladora pueda utilizar polinomios de Hermite de cualquier tipo sin necesidad de restringirse a las condiciones impuestas para las superficies de Ferguson. El hecho de imponer que las derivadas de segundo orden valgan 0 es una circunstancia que puede resultar desafortunada a la vista del gráfico anterior, sobre todo si esperamos que la relación existente entre las variables de análisis sea suave.

4.2.3.2.2 Condición a imponer para que una superficie interpoladora de Hermite sea 2-creciente

El desarrollo anterior nos servirá de base para conseguir construir una superficie interpoladora mediante polinomios cúbicos de Hermite cualesquiera sin necesidad de restringir a que su derivada segunda valga cero (superficie de Ferguson). No parece conveniente que la densidad cópula tome valor cero en cada uno de los nodos de unión de la rejilla unidad (y que definen el dominio de la subcópula que proporciona un valor óptimo del estadístico de Pearson) puesto que esta circunstancia implica que se sucedan cambios muy bruscos de comportamiento, algo que no parece adecuado para expresar una relación de dependencia entre variables y a partir de ella llevar a cabo una predicción. Las densidades cópulas son un fiel reflejo de la relación que existe entre los datos que representan y, salvo raras excepciones, dicha relación acostumbra a “moverse” de forma suave con mayor o menor velocidad (dada por las derivadas primeras o vectores tangentes a la superficie).

La razón que justifica esta falta de uniformidad no es el hecho de que la cópula no sea paramétrica, sino el empeño que estamos poniendo en forzar que nuestra cópula conserve el valor óptimo del estadístico de Pearson asociado a la subcópula a partir de la cual la estamos construyendo. Sirva como ejemplo el gráfico de la densidad cópula construida por interpolación bilineal (ilustración (4.2)) caracterizado por su estructura a saltos. Mediante la superficie de Ferguson, esta forma escalonada se ha transformado en una sucesión de montículos (ilustración (4.6)) que incluso podría resultar más errática si se utilizara con fines predictivos. Sólo en situaciones en las que la relación de dependencia entre las variables estuviera concentrada alrededor de varios valores modales (relaciones multimodales), el empleo de esta cópula podría ser más conveniente. Sin embargo, como hemos señalado, esta falta de regularidad no es una implicación del hecho de que la cópula sea no paramétrica. Según hemos visto, la densidad cópula de Bernstein presenta la suavidad y uniformidad que estamos echando en falta (ilustración (4.4)), y esto es porque dicha cópula no es una interpolación exacta que vaya a conservar el valor del estadístico de Pearson sino una interpolación aproximada del dominio de definición de la subcópula C^* .

Lo que vamos a hacer para incrementar la suavidad de la densidad asociada a la función cúbica que estamos construyendo, es permitir que el valor de la derivada segunda en los nodos que resultan de la partición del cuadrado unidad no sea necesariamente cero, evitando de esta manera las drásticas caídas que se muestran en la ilustración (4.6). Para tener una intuición del valor al que intentar aproximar estas derivadas de segundo orden, utilizaremos como referencia el valor que dichas derivadas toman en las dos cópulas anteriores (la obtenida por interpolación bilineal y la que emplea polinomios de Bernstein) y que, como veremos, está muy relacionado con el valor del volumen de los rectángulos que han resultado de hacer la partición. Así, vamos a transformar el problema de programación matemática (4.21) en otro con un objetivo adicional (multiobjetivo) consistente en que las derivadas segundas (que pasarán a ser junto con las de primer orden, variables de decisión) se aproximen lo más posible a estos valores de referencia. Esto nos conducirá inevitablemente a la definición de una función “distancia” a minimizar (objetivo) y, en consecuencia, a un problema de programación cuadrática. Habremos de ver qué restricciones es preciso añadir al problema (4.21) sobre las derivadas segundas para que la función resultante sea una cópula. Observemos que, al construir el problema de optimización como extensión del anterior, estaremos garantizando que si existe solución para (4.21), dado por una relación de valores para las derivadas primeras, existirá entonces también al menos una solución para este nuevo problema que tomará el mismo valor en dichas derivadas y el valor cero en las de segundo orden.

Una propuesta para los valores de referencia a los que se aproximarán las derivadas de segundo orden en los vértices de la partición se puede obtener simplemente conociendo el valor que en ellos presentan las dos cópulas interpoladoras que hemos definido, es decir, la bilineal y la densidad cópula de Bernstein.

- En el primer caso para la densidad de la cópula obtenida por interpolación bilineal (véase la ecuación (4.9)) se

$$\text{obtiene el valor } \frac{V_{C^*}(B)}{(u_{i+1}-u_i)*(v_{j+1}-v_j)} = \frac{V_{C^*}(B)}{\frac{1}{n} * \frac{1}{m}} \quad (4.27).$$

Así, éste podría ser un primer candidato al que aproximar el valor de $\frac{\partial^2 C'(u_i, v_j)}{\partial u \partial v}$.

- En el caso de la densidad cópula de Bernstein, de acuerdo a la definición que se propone en [SANSAT] la densidad cópula responde a las expresiones (4.12) que recordemos era:

$$c_B(u_1, u_2) = \sum_{v_1=0}^m \sum_{v_2=0}^m \beta\left(\frac{v_1}{m}, \frac{v_2}{m}\right) * \prod_{j=1}^2 \binom{m}{v_j} * u_j^{v_j} * (1-u_j)^{m-v_j} \quad \text{siendo}$$

$$\begin{aligned} \beta\left(\frac{v_1}{m}, \frac{v_2}{m}\right) &= (m+1)^2 * \Delta_{1,2} \alpha\left(\frac{v_1}{m+1}, \frac{v_2}{m+1}\right) = \\ &= (m+1)^2 * \left[\alpha\left(\frac{v_1+1}{m+1}, \frac{v_2+1}{m+1}\right) - \alpha\left(\frac{v_1+1}{m+1}, \frac{v_2}{m+1}\right) - \alpha\left(\frac{v_1}{m+1}, \frac{v_2+1}{m+1}\right) + \alpha\left(\frac{v_1}{m+1}, \frac{v_2}{m+1}\right) \right] \end{aligned}$$

Éste es el volumen a través de α de $B = \left[\frac{v_1}{m+1}, \frac{v_2}{m+1} \right] \times \left[\frac{v_1+1}{m+1}, \frac{v_2+1}{m+1} \right]$, rectángulo que surge de hacer la partición de $[0,1]^2$ en $(m+1) \times (m+1)$ rectángulos (no en $m \times m$). En nuestro caso el papel de la función α lo desempeña la subcópula C^* .

$$\text{Teniendo en cuenta que } \lim_{m \rightarrow \infty} \Delta \alpha\left(\frac{v}{m}\right) * m = \lim_{m \rightarrow \infty} \frac{\alpha\left(\frac{v}{m} + \frac{1}{m}\right) - \alpha\left(\frac{v}{m}\right)}{\frac{1}{m}} = \frac{\partial \alpha}{\partial v}\left(\frac{v}{m}\right) \quad (4.28)$$

$$\text{se tiene que } \lim_{m \rightarrow \infty} \Delta \alpha\left(\frac{v}{m+1}\right) * (m+1) = \lim_{m \rightarrow \infty} \frac{\alpha\left(\frac{v}{m+1} + \frac{1}{m+1}\right) - \alpha\left(\frac{v}{m+1}\right)}{\frac{1}{m+1}} = \frac{\partial \alpha}{\partial v}\left(\frac{v}{m+1}\right) \quad (4.29)$$

$$\text{y por tanto que } \lim_{m \rightarrow \infty} (m+1)^2 * \Delta_{1,2} \alpha\left(\frac{v_1}{m+1}, \frac{v_2}{m+1}\right) = \frac{\partial^2 \alpha}{\partial v_1 \partial v_2}\left(\frac{v_1}{m+1}, \frac{v_2}{m+1}\right) \quad (4.30)$$

Como $\beta\left(\frac{v_1}{m}, \frac{v_2}{m}\right) = (m+1)^2 * \Delta_{1,2} \alpha\left(\frac{v_1}{m+1}, \frac{v_2}{m+1}\right)$, podemos concluir que

$$\lim_{m \rightarrow \infty} \beta\left(\frac{v_1}{m}, \frac{v_2}{m}\right) = \frac{\partial^2 \alpha}{\partial v_1 \partial v_2}\left(\frac{v_1}{m+1}, \frac{v_2}{m+1}\right) = \frac{\partial^2 C^*}{\partial v_1 \partial v_2}\left(\frac{v_1}{m+1}, \frac{v_2}{m+1}\right) \quad (4.31)$$

En consecuencia β es una aproximación a la derivada segunda de la subcópula C^* en uno de los nodos que surgen al dividir el rectángulo $[0,1]^2$ en $(m+1) \times (m+1)$ partes (no en $m \times m$). Por tanto, la derivada segunda puede ser aproximada por

$$\beta\left(\frac{v_1}{m}, \frac{v_2}{m}\right) = (m+1)^2 * \Delta_{1,2} \alpha\left(\frac{v_1}{m+1}, \frac{v_2}{m+1}\right) = \frac{\Delta_{1,2} \alpha\left(\frac{v_1}{m+1}, \frac{v_2}{m+1}\right)}{\frac{1}{(m+1)} * \frac{1}{(m+1)}} = \frac{V_{C^*}(B)}{\frac{1}{(n+1)} * \frac{1}{(m+1)}} \quad (4.32)$$

La similitud con el candidato que surgió a través de la cópula que resultaba de hacer la interpolación bilineal es clara.

Supongamos que se realiza una partición del cuadrado unidad en $n \times m$ rectángulos y obtenemos el valor que tomaría la derivada segunda de la cópula obtenida por interpolación bilineal y la que tomaría la derivada segunda de la cópula de Bernstein. Estableciendo la comparativa entre ambas expresiones ((4.27) y (4.32)) podemos observar que:

- El numerador de ambas expresiones es el volumen de un rectángulo. En el primer caso, se trata del rectángulo de la partición cuyo vértice inferior izquierdo es $(\frac{i}{n}, \frac{j}{m})$; en el segundo se trata del rectángulo (de una partición de mayor dimensión) cuyo vértice inferior izquierdo es $(\frac{i}{n+1}, \frac{j}{m+1})$, pudiendo interpretarse como el “rectángulo del mismo orden” que se obtiene al aumentar en una unidad la partición del cuadrado unidad. La ventaja de esta forma de proceder es la posibilidad de asignar un valor propio a los vértices fronterizos del cuadrado unidad.

Así en el punto $(\frac{n-1}{n}, \frac{1}{m})$, aproximariamos la derivada segunda por el volumen del rectángulo

$$[\frac{n-1}{n+1}, \frac{n}{n+1}] \times [\frac{1}{m+1}, \frac{2}{m+1}]$$

, y en el punto $(\frac{n}{n}, \frac{1}{m}) = (1, \frac{1}{m})$ haríamos la aproximación por el

$$\text{volumen del rectángulo } [\frac{n}{n+1}, \frac{n+1}{n+1}] \times [\frac{1}{m+1}, \frac{2}{m+1}] = [\frac{n}{n+1}, 1] \times [\frac{1}{m+1}, \frac{2}{m+1}]$$

El siguiente gráfico permite entender mejor esta explicación. Se realizan dos particiones sobre el cuadrado unidad, una de dimensión $n \times m$ y la otra de dimensión $(n+1) \times (m+1)$ y se muestra el rectángulo que iría asociado al cálculo de la derivada segunda en dos situaciones distintas: un vértice interior y otro fronterizo. Se marcan en rojo los rectángulos correspondientes a la aproximación a la cópula obtenida por interpolación bilineal y en azul aquellos que resultarían de la aproximación a la cópula de Bernstein.

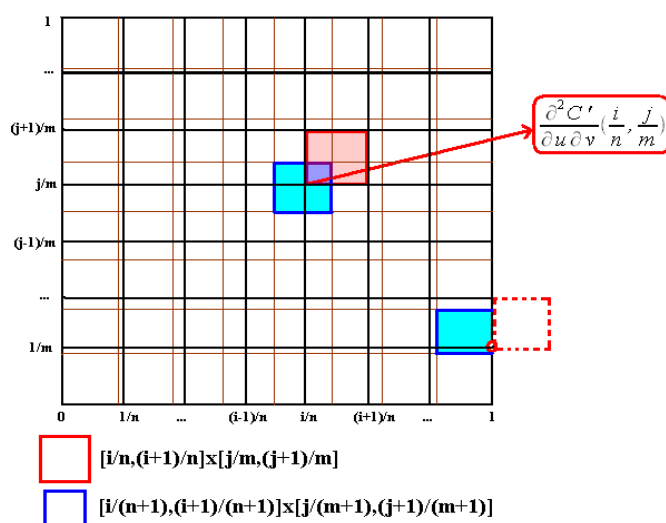


Ilustración 4.7: Particiones de dimensión $(n \times m)$ y $(n+1) \times (m+1)$ sobre el cuadrado unidad

De no realizar la aproximación mediante la densidad de Bernstein, la posibilidad que se antoja más coherente es aproximar las derivadas segundas asociadas a los vértices fronterizos por el mismo valor que se utiliza para aproximar las derivadas segundas asociadas a los vértices inmediatamente anteriores en la partición, es decir,

en el punto $(\frac{n-1}{n}, \frac{1}{m})$, aproximaríamos la derivada segunda al volumen del rectángulo

$[\frac{n-1}{n}, \frac{n}{n}] \times [\frac{1}{m}, \frac{2}{m}] = [\frac{n-1}{n}, 1] \times [\frac{1}{m}, \frac{2}{m}]$, y en el punto $(\frac{n}{n}, \frac{1}{m}) = (1, \frac{1}{m})$ haríamos la aproximación al

volumen de ese mismo rectángulo. Así, la derivada segunda en los vértices que definen un rectángulo de la partición (en $n \times m$ rectángulos) sería aproximada por el volumen del propio rectángulo.

- El denominador de ambas expresiones es la superficie de los rectángulos de la partición. En el caso de la aproximación a la cópula resultante de la interpolación bilineal esta cantidad es $\frac{1}{n} * \frac{1}{m}$, mientras que en la aproximación a la densidad cópula de Bernstein la cantidad va referida lógicamente a la partición de un orden mayor, siendo por tanto $\frac{1}{n+1} * \frac{1}{m+1}$.

Tenemos así dos tipos de candidatos para los valores de referencia $Ref_{V_b}(u_i, v_j)$:

- $$Ref_{V_b}(u_i, v_j) = \frac{V_c'(B)}{\frac{1}{n} * \frac{1}{m}} \quad (4.33)$$

siendo B el rectángulo de la partición en $n \times m$ rectángulos cuyo vértice inferior izquierdo es (u_i, v_j) .

- $$Ref_{V_b}(u_i, v_j) = \frac{V_c'(B)}{\frac{1}{n+1} * \frac{1}{m+1}} \quad (4.34)$$

siendo B el rectángulo de la partición en $(n+1) \times (m+1)$ rectángulos cuyo vértice inferior izquierdo es el que ocupa la posición (i, j) dentro del conjunto de todos los vértices de dicha partición.

El primero de los valores depende directamente del volumen de los rectángulos de la partición de orden $n \times m$, el cual a su vez depende de los valores asignados a cada vértice de dicha partición al tratar de optimizar el valor del estadístico de Pearson (véase ecuación (4.1)). Sin embargo, para el segundo no se dispone del volumen asociado a los rectángulos de la partición de orden $(n+1) \times (m+1)$ a través de la cópula C porque ésta todavía no ha sido construida (su expresión depende del valor de las derivadas segundas en los nodos de la partición, valor que estamos tratando de aproximar). El hecho de que sí conozcamos el volumen de los rectángulos de la partición de orden $n \times m$ a través de ella (de C) se debe a que va a ser el mismo que el que se obtenga a través de la subcópula C^* a la que interpola dado que esta interpolación se produce precisamente a partir del valor de los vértices de dichos rectángulos.

Esto provoca lógicamente una dificultad si tratamos de aproximar el valor de las derivadas segundas a la densidad cópula de Bernstein. La solución que planteamos a esta dificultad es resolver el problema (4.7) para una partición de dimensión mayor $(n+1) \times (m+1)$ y de esta manera obtener los valores asociados a los vértices que permiten calcular, para dicha partición, el mejor valor para el estadístico de Pearson. Dichos valores definirían la referencia a la que aproximar la densidad cópula ajustada a la partición de dimensión $n \times m$.

Hechas estas consideraciones, se trata nuevamente de aplicar la expresión de la superficie cúbica (ahora (4.19) en vez de (4.20)) a cada uno de los cuatro vértices de un rectángulo $B \subset I^2$ para obtener, en esta ocasión, otra en la que los elementos desconocidos van a ser, no sólo las derivadas primeras sino también las segundas respecto de U y V . El resultado que a continuación enunciaremos, permite establecer las restricciones que garantizan la “2-crecencia” de C proporcionando unos rangos de variación para las derivadas de primer y segundo orden. Estos rangos vienen determinados por la región factible asociada al siguiente problema de programación multiobjetivo no lineal y entera:

$$\text{Max} \text{Min} \sum_{i=0}^{n-1} \sum_{j=0}^{m-1} \left[\frac{\partial C'(u_i, v_j)}{\partial u} + \frac{\partial C'(u_i, v_j)}{\partial v} \right] \text{ y } \text{Min} \sum_{i=0}^{n-1} \sum_{j=0}^{m-1} \left[\frac{\partial^2 C'(u_i, v_j)}{\partial u \partial v} - \text{Ref}_{V_s}(u_i, v_j) \right]^2$$

Restricciones asociadas a las derivadas respecto de U

$$\begin{aligned} & \frac{2*(u_{i+1}-u_i)}{V_c'(B)} * \left[\frac{\partial C'(u_i, v_j)}{\partial u} - \frac{\partial C'(u_i, v_{j+1})}{\partial u} \right] + \frac{2*2*(u_{i+1}-u_i)}{V_c'(B)} * \left[\frac{\partial C'(u_{i+1}, v_j)}{\partial u} - \frac{\partial C'(u_{i+1}, v_{j+1})}{\partial u} \right] + 3 \leq 3*(1-\delta_1) \\ & -2 * \frac{2*(u_{i+1}-u_i)}{V_c'(B)} * \left[\frac{\partial C'(u_i, v_j)}{\partial u} - \frac{\partial C'(u_i, v_{j+1})}{\partial u} \right] - \frac{2*(u_{i+1}-u_i)}{V_c'(B)} * \left[\frac{\partial C'(u_{i+1}, v_j)}{\partial u} - \frac{\partial C'(u_{i+1}, v_{j+1})}{\partial u} \right] - 3 \leq \left(\frac{24+15*\sqrt{2}}{7} \right) * (1-\delta_1) \\ & 2 * \frac{2*(u_{i+1}-u_i)}{V_c'(B)} * \left[\frac{\partial C'(u_i, v_j)}{\partial u} - \frac{\partial C'(u_i, v_{j+1})}{\partial u} \right] + \frac{2*(u_{i+1}-u_i)}{V_c'(B)} * \left[\frac{\partial C'(u_{i+1}, v_j)}{\partial u} - \frac{\partial C'(u_{i+1}, v_{j+1})}{\partial u} \right] + 3 \leq 3*(1-\delta_2) \\ & - \frac{2*(u_{i+1}-u_i)}{V_c'(B)} * \left[\frac{\partial C'(u_i, v_j)}{\partial u} - \frac{\partial C'(u_i, v_{j+1})}{\partial u} \right] - \frac{2*2*(u_{i+1}-u_i)}{V_c'(B)} * \left[\frac{\partial C'(u_{i+1}, v_j)}{\partial u} - \frac{\partial C'(u_{i+1}, v_{j+1})}{\partial u} \right] - 3 \leq \left(\frac{24+15*\sqrt{2}}{7} \right) * (1-\delta_2) \\ & - \frac{2*(u_{i+1}-u_i)}{V_c'(B)} * \left[\frac{\partial C'(u_i, v_j)}{\partial u} - \frac{\partial C'(u_i, v_{j+1})}{\partial u} \right] - \frac{2*(u_{i+1}-u_i)}{V_c'(B)} * \left[\frac{\partial C'(u_{i+1}, v_j)}{\partial u} - \frac{\partial C'(u_{i+1}, v_{j+1})}{\partial u} \right] - 2 \leq 4*(1-\delta_3) \\ & \left\{ \frac{2*(u_{i+1}-u_i)}{V_c'(B)} * \left[\frac{\partial C'(u_i, v_j)}{\partial u} - \frac{\partial C'(u_i, v_{j+1})}{\partial u} \right] + \frac{2*(u_{i+1}-u_i)}{V_c'(B)} * \left[\frac{\partial C'(u_{i+1}, v_j)}{\partial u} - \frac{\partial C'(u_{i+1}, v_{j+1})}{\partial u} \right] + 3 \right\}^2 - \\ & \frac{2*(u_{i+1}-u_i)}{V_c'(B)} * \left[\frac{\partial C'(u_i, v_j)}{\partial u} - \frac{\partial C'(u_i, v_{j+1})}{\partial u} \right] * \frac{2*(u_{i+1}-u_i)}{V_c'(B)} * \left[\frac{\partial C'(u_{i+1}, v_j)}{\partial u} - \frac{\partial C'(u_{i+1}, v_{j+1})}{\partial u} \right] \leq 9*(1-\delta_4) \\ & \left[\frac{\partial C'(u_i, v_j)}{\partial u} - \frac{\partial C'(u_i, v_{j+1})}{\partial u} \right] \leq 0; \left[\frac{\partial C'(u_{i+1}, v_j)}{\partial u} - \frac{\partial C'(u_{i+1}, v_{j+1})}{\partial u} \right] \leq 0 \end{aligned}$$

Restricciones asociadas a las derivadas respecto de V

$$\begin{aligned} & \frac{2*(v_{j+1}-v_j)}{V_c'(B)} * \left[\frac{\partial C'(u_i, v_j)}{\partial v} - \frac{\partial C'(u_{i+1}, v_j)}{\partial v} \right] + \frac{2*2*(v_{j+1}-v_j)}{V_c'(B)} * \left[\frac{\partial C'(u_i, v_{j+1})}{\partial v} - \frac{\partial C'(u_{i+1}, v_{j+1})}{\partial v} \right] + 3 \leq 3*(1-\gamma_1) \\ & -2 * \frac{2*(v_{j+1}-v_j)}{V_c'(B)} * \left[\frac{\partial C'(u_i, v_j)}{\partial v} - \frac{\partial C'(u_{i+1}, v_j)}{\partial v} \right] - \frac{2*(v_{j+1}-v_j)}{V_c'(B)} * \left[\frac{\partial C'(u_i, v_{j+1})}{\partial v} - \frac{\partial C'(u_{i+1}, v_{j+1})}{\partial v} \right] - 3 \leq \left(\frac{24+15*\sqrt{2}}{7} \right) * (1-\gamma_1) \\ & 2 * \frac{2*(v_{j+1}-v_j)}{V_c'(B)} * \left[\frac{\partial C'(u_i, v_j)}{\partial v} - \frac{\partial C'(u_{i+1}, v_j)}{\partial v} \right] + \frac{2*(v_{j+1}-v_j)}{V_c'(B)} * \left[\frac{\partial C'(u_i, v_{j+1})}{\partial v} - \frac{\partial C'(u_{i+1}, v_{j+1})}{\partial v} \right] + 3 \leq 3*(1-\gamma_2) \\ & - \frac{2*(v_{j+1}-v_j)}{V_c'(B)} * \left[\frac{\partial C'(u_i, v_j)}{\partial v} - \frac{\partial C'(u_{i+1}, v_j)}{\partial v} \right] - \frac{2*2*(v_{j+1}-v_j)}{V_c'(B)} * \left[\frac{\partial C'(u_i, v_{j+1})}{\partial v} - \frac{\partial C'(u_{i+1}, v_{j+1})}{\partial v} \right] - 3 \leq \left(\frac{24+15*\sqrt{2}}{7} \right) * (1-\gamma_2) \\ & - \frac{2*(v_{j+1}-v_j)}{V_c'(B)} * \left[\frac{\partial C'(u_i, v_j)}{\partial v} - \frac{\partial C'(u_{i+1}, v_j)}{\partial v} \right] - \frac{2*(v_{j+1}-v_j)}{V_c'(B)} * \left[\frac{\partial C'(u_i, v_{j+1})}{\partial v} - \frac{\partial C'(u_{i+1}, v_{j+1})}{\partial v} \right] - 2 \leq 4*(1-\gamma_3) \\ & \left\{ \frac{2*(v_{j+1}-v_j)}{V_c'(B)} * \left[\frac{\partial C'(u_i, v_j)}{\partial v} - \frac{\partial C'(u_{i+1}, v_j)}{\partial v} \right] + \frac{2*(v_{j+1}-v_j)}{V_c'(B)} * \left[\frac{\partial C'(u_i, v_{j+1})}{\partial v} - \frac{\partial C'(u_{i+1}, v_{j+1})}{\partial v} \right] + 3 \right\}^2 - \\ & \frac{2*(v_{j+1}-v_j)}{V_c'(B)} * \left[\frac{\partial C'(u_i, v_j)}{\partial v} - \frac{\partial C'(u_{i+1}, v_j)}{\partial v} \right] * \frac{2*(v_{j+1}-v_j)}{V_c'(B)} * \left[\frac{\partial C'(u_i, v_{j+1})}{\partial v} - \frac{\partial C'(u_{i+1}, v_{j+1})}{\partial v} \right] \leq 9*(1-\gamma_4) \\ & \left[\frac{\partial C'(u_i, v_j)}{\partial v} - \frac{\partial C'(u_{i+1}, v_j)}{\partial v} \right] \leq 0; \left[\frac{\partial C'(u_i, v_{j+1})}{\partial v} - \frac{\partial C'(u_{i+1}, v_{j+1})}{\partial v} \right] \leq 0 \end{aligned}$$

Restricciones asociadas a las derivadas segundas respecto de U y V

$$\begin{aligned} 0 & \leq (v_{j+1}-v_j) * \left[\frac{\partial^2 C'(u_i, v_j)}{\partial u \partial v} \right] \leq -3 * \left[\frac{\partial C'(u_i, v_j)}{\partial u} - \frac{\partial C'(u_i, v_{j+1})}{\partial u} \right] \\ 0 & \leq (u_{i+1}-u_i) * \left[\frac{\partial^2 C'(u_i, v_j)}{\partial u \partial v} \right] \leq -3 * \left[\frac{\partial C'(u_i, v_j)}{\partial v} - \frac{\partial C'(u_{i+1}, v_j)}{\partial v} \right] \\ 0 & \leq (v_{j+1}-v_j) * \left[\frac{\partial^2 C'(u_i, v_{j+1})}{\partial u \partial v} \right] \leq -3 * \left[\frac{\partial C'(u_i, v_j)}{\partial u} - \frac{\partial C'(u_i, v_{j+1})}{\partial u} \right] \\ 0 & \leq (u_{i+1}-u_i) * \left[\frac{\partial^2 C'(u_i, v_{j+1})}{\partial u \partial v} \right] \leq -3 * \left[\frac{\partial C'(u_i, v_{j+1})}{\partial v} - \frac{\partial C'(u_{i+1}, v_{j+1})}{\partial v} \right] \\ 0 & \leq (v_{j+1}-v_j) * \left[\frac{\partial^2 C'(u_{i+1}, v_j)}{\partial u \partial v} \right] \leq -3 * \left[\frac{\partial C'(u_{i+1}, v_j)}{\partial u} - \frac{\partial C'(u_{i+1}, v_{j+1})}{\partial u} \right] \\ 0 & \leq (u_{i+1}-u_i) * \left[\frac{\partial^2 C'(u_{i+1}, v_j)}{\partial u \partial v} \right] \leq -3 * \left[\frac{\partial C'(u_i, v_j)}{\partial v} - \frac{\partial C'(u_{i+1}, v_j)}{\partial v} \right] \\ 0 & \leq (v_{j+1}-v_j) * \left[\frac{\partial^2 C'(u_{i+1}, v_{j+1})}{\partial u \partial v} \right] \leq -3 * \left[\frac{\partial C'(u_{i+1}, v_j)}{\partial u} - \frac{\partial C'(u_{i+1}, v_{j+1})}{\partial u} \right] \\ 0 & \leq (u_{i+1}-u_i) * \left[\frac{\partial^2 C'(u_{i+1}, v_{j+1})}{\partial u \partial v} \right] \leq -3 * \left[\frac{\partial C'(u_i, v_{j+1})}{\partial v} - \frac{\partial C'(u_{i+1}, v_{j+1})}{\partial v} \right] \end{aligned}$$

$$\delta_1 + \delta_2 + \delta_3 + \delta_4 = 1; \gamma_1 + \gamma_2 + \gamma_3 + \gamma_4 = 1; \delta_1, \delta_2, \delta_3, \delta_4 \in \{0, 1\}; \gamma_1, \gamma_2, \gamma_3, \gamma_4 \in \{0, 1\} \quad \forall i \in \{0, 1, \dots, n-1\} \quad j \in \{0, 1, \dots, m-1\}$$

(4.35)

Dando solución a este problema, se dispondrá de todos los elementos que componen la expresión (4.19) pues, además de los valores que toma la cópula en los puntos del dominio de definición de la subcópula óptima C^* (aquellos que minimizan el estadístico de Pearson), se tendrán también los vectores tangentes a la superficie interpoladora en dichos puntos y los valores de la densidad cópula en los mismos (variables de decisión de (4.35)). Formalmente, el resultado es el siguiente:

Lema 4.6: Condición suficiente para que se cumpla la tercera de las propiedades de las funciones cópula en una superficie interpoladora de Hermite

“Sea una partición del cuadrado unidad en $n \times m$ rectángulos del mismo área y sea la solución óptima al problema de programación fraccional (4.7) dada por $c_{ij}^* = C^*\left(\frac{i}{n}, \frac{j}{m}\right) \forall i \in \{0, 1, 2, \dots, n\} \quad \forall j \in \{0, 1, 2, \dots, m\}$ que define una subcópula C^* de dominio los puntos de intersección entre los rectángulos que originan la partición.

Sea $C_{HERMITE}(u, v)$ la función que se obtiene por interpolación cúbica de una subcópula C^* mediante polinomios de Hermite.

Si $\frac{\partial C_{HERMITE}(u_i, v_j)}{\partial u}$, $\frac{\partial C_{HERMITE}(u_i, v_j)}{\partial v}$ y $\frac{\partial^2 C_{HERMITE}(u_i, v_j)}{\partial u \partial v} \quad \forall i \in \{0, 1, \dots, n\} \quad \forall j \in \{0, 1, \dots, m\}$ proporcionan

junto con una relación de δ 's y γ 's (variables binarias) una solución al problema de programación multiobjetivo no lineal entera (4.35), entonces la función $C_{HERMITE}(u, v)$ cumple la tercera de las condiciones de las funciones cópula:

$$V_{C_{HERMITE}}([u_1, u_2] \times [v_1, v_2]) = C_{HERMITE}(u_2, v_2) - C_{HERMITE}(u_2, v_1) - C_{HERMITE}(u_1, v_2) + C_{HERMITE}(u_1, v_1) \geq 0 \quad ,$$

$$\forall u_1, u_2, v_1, v_2 \text{ t.q. } u_1 \leq u_2, v_1 \leq v_2$$

La demostración de este resultado puede ser consultada en el Anexo VI. En ella, el desarrollo no se plantea a partir del problema MINLP (4.35) sino de un caso particular del mismo en el que la región factible es un subconjunto de esta última que da al problema el carácter lineal y continuo (no entero) que lo hace fácilmente resoluble por la mayor parte del software de optimización disponible. Por ello, de igual forma que hemos propuesto el planteamiento (4.26) como alternativo al (4.21), proponemos ahora considerar el siguiente modelo como simplificación de (4.35):

$$\text{Max/Min} \sum_{i=0}^{n-1} \sum_{j=0}^{m-1} \left[\frac{\partial C'(u_i, v_j)}{\partial u} + \frac{\partial C'(u_i, v_j)}{\partial v} \right] \text{ y } \text{Min} \sum_{i=0}^{n-1} \sum_{j=0}^{m-1} \left[\frac{\partial^2 C'(u_i, v_j)}{\partial u \partial v} - \text{Ref}_{V_B}(u_i, v_j) \right]^2$$

$$\text{o bien } \text{Min} \left\{ \lambda * \left[\sum_{i=0}^{n-1} \sum_{j=0}^{m-1} \left(\frac{\partial C'(u_i, v_j)}{\partial u} + \frac{\partial C'(u_i, v_j)}{\partial v} \right) \right] + (1-\lambda) * \left[\sum_{i=0}^{n-1} \sum_{j=0}^{m-1} \left(\frac{\partial^2 C'(u_i, v_j)}{\partial u \partial v} - \text{Ref}_{V_B}(u_i, v_j) \right) \right]^2 \right\}$$

Restricciones asociadas a las derivadas respecto de U

$$\begin{aligned} -3 &\leq \frac{2*(u_{i+1}-u_i)}{V_C'(B)} * \left[\frac{\partial C'(u_i, v_j)}{\partial u} - \frac{\partial C'(u_i, v_{j+1})}{\partial u} \right] \leq 0 \\ -3 &\leq \frac{2*(u_{i+1}-u_i)}{V_C'(B)} * \left[\frac{\partial C'(u_{i+1}, v_j)}{\partial u} - \frac{\partial C'(u_{i+1}, v_{j+1})}{\partial u} \right] \leq 0 \end{aligned}$$

Restricciones asociadas a las derivadas respecto de V

$$\begin{aligned} -3 &\leq \frac{2*(v_{j+1}-v_j)}{V_C'(B)} * \left[\frac{\partial C'(u_i, v_j)}{\partial v} - \frac{\partial C'(u_{i+1}, v_j)}{\partial v} \right] \leq 0 \\ -3 &\leq \frac{2*(v_{j+1}-v_j)}{V_C'(B)} * \left[\frac{\partial C'(u_i, v_{j+1})}{\partial v} - \frac{\partial C'(u_{i+1}, v_{j+1})}{\partial v} \right] \leq 0 \end{aligned}$$

Restricciones asociadas a las derivadas segundas respecto de U y V

$$\begin{aligned} 0 &\leq (v_{j+1}-v_j) * \left[\frac{\partial^2 C'(u_i, v_j)}{\partial u \partial v} \right] \leq -3 * \left[\frac{\partial C'(u_i, v_j)}{\partial u} - \frac{\partial C'(u_i, v_{j+1})}{\partial u} \right] \\ 0 &\leq (u_{i+1}-u_i) * \left[\frac{\partial^2 C'(u_i, v_j)}{\partial u \partial v} \right] \leq -3 * \left[\frac{\partial C'(u_i, v_j)}{\partial v} - \frac{\partial C'(u_{i+1}, v_j)}{\partial v} \right] \\ 0 &\leq (v_{j+1}-v_j) * \left[\frac{\partial^2 C'(u_i, v_{j+1})}{\partial u \partial v} \right] \leq -3 * \left[\frac{\partial C'(u_i, v_j)}{\partial u} - \frac{\partial C'(u_i, v_{j+1})}{\partial u} \right] \\ 0 &\leq (u_{i+1}-u_i) * \left[\frac{\partial^2 C'(u_i, v_{j+1})}{\partial u \partial v} \right] \leq -3 * \left[\frac{\partial C'(u_i, v_{j+1})}{\partial v} - \frac{\partial C'(u_{i+1}, v_{j+1})}{\partial v} \right] \\ 0 &\leq (v_{j+1}-v_j) * \left[\frac{\partial^2 C'(u_{i+1}, v_j)}{\partial u \partial v} \right] \leq -3 * \left[\frac{\partial C'(u_{i+1}, v_j)}{\partial u} - \frac{\partial C'(u_{i+1}, v_{j+1})}{\partial u} \right] \\ 0 &\leq (u_{i+1}-u_i) * \left[\frac{\partial^2 C'(u_{i+1}, v_j)}{\partial u \partial v} \right] \leq -3 * \left[\frac{\partial C'(u_i, v_j)}{\partial v} - \frac{\partial C'(u_{i+1}, v_j)}{\partial v} \right] \\ 0 &\leq (v_{j+1}-v_j) * \left[\frac{\partial^2 C'(u_{i+1}, v_{j+1})}{\partial u \partial v} \right] \leq -3 * \left[\frac{\partial C'(u_{i+1}, v_j)}{\partial u} - \frac{\partial C'(u_{i+1}, v_{j+1})}{\partial u} \right] \\ 0 &\leq (u_{i+1}-u_i) * \left[\frac{\partial^2 C'(u_{i+1}, v_{j+1})}{\partial u \partial v} \right] \leq -3 * \left[\frac{\partial C'(u_i, v_{j+1})}{\partial v} - \frac{\partial C'(u_{i+1}, v_{j+1})}{\partial v} \right] \end{aligned}$$

$$\forall i \in \{0, 1, \dots, n-1\} \quad j \in \{0, 1, \dots, m-1\} \quad \lambda \in [0, 1]$$

(4.36)

De acuerdo a este planteamiento “relajado”, podemos enunciar un resultado equivalente al lema (4.6) que es el que a continuación se presenta:

Lema 4.7: Condición suficiente “relajada” para que se cumpla la tercera de las propiedades de las funciones cópula en una superficie interpoladora de Hermite

“Sea una partición del cuadrado unidad en $n \times m$ rectángulos del mismo área y sea la solución óptima al problema de programación fraccional (4.7) dada por $c_{ij}^* = C^* \left(\frac{i}{n}, \frac{j}{m} \right) \forall i \in \{0,1,2,\dots,n\} \forall j \in \{0,1,2,\dots,m\}$ que define una subcópula C^* de dominio los puntos de intersección entre los rectángulos que originan la partición.

Sea $C_{HERMITE}(u,v)$ la función que se obtiene por interpolación cúbica de una subcópula C^* mediante polinomios de Hermite.

Si $\frac{\partial C_{HERMITE}(u_i, v_j)}{\partial u}, \frac{\partial C_{HERMITE}(u_i, v_j)}{\partial v}$ y $\frac{\partial^2 C_{HERMITE}(u_i, v_j)}{\partial u \partial v} \forall i \in \{0,1,\dots,n\} \forall j \in \{0,1,\dots,m\}$ proporcionan

junto con una relación de δ 's y γ 's (variables binarias) una solución al problema de programación multiobjetivo no lineal entera (4.36), entonces la función $C_{HERMITE}(u,v)$ cumple la tercera de las condiciones de las funciones cópula:

$$V_{C_{HERMITE}}([u_1, u_2] \times [v_1, v_2]) = C_{HERMITE}(u_2, v_2) - C_{HERMITE}(u_2, v_1) - C_{HERMITE}(u_1, v_2) + C_{HERMITE}(u_1, v_1) \geq 0 \quad ,, \\ \forall u_1, u_2, v_1, v_2 \text{ t.q. } u_1 \leq u_2, v_1 \leq v_2$$

Este lema, permite hacer un tratamiento “relajado” del problema de programación matemática (4.35) motivado por la escasez de paquetes estadísticos que permiten resolver problemas de programación no lineal y entera (MINLP). En esta ocasión:

- La no linealidad viene dada no sólo por la elipse que participa en la definición de la región factible, sino además por la segunda función objetivo a minimizar que define la distancia (función cuadrática) entre las derivadas segundas en los nodos de la partición y los valores de referencia $Ref_V(u_i, v_j)$.
- La existencia de variables enteras viene dada por la necesidad de imponer restricciones disyuntivas para definir la región factible asociada a las derivadas de primer orden. La asociada a las de segundo orden es lineal pues es una acotación de éstas por el valor que toman las de primer orden.

Emplearemos una región factible lineal y conexas para las derivadas de primer orden ($[-3,0] \times [-3,0]$) que permite evitar el uso de variables enteras y simplificar así el problema. La no linealidad no desaparecería dado que, como hemos señalado, no afecta únicamente a la región factible sino también a la función multiobjetivo. Efectivamente, el primero de los objetivos es una función lineal de las derivadas de primer orden que viene a controlar, como ya señalamos, el apuntamiento de los montículos que se originan como consecuencia de la interpolación (ilustración (4.6)). Sin embargo, el segundo es una función de distancia que determina la proximidad a una relación de valores de referencia que podría tomar la densidad cópula en los nodos de la rejilla obtenidos al dividir el cuadrado unidad.

Según se explica en el capítulo 9 de [SIXTO], existen varios métodos para resolver un modelo de programación **continua** multiobjetivo. Recalamos el término “continua” para dar a entender que los métodos que se van a citar, únicamente podrían ser aplicados al problema (4.36) pero no así al problema original (4.35). Aquí hacemos una

exposición a grandes rasgos de algunos de los más conocidos invitándose al lector a consultar la obra de **Sixto** para tener una visión más detallada de ellos y de cómo pueden ser aplicados en algunos ejemplos reales.

La técnica que proponemos emplear por su simplicidad, es el “**método de las ponderaciones**” que consiste en construir una única función como suma ponderada de las diferentes funciones objetivo. Así, el problema se reduciría a una asignación de pesos en función de la importancia que considerásemos tiene cada una de ellas. En nuestro caso se trataría de maximizar o minimizar la función objetivo asociada a las derivadas de primer orden y minimizar la función objetivo asociada a las derivadas segundas. En el caso de la segunda función objetivo es evidente la intención de minimizarla pues se trata de aproximarse lo más posible a los valores de referencia que hemos mencionado. Respecto de la primera función objetivo, si la intención fuese la misma, el objetivo final sería claramente minimizar la suma ponderada de las dos; si por el contrario la idea fuera su maximización, bastaría cambiar el signo de dicha función para transformar el objetivo en otro de minimización. De esta forma, el objetivo final volvería a ser minimizar la suma ponderada de la primera función transformada (multiplicada por -1) y la segunda con su signo original. La ventaja del método de las ponderaciones es que el problema queda transformado en un problema de programación con un único objetivo (no lineal o, más concretamente, cuadrático).

Otro método de resolución que reduce el problema a un único objetivo es el **método de las ε -restricciones** el cual consiste en optimizar la función objetivo que se considere más importante e introducir para la otra (u otras) una relación de números reales “ ε ” que la acoten. En nuestro caso, si el objetivo fundamental fuera minimizar la distancia de las derivadas segundas a los valores de referencia, se trataría de minimizar la segunda de las funciones objetivo (no lineal) y restringir a que la suma de las derivadas primeras (primera función objetivo) fuera mayor (si el objetivo es de minimización) o menor (si es de maximización) que un cierto “ ε ”.

Un tercer método que permite la simplificación al caso uniobjetivo es la **programación por metas (Charnes y Cooper)** en la que el decisor establece una relación de metas a las que pretende llegar por objetivo y considera como solución óptima aquella que minimiza la suma de desviaciones a las metas propuestas. La función resultante es una distancia y como tal no lineal, si bien, como explica Sixto, se puede hacer la transformación a un formato lineal mediante la incorporación de pares de variables (desviación por exceso y por defecto) asociadas a cada uno de los objetivos.

La razón por la que proponemos emplear el método de las ponderaciones con respecto al método de las ε -restricciones o al de programación por metas (que también permiten tratar el problema como si tuviera un único objetivo), es no aumentar el número de restricciones y variables del problema que es ya suficientemente amplio ($12 \cdot n \cdot m$ restricciones para (4.36)).

Señalaremos finalmente que existen otras técnicas que permiten abordar el problema sin necesidad de transformarlo en otro uniobjetivo en el que el concepto de solución óptima tiene pleno sentido pues la optimalidad, sin lugar a dudas, se dará respecto del único objetivo existente. En contraposición, el **método del Simplex Multiobjetivo de Zeleny** aborda directamente todos los objetivos y surge entonces el concepto de “**solución eficiente**” dado que raramente la solución va a ser simultáneamente óptima respecto de todos los objetivos. Por definición, una solución es eficiente si no existe otra que mejore el valor de uno de los objetivos sin provocar un empeoramiento en el valor de alguno de los objetivos restantes.

El método Simplex multiobjetivo se basa fundamentalmente en tres etapas:

- La primera determina una solución inicial básica factible (como en el Simplex uniobjetivo) introduciendo las variables de holgura y artificiales necesarias para ello.
- La segunda determina un punto extremo eficiente.
- La tercera genera a partir del punto anterior el resto de puntos (soluciones) extremos eficientes.

El empleo de esta técnica multiobjetivo pura debe ser aplicada a funciones objetivo de carácter lineal. Esto no es necesariamente un problema, dado que la no linealidad de (4.36) viene dada por la función distancia y ésta puede definirse alternativamente a través de la función valor absoluto (en vez de como una medida cuadrática), para dar posteriormente al problema un formato lineal, algo que es posible según acabamos de comentar para el método de programación por metas. Con todo y con eso, el método de Zeleny presenta la dificultad de encontrar un software que tenga implementado su algoritmo y, aún así, deberá tenerse en cuenta la complejidad computacional (en tiempo) que pueden suponer los cálculos para particiones de altas dimensiones. Mayores dificultades tendríamos aún para encontrar un software que permitiera resolver el problema de programación multiobjetivo no lineal en su versión **binaria (entera)** planteado por (4.35). En este caso, la no-linealidad del problema no es evitable dado que no va asociada sólo a uno de los objetivos sino a la forma de la región factible. El primer artículo sobre programación multiobjetivo entera data de 1977 (técnica de **Passy**) y aborda objetivos bilineales. Algunos autores posteriores como **Bitran** (1979) o **Winkoffki** proponen técnicas de enumeración implícita relacionados con la teoría de grafos. En general, los métodos existentes se basan en algoritmos de ramificación y acotación o de planos de corte en programación por metas, siendo el libro de **Ignizio** (1976) el que recomienda Sixto al lector interesado en implementar un programa que permitiese resolver el problema (4.35). A modo de resumen podemos establecer el siguiente teorema a partir de los lemas (4.2), (4.3) y (4.6 o 4.7):

Teorema 4.5: Condición suficiente para poder generar una cópula por interpolación de una subcópula mediante una superficie de Hermite

“Sea una partición del cuadrado unidad en $n \times m$ rectángulos del mismo área y sea la solución óptima al problema de programación fraccional (4.7) dada por $c_{ij}^* = C^*\left(\frac{i}{n}, \frac{j}{m}\right) \forall i \in \{0,1,2,\dots,n\} \quad \forall j \in \{0,1,2,\dots,m\}$ que define una subcópula C^* de dominio los puntos de intersección entre los rectángulos que originan la partición. Sea $C_{\text{HERMITE}}(u,v)$ la función que se obtiene por interpolación cúbica de una subcópula C^* mediante polinomios de Hermite (dada por (4.19)).

Si $\frac{\partial C_{HERMITE}(u_i, v_j)}{\partial u}$, $\frac{\partial C_{HERMITE}(u_i, v_j)}{\partial v}$ y $\frac{\partial^2 C_{HERMITE}(u_i, v_j)}{\partial u \partial v} \forall i \in \{0, 1, 2, \dots, n\} \forall j \in \{0, 1, 2, \dots, m\}$ cumplen las siguientes condiciones:

$$1. \quad \frac{\partial C_{HERMITE}(u_k, 0)}{\partial u} = 0 \forall k \in \{0, 1, 2, \dots, n\} \quad y \quad \frac{\partial C_{HERMITE}(0, v_l)}{\partial v} = 0 \forall l \in \{0, 1, 2, \dots, m\}$$

$$2. \quad \frac{\partial C_{HERMITE}(u_k, 1)}{\partial u} = 1 \forall k \in \{0, 1, 2, \dots, n\} \quad y \quad \frac{\partial C_{HERMITE}(1, v_l)}{\partial v} = 1 \forall l \in \{0, 1, 2, \dots, m\}$$

3. Proporcionan una solución del problema de programación multiobjetivo no lineal entera (4.35) (o bien su relajación (4.36))

entonces la función $C_{HERMITE}$ es una cópula”.

Adjuntamos finalmente el gráfico que resultaría de dar solución al problema relajado (4.36) mediante el método de las ponderaciones para obtener una cópula por interpolación cúbica de Hermite. La muestra de partida es nuevamente la mostrada en la ilustración (4.1). Se ha asignado peso 0.5 a cada uno de los objetivos donde el primero de ellos consiste en minimizar la suma de las derivadas de primer orden mientras que el segundo minimiza la suma de las distancias a los valores de referencia que, en este caso, han sido establecidos en función de la densidad de Bernstein, es decir, tratando de aproximar los valores de las derivadas segunda a los volúmenes de los rectángulos obtenidos en una partición de dimensión mayor (6 x 6).

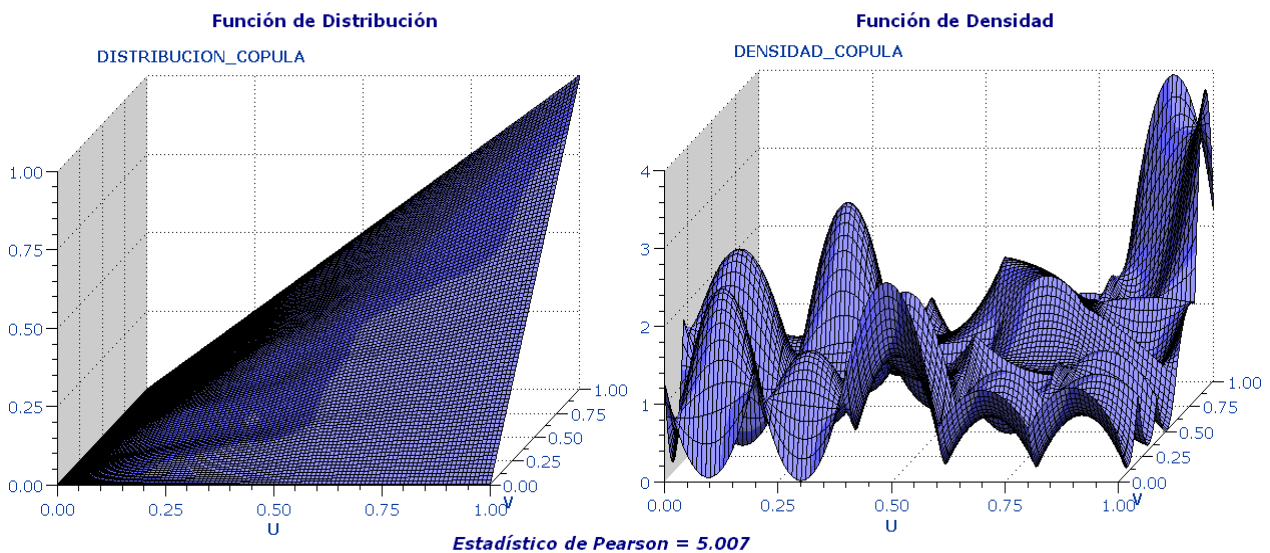


Ilustración 4.8: Cópula construida por interpolación cúbica mediante polinomios de Hermite, de la subcópula que optimiza el valor del estadístico de Pearson

Se puede establecer la comparativa entre la densidad obtenida con la superficie interpoladora de Ferguson y está última. Haremos esta comparativa mediante un **gráfico de contorno** (o térmico). Este tipo de gráficos es muy utilizados en teoría de cópulas dado que permite hacerse una idea muy buena de la forma de la densidad en todos sus frentes gracias a una perspectiva aérea mediante curvas de nivel. Cada color va asociado a un determinado nivel alrededor del cual se mueve la densidad cópula. Dicho nivel vendría a ser el valor del eje “z” en el gráfico anterior.

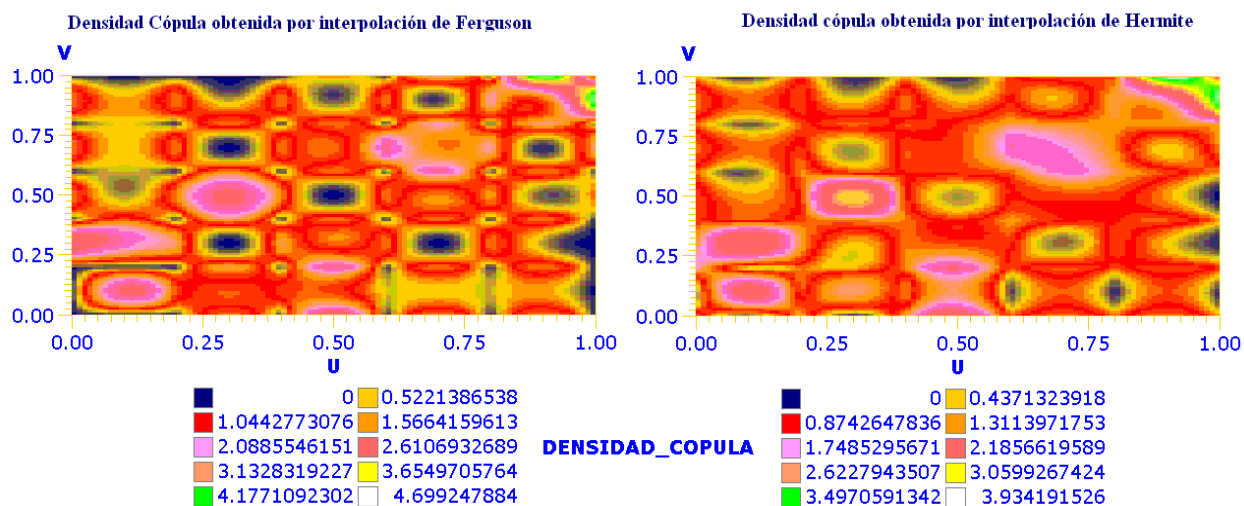


Ilustración 4.9: Comparativa entre la densidad cópula interpolada mediante superficies de Ferguson y la interpolada mediante polinomios de Hermite

Queda claramente puesta de manifiesto la mayor suavidad de la última cópula ajustada, pues presenta mayor continuidad en sus superficies coloreadas (no hay grandes cambios de nivel) y no sufre las bruscas caídas a cero en cada uno de los vértices de la partición (los cuales están representados en el gráfico de contorno asociado a la superficie de Ferguson por pequeños círculos en azul marino). Estas caídas resultan poco verosímiles en una relación de dependencia entre variables y más aún cuando se busca extrapolar ésta con fines predictivos. Obsérvese además el hecho de que, de igual manera que ocurría con el método de interpolación bilineal (ilustración (4.2)) y el que empleaba polinomios de Bernstein (ilustración (4.4)), existe una mayor densidad en la región asociada a las colas derechas (cercanas a 1) de las variables U y V , lo cual es indicio de que esta cópula, de igual manera que ocurre con otras conocidas como la HRT y la de Gumbel, enfatiza la relación entre valores extremos. Esto es gracias, como ya hemos señalado, al carácter empírico de las cópulas construidas, o sea, a la capacidad que tienen para ajustarse localmente a la muestra (ilustración (4.1)). Esto es un arma de doble filo dado que, si bien parece conveniente que se ajuste lo más posible a la muestra, puede ser que refleje relaciones espurias derivadas de la existencia de pares residuales en alguno de los rectángulos de la partición. No es tan grave la presencia de rectángulos “no informados” (sin datos) como la de rectángulos poco informados que puedan identificar relaciones inexistentes. En estos casos hay que garantizar que los rectángulos que lleven la parte más representativa de la relación tengan una densidad significativamente mayor. Así por ejemplo en el gráfico anterior puede observarse que la cópula pone un énfasis especial en el cuadrante superior derecho el cual está implicado en la ocurrencia conjunta de comportamientos extremos. De hecho, el número de puntos en el rectángulo superior derecho del gráfico mostrado en la ilustración (4.1) es superior al de cualquier otro rectángulo de la partición.

El último paso es establecer las condiciones bajo las cuales el problema (4.35) tiene solución factible, cuestión que habíamos dejado aparcada justo antes de empezar este desarrollo que ha permitido establecer las restricciones a satisfacer para que una superficie de Hermite genérica, construida por interpolación del dominio de una subcópula, sea una cópula. Esta factibilidad no está siempre garantizada y es la razón que provoca que no siempre sea posible conservar el valor óptimo del estadístico de Pearson obtenido sobre la clase de todas las subcópulas asociadas a la partición original.

4.2.3.3 Condiciones para construir cópulas por interpolación cúbica

Hemos establecido las restricciones que debe satisfacer la función $C_{HERMITE}$ (y $C_{FERGUSON}$) construida por interpolación cúbica de una subcópula para que pueda ser una cópula. Esto se ha traducido en encontrar solución a un problema de programación matemática (4.35) - (4.36) (y (4.21) - (4.26)) cuyas variables de decisión son las derivadas de primer y segundo orden de los puntos que definen el dominio de la subcópula. Sin embargo, faltaría conocer las condiciones que deben darse para que se pueda garantizar la existencia de una solución factible a dicho problema. Observemos que el verdadero problema es encontrar solución a (4.26) dado que:

- Si existe solución factible de (4.26) también existe para (4.21) dado que la región factible de este último contiene a la del primero.
- Como consecuencia, si existe solución factible de (4.21) también existe solución para (4.35) ya que al menos existe la solución en la que el valor de las derivadas primeras es el mismo que el que se haya obtenido para (4.21) y el valor para las derivadas segundas es 0.
- Del mismo modo, la existencia de solución para (4.26) garantiza la existencia de solución para (4.36) ya que al menos existe la solución en la que el valor de las derivadas primeras es el mismo que el que se haya obtenido para (4.26) y el valor para las derivadas segundas es 0.

Por ello centraremos nuestros esfuerzos en determinar las condiciones de factibilidad de (4.26) pero antes de ello, empezaremos proponiendo un ejemplo (más bien un contraejemplo) que muestra que, al construir la cópula interpoladora a partir de los lemas (4.2) y (4.3) y del problema (4.26), este último puede no tener solución factible.

Supongamos que $m = 2$, $n = 4$, y que la subcópula que proporciona el menor valor del estadístico de Pearson, y que es el resultado de resolver el problema (4.7), toma los siguientes valores en su dominio de definición:

$$\begin{aligned}
 C^*(0,1) &= 0, C^*\left(\frac{1}{4},1\right) = \frac{1}{4}, C^*\left(\frac{1}{2},1\right) = \frac{1}{2}, C^*\left(\frac{3}{4},1\right) = \frac{3}{4}, C^*(1,1) = 1 \\
 C^*\left(0,\frac{1}{2}\right) &= 0, C^*\left(\frac{1}{4},\frac{1}{2}\right) = \frac{1}{48}, C^*\left(\frac{1}{2},\frac{1}{2}\right) = \frac{1}{4}, C^*\left(\frac{3}{4},\frac{1}{2}\right) = \frac{3}{8}, C^*\left(1,\frac{1}{2}\right) = \frac{1}{2} \\
 C^*(0,0) &= 0, C^*\left(\frac{1}{4},0\right) = 0, C^*\left(\frac{1}{2},0\right) = 0, C^*\left(\frac{3}{4},0\right) = 0, C^*(1,0) = 0
 \end{aligned}$$

El siguiente gráfico ilustra perfectamente este ejemplo.

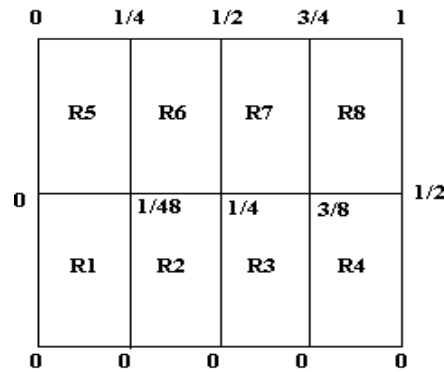


Ilustración 4.10: Contraejemplo. Un caso en el que el problema de programación matemática que permite construir una cópula por interpolación cúbica no tiene solución factible

Se puede comprobar que, de acuerdo a esta definición, C^* es efectivamente una subcópula pues se cumplen las propiedades (1.1) - (1.4). El dominio de definición es $S_1 \times S_2$ siendo $S_1 = \{0, \frac{1}{4}, \frac{2}{4}, \frac{3}{4}, 1\}$ y $S_2 = \{0, \frac{1}{2}, 1\}$, ambos contenidos en $[0,1]$ (propiedad (1.1)). Además, las dos condiciones frontera ((1.2) y (1.3)) se cumplen por la propia definición de C^* . Finalmente, para verificar (1.4), basta ver que el volumen de cualquiera de los rectángulos contenidos en el dominio de definición de C^* (véase el gráfico anterior) es mayor o igual que cero a través de la subcópula. Efectivamente,

$$V(R1) = \frac{1}{48} \geq 0, V(R2) = \frac{11}{48} \geq 0, V(R3) = \frac{1}{8} \geq 0, V(R4) = \frac{1}{8} \geq 0$$

$$V(R5) = \frac{11}{48} \geq 0, V(R6) = \frac{1}{48} \geq 0, V(R7) = \frac{1}{8} \geq 0, V(R8) = \frac{1}{8} \geq 0$$

Luego C^* es subcópula.

Sin embargo, vamos a ver que no es posible encontrar una relación de valores para las derivadas primeras respecto de V que satisfagan las restricciones del problema (4.26). Obsérvese previamente que $v_{j+1} - v_j = \frac{1}{2}$. Entonces, deberían cumplirse para las derivadas asociadas a $v_i = \frac{1}{2}$, las siguientes desigualdades asociadas a los rectángulos que quedan por encima y por debajo de dicha ordenada., es decir,

$$-3 \leq \frac{2*(v_{j+1} - v_j)}{V_{C^*}(B)} * \left[\frac{\partial C^*(u_i, v_j)}{\partial v} - \frac{\partial C^*(u_{i+1}, v_j)}{\partial v} \right] \leq 0 \Leftrightarrow -3 * V_{C^*}(B) \leq \left[\frac{\partial C^*(u_i, v_j)}{\partial v} - \frac{\partial C^*(u_{i+1}, v_j)}{\partial v} \right] \leq 0$$

En consecuencia, se tendría el siguiente conjunto de restricciones asociadas a cada uno de los rectángulos:

$$\begin{aligned}
R1 &\Rightarrow -3 * \frac{1}{48} \leq \left[\frac{\partial C^*(0, v_1)}{\partial v} - \frac{\partial C^*(u_1, v_1)}{\partial v} \right] \leq 0 \\
R2 &\Rightarrow -3 * \frac{11}{48} \leq \left[\frac{\partial C^*(u_1, v_1)}{\partial v} - \frac{\partial C^*(u_2, v_1)}{\partial v} \right] \leq 0 \\
R3 &\Rightarrow -3 * \frac{1}{8} \leq \left[\frac{\partial C^*(u_2, v_1)}{\partial v} - \frac{\partial C^*(u_3, v_1)}{\partial v} \right] \leq 0 \\
R4 &\Rightarrow -3 * \frac{1}{8} \leq \left[\frac{\partial C^*(u_3, v_1)}{\partial v} - \frac{\partial C^*(1, v_1)}{\partial v} \right] \leq 0 \\
R5 &\Rightarrow -3 * \frac{11}{48} \leq \left[\frac{\partial C^*(0, v_1)}{\partial v} - \frac{\partial C^*(u_1, v_1)}{\partial v} \right] \leq 0 \\
R6 &\Rightarrow -3 * \frac{1}{48} \leq \left[\frac{\partial C^*(u_1, v_1)}{\partial v} - \frac{\partial C^*(u_2, v_1)}{\partial v} \right] \leq 0 \\
R7 &\Rightarrow -3 * \frac{1}{8} \leq \left[\frac{\partial C^*(u_2, v_1)}{\partial v} - \frac{\partial C^*(u_3, v_1)}{\partial v} \right] \leq 0 \\
R8 &\Rightarrow -3 * \frac{1}{8} \leq \left[\frac{\partial C^*(u_3, v_1)}{\partial v} - \frac{\partial C^*(1, v_1)}{\partial v} \right] \leq 0
\end{aligned}$$

Se pueden eliminar aquellas restricciones que resultan redundantes, siendo el conjunto resultante:

$$\begin{aligned}
R1 &\Rightarrow -3 * \frac{1}{48} \leq \left[\frac{\partial C^*(0, v_1)}{\partial v} - \frac{\partial C^*(u_1, v_1)}{\partial v} \right] \leq 0 \\
R6 &\Rightarrow -3 * \frac{1}{48} \leq \left[\frac{\partial C^*(u_1, v_1)}{\partial v} - \frac{\partial C^*(u_2, v_1)}{\partial v} \right] \leq 0 \\
R7 &\Rightarrow -3 * \frac{1}{8} \leq \left[\frac{\partial C^*(u_2, v_1)}{\partial v} - \frac{\partial C^*(u_3, v_1)}{\partial v} \right] \leq 0 \\
R8 &\Rightarrow -3 * \frac{1}{8} \leq \left[\frac{\partial C^*(u_3, v_1)}{\partial v} - \frac{\partial C^*(1, v_1)}{\partial v} \right] \leq 0
\end{aligned}$$

Por otra parte, si para construir la cópula interpoladora imponemos que $\frac{\partial C^*(0, v_1)}{\partial v} = 0$ y $\frac{\partial C^*(1, v_1)}{\partial v} = 1$

(de acuerdo a los lemas (4.2) y (4.3)) para garantizar las propiedades frontera, se tiene que,

$$\begin{aligned}
R1 &\Rightarrow -3 * \frac{1}{48} \leq \left[\frac{\partial C^*(0, v_1)}{\partial v} - \frac{\partial C^*(u_1, v_1)}{\partial v} \right] \leq 0 \Rightarrow 0 \leq \frac{\partial C^*(u_1, v_1)}{\partial v} \leq \frac{3}{48} \\
R6 &\Rightarrow 0 \leq \frac{\partial C^*(u_1, v_1)}{\partial v} \leq \frac{\partial C^*(u_2, v_1)}{\partial v} \leq \frac{\partial C^*(u_3, v_1)}{\partial v} + \frac{3}{48} \leq \frac{3}{48} + \frac{3}{48} = \frac{6}{48} \\
R7 &\Rightarrow 0 \leq \frac{\partial C^*(u_1, v_1)}{\partial v} \leq \frac{\partial C^*(u_2, v_1)}{\partial v} \leq \frac{\partial C^*(u_3, v_1)}{\partial v} \leq \frac{\partial C^*(u_2, v_1)}{\partial v} + \frac{3}{8} \leq \frac{6}{48} + \frac{3}{8} = \frac{1}{2} \\
R8 &\Rightarrow \frac{5}{8} \leq \frac{\partial C^*(u_3, v_1)}{\partial v} \leq 1
\end{aligned}$$

Basta ver que las desigualdades asociadas a $R7$ y $R8$ generan intervalos disjuntos para $\frac{\partial C^*(u_3, v_1)}{\partial v}$, pues se debe

verificar que $\frac{\partial C^*(u_3, v_1)}{\partial v} \in [0, \frac{1}{2}]$ y que $\frac{\partial C^*(u_3, v_1)}{\partial v} \in [\frac{5}{8}, 1]$ lo cual es imposible.

Por tanto, no existe siempre solución factible al problema (4.26). La razón es el volumen tan bajo que se obtiene para los rectángulos $R1$ y $R6$ a través de C^* . Vamos a ver que la manera de resolver esta falta de factibilidad es imponer un volumen mínimo a través de C^* para los rectángulos contenidos en su dominio de definición.

Definición 4.2.- Se define una **2-subcópula- V_{MIN}** como aquella 2-subcópula C' tal que el volumen a través de ella de cualquier rectángulo B contenido en su dominio de definición es mayor o igual que V_{MIN} , es decir,

$$V_{c'}([u_1, u_2] \times [v_1, v_2]) = C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \geq V_{MIN} \quad (4.37)$$

$$\forall B = [u_1, u_2] \times [v_1, v_2] \subset \text{Dominio}(C') , V_{MIN} \in \mathbb{R}$$

Así, por ejemplo una 2-subcópula-0.25 verificaría que $V_{c'}(B) \geq \frac{1}{4} \quad \forall B \subset \text{Dominio}(C')$. El hecho de añadir el valor V_{MIN} como sufijo en vez de prefijo es por reservar esta posición para poder hacer referencia a la dimensión de la cópula (el número de variables sobre el que actúa). La definición se puede extender fácilmente a funciones cópula:

Definición 4.3.- Se define una **2-cópula- V_{MIN}** como aquella 2-cópula C tal que el volumen a través de ella de cualquier rectángulo B contenido en el rectángulo unidad es mayor o igual que V_{MIN} , es decir,

$$V_c([u_1, u_2] \times [v_1, v_2]) \geq 0 \quad \forall B = [u_1, u_2] \times [v_1, v_2] \subset I^2 , V_{MIN} \in \mathbb{R} \quad (4.38)$$

A partir de la definición (4.2), se puede establecer el siguiente teorema:

Teorema 4.6: Condición necesaria y suficiente para que el problema de programación lineal que permite determinar el valor de las derivadas en el dominio de la subcópula que optimiza el estadístico de Pearson, tenga solución

“Sea una partición del cuadrado unidad en $n \times m$ rectángulos del mismo área y sea la solución óptima al problema de programación fraccional (4.7) dada por $c_{ij}^* = C^*\left(\frac{i}{n}, \frac{j}{m}\right) \forall i \in \{0, 1, 2, \dots, n\} \quad \forall j \in \{0, 1, 2, \dots, m\}$ que define una subcópula C^* de dominio $\Delta = \{\text{los puntos de intersección entre los rectángulos que originan la partición}\}$.

La condición necesaria y suficiente para poder definir una cópula $C_{FERGUSON}$ por interpolación del dominio de C^* a través de las condiciones suficientes establecidas en los lemas (4.2), (4.3) y (4.5), es que C^* sea una 2-subcópula- V_{MIN}

siendo $V_{MIN} = \frac{2}{3 * m * n}$ ”.

La demostración de este resultado puede ser consultada en el Anexo VII.

El siguiente paso consiste en saber si sigue estando garantizada la existencia de solución factible a una modificación al problema de programación fraccional (4.7). La modificación consiste en imponer que las subcópulas obtenidas sean tales que, al aplicarlas sobre los rectángulos con vértices su propio dominio de definición, proporcionen un volumen

mayor o igual que $V_{MIN} = \frac{2}{3 * m * n}$. Queremos saber bajo qué condiciones dicho problema va a tener solución, algo que para el problema (4.7) quedaba asegurado a través del teorema (4.1).

El problema que resulta al hacer la modificación de (4.7) es el siguiente:

$$\begin{aligned} \text{Min} \sum_{i \in S_1} \sum_{j \in S_2} \frac{(N_{ij} - n * p_{ij})^2}{n * p_{ij}} &\Leftrightarrow \text{Min} \sum_{i \in S_1} \sum_{j \in S_2} \frac{(N_{ij} - n * (c'_{i,j} - c'_{i-1,j} - c'_{i,j-1} + c'_{i-1,j-1}))^2}{n * (c'_{i,j} - c'_{i-1,j} - c'_{i,j-1} + c'_{i-1,j-1})} \\ \text{s.a } c'_{i0} &= 0 \quad \forall i \in S_1; \quad c'_{0j} = 0 \quad \forall j \in S_2; \quad c'_{im} = \frac{i}{n} \quad \forall i \in S_1; \quad c'_{nj} = \frac{j}{m} \quad \forall j \in S_2 \\ p_{ij} &= c'_{i,j} - c'_{i-1,j} - c'_{i,j-1} + c'_{i-1,j-1} \geq V_{MIN} = \frac{2}{3 * m * n} \quad \forall i \in S_1, j \in S_2 \\ S_1 &= \{1, 2, \dots, n-1, n\} \quad S_2 = \{1, 2, \dots, m-1, m\} \end{aligned}$$

(4.39)

Basta observar que, de igual manera que ocurriera para el problema (4.7), la subcópula producto es una solución factible a este problema.

Teorema 4.7: Existencia de solución factible al problema de programación no lineal “modificado” que permite obtener una subcópula que optimiza el valor del estadístico de Pearson

“Sea una partición del cuadrado unidad en $n \times m$ rectángulos del mismo área. Entonces el problema de programación fraccional (4.39) tiene solución factible y va a proporcionar la relación de valores en cada uno de los puntos del dominio de definición de la subcópula que optimiza el estadístico de Pearson respecto del espacio de las

2-subcópulas- V_{MIN} siendo $V_{MIN} = \frac{2}{3 * m * n}$ ”.

demostración

Efectivamente, dada la subcópula producto $C'(\frac{i}{n}, \frac{j}{m}) = \frac{i}{n} * \frac{j}{m}$ se puede comprobar que es solución (no necesariamente la óptima) del problema (4.39) pues:

$$p_{ij} = c'_{i,j} - c'_{i-1,j} - c'_{i,j-1} + c'_{i-1,j-1} = \frac{i * j}{m * n} - \frac{(i-1) * j}{m * n} - \frac{i * (j-1)}{m * n} + \frac{(i-1) * (j-1)}{m * n} = \frac{1}{n * m} \geq \frac{2}{3 * n * m}$$

c.q.d

Obviamente, la región factible asociada a 2-subcópulas- V_{MIN} (problema (4.39)) está contenida en la que definen las 2-subcópulas (problema (4.7)) y por tanto, el valor que se obtenga para el estadístico de Pearson va a perder calidad. Así, al generar una cópula por interpolación con polinomios cúbicos de Hermite este valor va a ser mayor (y por tanto peor) que el que se obtendría si la interpolación se llevara a cabo con polinomios lineales (de orden 1 en U y V).

Teniendo en cuenta los lemas expuestos, podemos resumir que, dada una partición arbitraria del cuadrado unidad en $n \times m$ rectángulos del mismo área:

1. El problema de programación no lineal (4.39) tiene siempre solución. Este problema determina la relación de valores que debe tomar la 2-subcópula- V_{MIN} C^* en su dominio de definición ($V_{MIN} = \frac{2}{3 * m * n}$) para ser la que proporcione un valor más bajo del estadístico de Pearson dentro de la clase de las 2-subcópulas- V_{MIN} .
2. El problema (4.26) asociado a la subcópula C^* (y por tanto los problema (4.21),(4.35) y (4.36)) también tienen solución. En concreto, el problema (4.35) determina los valores que deben tomar las derivadas de primer y segundo orden de una función interpoladora de los valores de la subcópula C^* mediante polinomios cúbicos de Hermite para que dicha función interpoladora sea 2-creciente.
3. A dicha función es posible imponer de manera adicional las condiciones impuestas por los lemas (4.2) y (4.3) y obtener como resultado una función cópula $C_{HERMITE}$ que es una interpolación mediante polinomios cúbicos.

4.2.4 Algunas observaciones sobre las cópulas interpoladoras

Hemos visto tres formas distintas de interpolar los valores de la subcópula para la cual se obtiene un valor óptimo del estadístico de Pearson asociado al contraste de bondad de ajuste. De cada una de ellas, se puede hacer una síntesis:

- La interpolación mediante polinomios de orden 1 se puede llevar a cabo siempre a partir del conjunto de todas las subcópulas, es decir, sin necesidad de restringirse a la clase de las subcópulas- V_{MIN} . Sin embargo, la forma escalonada de la densidad cópula resultante se traduce en que no se hará discriminación respecto de todos los pares que pertenezcan a un mismo rectángulo. Su aspecto puede ser tan verosímil como el de la función de distribución empírica discreta (también escalonada) como aproximación de una distribución dada. El empleo de esta cópula requiere hacer particiones muy finas con una gran cantidad de rectángulos (siendo preciso entonces disponer de una muestra suficientemente grande) para poder esperar respuestas distintas de la variable V (y por tanto de la variable a predecir Y) a partir de un valor “ u ” del miembro conocido del par, U (del valor explicativo de X).

- La interpolación mediante polinomios de orden cúbico no se puede hacer a partir de una subcópula cualquiera, sino que precisa de que ésta pertenezca a una subclase que hemos definido como subcópulas- V_{MIN} donde

$V_{MIN} = \frac{2}{3 * m * n}$, las cuales son subcópulas tales que el volumen a través de ellas de cualquiera de los rectángulos contenidos en su dominio de definición es mayor o igual que V_{MIN} . Como consecuencia de ello, el valor del estadístico de Pearson asociado a la subcópula desde la cual se lleva a cabo la interpolación, deja de ser óptimo dentro de la clase de todas las posibles subcópulas asociadas a la partición y pasa a serlo dentro de

la clase de las subcópulas- V_{MIN} . Una vez hecha la interpolación bajo esta hipótesis, las cópulas obtenidas presentan una función de densidad caracterizada por una sucesión de montículos más o menos pronunciados pero que pueden ser suavizados en cierta medida mediante la aproximación del valor de la densidad en cada nodo de la partición (u_i, v_j) a un valor de referencia dado por (4.27) o por (4.32). Este aspecto característico de “valles” y “montañas” es una consecuencia directa de ajustar un polinomio bivalente cúbico a cada rectángulo de la partición y convierte a la cópula resultante en una buena candidata para reflejar relaciones multimodales. En este caso, la dimensión de la partición no tiene por qué ser necesariamente muy alto como pudiera ocurrir en el caso lineal. Obsérvese que, si bien particiones más finas pueden ser convenientes para detectar diferentes valores modales, en el caso más sencillo (cuando $n = m = 1$, es decir, en ausencia de partición), la cópula obtenida es, en cierto modo, una variante de las familias de cópulas con secciones cúbicas en U y V que se presentan en el apartado 3 de [NELQUERO] (expresión (4.3)) y que responden a una expresión general dada por:

$$C_{a,b}(u, v) = u*v + u*v*(1-u)*(1-v)*[a + b*(1-2u)*(1-2v)] \quad (4.40)$$

siendo $\rho = \frac{a}{3}$, $\tau = \frac{2*a}{9} + \frac{2*a*b}{225}$

Algunas de las variantes que se presentan en el artículo citado a modo de ejemplo son la familia de **Farlie-Gumbel-Morgenstern iterada de Kotz y Johnson** (véase (AI.50) en Anexo I), la familia de **Lin** (véase (AI.51)), la familia de **Kimeldorf y Sampson** (AI.35) y la familia de Sarmanov (AI.38). El siguiente gráfico pone de manifiesto el parentesco entre la densidad obtenida para uno de los representantes de estas familias y el que resulta de interpolar con polinomios de Hermite la subcópula que optimiza el valor del estadístico de Pearson en ausencia de partición (dimensión 1 x 1) y donde los valores de referencia respecto de los cuales se establece la aproximación de las derivadas de segundo orden son los de Bernstein.

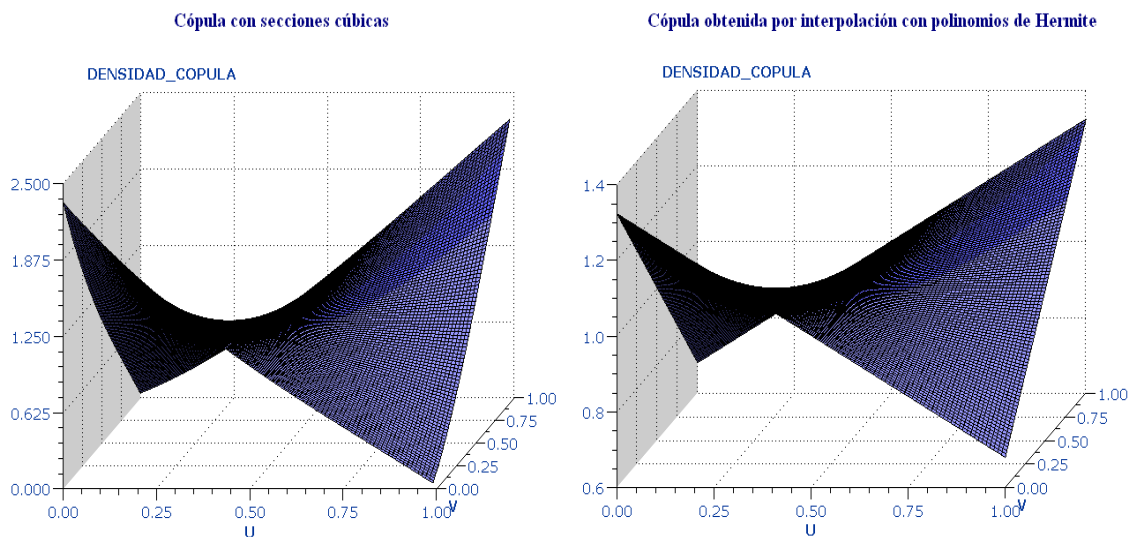


Ilustración 4.11: Comparativa entre la densidad de una cópula con secciones cúbicas en U y V y la obtenida por interpolación cúbica de la subcópula que optimiza el valor del estadístico de Pearson

De haber utilizado una superficie interpoladora de Ferguson, el valor de las derivadas segundas (y por tanto de la densidad cópula) en cada uno de los cuatro vértices del cuadrado unidad hubiese sido cero. En dicho caso, el aspecto de la densidad obtenida sin realizar partición alguna hubiese sido el que se muestra en el siguiente gráfico:

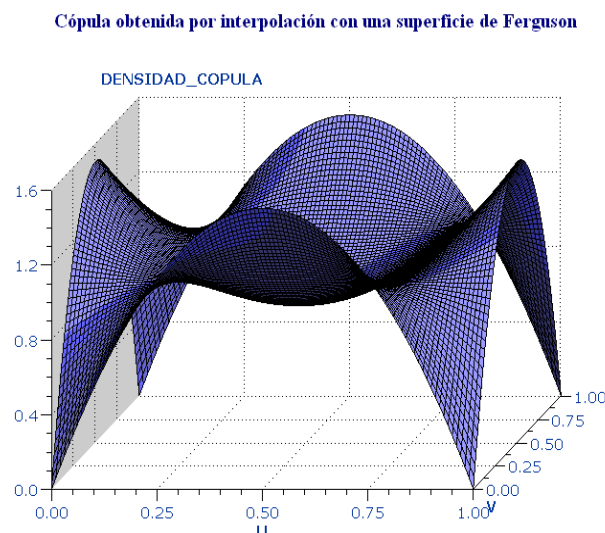


Ilustración 4.12: Densidad cópula construida mediante superficies de Ferguson en ausencia de partición

- La interpolación mediante polinomios de Bernstein trata de suplir las carencias de las cópulas anteriores desde el punto de vista de la uniformidad de su función de densidad. Si bien es verdad que las cópulas anteriores se adaptan localmente a los datos, el ajuste puede presentar demasiados cambios de comportamiento poco esperables en la mayor parte de las relaciones de dependencia entre variables. La regularidad puede ser una característica deseable en la relación de dependencia entre las variables de estudio y aun cuando pueda conseguirse mayor o menor suavidad de la superficie interpoladora cúbica mediante la “aproximación” a los valores de referencia, la aproximación mediante polinomios de Bernstein puede ofrecer una solución más satisfactoria. Hay que resaltar que, en contra de este ajuste, juega el hecho de que no se trata de una interpolación exacta del dominio de definición de la subcópula que optimiza el estadístico de Pearson, sino de una aproximación a los valores de la subcópula en dicho dominio.

- Para la partición de dimensión 5 x 5 mostrada en la ilustración (4.1), hemos visto que los estadísticos obtenidos para los representantes de diferentes familias de cópulas convencionales (tabla (4.1)), proporcionan p-valores a partir de los cuales no es posible rechazar la hipótesis del contraste de bondad de ajuste a la cópula de expresión desconocida. Pero también puede suceder todo lo contrario, es decir, que todas las cópulas sometidas al contraste conduzcan al rechazo del mismo, siendo en tales ocasiones, en las que las cópulas construidas empíricamente por interpolación del dominio de aquella que minimiza el valor del estadístico de Pearson, puedan suponer una buena alternativa de solución.

A modo de contraejemplo, podemos considerar otra nube de puntos asociada al mes de Marzo (la presentada en la ilustración (4.1) correspondía al mes de Noviembre).

Según podemos ver en el siguiente gráfico, a través de la cópula construida por interpolación bilineal de la subcópula que optimiza la expresión de Pearson, conseguiremos un valor de ésta suficientemente pequeño.

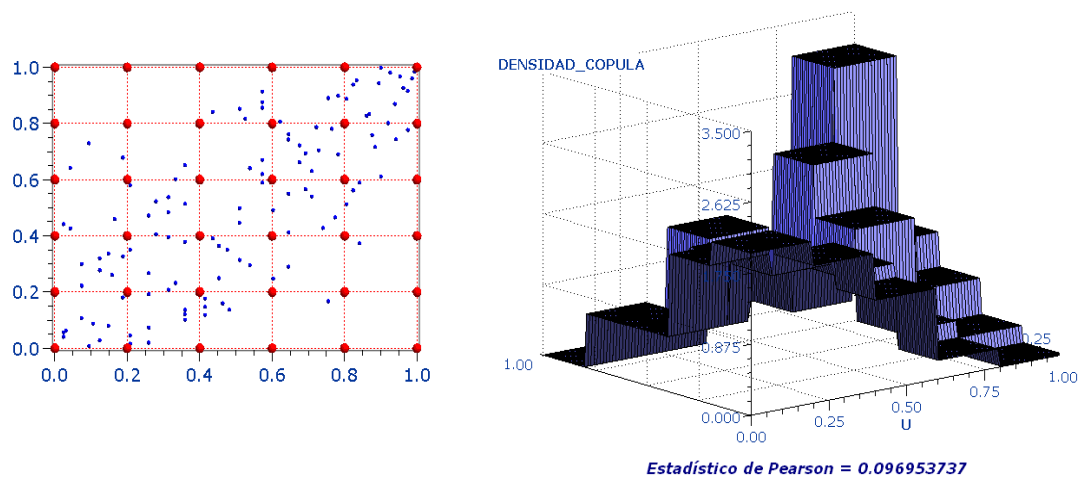


Ilustración 4.13: Densidad cópula construida por interpolación bilineal

Sin embargo, si calculamos los valores de los estadísticos de Pearson para las mismas cópulas candidatas que planteábamos en el ejemplo anterior, veremos que ninguno de dichos valores permite “aceptar” (no rechazar) el contraste de bondad de ajuste al 90% de confianza. Al 95%, la cópula gaussiana y la de Gumbel sí superarían dicho contraste. Estas observaciones pueden ser confirmadas en la siguiente tabla, la cual muestra los p-valores asociados a los estadísticos calculados para cada una de las cópulas utilizadas.

Contraste de bondad de ajuste a una Cópula

Mes = 03

COPULA	PARAMETRO_COPULA_CHAR	PEARSON	P_VALOR
NOR	0.7263	22.8813	0.086694
GUM	2.0728	22.9550	0.085099
F14	1.5728	26.0486	0.037515
PLA	13.469	29.4998	0.013859
F12	1.3819	31.1235	0.008457
HRT	0.4660	31.2796	0.008058
FRA	1.1503	42.8999	0.000163
ALI	0.4829	44.6184	0.000088
CLA	2.1457	54.2155	0.000002
PRO	.	68.2569	0.000000

La región crítica viene dada por $\{X|X > 24.99579014\}$

Tabla 4.2: Evaluación del estadístico de Pearson para una serie de cópulas

Con esta observación queremos hacer hincapié en que el empleo de las cópulas interpoladoras puede resultar interesante en aquellos casos en los cuales ninguna de las familias candidatas proporcione un ajuste suficientemente bueno, pero parece existir evidencia de dependencia entre las variables (el p-valor más bajo corresponde a la cópula producto).

Así, proponemos la construcción de estas cópulas interpoladoras en aquellas circunstancias en las cuales ninguna de las cópulas candidatas nos ofrezca ciertas garantías de que la distribución conjunta obtenida a través de cada una de ellas (mediante el teorema de Sklar), lleve implícita la verdadera relación de dependencia existente entre las variables.

Nuestro propósito final consiste en automatizar el proceso de selección de la función cópula que mejor refleje la relación entre un par de variables (*demanda de gas/electricidad y temperatura*), de cara a su implementación informática, disponiendo de un recurso adicional ante la posibilidad de que los p-valores asociados a las distintas candidatas, conduzcan al rechazo de la hipótesis planteada por el contraste de bondad de ajuste a distribuciones bidimensionales. Este recurso adicional viene dado por una cópula empírica y no paramétrica que respecto de la expresión de Pearson, que es a fin de cuentas una medida de proximidad entre lo observado en la muestra y lo teóricamente esperado a través de la cópula, presente un valor lo más bajo posible.

5. Simulación de valores para las cópulas construidas por interpolación

Como se explica en el apartado 2.5.6 las cópulas pueden ser utilizadas para hacer predicciones a partir de la simulación de valores de la distribución de sus condicionadas $C_1 = C_{V|U}$. Se trata de conocer el valor que se espera para la variable Y en función del valor conocido “ x ” de la variable X o lo que es lo mismo, la esperanza asociada a la variable $Y|X=x$. Para calcular esta esperanza, se emplea la función de distribución $H_{Y|X}$ de la cual se genera un conjunto suficientemente grande de valores mediante técnicas de simulación y se puede seleccionar entre ellos por ejemplo la media o la mediana. La simulación de valores de $H_{Y|X}$ se plantea a partir de la cópula condicionada $C_1 = C_{V|U}$ con la cual se encuentra relacionada a través del teorema de Sklar (véase expresión (1.11)). Esto es así puesto que son las funciones cópula las que se utilizan para calibrar la relación de dependencia entre X e Y (a través de U y V) y seleccionar la distribución conjunta con marginales X e Y que mejor se adapte a dicha relación. Dado que la simulación de valores se hace para $C_1 = C_{V|U}$, los valores simulados “ v ” deben ser posteriormente “destransformados” a través de la marginal de la variable Y para obtener valores simulados “ y ”. Como bien se resume en el apartado 2.4.2 de [LIU] la simulación se puede hacer en los siguientes pasos:

1. Generar u^s y ϵ^s valores aleatorios que se distribuyen según una $U(0,1)$. En nuestro caso “ u ” viene dado por $F_X(x)$.
2. Calcular $v^s = C_{2|1}^{-1}(\epsilon^s | u^s)$; entonces (u^s, v^s) es una simulación de la cópula condicionada C_1 ($C_{2|1} = C_1$).
3. Hacer $r_X^s = F_X^{-1}(u^s)$ y $r_Y^s = F_Y^{-1}(v^s)$ (en nuestro caso bastaría con calcular éste último) para obtener los puntos originales.

La técnica de simulación que se propone en el apartado al que hacemos referencia es la de la transformada inversa. Este método para variables aleatorias continuas X , se basa en lo siguiente:

Sea $X \stackrel{d}{=} F_X$ y $U \stackrel{d}{=} U(0,1) \Rightarrow Z = F_X^{-1}(U) \stackrel{d}{=} F_X$ pues $F_Z(z) = P(Z \leq z) = P(F_X^{-1}(U) \leq z)$. Como F_X es la función de distribución de una variable aleatoria continua, es monótona y estrictamente creciente, lo cual permite sacar en conclusión que $P(F_X^{-1}(U) \leq z) = P(U \leq F_X(z)) = F_X(z)$. En consecuencia, las funciones de distribución de X y Z coinciden y por tanto, para simular valores de X basta simularlos para $Z = F_X^{-1}(U)$, lo cual se traduce en igualar la función de distribución $C_1(u,v)$ a un valor “ p ” distribuido según una uniforme estándar y despejar “ v ” para obtener una expresión en función del valor “ u ” al cual se condiciona y en función también de “ p ”. Bien es verdad que la inversa de la cópula condicional no siempre puede calcularse por estos métodos convencionales (por ejemplo para la t-cópula o cópula de la t-student) siendo necesario en dichos casos recurrir a aproximaciones numéricas.

5.1 Simulación de valores para algunas familias de cópulas conocidas

En artículos como el de [LIU], el de [ARMGAL], o el de [VENTER], por poner algunos ejemplos, se citan familias de cópulas para las que es posible calcular la inversa de su condicionada y se presenta también la expresión a la que responde el valor generado para $V|U$. Algunos ejemplos son:

- Cópula de Cola Derecha Pesada (véase (AI.12) en Anexo I)

$$v = 1 - \left\{ 1 - (1-u)^{\frac{-1}{a}} + [1 - p * (1-u)^{1+\frac{-1}{a}}]^{-a} \right\}^{-a} \quad (5.1)$$

$$\text{siendo } a = \frac{1-\tau}{2 * \tau} \quad (5.2)$$

- Cópula de Clayton (véase (AI.11) en Anexo I)

$$v = \left\{ \left(\frac{u^{-1-a}}{p} \right)^{\frac{a}{1+a}} + 1 - u^{-a} \right\}^{\frac{-1}{a}} \quad (5.3)$$

$$\text{siendo } a = \frac{2 * \tau}{1-\tau} \quad (5.4)$$

- Cópula de Frank (véase (AI.13) en Anexo I)

$$v = \frac{-1}{a} * \ln \left[1 + \frac{p * g_1}{1 + g_u * (1-p)} \right] \quad (5.5)$$

$$\text{siendo } g_z = e^{-a * z} - 1 \quad (5.6)$$

$$\text{y donde "a" se puede calcular a partir de la expresión } \tau = 1 - \frac{4}{a} + \frac{4}{a^2} * \int_0^a \frac{t}{e^t - 1} dt \quad (5.7).$$

La integral de esta expresión (función de **Debye**) se puede resolver a través de la siguiente aproximación

$$\int_0^a \frac{t}{e^t - 1} dt = a * \left[1 - \frac{a}{4} + \frac{\sum_{k=1}^{\infty} B_{2k} * a^{2k}}{(2k+1) * (2 * k!)} \right] \quad (5.8)$$

donde B_n denota a los números de **Bernuolli** (véase el apéndice 7 de [MOLIFE]).

- Cópula de Farlie-Gumbel-Morgenstern (véase (AI.34) en Anexo I)

$$v = \frac{-1 + a * (2 * u - 1) + \sqrt{(1 - a * (2 * u - 1))^2 + 4 * p * a * (2 * u - 1)}}{2 * a * (2 * u - 1)} \quad (5.9)$$

$$\text{siendo } a = \frac{9 * \tau}{2} \quad (5.10)$$

- Cópula Normal (véase (AI.5) en Anexo I)

$$v = \Phi[\rho * \Phi^{-1}(u) + \sqrt{1 - \rho^2} * \Phi^{-1}(p)] \quad (5.11)$$

$$\text{siendo } \rho = \text{sen}\left(\frac{\pi * \tau}{2}\right) \quad (5.12)$$

y donde Φ es la función de distribución normal estándar univariante

Vamos a ver ahora cuál es el resultado de aplicar el método de la transformada inversa a las cópulas construidas por métodos de interpolación que hemos citado en el apartado 4.

5.2 Simulación de valores para la cópula construida por interpolación bilineal

Supongamos que “ x ” es el valor al que se condiciona (valor de la variable explicativa) y $u = F_X(x)$ el valor transformado a través de la función de distribución de X (que podemos suponer que es la empírica continua). La cópula condicionada asociada a la cópula construida por interpolación bilineal (4.8) se puede calcular derivando dicha expresión respecto de U . El resultado es:

$$C_1(u, v) = \frac{\partial C(u, v)}{\partial u} = \frac{v * C'(u_i, v_j) - C'(u_i, v_{j+1}) - C'(u_{i+1}, v_j) + C'(u_{i+1}, v_{j+1})}{(v_{j+1} - v_j) * (v_{i+1} - v_i)} + \frac{v_{j+1} * C'(u_{i+1}, v_j) - C'(u_i, v_j) + v_j * C'(u_i, v_{j+1}) - C'(u_{i+1}, v_{j+1})}{(v_{j+1} - v_j) * (v_{i+1} - v_i)} = a_j * v + b_j \quad \forall v \in [v_j, v_{j+1}] \quad (5.13)$$

$$\text{siendo } a_j = \frac{C'(u_i, v_j) - C'(u_i, v_{j+1}) - C'(u_{i+1}, v_j) + C'(u_{i+1}, v_{j+1})}{(v_{j+1} - v_j) * (v_{i+1} - v_i)} = \frac{V_C([u_i, u_{i+1}]x[v_j, v_{j+1}])}{(u_{i+1} - u_i) * (v_{j+1} - v_j)} \quad (5.14)$$

$$\text{y } b_j = \frac{v_{j+1} * C'(u_{i+1}, v_j) - C'(u_i, v_j) + v_j * C'(u_i, v_{j+1}) - C'(u_{i+1}, v_{j+1})}{(v_{j+1} - v_j) * (v_{i+1} - v_i)} \quad (5.15)$$

Observemos que la ecuación de la densidad cópula (véase (4.9)) es fácil de calcular a partir de (5.13) pues

$$c(u, v) = \frac{\partial C_1(u, v)}{\partial v} = a_j = \frac{V_C([u_i, u_{i+1}]x[v_j, v_{j+1}])}{(u_{i+1} - u_i) * (v_{j+1} - v_j)}$$

Sea ahora “ p ” un valor extraído de una distribución uniforme (0,1). Si “ p ” es tal que $a_j * v_j + b_j \leq p \leq a_j * v_{j+1} + b_j$, entonces igualando a “ p ” la expresión (5.13) y despejando “ v ” se llega a la expresión que permite hacer simulaciones de la cópula construida por interpolación bilineal:

$$v = \frac{p * (v_{j+1} - v_j) * (u_{i+1} - u_i)}{V_C(B_j)} + v_{j+1} * \frac{C'(u_i, v_j) - C'(u_{i+1}, v_j)}{V_C(B_j)} + v_j * \frac{C'(u_{i+1}, v_{j+1}) - C'(u_i, v_{j+1})}{V_C(B_j)} \quad (5.16)$$

verificándose que $v_j \leq v \leq v_{j+1}$

5.3 Simulación de valores para la cópula construida por interpolación cúbica

La cópula condicionada asociada a la cópula construida por interpolación cúbica mediante polinomios de Hermite (4.19) se puede calcular derivando dicha expresión respecto de U . El resultado es:

$$\frac{\partial C_{HERMITE}(u, v)}{\partial u} = A * \left(\frac{2 * (v - v_j)^3}{(v_{j+1} - v_j)^3} - \frac{3 * (v - v_j)^2}{(v_{j+1} - v_j)^2} + 1 \right) + B * \left(\frac{-2 * (v - v_j)^3}{(v_{j+1} - v_j)^3} + \frac{3 * (v - v_j)^2}{(v_{j+1} - v_j)^2} \right) + C * \left(\frac{(v - v_j)^3}{(v_{j+1} - v_j)^2} - \frac{2 * (v - v_j)^2}{(v_{j+1} - v_j)} + (v - v_j) \right) + D * \left(\frac{(v - v_j)^3}{(v_{j+1} - v_j)^2} - \frac{(v - v_j)^2}{(v_{j+1} - v_j)} \right) \quad \forall v \in [v_j, v_{j+1}]$$

(5.17)

siendo

$$A = \left[\left(\frac{6 * (u - u_i)^2}{(u_{i+1} - u_i)^3} - \frac{6 * (u - u_i)}{(u_{i+1} - u_i)^2} \right) * C'(u_i, v_j) + \left(\frac{-6 * (u - u_i)^2}{(u_{i+1} - u_i)^3} + \frac{6 * (u - u_i)}{(u_{i+1} - u_i)^2} \right) * C'(u_{i+1}, v_j) + \left(\frac{3 * (u - u_i)^2}{(u_{i+1} - u_i)^2} - \frac{4 * (u - u_i)}{(u_{i+1} - u_i)} + 1 \right) * \frac{\partial C'(u_i, v_j)}{\partial u} + \left(\frac{3 * (u - u_i)^2}{(u_{i+1} - u_i)^2} - \frac{2 * (u - u_i)}{(u_{i+1} - u_i)} \right) * \frac{\partial C'(u_{i+1}, v_j)}{\partial u} \right]$$

(5.18)

$$B = \left[\left(\frac{6 * (u - u_i)^2}{(u_{i+1} - u_i)^3} - \frac{6 * (u - u_i)}{(u_{i+1} - u_i)^2} \right) * C'(u_i, v_{j+1}) + \left(\frac{-6 * (u - u_i)^2}{(u_{i+1} - u_i)^3} + \frac{6 * (u - u_i)}{(u_{i+1} - u_i)^2} \right) * C'(u_{i+1}, v_{j+1}) + \left(\frac{3 * (u - u_i)^2}{(u_{i+1} - u_i)^2} - \frac{4 * (u - u_i)}{(u_{i+1} - u_i)} + 1 \right) * \frac{\partial C'(u_i, v_{j+1})}{\partial u} + \left(\frac{3 * (u - u_i)^2}{(u_{i+1} - u_i)^2} - \frac{2 * (u - u_i)}{(u_{i+1} - u_i)} \right) * \frac{\partial C'(u_{i+1}, v_{j+1})}{\partial u} \right]$$

(5.19)

$$C = \left[\left(\frac{6 * (u - u_i)^2}{(u_{i+1} - u_i)^3} - \frac{6 * (u - u_i)}{(u_{i+1} - u_i)^2} \right) * \frac{\partial C'(u_i, v_j)}{\partial v} + \left(\frac{-6 * (u - u_i)^2}{(u_{i+1} - u_i)^3} + \frac{6 * (u - u_i)}{(u_{i+1} - u_i)^2} \right) * \frac{\partial C'(u_{i+1}, v_j)}{\partial v} + \left(\frac{3 * (u - u_i)^2}{(u_{i+1} - u_i)^2} - \frac{4 * (u - u_i)}{(u_{i+1} - u_i)} + 1 \right) * \frac{\partial^2 C'(u_i, v_j)}{\partial u \partial v} + \left(\frac{3 * (u - u_i)^2}{(u_{i+1} - u_i)^2} - \frac{2 * (u - u_i)}{(u_{i+1} - u_i)} \right) * \frac{\partial^2 C'(u_{i+1}, v_j)}{\partial u \partial v} \right]$$

(5.20)

$$D = \left[\left(\frac{6 * (u - u_i)^2}{(u_{i+1} - u_i)^3} - \frac{6 * (u - u_i)}{(u_{i+1} - u_i)^2} + 1 \right) * \frac{\partial C'(u_i, v_{j+1})}{\partial v} + \left(\frac{-6 * (u - u_i)^2}{(u_{i+1} - u_i)^3} + \frac{6 * (u - u_i)}{(u_{i+1} - u_i)^2} \right) * \frac{\partial C'(u_{i+1}, v_{j+1})}{\partial v} + \left(\frac{3 * (u - u_i)^2}{(u_{i+1} - u_i)^2} - \frac{4 * (u - u_i)}{(u_{i+1} - u_i)} + 1 \right) * \frac{\partial^2 C'(u_i, v_{j+1})}{\partial u \partial v} + \left(\frac{3 * (u - u_i)^2}{(u_{i+1} - u_i)^2} - \frac{2 * (u - u_i)}{(u_{i+1} - u_i)} \right) * \frac{\partial^2 C'(u_{i+1}, v_{j+1})}{\partial u \partial v} \right]$$

(5.21)

La cópula condicionada asociada a la expresión (4.20) (interpolación realizada mediante una superficie de Ferguson), se obtendría dando valor 0 a las derivadas segundas.

Para la simulación de valores en uno u otro caso, bastaría igualar la expresión correspondiente a un valor “p” generado a partir de una uniforme estándar y resolver en “v”. Podemos reescribir (5.17) como una función de “v” de la siguiente forma:

$$\left\{ \frac{2 * (A - B)}{(v_{j+1} - v_j)^3} + \frac{(C - D)}{(v_{j+1} - v_j)^2} \right\} * v^3 + \left\{ \frac{-6 * (A - B) * v_j}{(v_{j+1} - v_j)^3} + \frac{3 * (B - A)}{(v_{j+1} - v_j)^2} - \frac{3 * (C + D) * v_j}{(v_{j+1} - v_j)^2} + \frac{(-2 * C - D)}{(v_{j+1} - v_j)} \right\} * v^2 + \left\{ \frac{6 * (A - B) * v_j^2}{(v_{j+1} - v_j)^3} - \frac{6 * (B - A) * v_j}{(v_{j+1} - v_j)^2} + \frac{3 * (C + D) * v_j^2}{(v_{j+1} - v_j)^2} - \frac{2 * (-2 * C - D)}{(v_{j+1} - v_j)} * v_j + C \right\} * v + \left\{ \frac{-2 * (A - B) * v_j^3}{(v_{j+1} - v_j)^3} + \frac{3 * (B - A) * v_j^2}{(v_{j+1} - v_j)^2} + A - \frac{(C + D) * v_j^3}{(v_{j+1} - v_j)^2} + \frac{(-2 * C - D)}{(v_{j+1} - v_j)} * v_j^2 - C * v_j \right\} = p$$

(5.22)

Esta ecuación se puede resolver mediante la fórmula de **Cardano-Tartaglia** pudiendo darse alguna de las siguientes posibilidades:

- Una de las soluciones a la ecuación (5.22) es real y las otras dos complejas. En dicho caso, la solución real “v” es la que deberemos tomar como valor simulado.
- Se obtiene una única raíz real (triple). En dicho caso, éste es el valor “v” simulado.
- Se obtiene una raíz doble y una raíz simple. Se tomarán como valores “v” simulados aquellas soluciones que pertenezcan al soporte de la variable V (deberán estar entre 0 y 1).
- Se obtienen tres raíces simples. De la misma forma que en el caso anterior, sólo interesan las soluciones “v” que pertenezcan al soporte de V (que se muevan entre 0 y 1).

Cualquiera que sea el caso, una vez obtenido el valor “v” simulado correspondiente a la distribución de V , será necesario “destransformar” a través de la función de distribución estimada para Y (habíamos considerado también la función de distribución empírica continua), para obtener el correspondiente valor “y” simulado condicionado a “x” conocido. Sucesivas simulaciones de valores “p” proporcionarán sucesivos “y”. La media o mediana de dichos valores se tomará como valor predicho de la variable Y condicionado a $X = x$.

5.4 Simulación de valores para la cópula construida por interpolación mediante polinomios de Bernstein

La cópula condicionada asociada a la cópula construida por interpolación mediante polinomios de Bernstein (4.11) se puede calcular derivando dicha expresión respecto de U . El resultado es:

$$C_1^B(u, v) = \frac{\partial C^B(u, v)}{\partial u} = \sum_{k=0}^m \sum_{l=0}^m \alpha\left(\frac{k}{m}, \frac{l}{m}\right) * \binom{m}{k} * P_{l,m}(v) * [k * u^{k-1} * (1-u)^{m-k} - u^k * (m-k) * (1-u)^{m-k-1}] \quad (5.23)$$

siendo $P_{l,m}(v) = \binom{m}{l} * v^l * (1-v)^{m-l} \forall l \in \{0, 1, 2, \dots, m\}$ (5.24)

La inversión de esta condicionada es complicada. Habríamos de igualar la expresión (5.23) a un valor “p” generado a partir de una $U(0,1)$ y despejar “v”. En su lugar se puede resolver la ecuación resultante mediante aproximaciones numéricas.

6. Empleo de funciones cópula para predecir

En el apartado 2.5.6 explicábamos cómo se pueden realizar predicciones utilizando una función cópula mediante la simulación de valores de la derivada de dicha función (cópula condicionada). El empleo de técnicas de simulación suele estar orientado a hacer predicciones cuándo se desconoce el valor de los factores que intervienen en el modelo (variables explicativas). En estos casos, el analista tiende a configurar escenarios en función de ciertos valores propuestos para dichos factores y realiza simulaciones del comportamiento que cabría esperar para la variable a predecir bajo dichos condicionantes. A través de la distribución de los valores simulados se puede proponer una predicción dada por el valor esperado (por ejemplo la media o la mediana) bajo las circunstancias que definen el escenario e incluso evaluar el riesgo asociado a dicha predicción.

6.1 Empleo de funciones cópula como modelos de función de transferencia

El objetivo que vamos a perseguir con las cópulas es hacer una predicción para una serie temporal, fin al que está destinado el bloque práctico de esta tesis. Si el horizonte que se plantea es medio/largo, los factores explicativos serán probablemente desconocidos y difíciles de estimar, por lo que no parece extraño el utilizarlas como método de simulación. Sin embargo, pretendemos ir más allá y utilizarlas también para hacer una predicción a corto plazo y supuestos conocidos los valores de los agentes que intervienen en el modelo. En concreto, lo que se persigue es partir de la predicción proporcionada por un modelo ARIMA ajustado con la metodología **Box-Jenkins**, y suplir con funciones cópula el papel desempeñado por los modelos de función de transferencia cuando éstos no consiguen ajustar bien la relación existente entre la variable a predecir y la variable explicativa. La metodología Box-Jenkins, que a continuación exponemos a grandes rasgos, permite el ajuste de modelos ARIMA para hacer una predicción a corto plazo, proporcionando buenos resultados si la serie reúne ciertas condiciones de estabilidad, disponibilidad de histórico, etc. En primer lugar, y esto debe ser un paso previo a cualquier estudio de predicción, conviene reservar parte del histórico para hacer predicciones fuera de muestra (*out-of-sampling*) y emplear el resto para ajustar el modelo (véase el punto I en el apartado 2.5.6). Si se trata de realizar una predicción diaria fuertemente influenciada por el calendario, sería suficiente reservar un año completo como **conjunto de validación** y disponer de al menos 2 o 3 años para ajustar el modelo (**conjunto de entrenamiento**). El conjunto de validación sirve para “validar” si se puede esperar que el modelo proporcione buenos resultados a futuro, utilizando un conjunto que no participa en el ajuste del modelo.

Como se explica en [AZTRI], la metodología Box-Jenkins consta de las siguientes fases:

- I. FASE DE IDENTIFICACIÓN.- que conlleva los siguientes pasos:
 1. Determinar si es necesario o no transformar los datos (logaritmo, raíz cuadrada, etc.) con el fin de disminuir (“explicar”) la variabilidad de los datos: parámetro λ de la transformación de **Box-Cox**.
 2. Evaluar la necesidad de hacer diferencias regulares con el fin de que el modelo recoja (“explique”) la

posible tendencia de la serie: parámetro d del modelo $ARIMA(p, d, q)$.

3. Evaluar la necesidad de hacer diferencias estacionales que reflejen (“expliquen”) posibles periodicidades: parámetro D del modelo $SARIMA(p, d, q) \times (P, D, Q)$ (en caso de que la tendencia se incorpore al modelo de forma multiplicativa).

Estos tres primeros pasos van orientados a hacer que la serie a modelizar mediante un $ARIMA$, sea un proceso estacionario en varianza (1) y en media (2 y 3) dado que esta es una hipótesis que se hace cuando se ajustan este tipo de modelos: “la distribución de los datos no cambia con el tiempo”.

4. Identificar otros órdenes del modelo que determinen la dependencia del valor a predecir en función del histórico de datos reales (p, P), y otros que hagan que la predicción se corrija en función de los errores cometidos por el propio modelo (q, Q).
5. Realizar un análisis de intervenciones para que queden explicados en el modelo determinados eventos aislados identificados a partir del propio histórico.
6. Incluir modelos de función de transferencia (variables explicativas).

II. FASE DE ESTIMACIÓN.- supone estimar la relación de parámetros asociadas a cada una de las variables que surgen de los pasos 4,5 y 6. Los métodos convencionales a tal fin son: el método de la **máxima verosimilitud** (ML), el método de los **mínimos cuadrados condicionados** (CLS) y el método de los **mínimos cuadrados no condicionados** (ULS).

III. FASE DE CONTRASTE.- supone contrastar la significatividad de los parámetros estimados, de las correlaciones entre ellos, la hipótesis de ruido blanco, etc.

Los parámetros básicos del modelo (determinados por los órdenes p y q), así como los asociados a las variables de intervención y a los modelos de función de transferencia deben ser incluidos poco a poco e ir contrastando continuamente su significatividad y correlación. Así, las fases I-II-III deben repetirse paulatinamente hasta que el proceso de error resultante sea ruido blanco y todos los parámetros incluidos en el modelo sean significativos y no presenten altas correlaciones entre ellos.

IV. FASE DE PREDICCIÓN.- consiste en aplicar el modelo sobre una muestra que no ha sido utilizada para el ajuste (conjunto de validación) y determinar si la calidad de las predicciones a diferentes horizontes temporales es buena o no:

- Si se considera buena, utilizar el modelo para hacer predicciones hasta que éste se “desajuste” (parámetros que dejan de ser significativos, el error deja de ajustarse a un proceso de ruido blanco, etc.). Esto puede venir provocado por un cambio de comportamiento de la serie.

- Si no se estima satisfactoria, regresar a la fase de identificación, proponer otro modelo y volver a esta fase de predicción. Si tras varios ajustes ninguno resulta satisfactorio, puede ser que la metodología Box-Jenkins no sea apropiada para el tratamiento de la serie en cuestión siendo aconsejable en dicho caso, el empleo de otro tipo de metodología alternativa (posiblemente no lineal).

El siguiente esquema, extraído básicamente de [AZTRI], resume perfectamente la metodología Box-Jenkins:

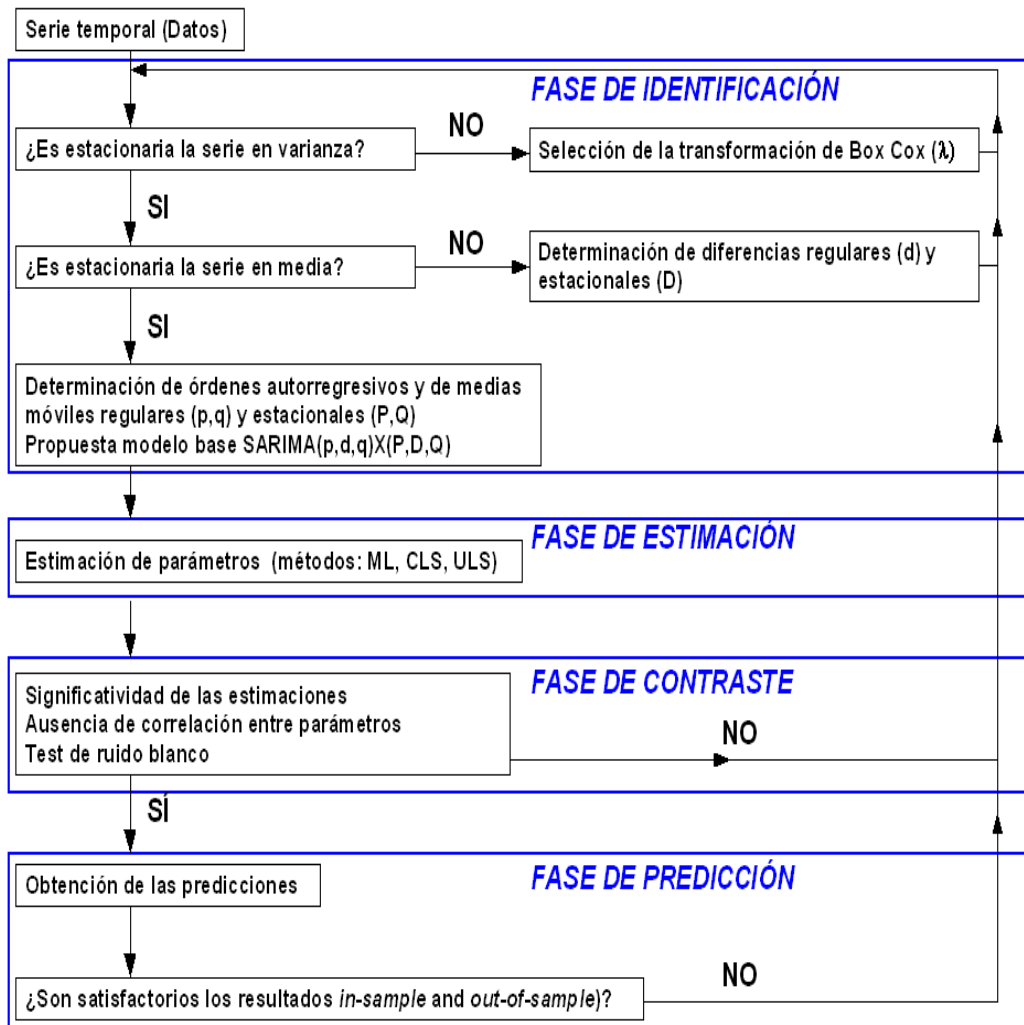


Ilustración 6.1: Metodología Box-Jenkins

Queremos destacar en qué consiste realmente el proceso iterativo de las fases I-II-III. Obsérvese que de lo que realmente se trata es de ir incorporando información en una fórmula buscando que aquello que quede sin explicar (proceso de error) sea ruido blanco, es decir, algo impredecible. Dicho de otra manera, cada vez que incorporamos una variable al modelo, estamos explicando parte del comportamiento de la serie para, posteriormente, tratar de justificar con otras variables lo que vaya quedando por explicar (error). El proceso de error va siendo relacionado en cada iteración con la información disponible hasta que no quede nada más por explicar (ruido blanco) o bien hasta que no

tengamos más información para intentar hacerlo. En esta línea, podría pensarse en un método de ajuste secuencial del proceso de error mediante funciones cópula. Concretamente, vamos a utilizar estas funciones para buscar la relación de la variable a predecir con posibles variables explicativas de igual manera que hacen los modelos de función de transferencia.

Supongamos por ejemplo que nos planteamos predecir la variable X a partir de su propio histórico y de “ m ” variables explicativas $Y\{1\}, Y\{2\}, \dots, Y\{m\}$. Para referirnos al valor específico de cada una de estas variables en el instante “ t ”, utilizaremos la notación $Y\{j\}(t)$.

Sea $Error\{1\}$ el proceso de error “porcentual” asociado a un modelo SARIMA que ha sido ajustado únicamente a partir del histórico de X (también se ha realizado el correspondiente análisis de intervenciones) sin emplear ninguna de las variables explicativas (sin usar modelos de función de transferencia). Si $P\{1\}$ identifica la predicción hecha para X ,

$$Error\{1\}(t) = \frac{100 * (X(t) - P\{1\}(t))}{P\{1\}(t)} \quad (6.1)$$

Mediremos el error de forma porcentual con la idea de tratar de evitar la posible dependencia temporal de las unidades en las que se mide el error (de igual forma que se emplea la correlación en vez de la covarianza) pues esto puede ser problemático si la variabilidad de los datos cambia con el tiempo (véase el apartado siguiente, 6.2).

Partiendo de esta base, el algoritmo de ajuste secuencial mediante cópulas que proponemos, presentado también gráficamente en la ilustración (6.2), sería el siguiente:

Algoritmo de ajuste iterativo mediante funciones cópula.

1. Sea $j = 1, i = 1$ y $ME^* = ME\{1\}$ el error medio o mediano asociado al valor absoluto de la variable $Error\{1\}$.
2. Determinar a partir del histórico, la cópula $C\{i, j\}$ que mejor refleje la relación entre el proceso $Error\{i\}$ y la variable explicativa $Y\{j\}$:
 - 2.1 Si la mejor cópula es la cópula producto ($\Pi(u, v) = u * v$), no existe relación entre $Error\{i\}$ e $Y\{j\}$ (la cópula que mejor se ajusta es aquella que refleja la independencia entre las variables). Ir al paso 4.
 - 2.2 En caso contrario, ir al paso 3.

Nota: Para seleccionar la cópula que mejor refleje la relación, proponemos utilizar por ejemplo el contraste de bondad de ajuste descrito en el apartado 2.5.5, esto es, asociar a cada cópula representante (seleccionada como la mejor dentro de cada familia candidata) un valor del estadístico de Pearson (y a través de él un p-valor) y tomar aquella que menor valor proporcione (mayor p-valor).

3. Realizar predicciones mediante la cópula condicionada de $C\{i, j\}$ a través del método de la transformada inversa o bien por aproximaciones numéricas. Para ello, simular para cada valor del histórico $Y\{j\}(t)$, valores de $Error\{i\}(t)$, es decir, simular valores de $Error\{i\}$ condicionado a cada $Y\{j\}(t)$. Proponer como valor esperado para $Error\{i\}(t)$ (sobre el histórico) la media o mediana ($Mediana_Error\{i\}$) de dichas simulaciones.

Sea $P'(t)$ la predicción de $X(t)$ corregida por la predicción del error esperado, es decir,

$$P'(t) = \frac{\text{Mediana}_{Error\{i\}} * P\{i\}(t)}{100} + P\{i\}(t) \quad (6.2)$$

Considérese el nuevo proceso de error porcentual que determina lo que queda por explicar a raíz de esta predicción corregida,

$$Error'(t) = \frac{100 * (X(t) - P'(t))}{P'(t)} \quad (6.3)$$

Calcular ME' como la media o mediana de estos nuevos errores pero medidos en términos absolutos.

3.1 Si $ME' < ME^*$ (el error medio/mediano de predicción mejora en valor absoluto) hacer $P\{i+1\}(t) = P'(t)$, $Error\{i+1\}(t) = Error'(t)$, $ME^* = ME'$, $i = i+1$ e ir al paso 4.

3.2 Si no, ir directamente al paso 4. (El proceso de error no se modifica pues la explicación proporcionada por $Y\{j\}(t)$ no mejora la calidad de las predicciones. Por ello, la "i" se mantiene fija)

4. Hacer $j = j+1$.

(Si se viene del paso 3, opcionalmente se puede dejar "j" fija y tratar de seguir explicando el proceso de error con la misma variable mediante otra cópula, hasta que aquella que mejor se ajuste sea la cópula producto).

4.1 Si $j \leq m$ volver al paso 2.

4.2 En caso contrario, FIN.

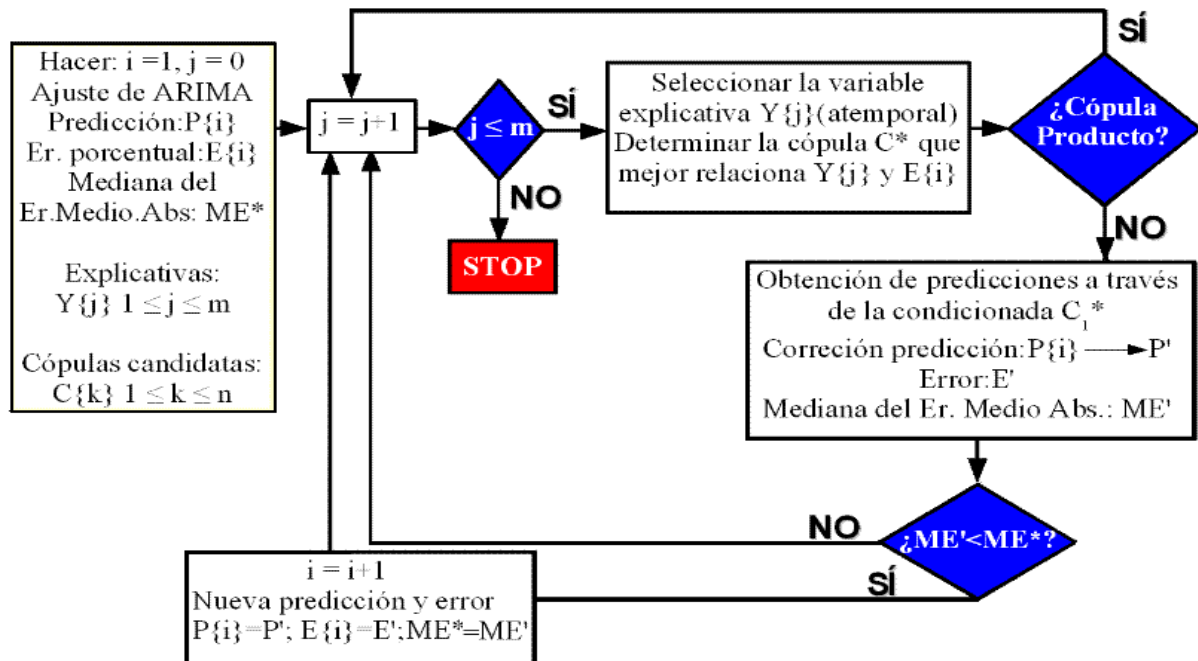


Ilustración 6.2: Metodología de predicción basada en el tratamiento del proceso de error de un modelo ARIMA mediante funciones cópula

Observaciones

1. De acuerdo al paso 2.1 del algoritmo, si la cópula que mejor refleja la relación de dependencia entre las variables $Y_{\{j\}}$ y $Error_{\{i\}}$ es Π , podemos descartar la posible vinculación entre ellas. Esto es una consecuencia derivada del teorema 2.4.2 que encontramos en [NELSEN], según el cual, la independencia entre variables aleatorias queda totalmente caracterizada a partir de la cópula producto (“Dos variables aleatorias continuas X e Y son independientes si y sólo si $C_{XY} = \Pi$ ”).

Esta es una propiedad diferencial de las funciones cópula, su posibilidad para detectar independencia entre variables frente a otras medidas clásicas que no lo permiten como el coeficiente de correlación lineal (que en caso de ser cero, únicamente permite confirmar la ausencia de dependencia lineal entre las variables), e incluso las propias **medidas de concordancia** a las que ya nos hemos referido en alguna ocasión en esta tesis: la Tau de Kendall ((1.19)) y el coeficiente de correlación de Spearman ((1.20)). Efectivamente, de acuerdo a la definición 5.1.7 que se presenta en la obra de Nelsen, el hecho de que X e Y sean independientes implica que dichas medidas valgan cero, sin que la implicación recíproca sea necesariamente cierta. Sólo para medidas de asociación que respondan a la definición de **medida de dependencia** (definición 5.3.1 de [NELSEN], y que viene a ser una adaptación del conjunto de propiedades discutidas en [Rényi (1959), Schweizer y Wolff (1981), Jogdeo (1982), Lancaster (1982)]), esta reciprocidad está garantizada.

2. En segundo lugar observemos que, en cada iteración del algoritmo se obtiene un nuevo proceso de error medido respecto de la variable X (a predecir) que intenta ser explicado por las variables Y que van quedando. Otra posibilidad alternativa sería medir el error respecto del propio proceso de error que se va obteniendo. Es decir, en una primera iteración tratar de predecir el error E resultante del ajuste con el modelo ARIMA. En la segunda iteración tratar de predecir el error E' en el que se vuelve a incurrir cuando se hace la predicción de E mediante cópulas, pero donde E' está medido respecto de E y no respecto de X .
3. Finalmente comentaremos que el método de ajuste implica ir midiendo constantemente relaciones entre pares de variables: un proceso residual y una variable explicativa candidata. Para ello, se utiliza una 2-cópula que permite evaluar dicha relación. Sin embargo, también sería posible incluir simultáneamente varias variables explicativas (“ m ”) para tratar de explicar el error, siendo necesario en dicho caso utilizar una m -cópula para estudiar la relación. La razón por la que se propone hacer este ajuste 2 a 2 es, como ya hemos señalado en alguna ocasión, la amplia literatura que existe para cópulas bivariantes y, en consecuencia, el amplio abanico de funciones de este tipo de las que puede disponerse.

6.2 Empleo de funciones cópula para explicar dependencia dinámica

Se ha propuesto emplear este método para suplir el empleo de modelos de función de transferencia con funciones cópula, pero se parte de una predicción inicial (y de un error inicial) proporcionada por un modelo ARIMA. La posibilidad de ajustar la serie temporal desde el principio (desde los datos reales originales) mediante dichas

funciones cópula sirviéndonos del algoritmo propuesto puede presentar algunos inconvenientes.

Las funciones cópula nos van a proporcionar buenos resultados si los pares de variables (U, V) entre las que intenta buscar relación son “ruidosas”. Es decir, conviene que no exista ningún tipo de estructura dentro de la propia variable y por tanto que no exista dependencia temporal, pues en dicho caso, la relación no quedaría establecida únicamente entre U y V , sino que participaría una tercera variable T , el tiempo.

En el capítulo 7.3 abordaremos con todo lujo de detalles el tratamiento de la relación *demanda de gas/temperatura* mediante funciones cópula. En concreto, analizaremos la dependencia entre la variable de error de un modelo ARIMA ajustado a la demanda de gas en Madrid y la temperatura máxima registrada en uno de los observatorios de la provincia. Según acabamos de ver, habremos de evitar que cada una de estas variables dependa de su pasado, es decir:

- El proceso de error no debe estar correlado con su pasado. Por ello, tomaremos el proceso de error obtenido una vez ajustada la parte de medias móviles del modelo ARIMA que elimina esta posible correlación. Además, la magnitud del error no debe depender de la tendencia: a medida que pasa el tiempo el valor de la serie crece y así también la magnitud del error (pero no en porcentaje). Por ello, utilizaremos errores medidos en términos porcentuales.
- Es preciso eliminar la doble dependencia temporal asociada a la variable climática. Por un lado, la temperatura de dos días consecutivos acostumbrará a ser parecida, es decir, la de un día será similar en líneas generales a la del día anterior (AR(1)). Por otro lado, la temperatura dependerá de la estación del año en la haya sido medida (en invierno serán bajas y en verano altas). Para eliminar dichas dependencias se propone utilizar por ejemplo el incremento climático de un día a otro. El hecho de que la temperatura bajara un determinado número de grados de un día al siguiente no justificaría que se produjese otro descenso de la misma magnitud entre los dos días posteriores. Realmente, ni siquiera existiría motivo para suponer que volviera a bajar.

De esta manera, la función cópula trataría de ajustarse a la relación que existe entre las variables “*error porcentual*” e “*incremento climático*”, ambas dos “ruidosas”.

Con esta explicación pretendemos justificar que puedan encontrarse problemas al tratar de ajustar una serie temporal desde los datos originales ya que, si bien siempre es posible considerar una variable de error porcentual que trataríamos de ir explicando, las posibles variables explicativas sí podrían tener una dependencia temporal.

Sí se piensa en un modelo ARIMA, los agentes que en éste intervienen y que podrían intervenir como variables explicativas son: el término constante (μ), los términos autorregresivos (parte AR) y los términos de medias móviles (parte MA). Podría comenzarse al ajuste tratando de explicar la serie con el término constante (μ) que sería la media de la propia serie.

Automáticamente, se podría obtener un primer proceso de error porcentual asociado a este ajuste que podríamos intentar relacionar mediante una cópula con la variable $X(t-1)$ (o $X(t-p)$ para un AR(p)) para, de esta forma, identificar un posible efecto de tendencia. Posteriormente, para reflejar una posible estacionalidad de periodo “ s ”, se buscaría la relación entre el proceso de error porcentual obtenido tras el ajuste anterior y la variable $X(t-s)$ y así sucesivamente.

Sin embargo, las variables $X(t-1), \dots, X(t-s)$ están relacionadas con su propia historia, tienen estructura temporal, no son

ruidosas, y por ello el resultado del ajuste mediante cópulas no resultaría adecuado.

Algunos artículos en los que se ha analizado la dependencia dinámica mediante el empleo de funciones cópula son [PATTON], [FERSCA] y [BOGASA]. En este último, los autores definen la **función de autoconcordancia** como alternativa a la función de autocorrelación para medir la dependencia dinámica, dado que el **autocorrelograma** es diseñado únicamente para detectar un proceso lineal autorregresivo. La función de autoconcordancia podría permitir en cambio detectar una estructura dinámica no lineal. Para el cálculo de cada uno de los coeficientes de esta función, basta hallar el índice de Kendall o el de Spearman entre la variable y sus propios retardos. Parece lógico que si los parámetros asociados a funciones cópula pueden ser estimados a partir del valor muestral de la Tau de Kendall o del coeficiente de correlación de Spearman, se utilicen estos mismos indicadores para medir posibles dependencias entre las variables que relacionan, en vez de emplear el coeficiente de correlación lineal que es a fin de cuentas, una medida de linealidad.

Por analogía, podríamos pensar en definir una función de concordancia cruzada, calculada como el índice de Kendall o Spearman entre la variable a explicar (por ejemplo el proceso de error porcentual) y los diferentes retardos de la variable explicativa (el incremento de la temperatura máxima). Esta función permitiría detectar no sólo la posible dependencia entre las dos variables a relacionar mediante la cópula, sino posibles relaciones entre el valor contemporáneo de una de ellas (el error) y el pasado de la otra (los últimos incrementos climáticos): “si la temperatura ha estado bajando durante la última semana, el error que comete el modelo es mayor”. Obsérvese que esta situación puede producirse sin necesidad de que ninguna de las dos variables presente estructura.

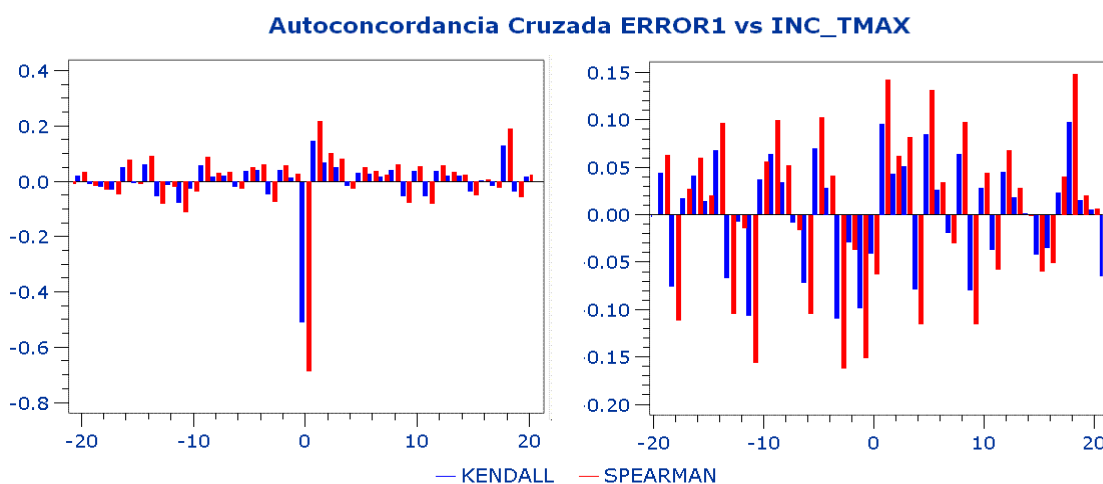


Ilustración 6.3: Gráfico de autoconcordancia cruzada asociado a un par de variables

En el gráfico de la izquierda de la ilustración anterior se puede observar la dependencia entre el proceso de error porcentual resultante de un modelo ARIMA para ajustar la demanda de gas y el incremento climático en el mes de Enero. El alto valor (en términos absolutos) de los índices de Kendall y Spearman en el retardo 0 pone de manifiesto que cuanto mayor es la bajada de temperatura en invierno (incremento de temperatura negativo), mayor es el error que comete el modelo (el dato real se queda por encima de la predicción). Una vez realizado el ajuste mediante una función cópula, los nuevos valores obtenidos para dichos índices en ese mismo retardo son de una magnitud considerablemente inferior, como se puede apreciar en el gráfico de la derecha.

7. Aplicación práctica: Predicción de la demanda de gas natural

7.1 Introducción

Hace cientos de millones de años, existieron en los océanos pequeñas plantas y animales marinos que, tras completar su ciclo de vida, se fueron descomponiendo con el tiempo para terminar siendo sepultados por una capa de arena y sedimentos. Como consecuencia de los cambios de presión y temperatura del planeta, parte de esta materia orgánica putrefacta, filtrada entre las rocas submarinas, acabó transformándose en carbón, petróleo y gas natural.

El gas natural se define por ello como un combustible fósil. Está compuesto en más de un 95% por metano, un gas integrado por un átomo de carbono y cuatro átomos de hidrógeno. El etano, el propano y otros elementos más pesados completan su composición (véase ilustración (7.1)).

Las ventajas principales del uso de gas natural desde el punto de vista medioambiental son su bajo índice contaminante para la atmósfera gracias al reducido porcentaje de dióxido de carbono que produce en combustión, su seguridad para ser transportado, el hecho de no ser corrosivo o su contribución a abatir el efecto invernadero por citar algunos ejemplos. Económicamente destaca frente a otras fuentes de combustible como el carbón y el petróleo por su precio competitivo, los reducidos costes de mantenimiento de los equipos de combustión y su alto grado de eficiencia desde su extracción hasta su consumo final (9 de cada 10 unidades extraídas del suelo son utilizadas).

Desde profundos depósitos subterráneos formados por roca porosa, el gas natural impulsado por su propia presión es extraído hoy día mediante pozos de perforación y conducido a la superficie por medio de ductos. Una vez procesado por centrales de recolección el gas es comprimido y preparado para su distribución.

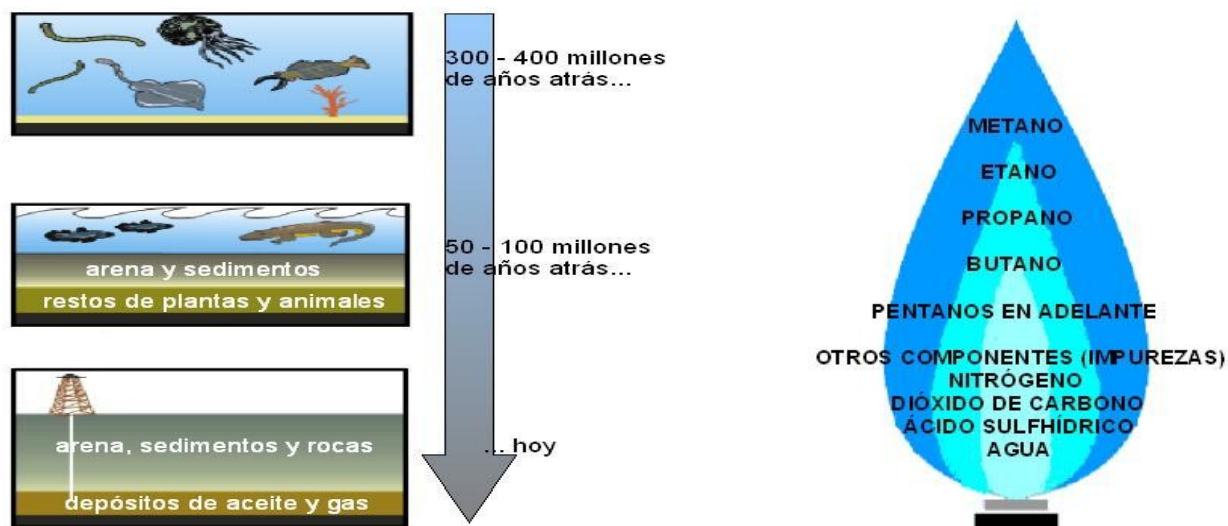


Ilustración 7.1: Origen del gas natural y composición

El gas circula por un entramado de tuberías de acero denominado gasoducto, manteniendo unos altos niveles de presión que suelen ser normalmente de 72 bares para redes de transporte y de 16 para redes de distribución siendo en las estaciones de regulación y medida en las que se lleva a cabo este cambio de presiones. Cuando la distancia a recorrer por el gasoducto es muy grande, se intercalan en su recorrido estaciones de bombeo que pueden ser bien yacimientos o bien puertos de mar a los que llegan buques denominados metaneros (véase ilustración (7.39)) que lo transportan de forma licuada a muy alta presión.

Con el objetivo de construir una red de gasoductos que recorra toda la península ibérica (véase ilustración (7.2)), el Ministerio de Industria publicó un decreto el 23 de Marzo de 1972 creando Enagás, la empresa nacional de gas. El gas que llega a la península procede de Noruega y desde Argelia a través del gasoducto Magreb-Europa (véase ilustración (7.3)) siendo este último, de acuerdo a los datos presentados en [PEMEX], uno de los diez países con mayores reservas del mundo (véase ilustración (7.2)).

En la actualidad, Enagás es la principal empresa transportista española de gas y entre sus activos, cuenta con 7.538 km de gasoductos de alta presión, tres plantas de regasificación (Barcelona, Cartagena y Huelva) y la propiedad o gestión de los almacenamientos de gas natural (véase [ENAGAS]). Una de sus principales funciones es diseñar un plan de operaciones a través del cual pueda garantizar la continuidad y seguridad del suministro de gas natural y la correcta coordinación entre los puntos de acceso, los almacenamientos, el transporte y la distribución teniendo en cuenta las necesidades y características de cada uno de sus tipos de clientes tanto dentro del ámbito industrial como del doméstico. En el sector industrial los clientes pueden ser desde grandes consumidores como son las centrales de generación eléctrica o las centrales de cogeneración, hasta empresas medianas y pequeñas que lo emplean en función de sus características (siderurgia, química, azulejeras, celulosa y papel, cemento, etc.). En el sector doméstico encontramos un grueso constituido por millones de clientes residenciales que utilizan el gas como combustible para la calefacción, para la cocina o para la obtención de agua caliente.

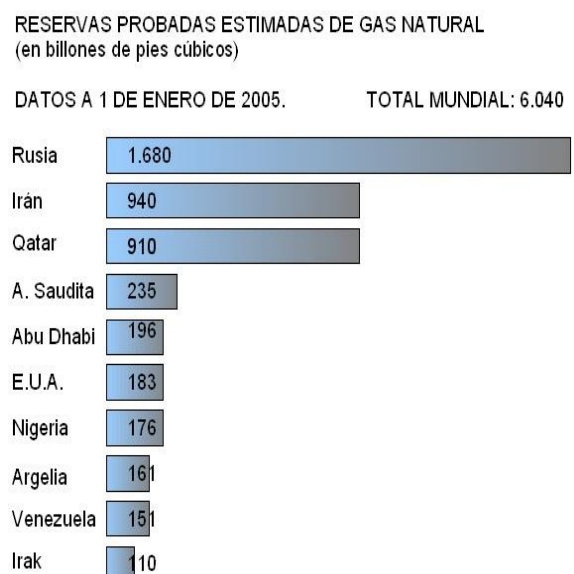
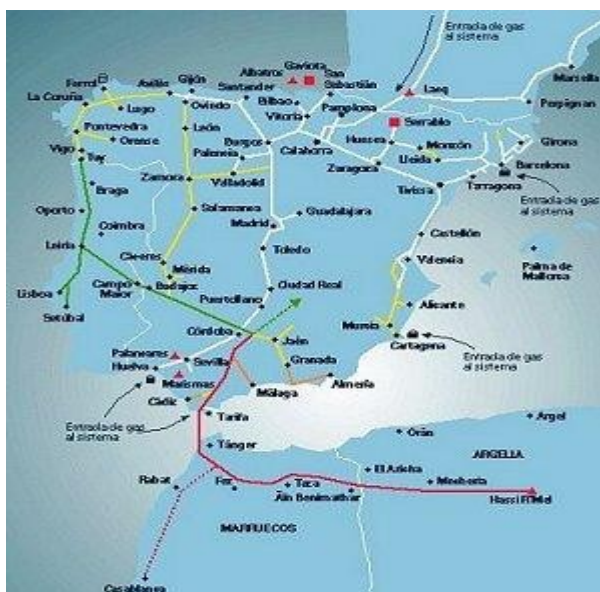


Ilustración 7.2: Red de gasoductos de Enagás (a Diciembre de 1996) y datos de reservas estimadas de gas natural

El comportamiento de la demanda de cada segmento es bastante diferente dado que uno y otro están influenciados por agentes distintos. El sector industrial tiene un patrón de consumo muy marcado por el efecto del calendario de acuerdo al cual, la actividad de las industrias acostumbra a ser mayor durante los días laborables que durante los días festivos, los puentes o los periodos vacacionales. En el caso doméstico, si bien este efecto está obviamente presente pues la presencia o no de las personas en sus hogares puede depender del carácter laboral del tipo de día, existe un claro condicionante adicional que es el efecto derivado del comportamiento climático durante los meses invernales. La influencia de estos agentes resulta además diferente en función de la zona geográfica en la que se ubica el usuario.

Para Enagás, el conocimiento de los perfiles que identifican a sus distintos clientes es un aspecto clave de cara a elaborar su plan de operaciones. Esto implica, además de saber por supuesto el número de ellos que tiene de cada tipo, conocer cómo va a variar este número en el futuro, el efecto de los planes de expansión de las distribuidoras que van a permitir el acceso al consumo doméstico a nuevas poblaciones, la variación de los hábitos de consumo en función del progreso y la mejora del nivel de vida, etc.

La gestión del plan de operaciones es una tarea muy compleja que se realiza no sólo de acuerdo a los recursos de gas disponibles sino también a la demanda prevista que hay que abastecer en función de las necesidades y características de cada tipo de cliente y en función además del cumplimiento de los compromisos contractuales tanto con las empresas suministradoras como con las distribuidoras y comercializadoras.

La aparición de las comercializadoras de gas ha tenido lugar a raíz de la reciente liberalización del mercado del gas en España (desde Enero de 2003). Según la nueva normativa, un cliente no tiene porqué mantenerse en el mercado regulado que se rige de acuerdo a las tarifas fijadas por el gobierno, sino que puede comprar directamente en dicho mercado, elegir una comercializadora a la que pague una tarifa de menor coste o establecer contratos bilaterales.

El crecimiento del número de comercializadoras, unido al esperado incremento del consumo y a otro factor determinante como es la capacidad limitada de la red actual hacen prever a Enagás dificultades en un futuro próximo a la hora de poder elaborar su programación. Resulta por ello imprescindible para el departamento de operaciones disponer de unas buenas herramientas de predicción de la demanda a distintos horizontes para poder adaptarse de manera adecuada a las necesidades de suministro a medio (1-3 años) y corto plazo (día/hora). Cuanto mayor sea la bondad de las predicciones que proporcionen estos sistemas, más fielmente se ceñirán a la realidad los programas y planes de gestión citados y por tanto más beneficiosos resultarán los resultados de la operación.



Ilustración 7.3: Gasoducto Magreb-Europa en su recorrido por Argelia, Marruecos y el estrecho

7.2 Predicción de la demanda de gas natural a medio plazo

7.2.1 Introducción al problema

Empezaremos tratando la predicción de la demanda de gas a medio plazo con detalle diario. El conocimiento de la demanda con la suficiente antelación va a permitir el diseño y la construcción de las infraestructuras que deben existir para poder prestar el servicio. La construcción de un gasoducto, de una estación de compresión que provee la energía necesaria para hacer llegar el gas natural a lo largo y ancho del territorio nacional, de una planta de regasificación (véase ilustración (7.4)), o de cualquiera de las instalaciones que participan en el suministro del gas, es un proceso largo y complejo condicionado por las exigencias del entorno social. Es necesario cumplir con unos compromisos medioambientales, una serie de requerimientos de las autoridades autonómicas y locales, tener en cuenta los derechos de los propietarios que se puedan ver afectados de manera temporal o definitiva, etc. La duración de este periodo en el que se concreta el proyecto, se hacen las correspondientes tramitaciones administrativas y se hace la ejecución definitiva suele ser de entre 3 y 4 años.

Aunque pueda pensarse que sólo tiene sentido medir la predicción a tan largo horizonte en unidades mensuales (demanda mensual), existe una razón que justifica conocer el detalle diario de la predicción. Se trata de evaluar si la capacidad del gasoducto es suficiente para soportar en un día cualquiera del futuro la demanda del colectivo industrial y doméstico asociado a una determinada población dado que en función de ello se planteará la necesidad de expandir la infraestructura existente. La herramienta de predicción a medio plazo debe permitir así no sólo conocer el valor diario de demanda que cabe esperar en cada uno de los dos o tres próximos años en función de la tendencia observada en el histórico, sino además valorar el impacto de una ola de frío que se prolongue durante varios días, pues es un condicionante extremo de este tipo bajo el que se podrá esperar un valor máximo (denominado “pico” en la terminología energética) para la demanda de gas. Es en este contexto de la simulación donde hacen su aparición las funciones cópula.



Ilustración 7.4: Plantas de Barcelona y Cartagena

7.2.2 Planteamiento de la solución

Existen diferentes planteamientos ante un problema de predicción de picos de demanda de gas o electricidad a medio plazo. Algunas buenas referencias las podemos encontrar en el capítulo 2 de la tesis de [GUTIER]. Entre ellas se cita por ejemplo el empleo de modelos econométricos en la línea de los desarrollados por Spann y Beauvais (1977) en los que el pico de demanda se expresa como una combinación lineal de una componente base y otra que refleja la influencia de las variables climatológicas:

$$PD = \beta + \alpha * T \quad (7.1)$$

siendo PD el pico de demanda, β el comportamiento básico de la demanda cuyo crecimiento viene justificado por la influencia de variables económicas y demográficas y donde T es una función de los posibles agentes meteorológicos que puedan condicionar el sistema.

Otra metodología comúnmente utilizada para predecir picos de demanda energética (electricidad o gas) son los modelos ARIMA y a modo de ejemplo se menciona en [GUTIER] el trabajo de Abraham, B. (1983) para la predicción de demanda eléctrica en el estado de Iowa (Estados Unidos). La complejidad de estos modelos, de los que hablaremos más adelante, dependerá normalmente de la manera de reflejar en ellos el efecto de las variables climáticas.

La metodología que vamos a proponer combina en cierto modo estas dos que hemos citado. Partiendo de la base de que un pico de demanda no es más que una desviación extrema del comportamiento básico de la demanda justificada por una desviación extrema del comportamiento básico de unos agentes climatológicos, plantearemos un modelo autorregresivo para predecir el consumo en condiciones normales de temperatura y posteriormente llevaremos a cabo un análisis mediante funciones cópula para simular posibles valores de la demanda de gas ante situaciones meteorológicas especialmente desfavorables. Es decir, en la línea del modelo de Spann y Beauvais, predeciremos el pico de consumo de acuerdo a una estimación del comportamiento básico o normal de la demanda y a otra estimación mediante cópulas de sus posibles desviaciones ante adversidades climáticas. Es precisamente la inclusión de este tipo de funciones la que aporta un carácter más innovador a la metodología propuesta.

Las cópulas se mueven habitualmente en el ámbito de la simulación, del condicionamiento a escenarios configurados por el valor supuesto para ciertas variables explicativas ante la imposibilidad de tener un conocimiento más preciso de ellas, como es el caso de los datos diarios de temperatura a 1,2 ó 3 años. Su utilización no es tan común en problemas de predicción a corto plazo en los que la fiabilidad de dichos valores es tal, que pueden ser considerados casi reales en vez de hipotéticos y donde tampoco existe margen de maniobra para plantearse posibles realidades alternativas. A pesar de ello, en el capítulo 7.3 sí abordamos el problema de la predicción a corto plazo apoyándonos en el empleo de cópulas, si bien consideramos que, dada la gran cantidad de técnicas que existen para tratar con ciertas garantías este problema, el potencial de estas funciones puede no resultar siempre tan determinante.

Como paso previo al empleo de las funciones cópula es necesario predecir el comportamiento básico de la demanda. La técnica que utilizaremos viene a ser una extensión de la que habitualmente se utiliza en la modelización de este tipo de series a corto plazo: la metodología Box-Jenkins de la que ya hemos hablado en el apartado 6.1. La experiencia demuestra que esta metodología fundamentada en el empleo de modelos ARIMA es la más adecuada para predecir a un

horizonte de pocos días frente a otras alternativas como por ejemplo las redes neuronales, las cuales, salvo raras excepciones, no conducen a mejores resultados. La razón fundamental es que los modelos ARIMA permiten identificar fácilmente rasgos muy característicos de una serie de demanda de gas (o electricidad) como son su tendencia, su variabilidad, la influencia del calendario a través de un análisis de intervención o el efecto de variables climáticas a través de modelos de función de transferencia (véase apartado 7.3.2.1).

A medio plazo, 1 o 2 años, se consideraría impensable plantear una predicción **diaria** con modelos ARIMA ya que éstos son útiles para hacer predicciones para un horizonte temporal corto, entre otras cosas porque el orden asociado a la parte MA raramente será superior a 7 (si tiene una estacionalidad semanal), con lo que la capacidad de que la predicción sea corregida en función de los errores que el modelo comete (una de las principales características de un ARIMA reflejada por sus componentes de medias móviles) se perderá a partir de dicho horizonte. Otra razón obvia es la imposibilidad de disponer de previsiones diarias de temperatura fiables a tan largo plazo: el INM no proporciona más de 10 (y no garantiza la calidad de las previsiones a partir del cuarto día), y algunos otros proveedores (*Accuweather.com*) facilitan un máximo de 15.

Lo más común para un horizonte anual es realizar la predicción mensual en función de un histórico construido a ese nivel de detalle, es decir, por meses. De esta forma, el número de unidades temporales a predecir se sitúa entre 12 y 24, una cantidad razonable pues el posible término MA(12), en caso de que exista una estacionalidad anual, garantizaría la corrección de las predicciones realizadas para los doce primeros meses, algo más sensato que una semana. En esta caso, el efecto del calendario, de la Semana Santa móvil (que puede aparecer en distintos meses) y de los posibles años bisiestos son recogidos mediante variables que cuantifican el número de días laborables del mes, la longitud del mismo (para los bisiestos), etc. Los modelos X11 y X12 que permiten hacer referencia a este tipo de efectos son realmente ARIMA's especialmente diseñados para el tratamiento de series mensuales.

Sin embargo, de acuerdo a la problemática que hemos planteado, el interés de la compañía es disponer de un patrón diario de la demanda para los dos próximos años pues éste le permite intuir la época del año (quincena o mes) en la que puede esperar un pico de demanda y sobre él, suponer unas condiciones climatológicas especialmente adversas para conocer cuál es el máximo diario que podría llegar a alcanzarse en el futuro planteado. Si bien una posibilidad podría ser desagregar la predicción mensual en unidades diarias (véase [GIDEDU]), nos proponemos aprovechar un planteamiento de predicción a corto plazo y hacer una extrapolación a un horizonte mayor. En la metodología que vamos a desarrollar, distinguiremos las siguientes fases:

1. **Construcción de un histórico de la demanda en condiciones normales de temperatura.**- Para ello, se establecerán unos valores de temperaturas que se consideren representativos de cada uno de los meses del año (temperaturas normales de un mes). En función de ellos, haremos una depuración del histórico de demanda, eliminando picos y caídas que puedan venir justificadas por desviaciones de esos valores. Destinaremos parte de los datos disponibles a la construcción de este patrón de demanda y otra parte (de 1 a 2 años en función del horizonte al que se desee hacer la predicción) a contrastar la calidad de las predicciones que se harán tomando como base dicho histórico (véase el punto I en el apartado 2.5.6).

2. **Construcción de un modelo autorregresivo (AR) que permita hacer una predicción a medio plazo en condiciones normales de temperatura.**- Dicho modelo debe identificar y extrapolar a un horizonte de 1 o 2 años los rasgos más característicos de este tipo de series, que son:

- La estacionalidad es una propiedad presente a lo largo de todo el año y por tanto extrapolable. Bien es cierto que el distanciamiento entre lo que llamaremos días “intrasemanales” y fines de semana puede variar en función del mes o estación del año.
- El efecto “calendario” referido a festividades aisladas y periodos vacacionales es identificable a partir del histórico, conocido de antemano, y por tanto también transferible a cualquier día del horizonte de 2 años que se contempla.
- El efecto climático es despreciable puesto que al haber construido un histórico en condiciones normales de temperatura, podemos suponer que la predicción estará hecha también en esas mismas circunstancias.

Hemos dejado para el final dos de los aspectos más importantes de la serie, la tendencia y la variabilidad que, a diferencia de las componentes anteriores, no pueden ser tratados por extrapolación directa de un modelo diseñado para una predicción a corto plazo.

- La tendencia no puede ser establecida únicamente a partir de los últimos datos pues las predicciones quedarían estancadas en un nivel fijo (aquél en el que se muevan los últimos datos). Por ello, se utilizará como “input” del modelo una variable que permita recoger la periodicidad anual de la serie. Es este “ciclo anual” el que va a permitir expandir las propiedades citadas (calendario, estacionalidad, etc.) a un horizonte mayor. La construcción de esta variable *Ciclo* se hará extrayendo la señal de baja frecuencia asociada a la serie mediante una técnica de suavizado de curvas como por ejemplo las wavelets o los splines. El comportamiento suave de estas señales es fácilmente extrapolable a varios años, pudiéndose utilizar por tanto como una variable explicativa del modelo autorregresivo. Es sobre esta curva sobre la que van a incorporarse la estacionalidad y el efecto del calendario.
- La variabilidad puede venir inducida como hemos comentado por diferencias en la dispersión entre días laborables y festivos pertenecientes a una misma semana en distintas épocas del año. De igual manera que el modelo autorregresivo en ausencia de la componente *Ciclo* seguiría moviéndose en el nivel marcado por los últimos datos reales del histórico, así lo haría también su variabilidad. Para corregir esta circunstancia y permitir que la variabilidad del modelo cambie a lo largo del año, introduciremos variables

dummy's que sean capaces de reflejar los rangos de variación diferentes que pueden esperarse para cada mes entre los días laborables y los festivos, puentes y fines de semana.

El modelo quedará ajustado cuando consigamos reflejar en una ecuación todas estas componentes. A partir de ellas, será posible realizar la predicción diaria a 1 o 2 años en condiciones normales de temperatura.

Obsérvese que en todo momento hemos hablado de modelos autorregresivos (AR), no teniendo sentido la inclusión de una parte de medias móviles que pudiera corregir la predicción en función de los errores cometidos. La parte MA del modelo es la que justifica en mayor medida el empleo de los modelos ARIMA para el tratamiento a corto plazo pues una predicción tiene en cuenta el error que cometió recientemente (el último día o el mismo tipo de día de la semana pasada) y en función de él se corrige de manera que, si la predicción se iba quedando por debajo del dato real, se hace una corrección a la alza y viceversa. A medio plazo, como decimos, esta corrección carece de sentido.

3. **Simulación del valor esperado para la demanda ante diferentes escenarios configurados por el valor de las variables de temperatura.**- El último paso es simular la desviación esperada para la demanda construida, ante posibles incrementos y decrementos de la temperatura sobre el nivel establecido. En este contexto, utilizaremos las funciones cópula.

Construiremos una muestra de pares cuya primera componente (X) sea la diferencia entre el dato de temperatura real y el que ha sido supuesto como “normal” (para el mes al que pertenece el dato); la segunda componente (Y) será la diferencia porcentual entre el dato histórico real de demanda y el que forme parte del patrón construido en condiciones estándares. La razón de considerar la diferencia porcentual en lugar de absoluta, es hacer que el par de variables no dependa del tiempo, una circunstancia que como ya hemos señalado y más adelante recalcaremos, no es aconsejable a la hora de emplear con libertad las cópulas.

Así, buscaremos la cópula C que mejor refleje la relación de dependencia entre X e Y , es decir, la que mejor se ajuste a los pares $(F(X), G(Y)) = (U, V)$ siendo F y G las distribuciones marginales de X e Y respectivamente. La relación entre ambas variables no es lineal y además presenta un comportamiento difícil de capturar en situaciones extremas. Veremos que, aquellas cópulas que permiten enfatizar la dependencia entre sucesos “cola”, son especialmente apropiadas para capturar este tipo de relación.

Una vez determinada C , su condicional C_I , nos posibilitará hacer simulaciones de valores de V (de Y) a partir de valores propuestos para U (para X). Esto nos va a permitir conocer el incremento porcentual que se puede esperar sobre la demanda cuando se supone una variación en grados del valor normal de temperatura para un mes concreto. Gracias a ello y teniendo en cuenta el mes en el que se presentará el pico invernal de acuerdo a la predicción realizada en condiciones normales de temperatura, podremos simular desviaciones porcentuales de ese valor ante una bajada extrema de temperatura (la máxima caída recogida sobre el histórico). Tomando la media o mediana de dichas simulaciones y aplicándola sobre el pico invernal, podremos establecer el valor que cabría esperar para la demanda ante una bajada de temperatura extrema. Pero además, el conocimiento de la

distribución de las simulaciones realizadas y por tanto de sus colas, nos dará una idea del máximo valor que se puede llegar a esperar y evaluar el riesgo de que éste se presente.

De acuerdo al criterio establecido en el apartado 2.5.5, seleccionaremos aquella cópula que proporcione, respecto de una partición determinada, un menor valor del estadístico de Pearson (veremos que es una cópula que enfatiza la relación entre sucesos extremos) y compararemos los resultados que proporcione con aquellas otras que son resultado de interpolar la subcópula que para dicha partición optimiza el valor de dicho estadístico (y que son objeto de estudio de esta tesis). El decantarnos por la familia conocida o por la construida mediante técnicas de interpolación dependerá de la calidad de las predicciones realizadas (véase apartado 2.5.6) contrastando para ello, la distribución de la demanda real y esperada (propuesta por cada cópula) ante una misma situación climática extrema.

7.2.3 FASE I: Construcción de un histórico de demanda en condiciones normales de temperatura

Ya hemos mencionado que enfocaremos el tratamiento a medio plazo de este tipo de series como una extensión de un estudio a corto plazo del que hablaremos en el apartado 7.3 y para el que la mejor solución viene dada por el empleo de modelos ARIMA. En este tipo de modelos, el efecto del calendario es reflejado mediante variables de intervención y la influencia de la meteorología a través de funciones de transferencia, siendo el ajuste de estas últimas el que ha permitido determinar las variables que definirán el concepto de “**normalidad climática**”. El hecho de que un alto porcentaje de la demanda de gas en Madrid proceda del segmento residencial (por encima del 90%), justifica que se haya contemplado en principio su relación con diferentes tipos de agentes climáticos. Los históricos de éstos han sido proporcionados por la red SYNOP del Instituto Nacional de Meteorología (INM) para un total de cuatro estaciones de medida: Barajas, Cuatro Vientos, Getafe y el Puerto de Navacerrada. Estas variables son:

- Temperatura máxima y mínima diaria medida en grados centígrados.
- Precipitación acumulada en 24 horas (de 7:00 a 7:00 horas) medida en litros por metro cuadrado.
- Nubosidad (a las 6:00, a las 12:00 y a las 18:00 horas) medida como el número de octavos de cielo cubierto.
- Velocidad del viento (a las 6:00, a las 12:00 y a las 18:00 horas) medida en metros por segundo.
- Humedad relativa (a las 6:00, a las 12:00 y a las 18:00 horas) medida porcentualmente.

Tras contrastar los resultados obtenidos con diferentes modelos ARIMA, en los que el efecto de la meteorología ha sido reflejado de maneras distintas, hemos llegado a estas conclusiones:

1. No es conveniente incluir de manera simultánea en el modelo, variables de un mismo tipo (*temperatura, nubosidad,...*) asociadas a observatorios distintos (por ejemplo Barajas y Getafe). El alto grado de correlación que entre ellas se establece proporciona al modelo un carácter peligrosamente inestable. De esta forma, no es conveniente por ejemplo, incluir a la vez en el modelo, variables asociadas a las temperaturas máximas de Barajas y Getafe.
2. No proporciona mejores resultados la inclusión de una variable (*temperatura, nubosidad,...*) obtenida como agregación ponderada de las variables de su mismo tipo recogidas en cada uno de los observatorios. Es decir, la consideración de la variable *temperatura máxima de Barajas* proporciona un mejor ajuste del modelo y unos errores de predicción más bajos que la variable de temperatura máxima que surge como resultado de la ponderación (en función de su grado de dependencia de la variable *demanda*) de las temperaturas máximas de cada uno de los 4 observatorios.
3. La incidencia de las variables *precipitación acumulada, nubosidad, velocidad del viento y humedad relativa* resulta ser insignificante para el modelo matemático de predicción a corto plazo, ya que el efecto de estas variables queda absorbido, casi en su totalidad, por las variables de temperatura. Dada la escasa mejora que sobre el error medio de predicción aportan estas variables y de cara a simplificar el modelo (el principio de parsimonia establece la conveniencia de modelos sencillos), se ha despreciado la inclusión de las mismas.
4. Entre todas las variables de temperatura contempladas, han sido la máxima de Barajas y la mínima de Cuatro Vientos las que, por su mayor grado de significatividad y por proporcionar mejores resultados de predicción, han sido incluidas finalmente en el modelo.

De esta manera la influencia de la meteorología ha sido reflejada en el modelo ARIMA, del que como decimos hablaremos en el apartado 7.3, a través de funciones de transferencia asociadas a las temperaturas de Barajas y Cuatro Vientos. Es éste el motivo por el cual hemos establecido el concepto de “climatología normal” en función de los valores recogidos para dichas estaciones. De ellas, es la variable de temperatura máxima la que presenta una relación más clara con la demanda y la que posee mayor poder predictivo puesto que el valor de temperatura mínima suele darse a altas horas de la madrugada lo cual no suele condicionar de manera determinante el mayor uso de la calefacción. Por ello, tampoco perderemos mucha rigurosidad en el estudio si simplificamos aún más y establecemos el concepto de “normalidad meteorológica” en función únicamente de la temperatura máxima de Barajas. Una de las razones fundamentales que nos ha llevado a tomar esta decisión es facilitar la labor de simulación de valores de demanda a partir de una única variable en vez de dos, sirviéndonos para ello de una 2-cópula y no de una 3-cópula para las que

existe mucha menos literatura al respecto.

Fijaremos el periodo de un mes como la unidad temporal asociada a la cual definiremos una temperatura “normal” y así, hablaremos de la temperatura normal de Enero, Febrero, etc. El concepto de “estación del año” parece demasiado amplio para ser identificado con un rango de temperaturas y el de “semana” o “quincena” tal vez demasiado reducido. De los 9 años de datos de los que disponemos (desde el 1 de Abril de 1997 al 31 de Marzo de 2006), utilizaremos 7 de ellos (del 1 de Abril de 1997 al 31 de Marzo de 2004) para construir el histórico de demanda no influenciado por la climatología y reservaremos los otros dos para contrastar la calidad de las predicciones que hagamos. Así, tendremos una muestra por encima de los 200 datos diario por mes ($7 \times 30 = 210$) para establecer el estándar climatológico. Podemos considerar por ejemplo como temperatura normal del mes la mediana de esos 200 valores. De esta forma, los datos de demanda “normales” van a ser aquellos que lleven asociado ese valor de referencia o al menos estén suficientemente próximos a él, donde el concepto de proximidad podría venir dado por un intervalo delimitado por ejemplo por los cuartiles 1 y 3. La siguiente tabla presenta esta relación de valores.

MES	Cuartil1	Mediana	Cuartil3
1	8.9	10.4	11.9
2	10.9	13.7	16.4
3	14.0	16.8	19.8
4	15.1	18.4	21.6
5	19.3	23.4	26.8
6	27.0	31.0	33.7
7	30.2	33.0	35.2
8	30.5	32.8	35.2
9	24.8	28.0	30.1
10	17.8	20.7	23.3
11	11.5	13.4	16.2
12	8.8	10.4	11.9

Tabla 7.1: Parámetros de localización de la temperatura máxima de Barajas por meses: “temperaturas normales”

Llegados a este punto vamos a hacer una observación que nos llevará a hablar de datos de demanda normales asociados a diferentes tipos de día. Como ya hemos comentado en alguna ocasión, la serie tiene una acentuada estacionalidad semanal. Esta periodicidad a corto plazo viene marcada fundamentalmente por la caída brusca de demanda que se produce el sábado respecto al viernes, el domingo respecto al sábado y el incremento del domingo al lunes que le permite recuperar su nivel. Dicho de otra manera, si los sábados y domingos estuvieran al nivel del resto de los días de la semana, no existiría posiblemente componente estacional. Esta apreciación queda patente en el siguiente gráfico.

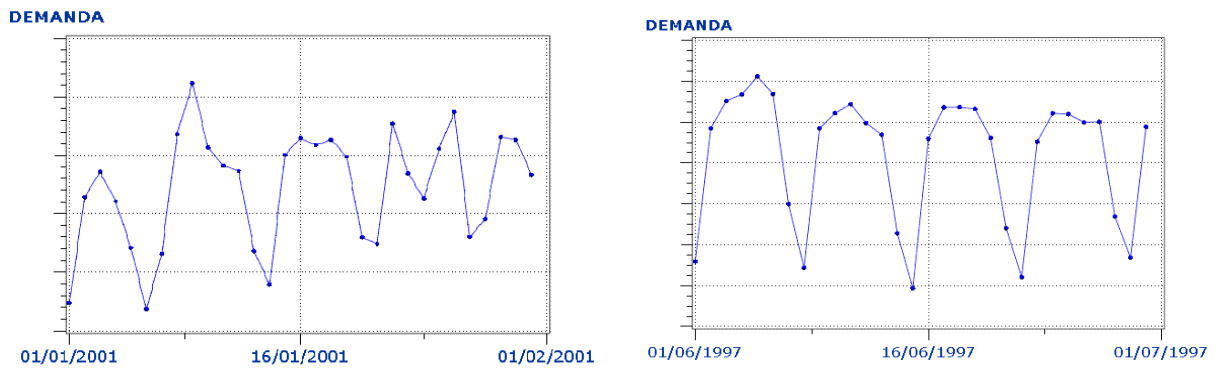


Ilustración 7.5: Estacionalidad semanal característica de una serie de demanda de gas

El gráfico de la izquierda muestra datos de demanda en Enero pudiendo reconocerse el patrón estacional que está algo distorsionado por la existencia de “picos” (9 de Enero) y “caídas” (23 y 24 de Enero) que son reflejo del efecto de la temperatura (frío y calor respectivamente) y otras “caídas” (1,5 y 6 de Enero) justificadas a partir del efecto del calendario (Año Nuevo y Reyes). Por otro lado, el gráfico de la derecha muestra datos de demanda en Junio donde a priori no es de esperar que influyan las temperaturas. El consumo residencial es muy regular a lo largo de todo el verano pues está sujeto a hábitos cotidianos como cocinar o ducharse y no a la mayor o menor utilización de sistemas de calefacción en función del frío que haga.

Veamos a continuación cómo la deformación del patrón estacional está condicionada por los valores de temperatura en función de la estación del año. Dado que ambas variables se mueven en escalas claramente diferentes, se han normalizado sus valores con vistas a poder compararlas de forma simultánea. Así, mientras que en el periodo estival la estructura semanal es fácil de identificar con independencia del año que consultemos, en el caso invernal podremos llegar a encontrarnos con situaciones en las que esta estructura queda prácticamente irreconocible, si bien cabe suponer que en condiciones normales de temperatura, debería haberse presentado un patrón estacional que es el que perseguimos construir. Los primeros gráficos van asociados a periodos invernales.

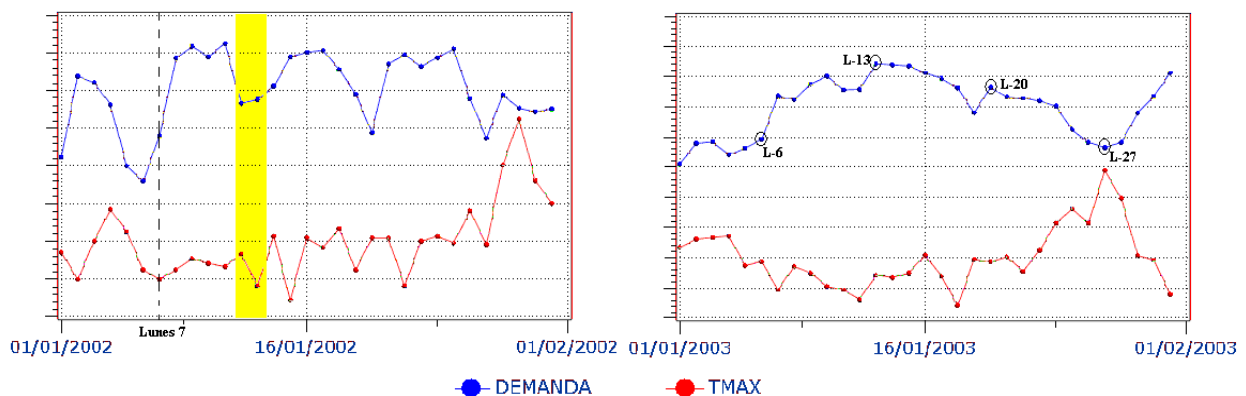


Ilustración 7.6: Distorsión del patrón estacional de demanda en función del efecto de la temperatura en invierno

El gráfico de la izquierda muestra un fin de semana (12 y 13 de Enero) cuyo nivel es demasiado alto. Esto puede venir justificado porque el periodo anterior era vacacional (Navidad) y en consecuencia el número de habitantes en Madrid se podría esperar que fuera menor (y por tanto el consumo doméstico). Sin embargo, lo que más llama la atención es que el valor de demanda del domingo día 13 se sitúa incluso por encima del día anterior (sábado 12) cuando por lo general siempre será inferior la demanda de un domingo (por la repercusión industrial de un día festivo) a la de un sábado. A partir de los valores de temperatura resulta más sencillo entender esta circunstancia dada la caída de temperatura que se produce el domingo. También es significativo el descenso climático de la semana del 7 con respecto a los últimos días de la anterior.

El gráfico de la derecha es aún más explícito. El patrón estacional es casi imposible de identificar y se aprecia muy claramente la relación de dependencia entre demanda y temperatura pues, para un mismo periodo, el abombamiento de ambas curvas se “abre” en sentidos opuestos. Obsérvese también que el valor del lunes 27 es incluso inferior al del domingo anterior, algo que puede llegar a entenderse a la vista del calor que llegó a hacer (“pico” de temperatura) en comparación con las semanas previas.

Los siguientes gráficos van asociados a periodos veraniegos. En ellos se puede reconocer fácilmente la estacionalidad semanal con independencia de las posibles inclemencias del tiempo, quedando los días laborables claramente diferenciados de los fines de semana.

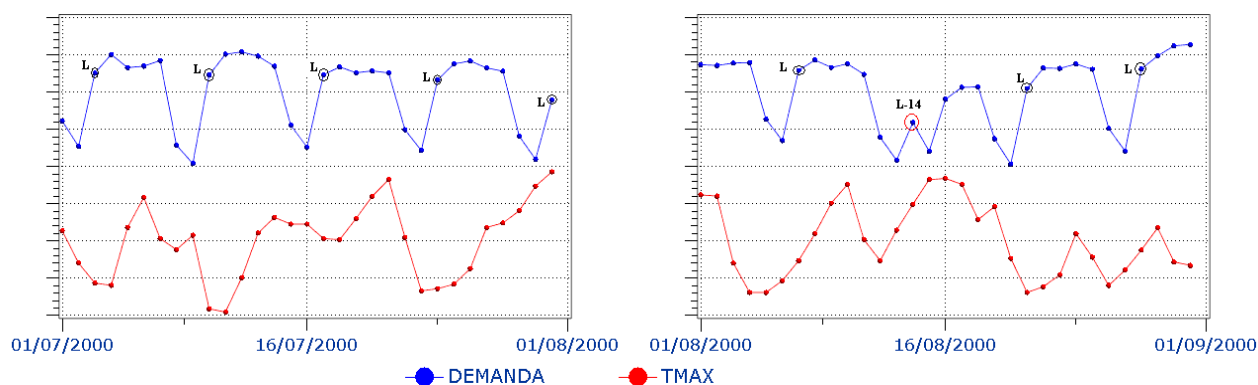


Ilustración 7.7: Independencia del patrón estacional de demanda respecto del efecto de la temperatura en verano

En el gráfico de la izquierda, vemos que el último lunes de Julio (día 31) la demanda cae un poco por debajo de la que cabría esperar para un lunes a tenor de lo que se observa en otros de ese mismo mes. En esta ocasión el motivo es el comienzo del periodo vacacional de Agosto que pudo provocar una salida masiva de personas de la capital, repercutiendo tanto en el sector industrial como en el doméstico (uso regular del gas para cocina, higiene, etc.).

El gráfico de la derecha es la prolongación del anterior (va asociado al mismo año) y muestra por tanto el periodo vacacional del que hablamos. A pesar de que se respeta el patrón estacional, podemos encontrar un par de anomalías los días 14 y 15 de Agosto, lunes y martes en que el nivel de la demanda no es el propio de un día laborable, rompiéndose el patrón estacional. Este hecho queda justificado si tenemos en cuenta que el día 15 de Agosto es festivo (día de la Virgen) con lo que la actividad industrial disminuyó. Su situación en la semana hace comprensible que se produzca el

efecto puente (lunes 14) siendo éste el motivo de que no se recupere dicho día el nivel característico de los lunes.

En resumen, el nivel “normal” de la demanda para un mes estará condicionado por el tipo de día al que nos estemos refiriendo: los domingos, los sábados o los días laborables (por lo general de lunes a viernes). Desde un punto de vista analítico se puede confirmar esta intuición, estableciendo un contraste de igualdad de medias de la variable *demanda* respecto de siete clases que van a ser los días de la semana. Vamos a plantear el contraste de **Kruskal-Wallis** (véase Anexo IX) que es una alternativa no paramétrica al método **ANOVA** y que no precise de validar las hipótesis de homocedasticidad y normalidad que presupone este último. Así, no es necesario ver que el comportamiento de la demanda se reparte por tipo de día de acuerdo a una distribución normal, ni tampoco importa el hecho de que la variabilidad de los datos cambie con el tiempo (heterocedasticidad que, como veremos, si es una de las propiedades de nuestra serie). Respecto de la tendencia de los datos, si bien es verdad que la media cambia con el tiempo, lo hace de manera compensada para todas las clases, es decir, al crecer la media de los lunes de un año a otro también lo hace para el resto de días de la semana y no existe por tanto necesidad de eliminar la tendencia implícita a la serie a la hora de llevar a cabo el contraste.

Plantaremos el test respecto de los 7 niveles que puede tomar la variable día de la semana (*L, M, X, J, V, S, D*).

$$H_0: \mu_L = \mu_M = \mu_X = \mu_J = \mu_V = \mu_S = \mu_D$$

$$H_1: \mu_i \neq \mu_j \text{ para algunos } i, j \in \{L, M, X, J, V, S, D\}$$

El primer resultado al que llegamos es:

Wilcoxon Scores (Rank Sums) for Variable CONSUMO Classified by Variable DIA_C					
DIA	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
M	366	507606.00	468114.0	13074.3869	1386.90164
X	366	508042.50	468114.0	13074.3869	1388.09426
J	365	501818.50	466835.0	13059.4928	1374.84521
V	365	493550.50	466835.0	13059.4928	1352.19315
S	365	404411.00	466835.0	13059.4928	1107.97534
D	365	370712.00	466835.0	13059.4928	1015.64932
L	365	484262.50	466835.0	13059.4928	1326.74658
Average scores were used for ties.					

Kruskal-Wallis Test	
Chi-Square	93.1038
DF	6
Pr > Chi-Square	<.0001

Tabla 7.2: Contraste de igualdad de medias de la variable demanda respecto de los 7 días de la semana

El contraste lleva a rechazar la igualdad entre las medias respecto de los 7 tipos de día. Se puede intuir por las puntuaciones obtenidas que son el tipo de día *sábado* y el tipo *domingo* (puntuaciones más bajas) los que están provocando esta desigualdad. Efectivamente se puede comprobar que si establecemos el contraste entre los sábados o los domingos y cualquiera de los días “*intrasemanales*” el resultado del contraste volvería a ser de rechazo. Por el contrario, cuando se plantea el contraste exclusivamente de estos últimos entre sí, no se puede rechazar la hipótesis de igualdad.

Wilcoxon Scores (Rank Sums) for Variable CONSUMO Classified by Variable DIA					
DIA	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
H	366	426197.00	401319.00	11051.4579	1164.47268
X	366	426490.00	401319.00	11051.4579	1165.27322
J	365	421029.00	400222.50	11039.3715	1153.50411
V	365	413940.50	400222.50	11039.3715	1134.08356
D	365	309811.50	400222.50	11039.3715	848.79863
L	365	406060.00	400222.50	11039.3715	1112.49315
Average scores were used for ties.					

Wilcoxon Scores (Rank Sums) for Variable CONSUMO Classified by Variable DIA					
DIA	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
H	366	421827.00	401319.00	11051.4577	1152.53279
X	366	422160.50	401319.00	11051.4577	1153.44399
J	365	416687.50	400222.50	11039.3714	1141.60959
V	365	409478.00	400222.50	11039.3714	1121.85753
S	365	332086.50	400222.50	11039.3714	909.82603
L	365	401288.50	400222.50	11039.3714	1099.42055
Average scores were used for ties.					

Wilcoxon Scores (Rank Sums) for Variable CONSUMO Classified by Variable DIA					
DIA	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
H	366	340418.0	334524.0	9025.32936	930.103825
X	366	340608.0	334524.0	9025.32936	930.622951
J	365	335898.0	333610.0	9016.07524	920.268493
V	365	329868.0	333610.0	9016.07524	903.747945
L	365	323086.0	333610.0	9016.07524	885.167123
Average scores were used for ties.					

Kruskal-Wallis Test	
Chi-Square	68.9288
DF	5
Pr > Chi-Square	<.0001

Kruskal-Wallis Test	
Chi-Square	40.0304
DF	5
Pr > Chi-Square	<.0001

Kruskal-Wallis Test	
Chi-Square	1.9841
DF	4
Pr > Chi-Square	0.7387

Tabla 7.3: Contrastes de igualdad de medias de la variable demanda respecto de diferentes grupos de días

De igual manera se establece la desigualdad entre sábados y domingos. El contraste al 95% a ellos asociados, también conduce a rechazar la igualdad entre las medias.

Wilcoxon Scores (Rank Sums) for Variable CONSUMO Classified by Variable DIA					
DIA	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
S	365	139119.50	133407.50	2848.79227	381.149315
D	365	127695.50	133407.50	2848.79227	349.850685
Average scores were used for ties.					

Kruskal-Wallis Test	
Chi-Square	4.0203
DF	1
Pr > Chi-Square	0.0450

Tabla 7.4: Contrastes de igualdad de medias de la variable demanda entre sábados y domingos

Esto nos lleva a considerar en principio 3 tipos de días: *domingos*, *sábados* e “*intrasemanales*” (resto).

Aún así, hemos considerado conveniente diferenciar el comportamiento de los lunes y viernes del resto de los días laborables. La justificación viene a ser que, en dichos días, el consumo industrial suele ser ligeramente inferior (maquinaria que empieza a arrancar el lunes o deja de funcionar el viernes a mediodía) e incluso el doméstico acostumbra a caer por debajo de su nivel habitual (usuarios que desde el viernes al mediodía abandonan su residencia de trabajo para marchar a su pueblo o a su casa de la playa o la sierra, o estos mismos que vuelven el propio lunes directamente a trabajar sin pasar por su hogar). Además, ambos tipos de días están sujetos a posibles efectos “puente” (si el martes o el jueves eran festivos). Tras los correspondientes contrastes analíticos se puede comprobar que, efectivamente, la inclusión de los lunes o los viernes junto con los martes, miércoles y jueves proporciona un valor menor del p-valor (0.61 y 0.89 respectivamente) que si se contemplaran únicamente estos tres últimos (0.95).

Wilcoxon Scores (Rank Sums) for Variable CONSUMO Classified by Variable DIA					
DIA	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
M	366	271729.00	267729.00	6993.22146	742.428962
X	366	271836.50	267729.00	6993.22146	742.722678
J	365	268064.50	266997.50	6986.84659	734.423288
L	365	257823.00	266997.50	6986.84659	706.364384
Average scores were used for ties.					

Wilcoxon Scores (Rank Sums) for Variable CONSUMO Classified by Variable DIA					
DIA	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
M	366	270357.50	267729.00	6993.22147	738.681694
X	366	270447.00	267729.00	6993.22147	738.926230
J	365	266742.50	266997.50	6986.84661	730.801370
V	365	261906.00	266997.50	6986.84661	717.550685
Average scores were used for ties.					

Wilcoxon Scores (Rank Sums) for Variable CONSUMO Classified by Variable DIA					
DIA	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
M	366	201668.50	200934.0	4947.77197	551.006831
X	366	201675.50	200934.0	4947.77197	551.025956
J	365	198909.00	200385.0	4944.38656	544.956164
Average scores were used for ties.					

Kruskal-Wallis Test	
Chi-Square	1.8152
DF	3
Pr > Chi-Square	0.6116

Kruskal-Wallis Test	
Chi-Square	0.6186
DF	3
Pr > Chi-Square	0.8922

Kruskal-Wallis Test	
Chi-Square	0.0891
DF	2
Pr > Chi-Square	0.9564

Tabla 7.5: Contrastes de igualdad de medias de la variable demanda respecto de grupos de días intrasemanales

Además, el test planteado únicamente sobre las clases *lunes* y *viernes* no permite rechazar la igualdad entre este tipo de días, lo cual nos permite considerar la nueva agrupación sugerida.

Wilcoxon Scores (Rank Sums) for Variable CONSUMO Classified by Variable DIA					
DIA	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
V	365	134757.0	133407.50	2848.79223	369.197260
L	365	132058.0	133407.50	2848.79223	361.802740
Average scores were used for ties.					

Kruskal-Wallis Test	
Chi-Square	0.2244
DF	1
Pr > Chi-Square	0.6357

Tabla 7.6: Contrastes de igualdad de medias de la variable demanda entre lunes y viernes

Finalmente, obsérvese que hemos hablado indistintamente de días laborables e “intrasemanales” cuando algunos de éstos, como hemos visto presentan un comportamiento completamente diferente si caen en puente o en festivo. En el gráfico de la derecha de la ilustración (7.7) podemos ver que el nivel de una festividad es equiparable al de un domingo (de hecho éste en sí es un festivo) y el de un puente (no todo el mundo lo hace y por tanto no debe ser englobado en la clase de los festivos) al de un sábado. De acuerdo a esta consideración final, vamos a distinguir cuatro tipos de días:

- **Tipo I:** Domingos y festivos.
- **Tipo II:** Sábados y puentes.

- **Tipo III:** Lunes y Viernes que no sean ni festivos ni puentes.
- **Tipo IV:** Martes, Miércoles y Jueves que no sean ni festivos ni puentes (estos días sólo podrían dar lugar al puente de la **Constitución** comprendido entre los festivos 6 y 8 de Diciembre).

Para construir el histórico de demanda en condiciones normales de temperatura necesitaremos para cada mes de cada año del histórico, algunos datos asociados a cada uno de los tipos de día que se hayan dado en dichas condiciones. Se puede entender entonces el riesgo que se corre al establecer como “normal” todos aquellos datos que se encuentren en un determinado intervalo alrededor del valor de referencia (dado por la mediana de los 7 años de datos diarios de temperatura de los que se dispone). Si por ejemplo la temperatura máxima mediana de Febrero fuese de 13.6°C, y considerásemos como normal todo dato que estuviese entre $Q_1=10.9^\circ\text{C}$ y $Q_3=16.4^\circ\text{C}$, la mala suerte nos podría deparar un mes de Febrero atípico en nuestro histórico que presentase pocos registros en dichas condiciones y en concreto, ninguno para alguno de los tipos de día más reducidos. Por ejemplo, la ausencia de festivos en Febrero simplificaría el grupo I a los cuatro domingos pudiendo ser que ninguno de ellos registrase datos en condiciones normales. En consecuencia, no tendríamos muestra de datos festivos en dichas condiciones para estimar cuál fue la demanda que debió esperarse un “festivo normal de Febrero de un año concreto”. Por ello, vamos a establecer el dato de demanda normal asociado a un tipo de día de un mes y año determinado como la media ponderada de todos los datos de ese mismo mes y año que comparten el mismo grupo de día. La ponderación se establecerá en función de la distancia entre el dato real de temperatura máxima y la mediana que sirve de valor representativo del mes. Datos a menor distancia del valor mediano del mes (asociados por tanto a temperaturas más “normales”) tendrán mayor peso en el cálculo de la media que datos más alejados. El valor obtenido para cada grupo de día como resultado de esa media ponderada es propuesto como valor “normal” de la demanda para todos los días del mismo tipo de ese año y mes. Se harán algunas correcciones a posteriori de acuerdo al conocimiento que se tiene de las propiedades que deben guardarse entre los distintos tipos de día:

1. Si en un mismo año y mes, el nivel medio del grupo III (lunes y viernes) es superior al del grupo IV (martes, miércoles y jueves), se sustituye el valor del tercero por el del cuarto.
2. Si en un mismo año y mes, el nivel medio del grupo I (domingos y festivos) es superior al del grupo II (sábados y puentes), se sustituye el valor del primero por el del segundo.

No ha sido preciso corregir ningún otro tipo de circunstancia más difícil de presentarse como por ejemplo que el nivel medio del grupo I o II estuviese por encima del de alguno de los otros grupos (III o IV) si bien en dicho caso, se hubiera hecho igualmente la correspondiente modificación de valores.

Vamos a presentar un gráfico por cada uno de los tipos de meses en el que comparamos el dato de demanda así

construido (en rojo) con el dato original (en azul). Respecto de los meses estrictamente veraniegos, dado que los datos no se encuentran condicionados por las variables de temperaturas, podríamos considerar que ellos mismos son los que se presentaron en condiciones normales. Sin embargo, aplicaremos también a estos meses el método de normalización del histórico y comprobaremos que ambos datos (el real y el que supuestamente se hubiera presentado en condiciones normales) están muy próximos.

No es sólo una cuestión de consistencia con la forma de proceder para el resto de los meses, sino también como medida de precaución ante la existencia de alguna anomalía climática que pudiese justificar contra todo pronóstico el mayor uso de gas durante estos periodos (un final de Septiembre muy frío que “anticipa” la llegada del invierno). A modo de ejemplo, haremos referencia a la existencia de una ola de frío inesperada en Mayo de 2003 que se manifestó entre los días 6 y 7 con máximas alrededor de 13°C y mínimas entorno a los 6°C (frente a los valores normales para dicho mes de 23°C y 11°C respectivamente). Como consecuencia, se produjo un incremento descomunal de la demanda en un mes en el que el uso de la calefacción parecería del todo injustificado.

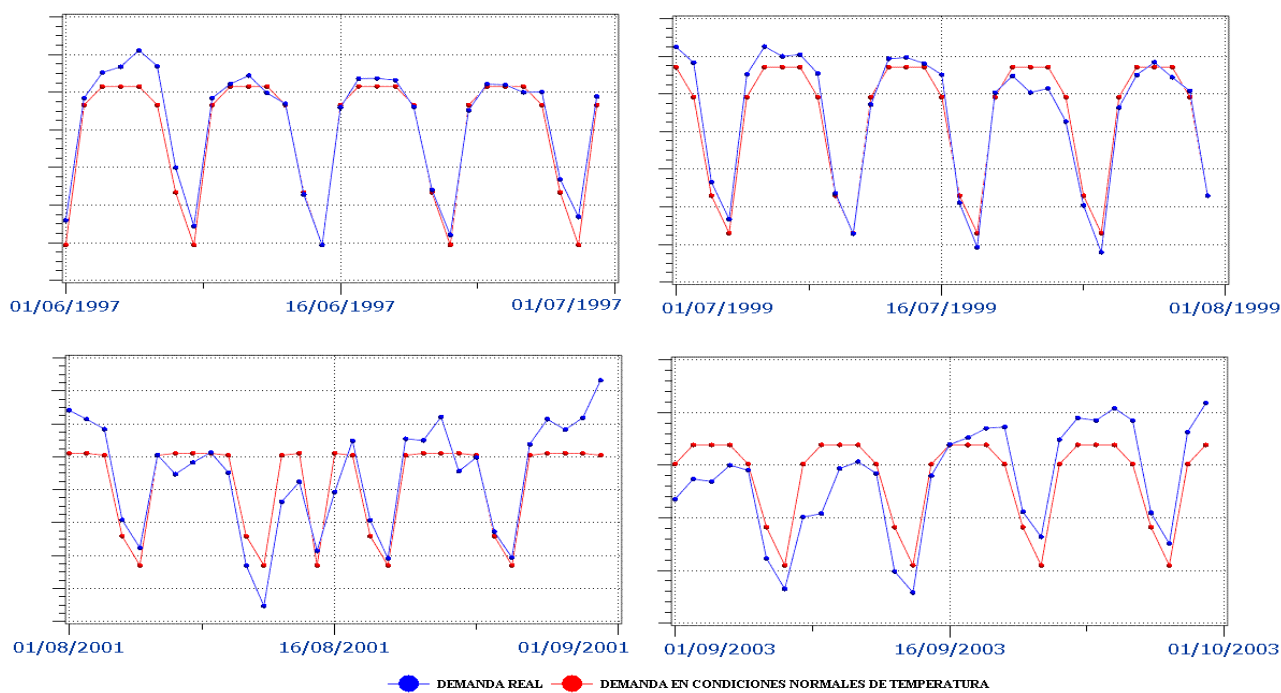


Ilustración 7.8: Comparativa entre el histórico real de demanda y el construido en condiciones normales de temperatura durante los meses de verano

Para los meses asociados al resto de las estaciones del año (de Octubre a Abril), se han tomado también distintos tipos de meses y años, con la intención de dar una idea evolutiva desde el mes que marca el comienzo del histórico (Abril de 1997) hasta el último que sirve de entrada al conjunto de entrenamiento (Marzo de 2004). El resultado es:

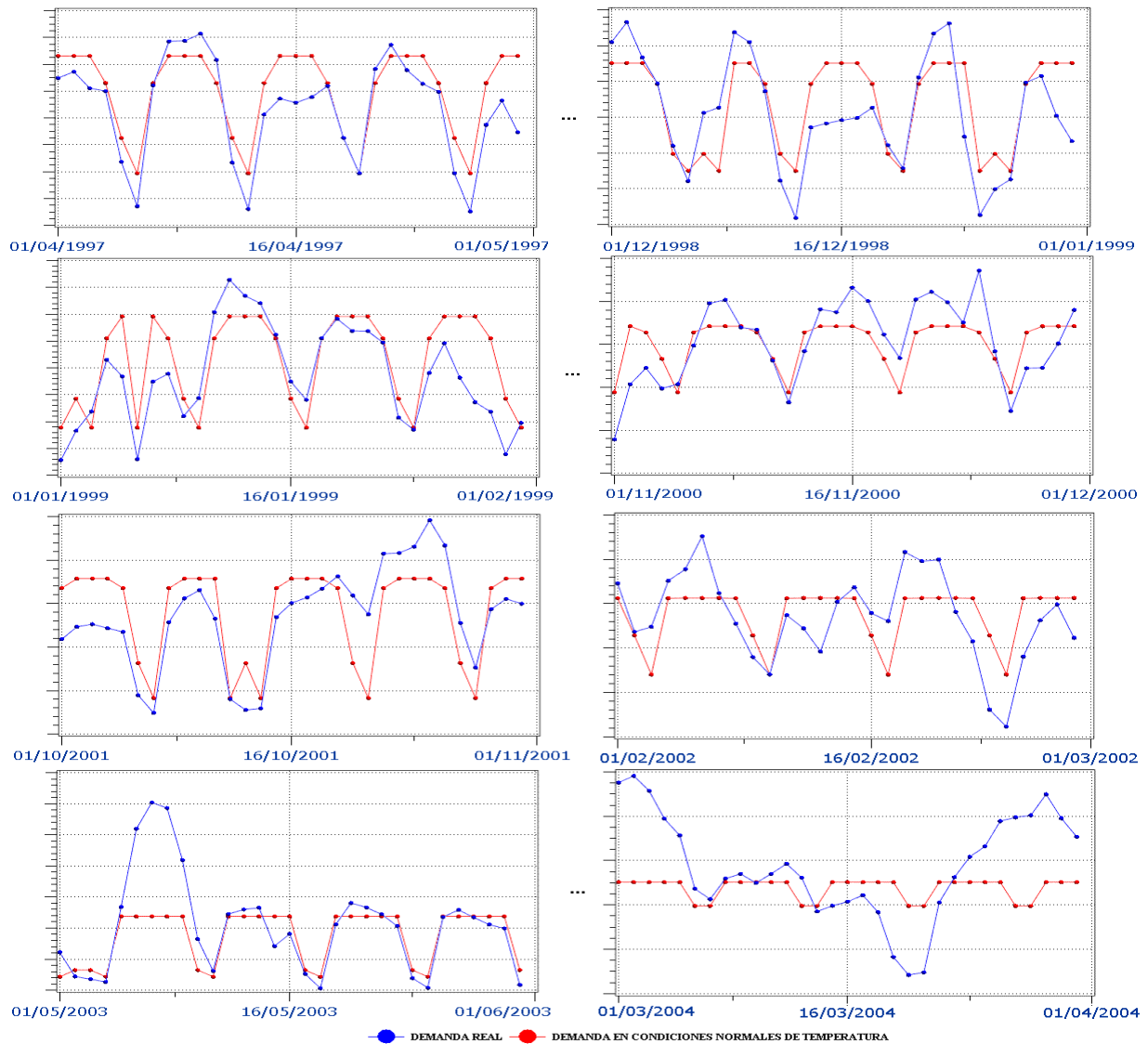


Ilustración 7.9: Comparativa entre el histórico real de demanda y el construido en condiciones normales de temperatura durante los meses de invierno

A la vista de ellos, podemos hacer una serie de observaciones:

- La diferencia entre el dato real y el que consideramos que se presenta en condiciones normales de temperatura es mayor que en los meses de verano. El patrón estacional real tiende a perderse con frecuencia.
- Aunque parezca contradictorio, estas diferencias no son mayores en los meses más fríos (Noviembre, Diciembre y Enero de acuerdo a la tabla (7.1)) y de hecho el patrón estacional es más fácilmente identificable en dichos meses. La razón es que al ser las bajas temperaturas un rasgo común a todos los días, el incremento de la demanda sobre el supuesto nivel de normalidad es constante. Los clientes utilizan el gas todos los días al máximo rendimiento (efecto saturación) de manera regular y es esa demanda regular la que se incorpora sobre

el patrón estacional característico de la demanda industrial. De esta forma, la estacionalidad no se pierde.

- Uno de los meses más inestables es Marzo por servir de transición entre el invierno y la primavera. La alternancia de periodos con buenas temperaturas y otros en los que se presentan repentinas olas de frío provoca la irregularidad en el consumo doméstico con la consecuente deformación del patrón estacional.
- Como advertimos anteriormente, podemos encontrar meses más próximos al verano que al invierno que son sorprendidos por una drástica caída de temperaturas. El resultado salta a la vista en la primera semana del mes de Mayo representado.

En cualquiera de los meses el patrón construido es puramente estacional y respeta las condiciones impuestas de antemano en cuanto a las relaciones que se establecen entre los días de una misma semana (grupos I-IV). Sin embargo, el hecho de proceder de esta manera conlleva que se produzcan saltos bruscos del nivel de la demanda al pasar del último día de un mes al primero del siguiente. Dado que el valor de la demanda es el mismo para los días centrales de una semana (Martes, Miércoles y Jueves), el patrón ofrece un aspecto aún menos creíble si la transición entre meses se produce a mitad de semana (véase ilustración (7.10)). Para mejorar esta falta de suavidad, se ha procedido de la siguiente manera:

1. Las semanas de un mes han sido numeradas de 1 a 4 entendiéndose por una **semana** el periodo que empieza en lunes y acaba en domingo. Así, la primera semana de un mes MM (que denominaremos como $S_{1_M_MM}$ = semana 1 del mes MM , con $MM \in \{1, 2, \dots, 12\}$) no empieza necesariamente en día 1, sino en el primer lunes de dicho mes. Las semanas sucesivas de dicho mes ($S_{2_M_MM}$, $S_{3_M_MM}$, ...) se van construyendo de manera sucesiva a partir de la primera. Obsérvese que puede ser que no exista semana cuarta ($S_{4_M_MM}$) por no poder incluir dentro del mismo mes una cuarta sucesión de siete días en caso de que la primera semana ($S_{1_M_MM}$) hubiese empezado demasiado tarde (un día 5 o 6 de dicho mes).
2. Definiremos además como **semana de transición** asociada a un mes MM (ST_{M_MM}) aquella que comience un lunes del mes MM y concluya un domingo del mes $MM+1$. Obsérvese que puede darse el caso de que no exista semana de transición en caso de que la última semana del mes MM ($S_{4_M_MM}$) acabe justamente en domingo (la semana siguiente en este caso sería $S_{1_M_}\{MM+1\}$).
3. A partir de variables de este tipo ha resultado fácil proceder al suavizado de los periodos de transición entre meses. De esta manera, dados dos meses consecutivos A y B ($B = A+1$), se ha considerado la variable $S_{4_M_A}$ (o en su ausencia $S_{3_M_A}$), asociada a la última semana completa del mes A , y la variable

$S_{I_M_B}$ asociada a la primera semana completa del mes B . Posteriormente se ha tomado un representante de cada grupo de días de cada una de estas dos semanas, por ejemplo un día del grupo IV (martes, miércoles o jueves que no sea puente ni festivo) de la semana $S_{4_M_A}$ (o, en caso de no existir ésta, $S_{3_M_A}$) y otro de ese mismo grupo de la semana $S_{I_M_B}$ y se ha calculado el valor medio de ellos. Se ha procedido de la misma forma para cada uno de los cuatro tipos de día, obteniéndose cuatro representantes “medios” o “de transición” que han sido utilizados para sustituir a las semanas $S_{4_M_A}$ (o $S_{3_M_A}$), $S_{I_M_B}$ y ST_M_A . En caso de no existir ST_M_A , las semanas sustituidas por estas “semanas medias” serían $S_{3_M_A}$, $S_{4_M_A}$ y $S_{I_M_B}$.

El siguiente gráfico muestra la transición entre los meses de Septiembre y Octubre de 1998 para los datos reales de demanda (en azul), los construidos supuesto que se dieron condiciones normales de temperatura (en rojo) y los finalmente propuestos una vez hecho el suavizado (en negro). El salto se produce el día 1 de Octubre que, por estar localizado entre semana (jueves), deforma de manera significativa la estacionalidad respecto del mismo día de la semana anterior generando un incremento repentino, desmesurado y poco verosímil. A través del patrón suavizado no sólo se solventa esta irregularidad sino que además se consigue una mayor aproximación a la curva original.

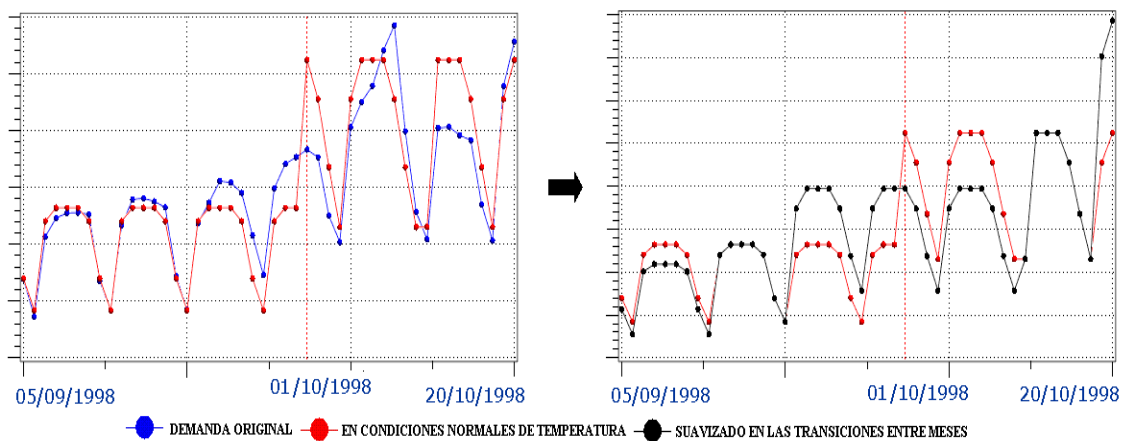


Ilustración 7.10: Suavizado “intramensual” del histórico de demanda construido en condiciones normales de temperatura

4. Finalmente se valora la subida (o bajada) que han supuesto los niveles de demanda de estas nuevas semanas (entre ellas coincidentes) con respecto a aquellas a las que han sustituido. Si la semana $S_{4_M_A}$ (o $S_{3_M_A}$) ha experimentado una subida del $X_a\%$ y la semana $S_{I_M_B}$ un descenso del $X_b\%$, con vistas a que el nivel de todas las semanas intermedias no sea el mismo y dotar de mayor suavidad aún al histórico, se ha reajustado la semana $S_{4_M_A}$ (o $S_{3_M_A}$) rebajando su nivel en un $\frac{X_a}{2}\%$ y la semana $S_{I_M_B}$ incrementando su nivel en un $\frac{X_b}{2}\%$. La semana de transición ST_M_A (o en caso de no existir, $S_{4_M_A}$)

conservará el valor medio estimado, previamente a este reajuste, como media de las semanas S_4M_A (o S_3M_A) y S_1M_B . El resultado de este proceso es el siguiente:

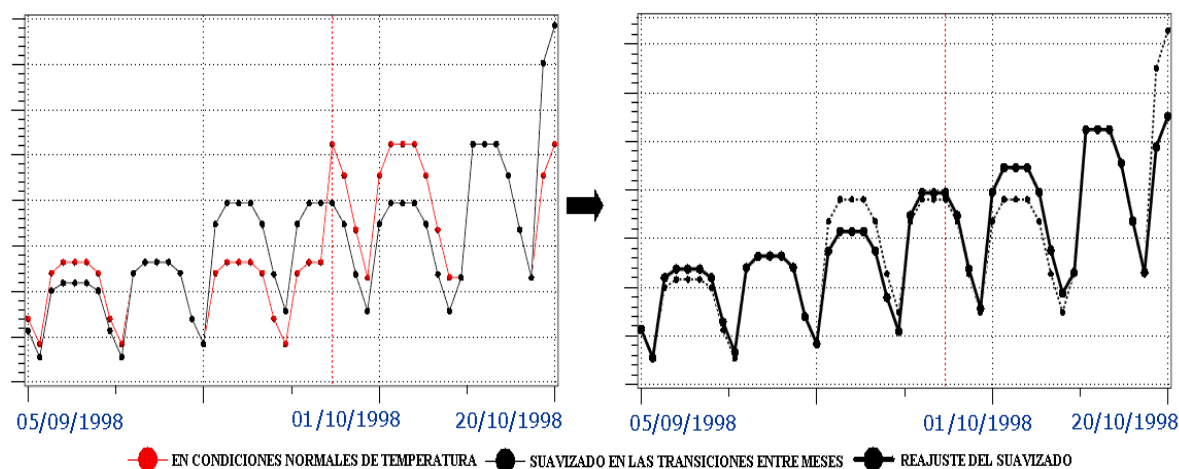


Ilustración 7.11: Reajuste del suavizado “intramensual” del histórico de demanda construido en condiciones normales de temperatura

En resumen, el procedimiento consiste en redistribuir el volumen de demanda acumulado durante la última semana de un mes, la primera del siguiente y la intermedia entre ambos para dar al patrón un aspecto más suave y menos escalonado. Bastaría proceder de esta manera sobre cada uno de los meses de transición para obtener el hipotético histórico en condiciones normales de temperatura (máxima).

A continuación se muestra el resultado de aplicar esta técnica a un año completo, pues éste es el periodo máximo que parece razonable para poder superponer en un mismo gráfico la curva original de demanda (en azul) frente a la que asume un valor constante para todos los días de un mismo grupo dentro de un mes (en rojo). En un gráfico posterior establecemos la comparativa entre esta última y la curva de demanda que finalmente se obtiene como suavizado de ella (en negro).

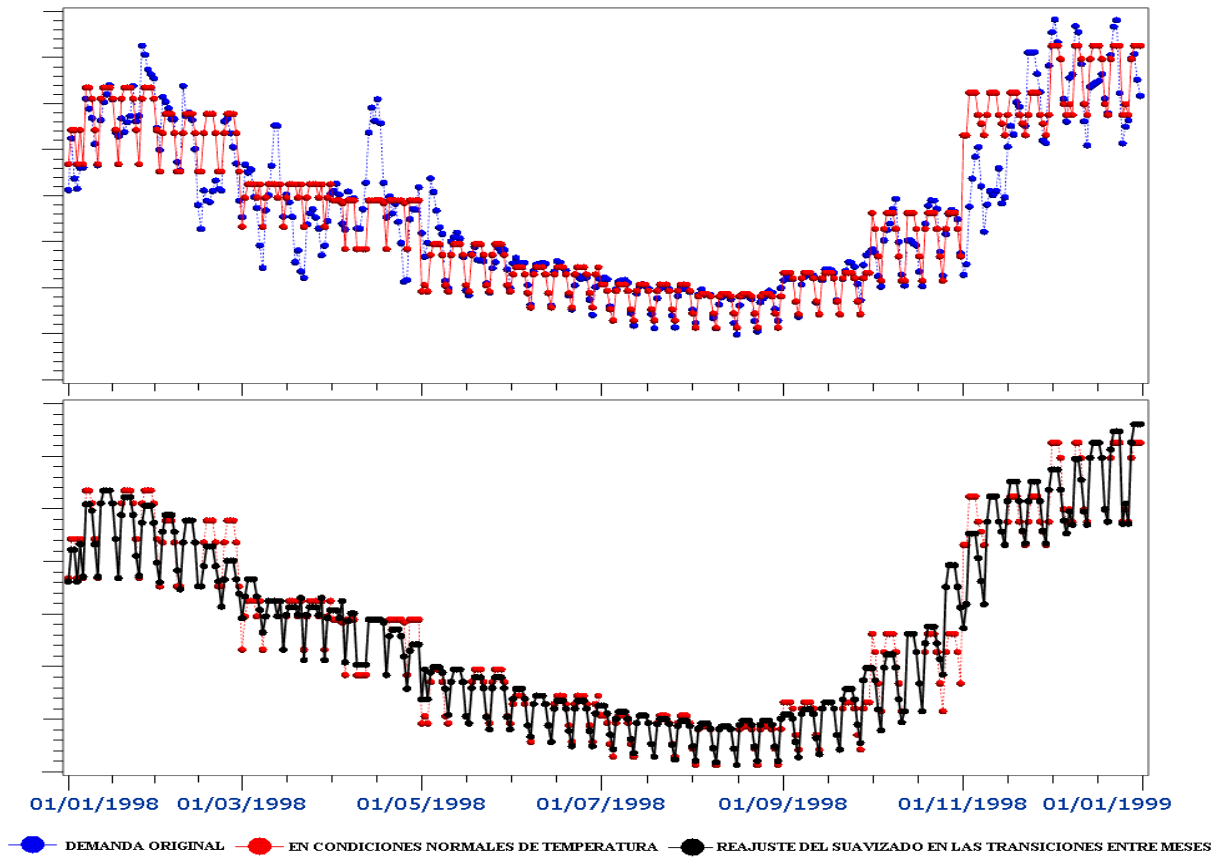


Ilustración 7.12: Comparativa, sobre el periodo de un año, entre el histórico real de demanda y el construido en condiciones normales de temperatura

Para terminar, presentamos el patrón de demanda que resultaría de utilizar este método sobre el histórico completo.

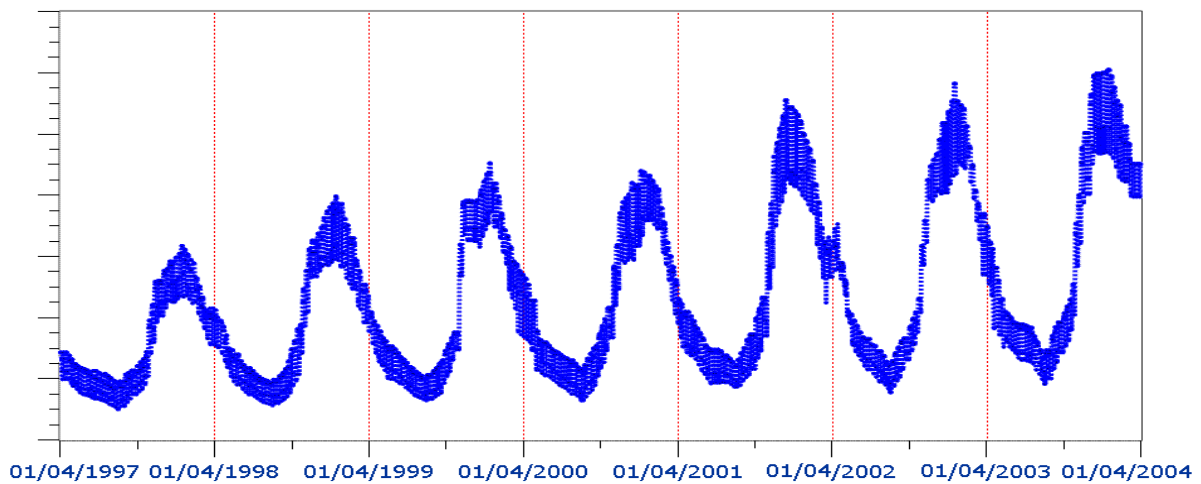


Ilustración 7.13: Patrón de demanda de gas supuestas unas condiciones normales de temperatura máxima

Vamos a contrastar si el patrón así construido responde a las condiciones meteorológicas normales que estamos asumiendo. Para ello vamos a generar a partir de este resultado y de los datos de partida, una muestra bidimensional donde la primera componente “x” de cada par refleje la diferencia entre el dato de temperatura máxima que se ha establecido como normal para el mes al que pertenece el dato y el de temperatura máxima real; la segunda componente “y” será la diferencia porcentual entre el dato real de demanda y el que forma parte de este patrón. Es decir,

$$X(t) = \text{Temperatura_Maxima_Normal}_{MES} - \text{Temperatura_Maxima_Real}(t) \quad (7.2)$$

$$Y(t) = \frac{100 * (\text{Demanda_Real}(t) - \text{Demanda_en_condiciones_normales_de_temperatura}(t))}{\text{Demanda_en_condiciones_normales_de_temperatura}(t)} \quad (7.3)$$

Si ambas componentes son positivas, podremos interpretar que cuando la temperatura real se sitúa por debajo del valor tomado como normal, el valor real de la demanda estará por encima del nivel marcado por el patrón. Coloquialmente hablando podemos decir que “si hiciera más frío de lo normal la demanda subiría por encima del patrón establecido”. En el caso en el que las dos componentes sean negativas podríamos igualmente decir que “si hiciera más calor de lo normal la demanda caería por debajo del patrón establecido”. En uno y otro caso es la relación de dependencia que cabría esperar, es decir, la lógica conduciría a pensar que el proceso de generación del patrón ha funcionado bien si estos pares se concentran en los cuadrantes primero y tercero del plano.

El siguiente gráfico viene a confirmar este hecho. Únicamente vamos a representar los puntos asociados a los meses comprendidos entre Noviembre y Marzo (**invierno gasista**) en los que la influencia de la temperatura se puede considerar manifiesta y en los que por tanto tiene sentido medir esta relación. Se representa con un color distinto los pares correspondientes a cada uno de los meses, dando a entender la necesidad de evaluar de manera diferente la relación en cada caso, puesto que por ejemplo, el efecto de una ola de frío no es el mismo en Diciembre y en Marzo.

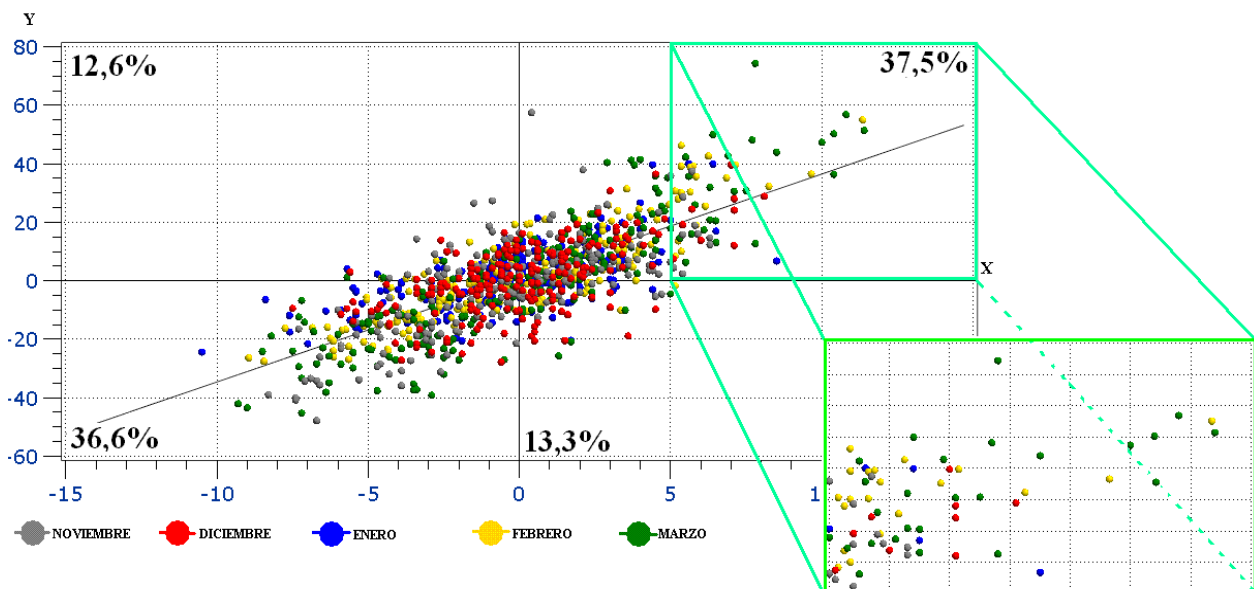


Ilustración 7.14: Relación entre el incremento de demanda y el incremento de temperatura

Casi el 75% de los puntos (3 de cada 4) se sitúan en los cuadrantes que podríamos considerar “lógicos”. Además, para aquéllos que caen en el primero y en el cuarto, el valor de ambas componentes es próximo a cero, circunstancia que también juega a favor del proceso de construcción del histórico dado que parece razonable esperar que cuando la temperatura experimente cambios muy pequeños respecto de su nivel normal, el valor de la demanda se mantenga más o menos constante. La siguiente tabla detalla por meses, el número de puntos que cae en cada uno de los cuadrantes y la proporción que dicha cantidad representa del total de los pares asociados a cada mes:

Table of CUADRANTE by MES						
CUADRANTE	MES					Total
	1	2	3	11	12	
1	84 38.71	76 38.38	94 43.32	83 39.52	75 34.56	412
2	37 17.05	17 8.59	15 6.91	28 13.33	40 18.43	137
3	70 32.26	82 41.41	89 41.01	76 36.19	67 30.88	384
4	26 11.98	23 11.62	19 8.76	23 10.95	35 16.13	126
Total	217	198	217	210	217	1059

Tabla 7.7: Distribución por mes y cuadrante del plano de los pares (incremento_temperatura, incremento_demanda)

El patrón de demanda que hemos construido nos va a servir de punto de partida para las dos fases siguientes:

- Por un lado, disponemos de un histórico en condiciones normales de temperatura con base en el cual podremos realizar predicciones en dichas condiciones (Fase II).
- Por otro lado, tenemos una serie de pares que reflejan el incremento porcentual que se puede esperar de la demanda ante posibles desviaciones de los valores estándares establecidos para la variable de temperatura (Fase III).

Antes de pasar a la segunda fase, vamos a hacer una observación que salta a la vista a partir de la ilustración (7.14). Nótese que, si bien la relación entre las variables parece lineal, el estudio está orientado al análisis de la dependencia en situaciones extremas para las que esta linealidad se pierde según se intuye en la región ampliada de dicho gráfico (descensos superiores a 5°C respecto del valor normal de la temperatura). Para comprobarlo basta ajustar un modelo de regresión lineal simple para cada uno de los meses a partir de dicho umbral climático, el resultado del cual se muestra en la siguiente tabla:

MES	ECUACION	T_RATIO_B0	T_RATIO_B1	R_SQUARE
01	$Y=65.90-6.41*X$	1.94	-1.23	0.3339
02	$Y=8.20+3.72*X$	0.85	2.54	0.2637
03	$Y=-5.72+5.36*X$	-0.51	3.60	0.3600
11	$Y=18.77-0.22*X$	0.40	-0.03	0.0001
12	$Y=-14.39+5.58*X$	-0.70	1.79	0.3133

Tabla 7.8: Rectas de regresión ajustadas a los pares (incremento_temperatura, incremento_demanda) para valores extremos de “incremento_temperatura”

Parece poco aconsejable el empleo de estas ecuaciones para predecir el incremento/decremento del consumo de gas supuesto una bajada o subida de las temperaturas. Vamos a enumerar las razones que hacen más aconsejable la utilización de funciones cópula a tal fin:

1. De entrada, las ecuaciones de regresión ofrecen poca fiabilidad. Para un nivel de confianza del 95%, únicamente en 2 de los 5 casos la pendiente de la recta se puede considerar significativa y en cuanto a los valores de los R^2 tampoco son representativos de un buen ajuste.
2. Estos modelos están ajustados a muestras de datos relativamente pequeñas de entre 10 y 20 datos, dado que van asociadas a comportamientos extremos respecto de la variable de temperatura. Esta circunstancia nos puede llevar a considerarlas poco representativas para hacer una predicción. En su lugar, las funciones cópula actuarán, en cada mes, sobre todos los pares localizados en el primer cuadrante, es decir, sobre conjuntos de entre 70 y 90 datos (véase tabla (7.7)). Pero además, las propiedades de las familias de cópulas que vamos a utilizar (HRT, las familias de valor extremo, o las ajustadas localmente por interpolación) van a permitir dar un tratamiento especial a los pares situados en la ventana de riesgo sin necesidad de centrar el ajuste exclusivamente en ella tal y como hacen los modelos de regresión.
3. La simetría de los intervalos de confianza asociados a las predicciones que ofrecen los modelos de regresión se antojan poco verosímiles. Por el contrario, las cópulas permitirán presentar las predicciones a través de distribuciones asimétricas, el análisis de las cuales (de sus colas), permitirá evaluar la probabilidad de que se alcance un récord histórico de demanda.

7.2.4 FASE II: Construcción de un modelo autorregresivo para hacer una predicción a medio plazo

La siguiente etapa de esta metodología consiste en determinar la ecuación generadora del histórico que hemos supuesto que debió presentarse en condiciones normales de temperatura. En esta ecuación deberán quedar reflejados sus rasgos más característicos que son:

- Su marcada **estacionalidad semanal** ya patente en los datos reales de demanda pero aún más señalada en el patrón que hemos construido (véase ilustración (7.11)). Esta estacionalidad sufrirá alteraciones en presencia de días atípicos conocidos de antemano como son las festividades y puentes que habitualmente son tratados mediante un análisis de intervenciones.
- Su **variabilidad cambiante** con mayores y menores diferencias entre días laborables y festivos en función del mes del año en el que oscilen los datos. Esta propiedad también es propia del histórico original, pero nuevamente se acentúa aún más en el patrón debido al tratamiento mensual que hemos llevado a cabo.
- Su también destacada **periodicidad anual** (véase ilustración (7.13)) que una vez identificada (variable *Ciclo*) nos va permitir, como veremos, extender el horizonte de predicción de corto a medio plazo. En este *Ciclo* irán implícitos cambios de tendencia y caídas justificadas del nivel de la demanda durante periodos vacacionales como Navidad o Agosto.

Vamos a ir incorporando en un modelo lineal de regresión cada una de estas componentes de manera secuencial.

7.2.4.1 Identificación del patrón estacional semanal

Una forma habitual de reflejar comportamientos periódicos en un modelo es incluyendo como variable explicativa el valor retardado de la propia variable a predecir tantas unidades como longitud tenga el periodo estacional. Este tipo de modelos en los que la propia variable actúa como regresor recibe el calificativo de “autorregresivo” (AR). El más popular y posiblemente el más utilizado para el tratamiento a corto plazo de series estacionales es el SAR(1)x(1)_s el cual responde a la ecuación:

$$(1 - \phi_1 * B) * (1 - \Phi_1 * B^s) X_t = \epsilon_t \Rightarrow X_t = \phi_1 * X_{t-1} + \Phi_1 * X_{t-s} - \phi_1 * \Phi_1 * X_{t-s-1} + \epsilon_t \quad (7.4)$$

siendo X_t la serie a predecir (*CONSUMO_MEDIO*), B el operador retardo $B^k X_t = X_{t-k}$, ϵ el término de error y “ s ” el periodo de estacionalidad ($s = 7$ que determina una estacionalidad semanal).

Este modelo establece la dependencia entre el dato real de una unidad temporal (día, mes, año, etc.), el dato real más reciente al que hace referencia el primer orden del modelo (ayer, el mes pasado, el año pasado, etc.) y el dato que se presentó hace “ s ” instantes de tiempo al que hace referencia el segundo orden (el mismo día de la semana pasada, el mismo mes del año pasado, etc.). La estacionalidad se incorpora de manera multiplicativa sobre la tendencia de acuerdo al resultado de los factores $(1 - \phi_1 * B)$ y $(1 - \Phi_1 * B^s)$.

Alternativamente, se puede incorporar de manera aditiva (SAR(s) con $\phi_2 = \dots = \phi_{s-1} = 0$) conforme a la expresión

$$(1 - \phi_1 * B - \phi_s * B^s) X_t = \epsilon_t \Rightarrow X_t = \phi_1 * X_{t-1} + \phi_s * X_{t-s} + \epsilon_t \quad (7.5)$$

En nuestro caso, de los dos términos explicativos únicamente nos va a interesar contemplar aquél que modeliza estacionalidad, dado que ésta, de acuerdo al patrón construido, se va a presentar a lo largo de las más de 700 unidades temporales (2 años en días) para las que realizaremos la predicción. Respecto del término tendencia, debe tenerse en cuenta que éste va más orientado a una predicción a corto plazo, de forma que el modelo fundamente la predicción en el último dato (“decir para mañana algo parecido a lo que sucedió hoy”) y utilice éste como referencia del nivel en el que se está moviendo y deberá moverse la serie. Sin embargo, no es realista suponer que este pasado inmediato vaya a condicionar el comportamiento diario de la demanda durante los dos próximos años y por ello, en principio, no tiene sentido que contemplemos dicha componente en la fórmula. En su lugar, incluiremos una variable *Ciclo* (apartado 7.2.4.3) que sirva de guía al modelo durante esos dos años y le informe sobre los cambios de nivel que cabe esperar en función del mes en el que se “mueva” la serie. Veremos que, de acuerdo a la forma en que definamos dicha variable, la componente estacional se incorporará sobre ella de manera aditiva o multiplicativa. De momento, empezaremos planteando una primera ecuación para la demanda que identifique su estacionalidad:

$$(1 - \phi_7 * B^7) X_t = \epsilon_t \Rightarrow X_t = \phi_7 * X_{t-7} + \epsilon_t \quad (7.6)$$

Otra manera de plasmar la estacionalidad en la fórmula es a través de $(s-1)$ variables *dummy* asociadas a cada uno de los tipos de día (*lunes, martes,...*), meses (Enero, Febrero,...) o en definitiva, clases distintas, que comprendan un periodo estacional. Estas variables son binarias y toman valor 1 para todos los datos que pertenecen a la clase que identifican. Así por ejemplo la variable *Lunes* valdrá 1 en cada uno de los lunes del histórico y cero en el resto de días. La razón por la que se consideran $(s-1)$ variables en lugar de “ s ” es por evitar efectos multicolineales derivados de que una de ellas se pueda expresar como combinación lineal de las restantes $D(=Domingo) = 1 - (L + M + X + J + V + S)$ (7.7)

De acuerdo a estas variables *dummy*, la ecuación del modelo sería

$$X_t = \phi_L * Lunes_t + \phi_M * Martes_t + \phi_X * Miercoles_t + \phi_J * Jueves_t + \phi_V * Viernes_t + \phi_S * Sabado_t + \epsilon_t \quad (7.8)$$

Normalmente se prefiere la utilización de la fórmula autorregresiva (7.6) dado que de acuerdo al principio de parsimonia interesan los modelos que tengan menor número parámetros (según vimos en el apartado 2.5.4, el *AIC*, indicador de la calidad del ajuste de un modelo, penaliza la presencia de parámetros). Sin embargo en nuestro caso nos encontramos ante un tipo de estacionalidad muy peculiar. Según hemos construido el histórico, la similitud no se presenta exclusivamente entre dos días separados por el periodo de una semana (dos lunes, dos martes, etc.) sino entre dos días que sean del mismo tipo (I, II, III o IV). Es decir, el comportamiento por ejemplo de un miércoles no es sólo similar al del miércoles pasado sino también al de cualquier martes, miércoles o jueves de ese mismo mes. Podríamos por tanto considerar variables asociadas a cada uno de los tipos de días que permitiesen reflejar este tipo característico de estacionalidad. El número de parámetros no sería mucho mayor que en el caso autorregresivo pues se pasaría de 1 a 3 dado que, a pesar de que hay cuatro tipos de días, uno de ellos siempre se puede obtener a partir de los restantes. Respecto de estas variables el modelo sería:

$$X_t = \phi_{SP} * TipoI_t + \phi_{LV} * TipoII_t + \phi_{MXJ} * TipoIII_t + \epsilon_t \quad (7.9)$$

Pero además, aunque pudiera pensarse que el número de parámetros de la expresión (7.9) es mayor que en (7.6), esto tampoco es realmente así. Debe tenerse en cuenta que por definición, estas variables están recogiendo también el efecto asociado a días festivos y puentes lo cual posibilita evitar llevar a cabo un análisis de intervenciones que se traduciría en

la inclusión de nuevas variables en el modelo. Efectivamente, si planteáramos la ecuación (7.6), este tipo de días atípicos rompen la estacionalidad marcada por el parámetro autorregresivo de orden 7 y resultaría por tanto conveniente reflejar su influencia en la serie, lo cual incrementaría el número de variables y, en consecuencia, de parámetros. Bien es verdad que en este caso no sería necesario incluir demasiadas variables de intervención. Es cierto que el efecto de una festividad es diferente en función del día de la semana en el que caiga dado que en ocasiones podrá dar lugar a un puente (martes o/y jueves) , ser un puente en sí misma (lunes y viernes), o constituir una festividad de forma aislada (miércoles). También encontraremos diferencias entre festividades de ámbito local (que afectan a la comunidad) y de ámbito nacional (celebradas en todo el país) y dentro de estas últimas la repercusión de algunas (1 de Enero, 25 de Diciembre, etc.) puede ser más fuerte que la de otras (1 de Mayo, 1 de Noviembre, etc.). Sin embargo, para el problema a medio plazo que estamos planteando no es necesario ser tan meticuloso a la hora de cuantificar con exactitud el impacto sobre la demanda de un festivo de acuerdo a un cruce concreto de circunstancias (1 de Enero que cae en Jueves y origina un puente el día 2). Esto, como veremos en el apartado 7.3.2.1, tiene más sentido para un problema de predicción a corto plazo en el que el objetivo es afinar al máximo la predicción para cualquier tipo de día. En cambio, en el estudio a medio plazo vamos a hacer una predicción a partir de un histórico que de entrada, hemos construido nosotros y donde el objetivo es conocer el “pico” aproximado de demanda que puede presentarse en un día genérico (laborable o festivo) del periodo de dos años que comprende el horizonte de predicción. Por ello, para el estudio a medio plazo, bastaría con añadir un par de variables dicotómicas a la ecuación (7.6), una asociada a los festivos y otra a los puentes (dado que el efecto de ambos tipos de días es distinto). Con su incorporación esta ecuación incrementaría en dos el número de variables (y de parámetros), cantidad con la que igualaría en complejidad al modelo de expresión (7.9) y que implicaría que este último no sufriera una mayor penalización en el cálculo del índice AIC. Nos decantaremos entonces por el modelo (7.9) dado que su forma de recoger la estacionalidad parece más adecuada al histórico que hemos construido. Esta ecuación sin ser exactamente autorregresiva (por no incluir a la propia variable X entre sus regresores) identifica de forma correcta no sólo la estacionalidad, sino la relación de parentesco que se establece entre días del mismo tipo.

Una clase de modelos más sofisticada que los autorregresivos AR (SAR en el caso estacional), son los ARMA pero que, para un problema de predicción a medio plazo, tampoco aportan ventajas suficientes que justifiquen su utilización. El papel que desempeña en estos modelos los parámetros de la parte MA, o de medias móviles (*moving average*), es corregir la predicción en función de los últimos errores que ha cometido el modelo. Así por ejemplo un ARMA(1,1) es un clásico que en general funciona muy bien puesto que calcula la predicción a partir del último dato real (AR(1)) y del último error que cometió el modelo (MA(1)) a partir del cual se corrige (si el modelo se equivocó por defecto hará una predicción a la alza y viceversa):

$$(1 - \phi_1 * B)X_t = (1 - \theta_1(B))\epsilon_t \Rightarrow X_t = \phi_1 * X_{t-1} - \theta_1 * \epsilon_{t-1} + \epsilon_t \quad (7.10)$$

El orden de una y otra parte raramente es superior a 2 (o a $2*s$ si la serie es estacional) ya que, por ejemplo para una serie diaria, no parece tener sentido que el valor esperado un día dependa del dato real de hace más de dos semanas (y mucho menos del error que cometió entonces el modelo). Esta es la razón por la cual no son útiles los modelos ARMA para realizar predicciones a tan largo plazo. Obviamente el factor corrector MA sólo es aplicable mientras se disponga de un dato real del pasado al que haga referencia (orden de la parte MA) pues de otra manera no podría ser evaluado el error que cometió el modelo. A partir del horizonte identificado por dicho orden (7 o 14) el error será desconocido y en

consecuencia se asumirá que toma valor 0 (se espera que el modelo no se equivoque) con lo que la influencia de dicho factor desaparece para los más de 700 días restantes (la parte MA es como si no existiera). Por ello, no incluiremos términos de medias móviles en nuestro modelo.

En resumen, propondremos la ecuación (7.9) como base para realizar una predicción a medio plazo que tenga en cuenta la estacionalidad. El resultado del ajuste es el siguiente:

Dependent Variable		CONSUMO_MEDIO		
Ordinary Least Squares Estimates				
SSE	2541483.81	DFE	2554	
MSE	995.09938	Root MSE	31.54520	
SBC	24927.5582	AIC	24910.0184	
Regress R-Square	0.6768	Total R-Square	0.6768	
NOTE: No intercept term is used. R-squares are redefined.				
Variable	Estimate	Standard Error	t Value	Approx Pr > t
DIA_MXJ	51.8410	0.9680	53.56	<.0001
DIA_SP	44.1864	1.6182	27.31	<.0001
DIA_LV	50.2330	1.2062	41.65	<.0001

Tabla 7.9: Modelo para la predicción a medio plazo de la demanda de gas que identifica el efecto estacional

De acuerdo al valor de los t-ratios podemos concluir que las tres variables *dummy* son muy significativas y el valor del R^2 también es representativo de la buena calidad del ajuste. En el siguiente gráfico se muestra la predicción realizada por este primer modelo. En color azul se presenta el patrón de demanda construido a partir del histórico y en rojo, la predicción que se obtendría teniendo en cuenta únicamente las variables estacionales.

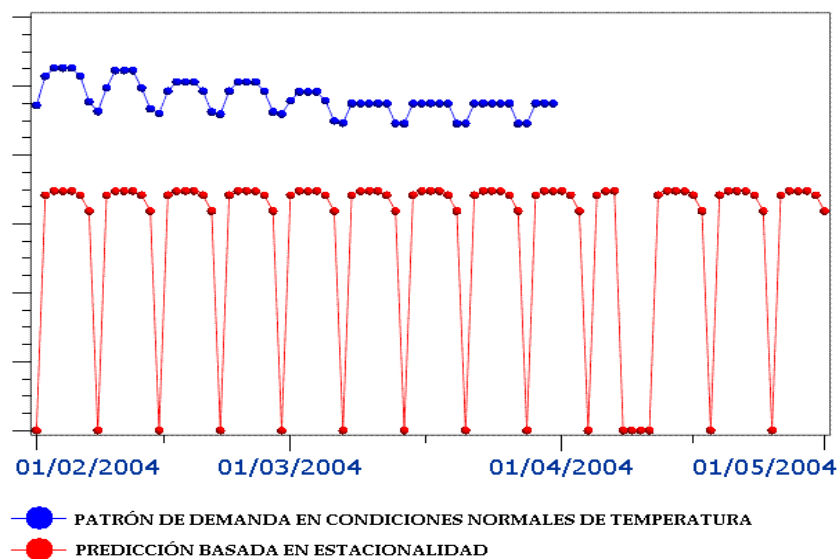


Ilustración 7.15: Predicción de la demanda de gas en condiciones normales mediante un modelo que atiende exclusivamente a la estacionalidad de la serie

Debe observarse que todavía no estamos reflejando en el modelo la información asociada al nivel en el que debe moverse la serie. En su ausencia, el patrón estacional semanal se repite de acuerdo a su comportamiento más reciente con independencia del mes o estación del año que vaya alcanzando el horizonte de predicción.

7.2.4.2 Identificación de la variabilidad

Por construcción, la variabilidad del patrón de demanda que hemos generado se mantiene constante a lo largo de todas las semanas que pertenecen a un mismo mes de un mismo año. Es decir, las diferencias entre días correspondientes a clases distintas (un *miércoles* frente a un *sábado*, un *lunes* frente a un *domingo*, etc.) dentro de un mes común son idénticas y en particular valen cero cuando van referidas a miembros de una misma clase. Sin embargo esta igualdad no tiene porqué presentarse entre semanas asociadas a meses distintos.

Cuando la varianza de la serie no es homogénea a lo largo del tiempo decimos que la serie es heterocedástica. Es conveniente corregir este fenómeno dado que nos interesa que el término de error resultante del ajuste sea algo impredecible y responda a las tres propiedades que caracterizan a un proceso de ruido blanco: de media cero, varianza constante e incorrelado con su pasado. De no corregir la posible heterocedasticidad de la serie esta propiedad puede conservarse en el proceso de error, dado que éste viene a ser a fin de cuentas lo que queda por explicar de la serie conforme el modelo va siendo ajustado y, en consecuencia, no se cumpliría la segunda de las hipótesis de ruido blanco. Para arreglar problemas de heterocedasticidad asociados a una serie temporal lo habitual es “aplanarla” a través de una transformación. En este contexto gozan de especial popularidad la familia de transformaciones de Box-Cox que responde a la expresión:

$$Y_t^{(\lambda)} = \begin{cases} \frac{X_t^\lambda - 1}{\lambda} & \text{si } \lambda \neq 0 \\ \ln(X_t) & \text{si } \lambda = 0 \end{cases} \quad (7.11)$$

El método consiste en ajustar modelos sencillos (por ejemplo. un $ARMA(p, q)$ con p y q pequeños) para una relación de valores de λ y utilizar un criterio que permita medir la bondad del modelo respecto de esa transformación. Los valores que normalmente se plantean para λ son 0 (logaritmo), 0.5 (raíz cuadrada) y 1 (variable sin transformar). Respecto del criterio de evaluación lo normal es emplear alguno de estos dos:

- **Escoger aquella transformación para la que se obtiene un mínimo AIC.**- este indicador (véase ecuación (2.6)) penaliza a los modelos con mayor número de parámetros a la vez que favorece a aquéllos que proporcionan un mejor error de predicción sobre el histórico. Para modelos estructuralmente iguales donde la única diferencia radica en la transformación aplicada sobre la serie, la penalización respecto del número de parámetros ($p+q$ para un $ARMA(p, q)$) es común, con lo cual elegir el modelo de menor AIC es sinónimo de tomar aquél que minimiza el error cuadrático medio (sobre el histórico).

- **Escoger aquella transformación que maximiza la función de verosimilitud.**- para un mismo modelo ARMA(p,q) se construye la función de verosimilitud $\Phi(\phi_1, \dots, \phi_p, \theta_1, \dots, \theta_q, \lambda)$ que representa la probabilidad de que la serie transformada quede bien representada por los parámetros, seleccionándose aquella transformación para la cual se maximiza dicha probabilidad.

El método de Box-Cox plantea modelos de diferentes órdenes respecto de cada una de las potencias de λ propuestas y termina seleccionando aquella que en media proporciona mejores resultados respecto del criterio de evaluación. En nuestro caso, respecto del primero de los criterios, la transformación recomendada es la raíz cuadrada según se muestra en la siguiente tabla:

LAMBDA	AIC
1.0	16349.74
0.5	15920.59
0.0	16328.94

Tabla 7.10: Evaluación del AIC para varias propuestas de transformación de la variable demanda

El hecho de aplicar la transformación a los datos, amortigua los cambios de variabilidad entre meses sobre el histórico y es conveniente desde el punto de vista de las propiedades asociadas al proceso de error resultante. Pero además estamos interesados en que esta volatilidad cambiante esté presente en cada uno de los 24 meses para los que nos proponemos hacer la predicción pues obviamente ésta debe ser fiel reflejo del comportamiento histórico o real de la serie. A falta de ajustar el nivel en el que se mueve la serie a través de la variable *Ciclo* cuya construcción detallaremos en el apartado siguiente, observaremos que, si la predicción se basara únicamente en la componente estacional descrita por las variables que hasta ahora hemos incluido, la variabilidad se mantendría constante a lo largo de todos los meses del horizonte (véase el gráfico de la izquierda en la ilustración (7.16)) dado que no hay nada que informe al modelo de que esta variabilidad debe cambiar. Para aportar este conocimiento vamos a generar 24 variables binarias *MES_TIPO*, donde *MES* puede ser cualquiera de los 12 del año y *TIPO* será *F* (festivo) si el día es de tipo I o II o bien *L* (laborable) si es de tipo III o IV. Si bien podrían generarse tantas variables como cruces entre tipos de meses (12) y tipos de días (4) aparte de que el número de ellas sería excesivo (48) tampoco aportaría nada en cuanto a la forma de recoger la variabilidad asociada a los meses. A fin de cuentas lo que nos interesa es identificar por meses una medida de dispersión de los datos y esto se puede conseguir a través de variables que marquen el día en que se dará el mínimo mensual (con seguridad un festivo) y el máximo (que se presentará un día laborable).

La expresión del modelo que contempla estas variables es la siguiente:

$$X(t) = \phi_{SP} * TipoI(t) + \phi_{LV} * TipoII(t) + \phi_{MXJ} * TipoIII(t) + \sum_{MES=1}^{MES=12} \sum_{TIPO \in \{F, L\}} \omega_{MES, TIPO} * I_{MES, TIPO}(t) + \epsilon(t) \quad (7.12)$$

A continuación se presenta el resultado del ajuste. La variable a predecir es la raíz cuadrada de la demanda en condiciones normales de temperatura (*RCONSUMO_MEDIO*). En cuanto a los nuevos “inputs” considerados (variables explicativas *MES_TIPO*) señalaremos que ha sido omitida una de ellas (*DICIEMBRE_L* = laborables de Diciembre) dado que se puede expresar como combinación lineal de las restantes (deben evitarse efectos multicolineales). Se puede apreciar la significatividad de todas ellas y la mejora considerable del valor del R^2 .

Dependent Variable		RCONSUMO_MEDIO		
Ordinary Least Squares Estimates				
SSE	1602.74032	DFE	2531	
MSE	0.63324	Root MSE	0.79577	
SBC	6266.03758	AIC	6114.02624	
Regress R-Square	0.9870	Total R-Square	0.9870	
NOTE: No intercept term is used. R-squares are redefined.				
Variable	Estimate	Standard Error	t Value	Approx Pr > t
DIA_MXJ	9.3372	0.0705	132.47	<.0001
DIA_SP	0.3160	0.0562	5.63	<.0001
DIA_LV	9.2397	0.0733	126.10	<.0001
ENERO_L	0.2369	0.0965	2.45	0.0142
FEBRERO_L	-0.4056	0.0962	-4.22	<.0001
MARZO_L	-1.4096	0.0944	-14.94	<.0001
ABRIL_L	-2.6573	0.0964	-27.58	<.0001
MAYO_L	-3.5300	0.0952	-37.07	<.0001
JUNIO_L	-3.9534	0.0949	-41.64	<.0001
JULIO_L	-4.2259	0.0938	-45.06	<.0001
AGOSTO_L	-4.4835	0.0955	-46.92	<.0001
SEPTIEMBRE_L	-3.9626	0.0948	-41.81	<.0001
OCTUBRE_L	-2.9829	0.0948	-31.47	<.0001
NOVIEMBRE_L	-0.6972	0.0960	-7.26	<.0001
ENERO_F	8.3910	0.0940	89.31	<.0001
FEBRERO_F	8.0777	0.1090	74.14	<.0001
MARZO_F	7.1626	0.1029	69.61	<.0001
ABRIL_F	5.8288	0.0972	59.96	<.0001
MAYO_F	4.9368	0.0989	49.89	<.0001
JUNIO_F	4.4346	0.1056	42.01	<.0001
JULIO_F	4.1191	0.1065	38.68	<.0001
AGOSTO_F	3.8385	0.0975	39.39	<.0001
SEPTIEMBRE_F	4.3390	0.1065	40.74	<.0001
OCTUBRE_F	5.2218	0.1008	51.78	<.0001
NOVIEMBRE_F	7.4692	0.1001	74.65	<.0001
DICIEMBRE_F	8.2636	0.0905	91.35	<.0001

Tabla 7.11: Modelo para la predicción a medio plazo de la demanda de gas que identifica el efecto estacional y la variabilidad de la serie

Con la inclusión de estas 23 de estas variables en el modelo se consigue un doble objetivo:

- Por un lado se recoge el comportamiento cíclico anual de la serie, que no es realmente el propósito de estas variables, ya que no queremos que informen del “orden de magnitud” en el que la predicción debe moverse

sino sólo de su “forma”. Además, el carácter mensual de las variables (una por mes) da un aspecto escalonado y poco verosímil a la predicción. Veremos cómo la inclusión de la variable *Ciclo* permite recoger no sólo el patrón periódico de los datos sino extrapolarlo de manera suave (no a saltos) a un nivel acorde con aquél en el que debería oscilar la serie.

- El segundo, y al que sí está orientado la inclusión de estas variables, es que la variabilidad de la predicción sea diferente en función del mes del año.

En los siguientes gráficos se puede comparar el resultado de incluir o no estas variables en el modelo. Nuevamente en color azul se presenta el patrón de demanda construido a partir del histórico y en rojo, la predicción que a partir de él se obtendría. En el de la izquierda se refleja una predicción basada sólo en la estacionalidad de la serie transformada a través de la raíz cuadrada. En el de la derecha se contempla el efecto adicional de las variables *MES_TIPO*.

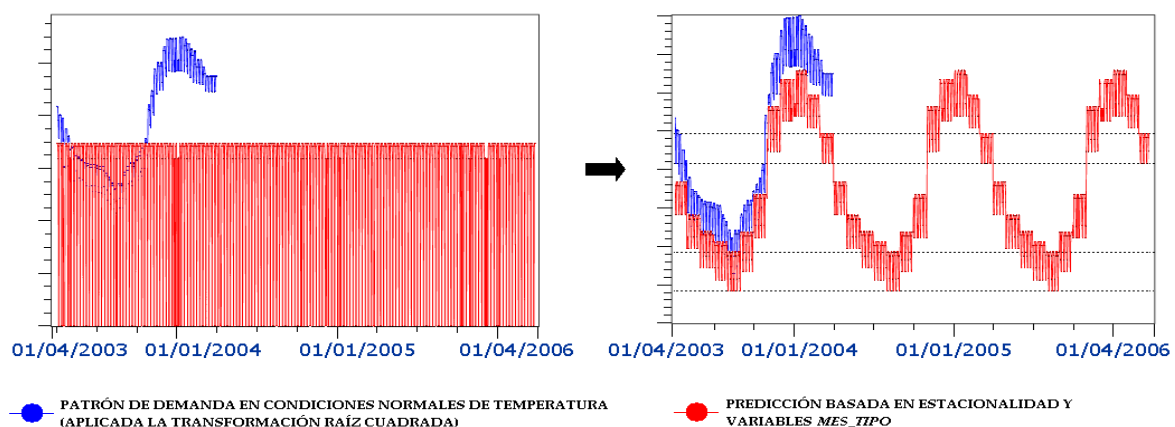


Ilustración 7.16: Predicción de la demanda de gas en condiciones normales mediante un modelo que atiende a la estacionalidad y variabilidad de la serie

Aun cuando lo que más llama la atención es la diferencia en lo que respecta el aspecto periódico de una y otra predicción, debemos insistir en el que la predicción del gráfico de la derecha está desubicada y presenta bruscos cambios de nivel, no es suave. Sin embargo, no olvidemos que el objetivo que perseguíamos con la inclusión de las variables *MES_TIPO* es una corrección de variabilidad. En ese sentido, el gráfico de la derecha sí refleja la distinta variabilidad para la predicción en función del mes del año, la cual se mantiene estática en el de la izquierda. Así, se puede ver por ejemplo que un mes de Abril (acotado por las referencias superiores del eje vertical) tiene un rango de variación que son 4/5 partes del correspondiente a un mes de Agosto (acotado por las referencias inferiores). Queremos reincidir también en el aspecto escalonado de la predicción, acentuado como hemos dicho en los meses que sirven de transición del invierno al verano (Marzo/Abril), o en otras palabras, del frío al calor, y viceversa (Octubre/Noviembre). Son meses que separan dos niveles de demanda claramente diferenciados: uno asociado al invierno gasista, que

comienza con el arranque de las calefacciones en la primera semana de Noviembre y termina con el comienzo de la primavera en la última semana de Marzo y otro asociado al resto del año que, salvo excepciones puntuales (olas de frío en Abril u Octubre y de manera excepcional en Mayo o Septiembre) es independiente de la climatología. Es ésta la razón que justifica el hecho de que la magnitud del salto sea mayor durante estas transiciones y también que únicamente hayamos tenido en cuenta el periodo Noviembre-Marzo para representar los pares de puntos que permitirán analizar la relación demanda-temperatura (véase ilustración (7.14)).

A raíz de esta observación, podremos encontrar problemas si planteamos la predicción desde cualquiera de estos dos meses. La presencia del escalón no hace fácil conseguir suavidad entre los datos reales de Marzo y las predicciones realizadas para Abril. Como veremos en el apartado siguiente, si se marca por ejemplo el día 31 de Marzo de 2006 como horizonte final, puede ser recomendable anticiparse un mes (desde el 1 de Marzo) o esperarse al siguiente (desde el 1 de Mayo) para realizar las predicciones diarias a dos años (25 o 23 meses respectivamente).

7.2.4.3 Identificación del ciclo anual de la serie

Ahora que disponemos de una predicción que presenta las dos estacionalidades más características de una serie de demanda, la semanal, generada por el efecto del calendario y cuya oscilación se amplifica o disminuye en función del mes del año, y la anual, que si bien presenta un aspecto algo abrupto, da respuesta a la influencia básica de las temperaturas (frío o calor) en las distintas estaciones (invierno o verano), el último paso consiste en nivelar dicha predicción, esto es, determinar el nivel en el que ésta deberá moverse.

Mediante técnicas de suavizado de curvas vamos a ser capaces de identificar un comportamiento básico anual subyacente bajo el histórico construido, el cual, de acuerdo a la terminología empleada en la **teoría de la señal**, vendrá

dado por una curva de baja frecuencia o equivalentemente largo periodo ($\omega = frecuencia = \frac{1}{periodo} = \frac{1}{T}$).

Esta curva, que pasaremos a denominar *Ciclo*, además de determinar el nivel medio alrededor del cual se distribuyen los datos, nos va a permitir recoger la periodicidad anual histórica de la serie de una manera suave, sin cambios bruscos de comportamiento. Si conseguimos que su extrapolación mantenga también esta suavidad, la variable *Ciclo* podrá participar en la expresión (7.12) como una variable explicativa más, informando al modelo de la base sobre la que deberán actuar las variables que reflejan la estacionalidad semanal y la variabilidad cambiante por meses. Para ser exactos, dado que la variable a predecir en el modelo es la raíz cuadrada de la demanda esperada en condiciones normales de temperatura, será también la raíz cuadrada de la variable *Ciclo* (*Rciclo*) la que finalmente sea incluida en el modelo. Con ella, el modelo quedará planteado de la siguiente manera:

$$X(t) = \phi_{SP} * TipoI(t) + \phi_{LI} * TipoII(t) + \phi_{MXJ} * TipoIII(t) + \sum_{MES=1}^{MES=12} \sum_{TIPO \in \{F, L\}} \omega_{MES, TIPO} * I_{MES, TIPO}(t) + \omega_{CICLO} * \sqrt{Ciclo(t)} + \epsilon(t) \quad (7.13)$$

El conocimiento de todos los términos que participan en la expresión anterior (calendario y *Ciclo*) permitirá que ésta sea utilizada para predecir el valor de la demanda en condiciones normales de temperatura.

Las dos metodologías que vamos a ver para extraer este patrón básico se fundamentan en conseguir expresar el histórico

de demanda construido en condiciones normales de temperatura, como una combinación lineal de los miembros de una base ortonormal compuesta por funciones suavizadoras, siendo dos las alternativas que para ello planteamos: las wavelets y los splines. A través de unas y otras se puede conseguir una aproximación suficientemente suave de la serie que refleje el comportamiento periódico anual de los datos y que pasaremos a definir como la variable *Ciclo* que participará en el modelo (7.13). La extrapolación de dicha variable puede ser llevada a cabo de dos maneras diferentes con independencia de la función suavizadora que a tal fin se utilice:

- Una primera forma de llevar a cabo la extrapolación de la componente *Ciclo* ajustada al histórico, es a través de un modelo sencillo que refleje exclusivamente la estacionalidad y tendencia de esta variable. Previamente a la aplicación de este modelo será preciso, obviamente, detectar el comportamiento de baja frecuencia. Para ello emplearemos una de las clases de funciones suavizadoras, por ejemplo, las wavelets. De ellas, así como de sus posibilidades para descomponer una señal como suma de procesos que se mueven a distintas frecuencias, se ofrece una breve descripción en el Anexo X.
- Otra forma alternativa de predecir el *Ciclo* es mediante el uso de las propias funciones suavizadoras como variables independientes de un modelo de regresión convenientemente extrapoladas. En esta ocasión será la estimación realizada por el propio modelo la que se utilice como variable *Ciclo* tanto a pasado (sobre el histórico) como a futuro (para el horizonte de dos años al que se plantea el estudio). Es decir, no es necesario dar un primer paso en el que se extraiga la componente de baja frecuencia y un segundo en el que se lleve a cabo su predicción a partir de otra ecuación diferente, sino que la detección del *Ciclo* y su extrapolación tienen lugar en una misma etapa. Según hemos dicho valdrían tanto las funciones wavelets como los splines como regresores del modelo que realiza la estimación y predicción del *Ciclo*. En este caso nos decantaremos por el uso de los splines, de cuya teoría se ofrece otra breve descripción en el Anexo XI.

7.2.4.3.1 Generación de la variable *Ciclo* a partir de wavelets

Las funciones wavelets van a permitir expresar la serie de demanda como suma de otras que miden sus variaciones a diferentes escalas o niveles de resolución: día, semana, año, etc. Este proceso de descomposición recibe el nombre de **análisis multirresolución de la señal** o **MRA** (véase Anexo X), siendo necesaria la especificación de la familia o base de funciones wavelets como combinación lineal de las cuales vamos a expresar nuestra serie. Serán las propiedades de esta última las que condicionen la selección de una familia u otra.

La primera de las propiedades a contemplar es la **continuidad** de la componente *Ciclo*. Esta propiedad lleva a descartar de entrada la wavelet de **Haar** dada la forma escalonada (a saltos) que caracteriza a las componentes que resultarían de realizar el análisis con este tipo de wavelets. En su lugar deben seleccionarse familias más suaves, viniendo determinado el grado de suavidad por el número de derivadas que estas funciones admiten.

En segundo lugar, de acuerdo a las características **periódicas** del estudio, una buena elección podría ser la clase

symmlet. Dentro de esta familia es necesario además especificar la amplitud del soporte de la wavelet (4, 6, 8,...,20,...) que viene a determinar el grado de suavidad del ajuste. Dado que con la curva *Ciclo* no tratamos de recoger características locales del histórico a un nivel de detalle máximo, nos van a interesar wavelets amplias puesto que la suavidad de estas funciones crecen de acuerdo a la amplitud de su soporte.

A continuación, se detalla el MRA llevado a cabo a través de *symmlets* de amplitud 6, 8, 10 y 12, es decir, de miembros de las clases *symmlet6*, *symmlet8*, *symmlet10* y *symmlet12* respectivamente.

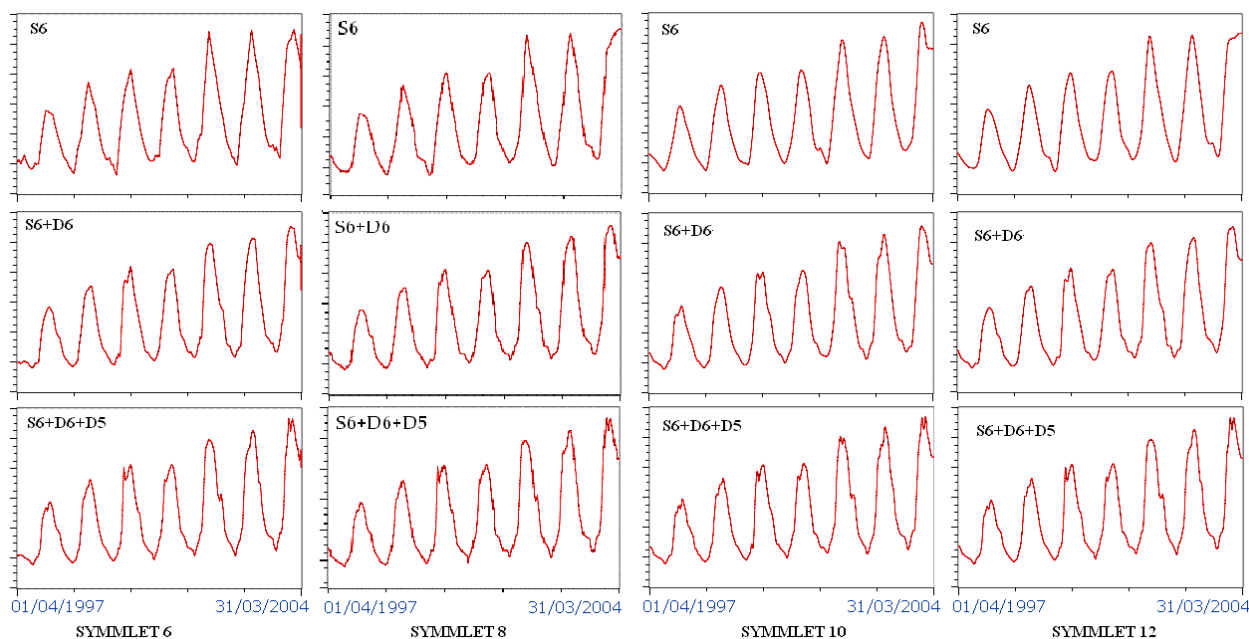


Ilustración 7.17: MRA del histórico de demanda en condiciones normales de temperatura llevado a cabo mediante *symmlets* de distintas amplitudes

A través de cada tipo de wavelet la serie original ha sido descompuesta en un total de 7 señales o niveles, cantidad que puede ser modificada pero que ha sido suficiente para la detección de la componente *Ciclo* que buscamos. Se presenta en cada caso la componente más suave (S_6) y el agregado que resulta de sumar ésta con una (D_6) o dos (D_6 y D_5) de las componentes siguientes de más baja frecuencia. La principal desventaja de la *symmlet* de amplitud 6 es el claro problema de ajuste que presenta en la frontera desde la cual se plantearía la predicción (obsérvese la irregularidad que se presenta en las fechas del histórico más cercanas al 31 de Marzo de 2004). Por otra parte las componentes obtenidas como resultado de los análisis asociados a *symmlets* de amplitudes grandes (10 y 12) no presentan grandes diferencias que hagan esperar mejoras sustanciales respecto de los resultados que se obtienen para el MRA llevado a cabo con una *symmlet8*. Es por ello que ha sido esta familia de wavelets la que finalmente se ha utilizado para descomponer nuestra serie, siendo el resultado el siguiente:

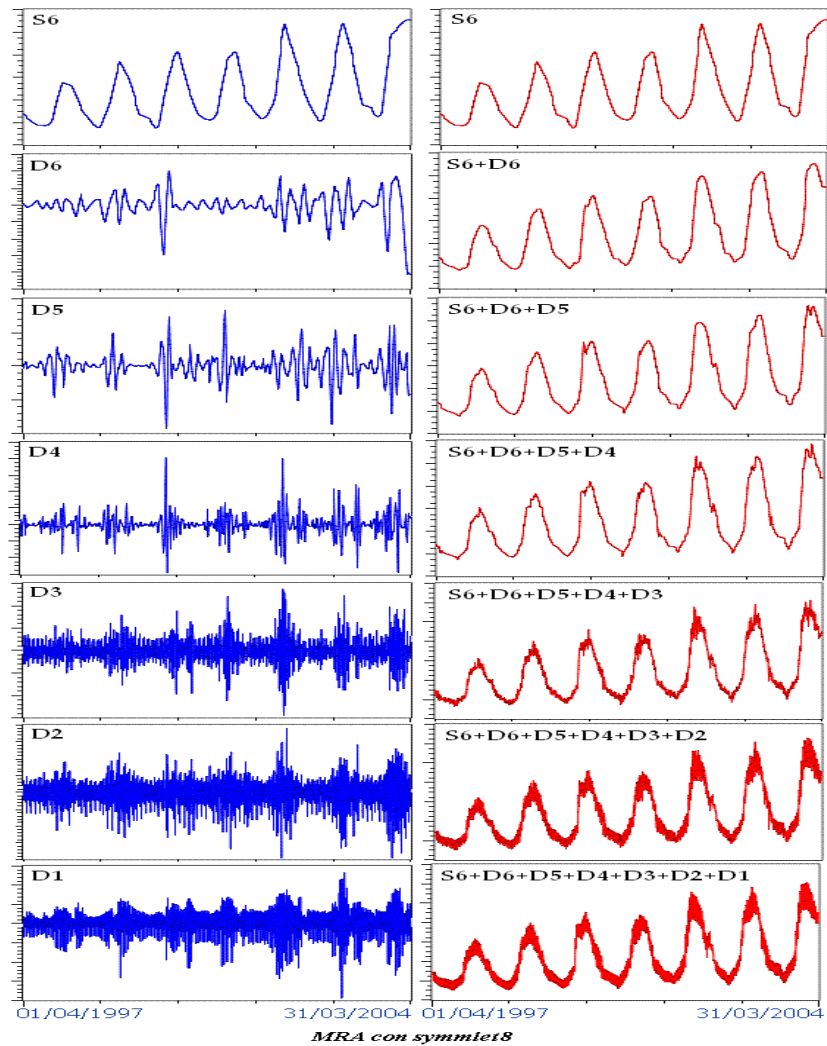


Ilustración 7.18: MRA a 7 niveles de resolución llevado a cabo con symmlets8 sobre el histórico de demanda en condiciones normales de temperatura

En el gráfico anterior se presentan en la columna de la izquierda cada una de las componentes resultantes del análisis multirresolución en orden creciente de frecuencias (desde S_6 hasta D_1). La señal más suave S_6 , caracteriza el comportamiento cíclico anual de la serie y las seis restantes, D_6 - D_1 recogen oscilaciones asociadas a periodos cada vez más cortos o frecuentes, es decir, correcciones cada vez más detalladas a la curva suave establecida por S_6 . La columna de la derecha muestra el resultado de ir sumando todas las señales de frecuencia más baja a una dada. En el caso extremo se obtiene la serie de partida como suma de todas ellas.

De todas las señales acumuladas sólo las más suaves van a servir a nuestros propósitos. Recordemos que nuestro objetivo es incluir una variable en la ecuación (7.12) que convenientemente extrapolada permita identificar el nivel en el que se moverá la serie en el futuro y sobre el que se va a incorporar la estacionalidad y variabilidad cambiante por meses recogida por los otros regresores de la ecuación. Es ésta la razón por la que la variable *Ciclo* debe ser suave, para limitarse a recoger la periodicidad anual y no otros efectos redundantes que ya puedan estar aportando otros miembros de la fórmula.

A la vista del gráfico anterior, se propone utilizar como *Ciclo* la componente S_6 o a lo sumo la adición de S_6 y D_6 ya que los acumulados posteriores empiezan a presentar irregularidades que no son comunes a todos los años. Por ejemplo, la adición de la componente D_5 acentúa demasiado un “extraño” abombamiento en Abril de 2002 (que empieza a contemplarse a raíz de la inclusión de D_6), y aporta pequeñas caídas asociadas a los periodos navideños de los años 1999-2000 y 2003-2004. Con la incorporación de la señal de detalle D_4 se incrementan aún más estas deformaciones y ya con la señal D_3 la curva se vuelve demasiado irregular para plantearse su posible extrapolación con fines predictivos.

En cualquier caso, para extender el *Ciclo* generado (S_6 o $S_6 + D_6$) a un horizonte de dos años y que así pueda ser utilizado por (7.13) como variable explicativa, vamos a ajustar a dicha variable el siguiente modelo:

$$\nabla^{365}Ciclo_t = (1 - B^{365}) * Ciclo_t = \mu + \epsilon_t \Rightarrow Ciclo_t = Ciclo_{t-365} + \mu + \epsilon_t \quad (7.14)$$

En éste, el operador diferencia ∇ permite controlar la periodicidad anual de la curva y μ es el único parámetro a estimar el cual controla el nivel en el que se mueve la serie. Una vez que dispongamos de la variable *Ciclo* ajustada al histórico de demanda y de su comportamiento futuro para los dos próximos años dado por (7.14), la ecuación (7.13) podrá ser utilizada para realizar predicciones para los dos próximos años en condiciones normales de temperatura.

En los gráficos que a continuación se presentan se marca con una referencia vertical el último día del conjunto de entrenamiento (31 de Marzo de 2004) y desde el cual se lanza la predicción a dos años. Se presentan los dos últimos años del histórico construido en condiciones normales de temperatura (en azul). Éste es utilizado para la construcción de la variable *Ciclo* (en rojo) que puede ser S_6 o bien la suma de ésta con una o varias señales de detalle (D_i). Dicho *Ciclo* es extrapolado a futuro (en rosa) mediante la ecuación (7.14). Finalmente, el *Ciclo* y su predicción proporcionada por (7.14) generan a través de (7.13) una predicción definitiva que se supone también en condiciones normales de temperatura (en celeste). El primero de los gráficos utiliza como variable *Ciclo* la componente S_6 . Se han marcado con un círculo los principales problemas que aporta esta predicción. Llama fundamentalmente la atención las dos grandes discontinuidades asociadas a las transiciones entre los meses de Marzo y Abril. Esto es consecuencia de que la curva S_6 no consigue recoger bien la tendencia descendente del patrón en dicho periodo. También se ha marcado el apuntamiento invernal que recoge la curva en 2003 y que no se repite al año siguiente. Estas observaciones no parecen aconsejar el uso de esta curva como regresor en la ecuación (7.13).

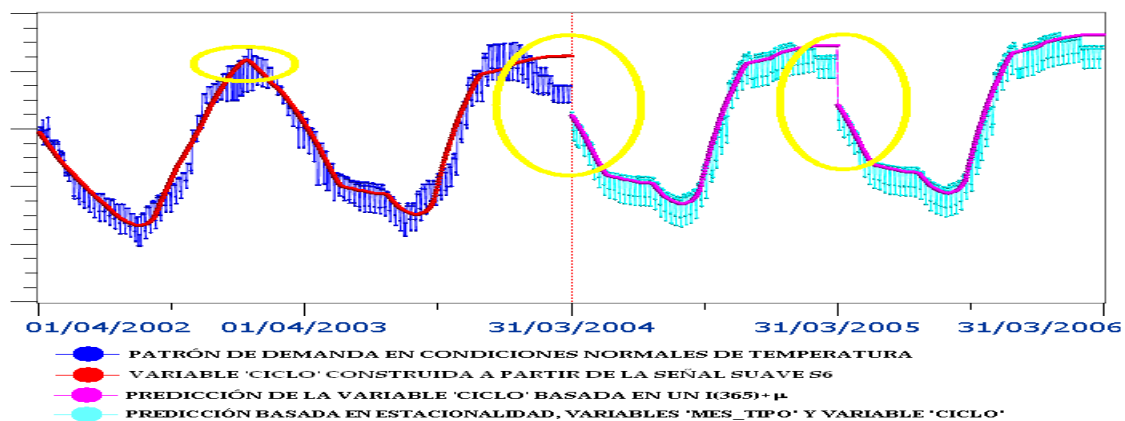


Ilustración 7.19: Predicción de la demanda de gas en condiciones normales mediante un modelo 1 que atiende a la estacionalidad semanal, variabilidad, tendencia y periodicidad anual de la serie identificada mediante wavelets

El siguiente gráfico utiliza como variable *Ciclo* la suma de S_6 y D_6 . Si bien la longitud del salto entre el patrón real y el predicho es menor y la curva se ajusta mejor al histórico, sigue siendo notoria la falta de continuidad. Se empieza a sospechar las dificultades que pueden surgir para conseguir suavidad en dicho periodo dado el brusco escalón que, aun a pesar del suavizado del histórico, se presenta entre dicho meses según señalábamos en el gráfico (7.16).

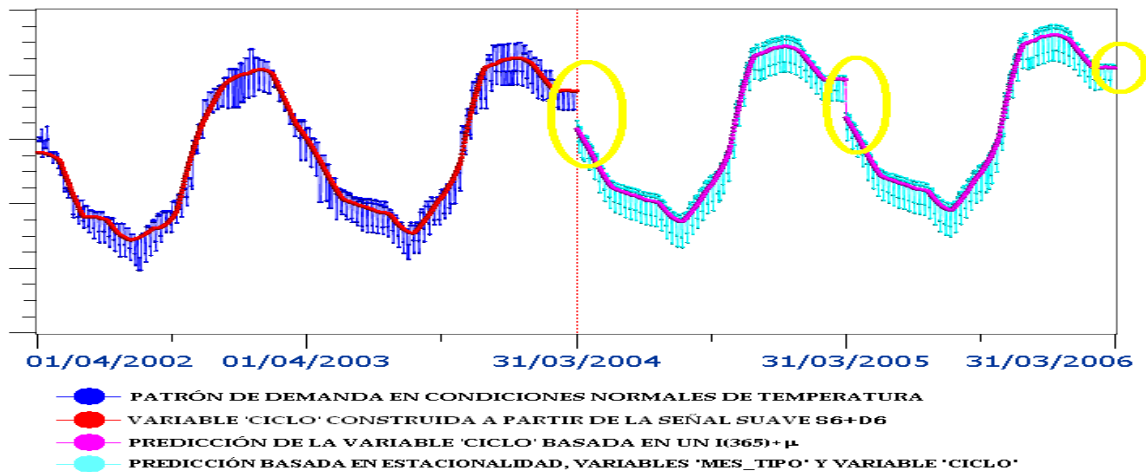


Ilustración 7.20: Predicción de la demanda de gas en condiciones normales mediante un modelo 2 que atiende a la estacionalidad semanal, variabilidad, tendencia y periodicidad anual de la serie identificada mediante wavelets

Una última tentativa con la variable que resulta de sumar las componentes S_6 , D_6 y D_5 consigue que la discontinuidad Marzo-Abril sea aún menor. Sin embargo comienzan a aparecer ciertas deformaciones que, si bien pueden estar justificadas por la influencia de la caída de la demanda durante el periodo de Navidad, anticipan para mayores niveles de acumulación, un grado de irregularidad poco deseable pensando en su extrapolación a futuro.

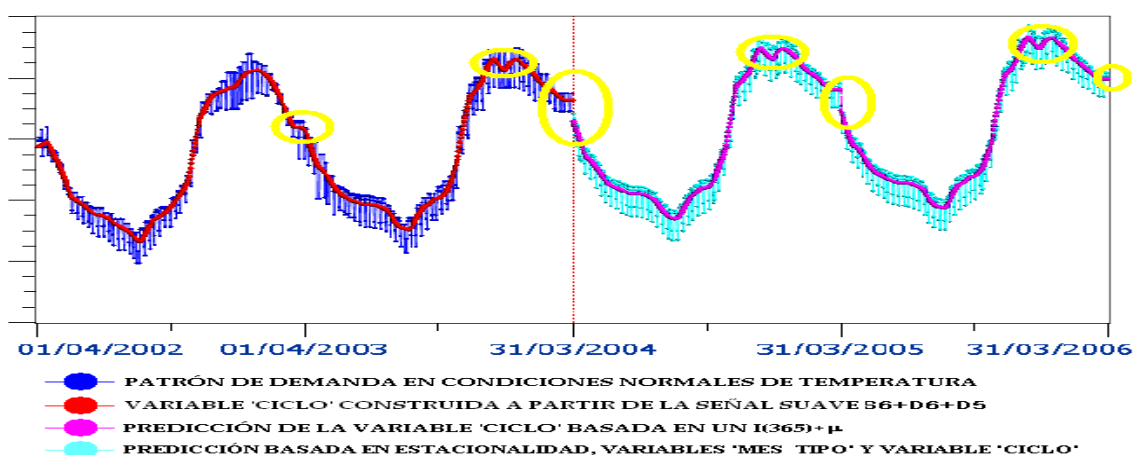


Ilustración 7.21: Predicción de la demanda de gas en condiciones normales mediante un modelo 3 que atiende a la estacionalidad semanal, variabilidad, tendencia y periodicidad anual de la serie identificada mediante wavelets

Dadas las dificultades que surgen para ajustar completamente el solape entre dato real y predicho y que la causa parece venir inducida por el cambio brusco que se produce del invierno a la primavera, una solución alternativa es ajustar el *Ciclo* hasta un periodo de transición menos conflictivo como es Febrero-Marzo y plantear su extrapolación mediante (7.14) desde ese instante. La curva extrapolada será finalmente utilizada para realizar la predicción final del patrón en condiciones normales de temperatura a través de (7.13). El nuevo resultado hace imperceptible la diferenciación entre pasado y futuro y además no presenta irregularidades difíciles de justificar. A lo único que estaríamos obligados es a hacer la predicción a 25 meses (en vez de 24) pero esto no parece ser un grave problema dado el largo horizonte que de entrada ya nos venimos planteando.

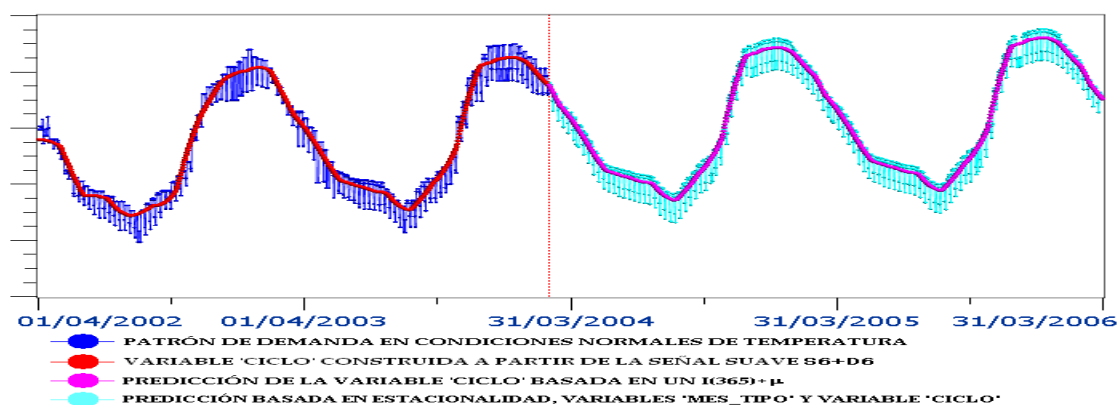


Ilustración 7.22: Predicción de la demanda de gas en condiciones normales mediante un modelo 4 que atiende a la estacionalidad semanal, variabilidad, tendencia y periodicidad anual de la serie identificada mediante wavelets

Así, esta sería una posible propuesta de predicción diaria en condiciones normales de temperatura para los dos próximos años utilizando el método de descomposición basado en wavelets. Veremos en el apartado 7.2.4.3.3 una corrección del mismo motivada por la necesidad de alcanzar mejores predicciones fuera de muestra.

7.2.4.3.2 Generación de la variable Ciclo a partir de splines de regresión

De igual manera que hemos expresado el histórico de demanda en condiciones normales de temperatura como una combinación de funciones wavelets a través de un análisis multirresolución (o MRA), vamos a expresar ese mismo histórico en función de otra clase de funciones suavizadoras denominadas splines mediante un modelo de regresión lineal. La estimación de los parámetros asociados a cada una de estas funciones permite la expresión de la demanda como combinación lineal de ellas y, a través de la ecuación resultante, predecir su comportamiento futuro. A tal fin utilizaremos **splines cúbicos**, que son polinomios de grado 3 y de clase C^1 . Con el fin de que estas funciones no tengan una finalidad puramente suavizadora sino que sirvan además para reflejar la estacionalidad anual del histórico (pues buscamos un patrón cíclico), vamos a considerar un subespacio del espacio de los splines en el que esta periodicidad quede puesta de manifiesto (véase Anexo XI).

$$\text{Sea } S_p = \{ s \in C^1 \text{ t.q. } s(t) = s(t+365), s|_{[t_{i-1}+365*k, t_{i+1}+365*k]} \in \mathbb{R}_3[t] \quad \forall i \in \{1, 2, \dots, n\}, k \in \mathbb{N} \} \quad (7.15)$$

El subespacio S_p define el conjunto de las funciones de clase C^1 (continuas y diferenciables) de periodo 365 (días) y definidas como polinomios cúbicos en la variable “ t ” sobre intervalos centrados en un instante de tiempo t_i dados por $[t_{i-1}+365*k, t_{i+1}+365*k]$

Los instantes $t_0, t_1, t_2, \dots, t_n$ determinan el número de trozos en que se divide el periodo de un año. Cuanto mayor sea este número, menor será la longitud de cada uno de los intervalos y por tanto la función se ajustará de forma local a un conjunto de puntos cada vez más reducido, proporcionando una curva sobreajustada que no resultaría conveniente para ser extrapolada. Por el carácter mensual del que estamos dotando al estudio, una primera alternativa sería por ejemplo dividir cada año en un total de 12 trozos (uno por mes), cantidad que en principio, no parece antojarse demasiado grande. Para expresar la demanda como combinación lineal de elementos de S_p bastaría considerar una base de dicho espacio dada por una relación de funciones ϕ, ψ 's (véase Anexo XI) que actuarán como variables independientes de un modelo de regresión (de ahí su denominación de splines de regresión) en el que la variable de respuesta es la demanda en condiciones normales de temperatura:

$$X(t) = \sum_{MES=1}^{MES=12} a_{0,MES} * \phi_{MES}(t) + \sum_{MES=1}^{MES=12} b_{0,MES} * \psi_{MES}(t) + \epsilon(t) \quad (7.16)$$

Obsérvese que resulta fácil utilizar los propios splines para predecir a través de esta ecuación dado que su valor a futuro es el mismo que a pasado, es decir, el comportamiento de los splines para los dos próximos años es conocido y por ello no es preciso predecirlo. La estimación proporcionada por este modelo, $\hat{X}(t)$, nos dará tanto el *Ciclo* anual de comportamiento sobre el histórico como el valor esperado en el futuro para el mismo. Para su obtención basta aplicar el método de mínimos cuadrados y realizar a través de él la estimación de los coeficientes de (7.16). El siguiente gráfico muestra el resultado del ajuste (en rojo) que en sí constituye un suavizado del histórico (en azul).

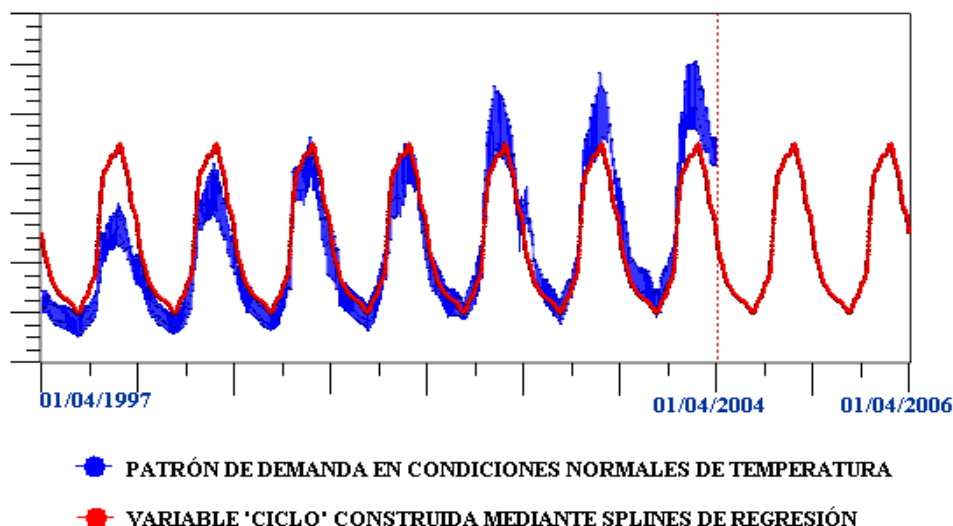


Ilustración 7.23: Suavizado del histórico de demanda en condiciones normales de temperatura mediante splines cúbicos de regresión

En esta ocasión, de forma opuesta al método de ajuste anterior en el que se utilizaban las wavelets como funciones suavizadoras, no se utiliza la expresión (7.16) para detectar el *Ciclo* sobre los datos históricos (como hacíamos a través del MRA) y otro modelo distinto (tipo (7.14)) para llevar a cabo la extrapolación al futuro del mismo, sino que es el propio modelo el que detecta el comportamiento de baja frecuencia y a partir de él plantea directamente la predicción. Un primera apreciación a la vista del gráfico anterior es que el modelo ajustado no está consiguiendo capturar el incremento de consumo que va teniendo lugar de un año a otro. Sin embargo solventar esta carencia resulta sencillo dado que basta incluir en la ecuación anterior una variable que modelice tendencia (una recta dada por ejemplo por la propia variable “*t*”, que determina la fecha a la que va referida el dato histórico) para que dicho efecto incremental sea tenido en cuenta.

$$X(t) = \omega * t + \sum_{MES=1}^{MES=12} a_{0,MES} * \phi_{MES}(t) + \sum_{MES=1}^{MES=12} b_{0,MES} * \psi_{MES}(t) + \epsilon(t) \quad (7.17)$$

A continuación se muestra el resultado asociado a este nuevo modelo:

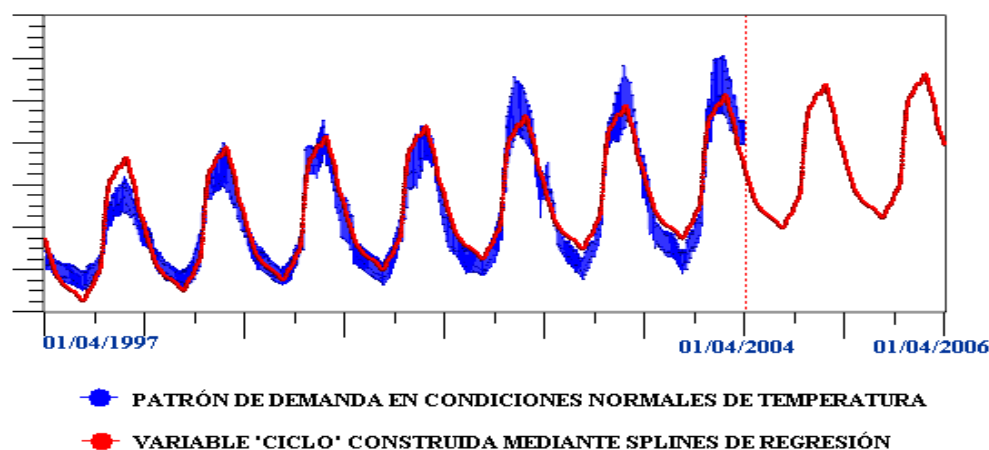


Ilustración 7.24: Suavizado del histórico de demanda en condiciones normales de temperatura mediante splines cúbicos de regresión teniendo en cuenta la tendencia del histórico

Aunque la mejora es sustancial con respecto al modelo anterior, la curva parece seguir sin adaptarse bien al histórico de datos sobre todo en lo que se refiere a las cotas anuales presentadas en el periodo invernal y durante los meses puramente vacacionales como son Julio y, sobre todo, Agosto. Obviamente, dado que nuestro estudio está orientado a la estimación del valor máximo diario que se puede esperar para la demanda durante los dos próximos años, debemos hacer todo lo posible por conseguir mejorar el resultado de este ajuste.

La consideración de otros 24 regresores adicionales que sean el producto de cada uno de los ya incluidos por una variable que refleje tendencia (nuevamente la propia variable “*t*”) permitirá que la variabilidad de los splines cambie con el tiempo, consiguiendo además, dado que cada uno de los splines va referido a un mes concreto, reflejar que el crecimiento de la demanda de un mes a otro no sea igual para todos los meses del año. La inclusión de estas variables permitirá la salida de la variable “*t*” de la ecuación (7.17) puesto que la información que esta última aporta viene ya recogida por ellas.

Así, de acuerdo a los nuevos regresores, el modelo quedaría planteado de la siguiente manera:

$$X(t) = \sum_{MES=1}^{12} a_{0,MES} * \phi_{MES}(t) + \sum_{MES=1}^{12} b_{0,MES} * \psi_{MES}(t) + \sum_{MES=1}^{12} a_{1,MES} * \phi_{MES}(t) * t + \sum_{MES=1}^{12} b_{1,MES} * \psi_{MES}(t) * t + \epsilon(t) \quad (7.18)$$

Conforme a esta expresión, el resultado del ajuste y de la consiguiente extrapolación sería la que ahora presentamos:

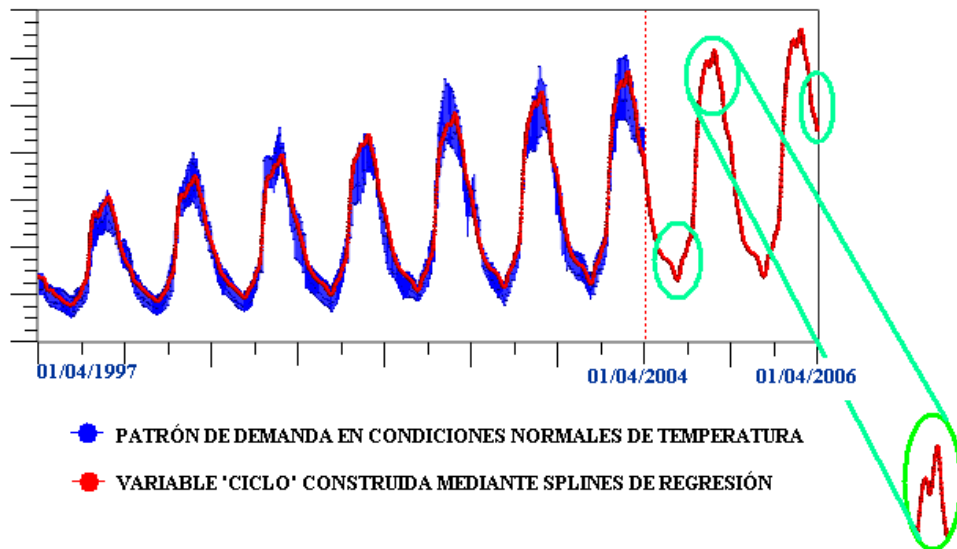


Ilustración 7.25: Suavizado del histórico de demanda en condiciones normales de temperatura mediante splines cúbicos de regresión teniendo en cuenta la tendencia y variabilidad del histórico

Visualmente el ajuste resulta mucho más satisfactorio pues la curva parece moverse en sintonía con el histórico. Una última apreciación que podríamos hacer se refiere al grado de suavidad del ajuste. Aun cuando hemos considerado lógico realizar la partición anual en periodos de 12 meses, tal vez no resulte conveniente extrapolar ciertas irregularidades observadas en el gráfico anterior como son los agudos picos observados durante los meses invernales y veraniegos derivados de las caídas de demanda producidas durante los periodos vacacionales de Navidad y Agosto o los sutiles abombamientos que se presentan en los meses de Abril. Debemos tener en cuenta que, conforme a la definición multiplicativa de los splines que intervienen en (7.18) estos “defectos” crecerán en magnitud de un año a otro.

Cuando llevábamos a cabo el análisis mediante funciones wavelets (MRA), estas anomalías comenzaban a aparecer a partir de la inclusión de la componente D_5 en la definición de la variable *Ciclo* (véanse las ilustraciones (7.18) y (7.21)), bastando omitir la presencia de esta componente en dicha definición para dejarla así planteada únicamente como la adición de S_6 y D_6 . En el caso que nos ocupa la forma de conseguir un mayor grado de suavidad consiste en reducir el número de trozos en la partición realizada sobre cada uno de los años. Las siguientes ecuaciones definen dos nuevos modelos asociados a una partición bimestral y trimestral respectivamente.

$$X(t) = \sum_{BIM=1}^6 a_{0,BIM} * \phi_{BIM}(t) + \sum_{BIM=1}^6 b_{0,BIM} * \psi_{BIM}(t) + \sum_{BIM=1}^6 a_{1,BIM} * \phi_{BIM}(t) * t + \sum_{BIM=1}^6 b_{1,BIM} * \psi_{BIM}(t) * t + \epsilon(t) \quad (7.19)$$

$$X(t) = \sum_{TRIM=1}^4 a_{0,TRIM} * \phi_{TRIM}(t) + \sum_{TRIM=1}^4 b_{0,TRIM} * \psi_{TRIM}(t) + \sum_{TRIM=1}^4 a_{1,TRIM} * \phi_{TRIM}(t) * t + \sum_{TRIM=1}^4 b_{1,TRIM} * \psi_{TRIM}(t) * t + \epsilon(t) \quad (7.20)$$

Los gráficos presentados muestran el resultado de aplicar dichos modelos sobre estas particiones de tamaño 6 y 4.

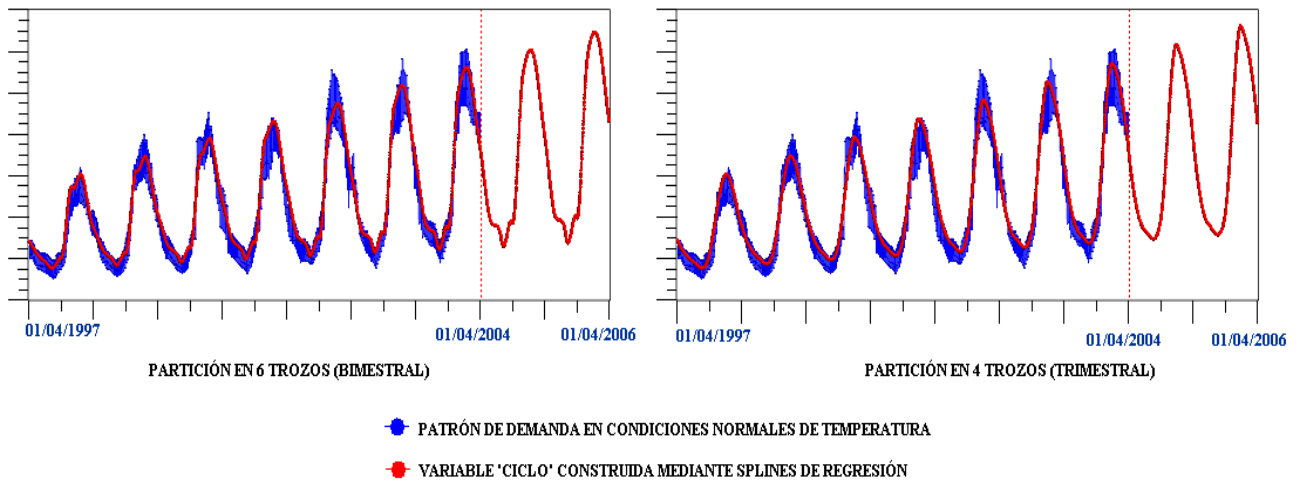


Ilustración 7.26: Suavizados del histórico de demanda en condiciones normales de temperatura mediante splines cúbicos de regresión teniendo en cuenta la tendencia y variabilidad del histórico

En el primero de los gráficos desaparecen por completo dos de las irregularidades a las que hacíamos referencia (vacaciones de Navidad y abultamiento en el mes de Abril) pero es en el segundo cuando se consiguen suavizar por completo las tres (vacaciones de Agosto). De esta manera tendríamos otra candidata diferente a participar en el modelo (7.13) como variable *Ciclo*, a la obtenida a través del ajuste mediante wavelets. Ésta constituirá otro posible soporte sobre el que “montar” las otras componentes reflejadas en dicho modelo, permitiendo en su conjunto, realizar la predicción diaria a dos años vista. Nuevamente se presentan los dos últimos años del histórico construido en condiciones normales de temperatura (en azul), la variable *Ciclo* sobre el histórico (en rojo) y su extrapolación a futuro (en rosa) generadas ambas por el modelo (7.20) y finalmente la predicción (en celeste) realizada por la ecuación (7.13) una vez conocidos todos sus integrantes.

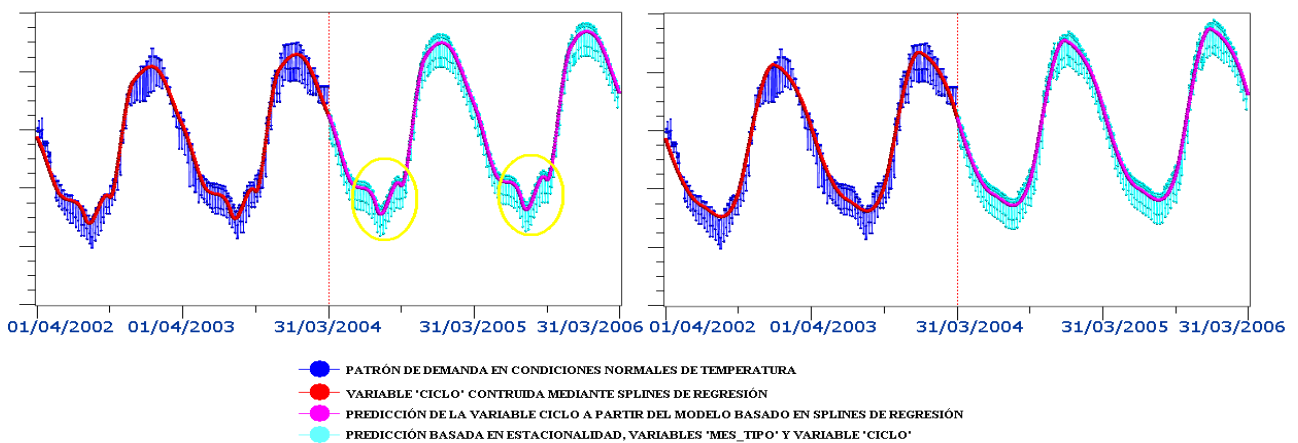


Ilustración 7.27: Predicción de la demanda de gas en condiciones normales mediante un modelo 5 que atiende a la estacionalidad semanal, variabilidad, tendencia y periodicidad anual de la serie identificada mediante splines

7.2.4.3.3 Comparativa entre los métodos planteados para la generación de la variable *Ciclo*

De acuerdo a lo expuesto en los dos apartados anteriores podemos enumerar una serie de aspectos que juegan claramente a favor del ajuste llevado a cabo mediante la segunda de las técnicas propuestas:

- La solución al problema se plantea en una **única etapa**: el ajuste y la extrapolación a futuro se lleva a cabo a partir de una única ecuación (expresión (7.20)). El ajuste mediante wavelets necesita realizar un análisis multirresolución previo (MRA) en función de una familia de wavelets, a determinar, para la obtención de una curva que es posteriormente extrapolada mediante un modelo autorregresivo (expresión (7.14)).
- El **grado de suavizado** de la curva depende de un único parámetro que es la dimensión de la partición anual. En el ajuste mediante wavelets existen demasiados elementos que controlan este factor: las propiedades de la wavelet seleccionada (continua o discreta, simetría, amplitud, etc.), el número de niveles a obtener a través del MRA y la determinación del número de niveles como suma de los cuales se genera la variable *Ciclo*.
- El modelo puede ser **aplicado en cualquier momento del tiempo** para aprovechar al máximo toda la información histórica disponible. El ajuste mediante wavelets puede presentar problemas de ajuste en las fronteras del histórico, es decir, en los primeros y últimos datos del histórico y en consecuencia, no debería realizarse la predicción a partir de estos últimos. De hecho, a diferencia del tratamiento llevado a cabo con splines, con las wavelets hemos tenido la necesidad de realizar la predicción a partir del día 1 de Marzo de 2004 y no desde el 1 de Abril de ese mismo año.

De manera adicional a todas estas ventajas existe un aspecto fundamental a tener en cuenta que es la calidad de las predicciones que han sido obtenidas por uno y otro método. Se ha establecido la siguiente fórmula de error para validar la calidad de las predicciones de uno y otro modelo tanto sobre el histórico como a futuro.

$$Error(t) = 100 * \left| \frac{Demanda_real(t) - Prediccion(t)}{Prediccion(t)} \right| \quad (7.21)$$

Sobre el histórico, parece lógico que el resultado conseguido mediante wavelets sea mejor dado que la curva se puede ajustar de manera particular a cada uno de los años del histórico, mientras que la obtenida mediante splines es una función periódica que responde a la misma forma en todos ellos, con la única diferencia de ver amplificada su variabilidad en el paso de un año a otro. Hemos comparado las predicciones que proporciona el patrón dentro y fuera de muestra con los datos reales, para aquellos días cuya diferencia de temperatura con respecto al patrón estándar no supere los 3°C. Es decir, hemos supuesto como datos normales de demanda (no influenciados por la temperatura) aquéllos que se presentaron en días en los que las temperaturas máximas reales no se desviaron más de 3°C del valor mediano mensual estimado en función del histórico. Hemos considerado como referencia esta diferencia de 3°C por ser

el número aproximado de grados que separa al valor mediano de la temperatura mensual de los valores térmicos asociados al primer y tercer cuartil (véase tabla (7.1)). Es decir, tomaremos como temperaturas normales el 50% de los datos de temperatura más centrados respecto a la mediana (porcentaje comprendido entre Q_1 y Q_3).

De acuerdo a esta consideración, la diferencia obtenida entre el error medio para uno y otro modelo y a un horizonte de 1 día *in-sampling*, está alrededor de un 1.5% a favor del modelo ajustado mediante wavelets (de un 4% frente a un 5.5% para el modelo basado en splines). Sin embargo, para predicciones fuera de muestra (*out-of-sampling*) la balanza se inclina claramente a favor del segundo de los métodos de ajuste. La predicción en condiciones normales de temperatura que genera el modelo que utiliza los wavelets no captura bien la tendencia creciente de los datos pues no se adapta bien a las cotas máximas de los inviernos. Por ello se ha reforzado el efecto tendencia de la curva ajustada mediante wavelets incluyendo un término autorregresivo de orden 1 en la ecuación (7.14) que de esta forma ha quedado planteada de la siguiente manera:

$$(1 - B^{365}) * (1 - \phi_1 * B) Ciclo_t = \mu + \epsilon_t \Rightarrow Ciclo_t = Ciclo_{t-365} + Ciclo_{t-1} - \phi_1 * Ciclo_{t-366} + \mu + \epsilon_t \quad (7.22)$$

Los resultados obtenidos por la combinación de modelos (7.22)-(7.13) mejoran durante el periodo invernal los conseguidos por la combinación (7.14)-(7.13) pero empeoran durante el periodo estival. Esto podría no ser un problema dado que nuestro estudio busca la obtención de buenas predicciones durante los meses comprendidos entre Noviembre y Marzo. El siguiente gráfico establece la comparativa entre ellos, pudiendo apreciarse este ligero crecimiento.

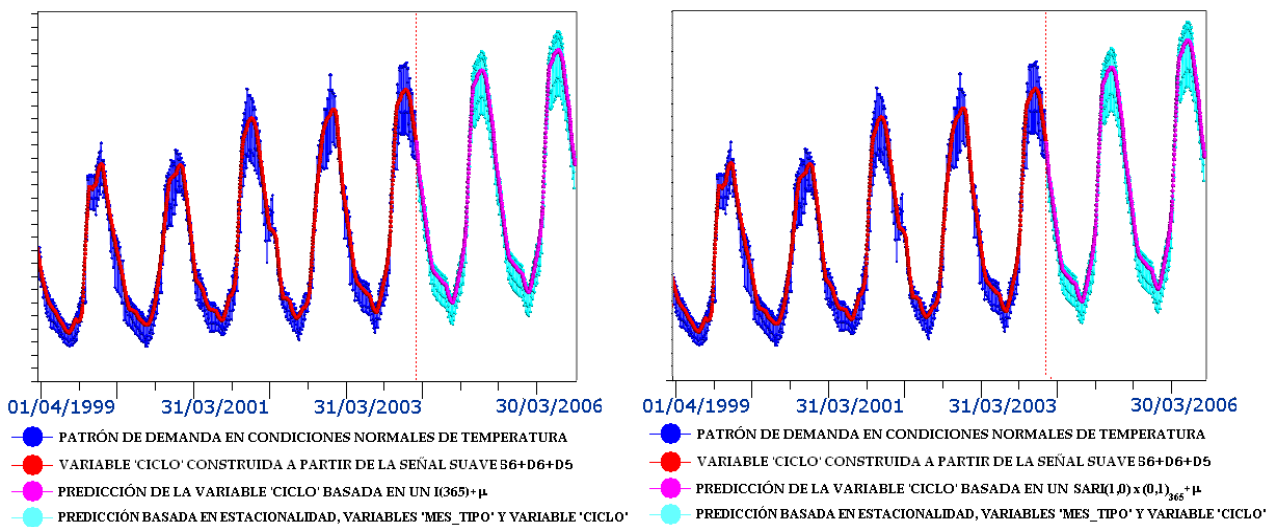


Ilustración 7.28: Comparativa entre las predicciones a medio plazo llevadas a cabo por dos modelos basados en la detección del comportamiento cíclico de la serie mediante funciones wavelets

Sin embargo, las predicciones que para los meses de invierno proporciona (7.22) - (7.13) tampoco han llegado a alcanzar la calidad conseguida mediante el modelo que emplea la curva ajustada por splines (7.20) - (7.13). Los errores medios y medianos de predicción para cada uno de los tres modelos se presentan en la siguiente tabla:

ANYO	MES	MEDIA_W1	MEDIA_W2	MEDIA_SP	MEDIANA_W1	MEDIANA_W2	MEDIANA_SP
2004	4	19.67	18.84	19.15	19.88	19.45	20.25
2004	5	5.18	3.27	5.68	4.65	2.39	4.37
2004	6	5.70	8.26	4.00	5.70	8.06	3.71
2004	7	9.74	12.66	4.19	9.96	12.76	4.51
2004	8	9.09	12.62	5.24	8.17	12.19	5.49
2004	9	9.18	12.35	4.78	10.28	13.27	4.65
2004	10	8.86	8.93	8.69	7.69	7.77	8.01
2004	11	9.02	8.45	7.90	7.35	7.91	7.12
2004	12	6.76	5.65	5.92	6.33	5.48	5.70
2005	1	18.12	16.36	14.16	19.10	17.42	14.08
2005	2	18.05	16.10	12.42	16.37	14.41	10.82
2005	3	16.63	15.96	14.74	14.22	13.35	11.36
2005	4	11.12	10.85	11.54	12.05	7.93	12.00
2005	5	11.93	16.90	9.01	8.36	14.51	3.09
2005	6	11.89	17.70	2.68	11.44	17.19	2.49
2005	7	17.33	23.62	4.36	16.26	23.06	3.01
2005	8	18.05	25.20	6.22	19.61	26.32	5.33
2005	9	14.93	21.24	2.82	14.89	21.41	2.16
2005	10	10.25	14.87	8.83	8.05	13.63	8.43
2005	11	6.83	6.09	6.22	5.71	5.65	5.54
2005	12	12.59	9.29	7.31	12.05	8.79	5.50
2006	1	8.22	6.61	6.74	9.25	6.69	5.63
2006	2	15.46	11.16	7.66	12.67	8.44	5.30
2006	3	9.53	10.69	10.92	7.95	10.35	11.07

Tabla 7.12: Comparativa entre los errores de predicción obtenidos por dos modelos que identifican el ciclo anual de la serie mediante funciones wavelets y uno que lo hace mediante splines

Se representan errores fuera de muestra para los modelos (7.14) - (7.13) identificados por *W1* (Wavelets1), (7.22) - (7.13) identificados por *W2* (Wavelets2) y (7.20) - (7.13) identificados por *SP* (Splines). A la vista de los resultados se ha optado finalmente por la inclusión de la variable *Ciclo* (más concretamente su transformada a través de la raíz cuadrada, *Rciclo*) resultante del ajuste por splines. En la tabla siguiente se presentan las estimaciones asociadas a los parámetros de la ecuación (7.13) así como su significatividad.

Dependent Variable		RCONSUMO_MEDIO	
Ordinary Least Squares Estimates			
SSE	163.727391	DFE	2530
MSE	0.06471	Root MSE	0.25439
SBC	440.683562	AIC	282.825633
Regress R-Square	0.9987	Total R-Square	0.9987
NOTE: No intercept term is used. R-squares are redefined.			

Variable	Estimate	Standard Error	t Value	Approx Pr > t
DIA_MXJ	0.3803	0.0642	5.93	<.0001
DIA_SP	0.3212	0.0180	17.89	<.0001
DIA_LV	0.2774	0.0645	4.30	<.0001
ENERO_L	0.1704	0.0309	5.52	<.0001
FEBRERO_L	0.0518	0.0309	1.67	0.0941
MARZO_L	-0.1116	0.0314	-3.56	0.0004
ABRIL_L	0.0443	0.0357	1.24	0.2156
MAYO_L	-0.0861	0.0382	-2.25	0.0243
JUNIO_L	-0.0305	0.0402	-0.76	0.4480
JULIO_L	-0.0286	0.0411	-0.69	0.4873
AGOSTO_L	-0.0773	0.0425	-1.82	0.0692
SEPTIEMBRE_L	0.1002	0.0407	2.46	0.0140
OCTUBRE_L	-0.2429	0.0354	-6.86	<.0001
NOVIEMBRE_L	0.3387	0.0315	10.76	<.0001
ENERO_F	-0.6028	0.0674	-8.95	<.0001
FEBRERO_F	-0.4506	0.0670	-6.73	<.0001
MARZO_F	-0.5276	0.0612	-8.63	<.0001
ABRIL_F	-0.4649	0.0524	-8.87	<.0001
MAYO_F	-0.6042	0.0488	-12.38	<.0001
JUNIO_F	-0.6291	0.0479	-13.14	<.0001
JULIO_F	-0.6479	0.0467	-13.87	<.0001
AGOSTO_F	-0.7181	0.0436	-16.46	<.0001
SEPTIEMBRE_F	-0.5963	0.0475	-12.56	<.0001
OCTUBRE_F	-0.9381	0.0524	-17.90	<.0001
NOVIEMBRE_F	-0.3528	0.0614	-5.74	<.0001
DICIEMBRE_F	-0.7351	0.0669	-10.98	<.0001
RCICLO	0.9837	0.006596	149.12	<.0001

Tabla 7.13: Modelo para la predicción a medio plazo de la demanda de gas que identifica el efecto estacional, la variabilidad, la tendencia y la periodicidad anual de la serie (las dos últimas componentes identificadas por splines)

Si bien el valor del R^2 mejora un poco respecto del último ajuste (véase la tabla (7.11)), lo que más llama la atención es que la nueva variable incluida (*Rciclo*) pasa a ser la que mayor significatividad presenta, modificando de manera sensible las estimaciones correspondientes al resto de los “inputs” de la ecuación. Esta variable además de cumplir su propósito fundamental de estimar el nivel en el que se moverá la serie durante los dos próximos años, viene a reforzar la periodicidad anual de la misma y la variabilidad cambiante por meses ya recogida en parte por las variables *MES_TIPO*. Recordemos que, al ir referida la definición de cada spline a un mes concreto, la variabilidad de éstos cambia con el tiempo permitiendo reflejar que el crecimiento de la demanda de un mes a otro no sea constante.

7.2.5 FASE III: Simulación del valor esperado de la demanda en situaciones climatológicas extremas

El resultado de las dos fases anteriores permite cuantificar el valor de la demanda diaria de gas que se espera en Madrid para los dos próximos años supuestas unas condiciones climatológicas normales. El paso final consiste en medir el impacto sobre dicho patrón, de una circunstancia meteorológica extrema como puede ser una ola de frío.

El punto de partida de esta última fase es la ilustración (7.14). En el primer cuadrante de ésta representábamos pares de puntos que reflejaban, durante los meses invernales, la relación existente entre el incremento de la demanda (respecto del valor que se estima debió presentarse en condiciones normales de temperatura) y un descenso del valor de la temperatura máxima (respecto de los valores medianos por meses presentados en la tabla (7.1)). La idea que perseguimos es determinar cuál es la distribución del incremento esperado para la demanda supuesto un descenso máximo de la temperatura, es decir, conocer la distribución de la variable $Y|X=x$ donde Y es el incremento porcentual de la demanda real respecto del patrón construido (véase ecuación (7.3)), X es la desviación entre el dato verdadero de temperatura y el establecido como estándar para un mes concreto (véase ecuación (7.2)) y “ x ” es la máxima desviación de temperatura respecto del estándar establecido para dicho mes que se tiene registrada en el histórico.

Una de las principales razones que justifican el tratamiento mensual que se ha dado a este estudio es distinguir la influencia de la temperatura en cada uno de los meses. Obviamente, no produce el mismo efecto una ola de frío en Diciembre o Enero, meses puramente invernales durante los cuales los niveles de calefacción acostumbran a estar al máximo, que en Marzo, Abril o Mayo, más primaverales, en los que porcentualmente el consumo de gas se incrementa muy por encima de la media mensual (véase ilustración (7.9)). Por ello, buscaremos la función de distribución que mejor refleje la dependencia entre los pares (x,y) en un mes determinado (de un color concreto en la ilustración (7.14)).

Es en este contexto de estimación de funciones de distribución y simulación de valores de sus condicionadas, en el que hemos considerado apropiado utilizar funciones cópula. Queremos resaltar un aspecto importante que es la independencia temporal que deben satisfacer las variables X e Y . Según se explica en [FERSCA], un error frecuente que se suele cometer al utilizar cópulas, es no garantizar la independencia temporal de las variables de análisis. Una cópula que se ajusta a los pares de valores (x,y) , busca medir la relación entre X e Y sin tener en cuenta la influencia de una posible tercera variable ($T = tiempo$). En los casos en los que esta dependencia temporal exista (por ejemplo, en la mayoría de las series financieras multivariantes), los autores consideran que no deben utilizarse las cópulas convencionales sino otras diseñadas específicamente a tal fin (consúltese el apartado 4 de [FERSCA] o los artículos de **Patton** que se referencian en esta tesis, [PATTON] y [PATTON2]).

En nuestro caso, no es preciso tener esta precaución dado que la independencia temporal de las variables X e Y es una consecuencia directa de su propia definición. La variable Y no mide la demanda de gas, la cual sí depende lógicamente del tiempo, sino los crecimientos o decrecimientos de demanda con respecto al patrón construido. Éstos, al estar medidos en términos porcentuales, no se ven fuertemente influenciados por sus instantes de ocurrencia “ t ”. De igual modo, la variable X no mide la temperatura que obviamente depende de la estación del año, sino el número de grados centígrados que ésta se desvía respecto de los valores medianos mensuales calculados a partir del histórico. Además debe tenerse en cuenta que, por el carácter mensual del estudio, vamos a ajustar una función cópula a los pares

muestrales de un mismo mes, siendo razonable pensar que la posible dependencia temporal entre las variables X e Y dentro de un mismo mes sea aún menor.

Aun cuando presentaremos los resultados finales para cada uno de los 10 meses invernales que se contemplan en los dos años del horizonte de predicción (dos meses de Noviembre, dos de Diciembre, dos de Enero, dos de Febrero y dos de Marzo), únicamente para uno de ellos, M , haremos un análisis más detallado y mostraremos los gráficos y resultados parciales que vayamos obteniendo en cada una de las etapas de esta última fase. Éstas son las siguientes:

- En primer lugar, determinaremos las distribuciones marginales F_X y G_Y de las variables X e Y en el mes M . Para ello, utilizaremos el estimador no paramétrico dado por la función de distribución empírica continua (véase apartado 2.2) construido a partir de los datos históricos de dicho “tipo” de mes (desde el 1 de Abril de 1997 hasta el 31 de Marzo de 2004).
- En segundo lugar, determinaremos aquella familia de cópulas $C_{XY,M}^*$, de entre todo un conjunto de candidatas (arquimedeanas, elípticas, de valor extremo, etc.) que mejor se aproxime a la verdadera distribución conjunta H_{XY} (de acuerdo a la relación establecida por el teorema de Sklar). El criterio de selección de la cópula por el que se ha optado, consiste en tomar aquélla que ofrece un mínimo valor respecto del estadístico de Pearson que evalúa el contraste de bondad de ajuste a una distribución desconocida respecto de una partición dada (véase apartado 2.5.5). La evaluación del estadístico se hará nuevamente sobre datos correspondientes al mes M localizados en el conjunto de entrenamiento (desde el 1 de Abril de 1997 hasta el 31 de Marzo de 2004).
- Finalmente realizaremos predicciones fuera de muestra (*out-of-sampling*) con la cópula seleccionada $C_{XY,M}^*$. Además, se contrastarán dichos resultados con aquéllos que proporcionen otras cópulas obtenidas por interpolación de la subcópula que optimiza el estadístico de Pearson y cuya generación se detalla en el capítulo 4 de esta tesis. En concreto, se propondrá una cópula obtenida por interpolación bilineal (véase apartado 4.2.1), otra mediante polinomios interpoladores de Bernstein (véase apartado 4.2.2) y una tercera obtenida por interpolación cúbica mediante polinomios de Hermite (véase apartado 4.2.3). El criterio de selección definitivo que se ha tomado consiste en elegir aquélla que mejores predicciones proporcione (véase apartado 2.5.6).

Vamos a empezar determinando el mes M para el que iremos detallando cada uno de estos pasos. Podemos considerar que tal vez, el que pueda ilustrar mejor el proceso completo es aquél para el cual, a futuro, se disponga de una mayor cantidad de descensos extremos de la temperatura. Esta consideración es importante pues perseguimos evaluar la calidad de las predicciones fuera de muestra en situaciones meteorológicas especialmente adversas y por ello es conveniente disponer de una muestra de validación que sea lo más amplia posible. La siguiente tabla muestra por filas, para cada uno de los 10 meses invernales fuera de muestra, el número de días para los que el valor real de la temperatura máxima se desvía un determinado número de grados (por columnas) de los valores medianos recogidos en la tabla (7.1).

AÑO	MES	DIF_1	DIF_2	DIF_3	DIF_4	DIF_5	DIF_6	DIF_7	DIF_8	DIF_9	DIF_10	NUM_DIF
2004	11	3	2	1	0	0	0	0	0	0	0	6
2004	12	1	5	0	1	1	0	0	0	0	0	8
2005	1	2	0	1	1	0	1	0	1	0	0	6
2005	2	3	3	3	1	4	3	3	0	0	2	22
2005	3	3	1	0	2	0	2	1	1	1	1	12
2005	11	4	2	0	3	3	3	0	0	0	0	15
2005	12	7	1	1	1	0	0	0	0	0	0	10
2006	1	2	5	4	1	2	1	0	1	0	0	16
2006	2	4	2	2	2	1	2	2	0	0	0	15
2006	3	3	4	2	2	3	0	0	0	1	0	15

Tabla 7.14: Conteo, sobre el conjunto de validación, del número de días por mes en los que la temperatura máxima desciende un determinado número de grados centígrados respecto del valor establecido como normal

A la vista de esta tabla y de acuerdo al criterio establecido, no existe duda en tomar el mes de Febrero de 2005 como referencia de estudio. Del mismo modo, si nos restringimos a las desviaciones de temperatura que hemos dejado de considerar como “normales” (a partir de 3°C) y a cuyo tratamiento está dirigido este análisis, el número de días que aparecen en dichas condiciones es muy superior (16) al de cualquier otro mes (Noviembre de 2005 y Enero y Febrero de 2006 como siguientes candidatos, sólo tienen 9).

En consecuencia, el soporte de las variables X e Y para el caso que vamos a detallar, estará conformado por datos referidos a los meses de Febrero comprendidos entre el 1 de Abril de 1997 y el 31 de Marzo de 2004, es decir, un total de 7 meses, que dan lugar, a una muestra de 198 puntos (véase tabla (7.7)) de los cuales 158 (casi un 80%) caen en los cuadrantes primero y tercero. Recordemos que son los pares ubicados en estos cuadrantes los que responden a una relación lógica entre las variables: “cuando la temperatura baja se produce un incremento de demanda y cuando sube se produce un decremento de la misma”. Sin embargo, dado que estamos únicamente interesados en posibles aumentos del consumo de gas a partir de bajadas de temperatura, vamos a restringir aún más el espacio de análisis al conjunto de pares que pertenecen al primer cuadrante. Éste se compone de un total de 76 puntos (casi la mitad de los bien clasificados) que presentamos en la ilustración (7.30). Este conjunto servirá de soporte para obtener las funciones de distribución, tanto de X e Y , como de la conjunta (X,Y) a través del ajuste de una cópula al espacio de pares transformados (u,v) por las respectivas marginales F_X y G_Y .

7.2.5.1 Generación de distribuciones marginales

Para la generación de las marginales F_X y G_Y , basta aplicar la definición de la función de distribución empírica continua que se presenta en [MATTEIS] y que ya dimos en el apartado 2.2. También podría contrastarse si X e Y responden a algunas de las distribuciones continuas univariantes conocidas (normal, uniforme, exponencial, etc.). Presentamos a continuación el resultado gráfico de esta aproximación empírica.

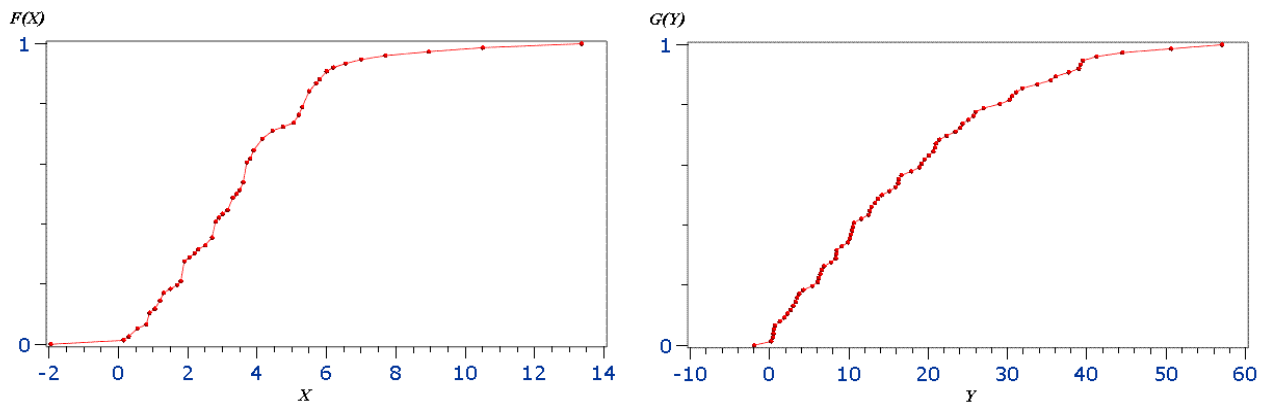


Ilustración 7.29: Funciones de distribución empíricas continuas asociadas al incremento de la temperatura máxima y al incremento de la demanda de gas en Febrero

Aplicando estas funciones obtenidas a las muestras respectivas de X e Y obtendremos el espacio transformado de pares $(u, v) = (F_X(x), G_Y(y))$ (véase el gráfico de la derecha en la ilustración (7.30)), que utilizaremos para determinar la cópula $C_{XY,M}^*$ que mejor representa la relación de dependencia entre las variables.

7.2.5.2 Determinación de una cópula a través del criterio de selección basado en el estadístico de Pearson

Dado que el criterio de selección que vamos a seguir está basado en determinar aquella cópula que respecto de una partición del cuadrado unidad presenta un mejor valor del estadístico de Pearson, es necesario en primer lugar determinar la dimensión de dicha partición. Si bien sería bueno que cada uno de los rectángulos que resultasen de esta partición tuviese un número mínimo de puntos (existen convenios empíricos que fijan en 5 este mínimo), proceder de esta manera dificultaría disponer de un rectángulo en el que se concentrasen claramente los valores extremos de ambas variables. De no disponer de él, por el carácter empírico de las densidad de las cópulas interpoladoras que vamos a ajustar, éstas no serían capaces de concentrar su probabilidad en esta región extrema y en consecuencia no serían capaces de capturar bien correlaciones entre sucesos extremos. Así, aunque la partición que asegura este mínimo de puntos por rectángulo es de dimensión 2×2 , se ha decidido tantear con particiones de diferentes dimensiones dentro de unos rangos razonables, para ver cuál de ellas termina ofreciendo mejores resultados en términos de errores de predicción. De todas ellas, ha sido la de dimensión 5×5 la que ha conseguido esta optimalidad. En el siguiente gráfico se presenta el espacio de análisis antes y después de la transformación llevada a cabo por las marginales F_X y G_Y , así como la configuración del reparto de pares en el cuadrado unidad respecto de esta partición.

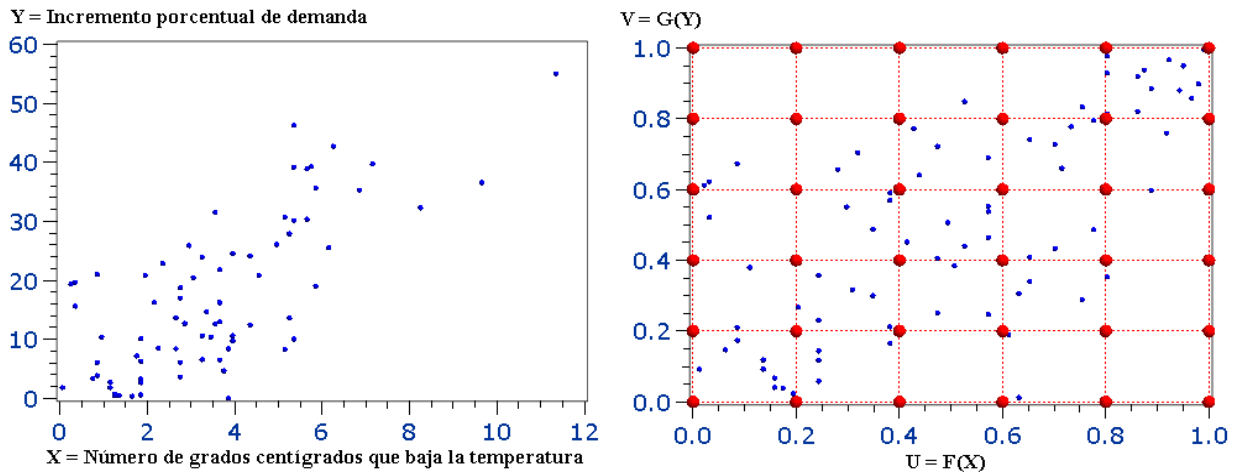


Ilustración 7.30: Espacio original de las variables incremento de temperatura e incremento de demanda en Febrero y espacio transformado mediante sus funciones de distribución empíricas continuas

Observemos que, como acabamos de decir, sería necesario reducir la dimensión del espacio hasta 2×2 para garantizar la presencia de 5 puntos por rectángulo.

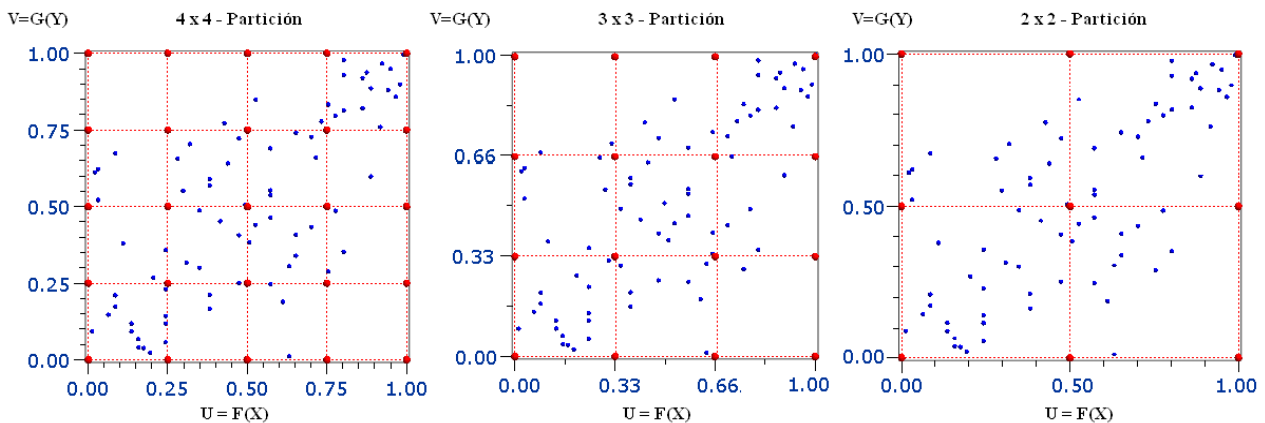


Ilustración 7.31: Tentativas de partición para el espacio transformado que sirve de soporte a una cópula

El siguiente paso nos lleva ahora a considerar una relación de familias de cópulas uniparamétricas candidatas a reflejar la relación entre nuestro par de variables. Son dos las condiciones que vamos a imponer para que una familia de cópulas pueda ser evaluada como posible alternativa:

- Que la construcción de la densidad cópula condicionada $Y|X=x$ no resulte compleja sino de expresión conocida y, si también es posible, que sea fácilmente invertible pues, a partir de dicha función, llevaremos a cabo la simulación de valores del incremento de demanda en función de una desviación propuesta para la temperatura máxima. En caso de existir problemas de invertibilidad, siempre podremos generar valores de la misma a través de métodos numéricos.

- Que el parámetro de la cópula se pueda calcular de una forma sencilla a partir de la Tau de Kendall (o del coeficiente de correlación de Spearman o incluso del de correlación lineal) para poder disponer así de un representante de la familia para el cual calcular el estadístico de Pearson asociado a la partición.

Conforme a este criterio, las familias propuestas, cuya expresión se detalla en el Anexo I, han sido:

- Familia de Ali-Mikhail-Haq, de nomenclatura ALI (véase (AI.10)).
- Familia de Clayton, de nomenclatura CLA (véase (AI.11)).
- Familia de Cola Derecha Pesada, de nomenclatura HRT (véase (AI.12)).
- Familia de Farlie-Gumbel-Morgenstern, de nomenclatura FGM (véase (AI.34)).
- Familia de Frank, de nomenclatura FRA (véase (AI.13)).
- Familia de Gumbel, de nomenclatura GUM (véase (AI.8) o (AI.15)).
- Familia Normal o gaussiana, de nomenclatura NOR (véase (AI.5)).
- Familia de Plackett, de nomenclatura PLA (véase (AI.37)).

Con el fin de reunir un total de 10 familias, hemos considerado otras dos arquimedianas (las número 13 (AI.22) y 15 (AI.24) dentro de la clase de las arquimedianas uniparamétricas presentadas en el Anexo I, de nomenclaturas F13 y F15 respectivamente) cuya distribución condicionada es también fácil de calcular y para las que el valor del parámetro es directamente calculable a través del estadístico de Kendall. Para conocer las condicionadas asociadas a las cópulas arquimedianas así como el cálculo del valor del parámetro se pueden consultar los Apéndices A,B y C de [MATTEIS]. Dentro de éste, estas dos últimas familias ((AI.22) y (AI.24)) responden a las numeraciones 12 y 14. También hemos considerado como candidata a la cópula producto (o independiente) para contrastar una posible falta de dependencia entre las variables.

Gracias a las condiciones impuestas es posible calcular el valor del estadístico de Kendall asociado al conjunto de pares (u,v) y, a partir de él, estimar el valor del parámetro correspondiente a cada una de las familias candidatas. La cópula cuyo parámetro tome dicho valor será considerada como la representante de la familia siempre y cuando el valor resultante se mantenga en el rango de variación que para ese parámetro tenga establecido la familia en cuestión. A modo ilustrativo proponemos como ejemplo seleccionar los representantes asociados a las familias de Gumbel y Farlie-Gumbel-Morgenstern.

Como paso previo calcularemos los valores de la Tau de Kendall y el coeficiente de correlación de Spearman respecto de la muestra (u,v) representada en la ilustración (7.30). Estas medidas de asociación, además de cuantificar la relación (no necesariamente lineal) entre las variables U y V , servirán para estimar los parámetros asociados a las dos familias consideradas.

Prob > r under H0: Rho=0 N = 76		
	Kendall Tau	Spearman Correlation
	U	U
V	0.49011 <.0001	0.66572 <.0001

Tabla 7.15: Valor de la Tau de Kendall y del coeficiente de Spearman para la muestra de entrenamiento transformada asociada al mes de Febrero

La clase de Gumbel responde a la expresión $C_{\theta}(u,v) = \exp(-[(-\ln u)^{\theta} + (-\ln v)^{\theta}]^{\frac{1}{\theta}})$ con $\theta \geq 1$

donde $\tau = 1 - \frac{1}{\theta}$ (véase Apéndice B de [MATTEIS]). En consecuencia, $\hat{\theta} = \frac{1}{1 - \hat{\tau}} = \frac{1}{1 - 0.49} = 1.96 \geq 1$ y así el representante de la familia de Gumbel sería

$$C_{GUMBEL}^*(u,v) = \exp(-[(-\ln u)^{1.96} + (-\ln v)^{1.96}]^{\frac{1}{1.96}}) \quad (7.23)$$

Por otra parte, la clase de Farlie-Gumbel-Morgenstern responde a la expresión

$$C_{\theta}(u,v) = u * v * [1 + \theta * (1-u) * (1-v)] \quad \text{con } -1 \leq \theta \leq 1$$

donde $\tau = \frac{2 * \theta}{9}$ (véase ejemplo 5.2 de [NELSEN]). En consecuencia, $\hat{\theta} = \frac{9 * \hat{\tau}}{2} = \frac{9 * 0.49}{2} = 2.25 \notin [-1, 1]$ y por tanto esta cópula es rechazada del conjunto de candidatas.

De todas las familias enumeradas, únicamente hemos rechazado la de Farlie-Gumbel-Morgenstern por no estar comprendida la estimación del parámetro dentro de su rango de variación. En la tabla (7.16) se presenta el valor asociado al parámetro de la cópula que se considerará representante de su respectiva familia. Posteriormente se calcula, para cada uno de estos representantes, el valor del estadístico de Pearson (expresión (4.1)) respecto de la partición planteada que permite contrastar la bondad de ajuste de cada cópula a la función de distribución verdadera, es decir, a la cópula que, en función del teorema de Sklar, está determinada de manera unívoca. El valor del estadístico es una

evaluación del contraste $H_0: C(u, v) = C(u, v; \theta)$ para algún $\theta \in \Theta$. Cuanto mayor sea la proximidad entre el volumen empírico de los rectángulos de la partición y el volumen esperado para ellos a través de la cópula representante de una clase, menores serán las cantidades como suma de las cuales se obtiene el estadístico de Pearson y en consecuencia, menor será el valor de éste. Así, conforme se hace más pequeño el valor de este estadístico, mayor es el grado de bondad del ajuste de la cópula a los datos. En la tabla siguiente se muestra para cada uno de los representantes de las familias de cópulas candidatas, los valores de los respectivos estadísticos de Pearson ordenados de forma creciente.

COPULA	PARAMETRO_COPULA	PEARSON	P_VALOR
HRT	0.5201	17.8778	0.26914
GUM	1.9611	17.8800	0.26903
PLA	10.625	22.1443	0.10408
F15	1.4611	24.0163	0.06482
NOR	0.6960	24.1406	0.06274
F13	1.3074	28.5944	0.01813
CLA	1.9223	45.8672	0.00006
FRA	1.1384	51.3273	0.00001
ALI	0.4755	53.1418	0.00000
PRO	.	75.9743	0.00000

**La región crítica viene dada por $\{X|X > 24.99579014\}$
Nivel de significación del contraste, $\alpha = 0.05$**

Tabla 7.16: Evaluación del estadístico de Pearson para una serie de cópulas ajustadas a la muestra de entrenamiento asociada al mes de Febrero

Dado que hemos supuesto que las marginales de X e Y son desconocidas y que hemos procedido a su estimación a partir de sus funciones de distribución empíricas, el estadístico de Pearson, según se expone en el capítulo 2 de [DOSCH], se distribuye según una Chi-cuadrado con $(m-1)*(n-1)-d=4*4-1=15$ grados de libertad, cuyo valor, a un nivel de significación del 0.05%, es 25.00, dando lugar a la región crítica $RC = \{X|X > 25.00\}$. Respecto de esta región la cópula gaussiana (NOR) marcaría el umbral desde el cual se rechazaría el contraste de bondad de ajuste. Si por el contrario, las funciones F_X y G_Y hubieran sido conocidas, el estadístico de Pearson se hubiera distribuido según una Chi-cuadrado con $m*n-1-d=5*5-1-1=23$ grados de libertad donde d sería el número de parámetros de la cópula. En esta caso, el valor del estadístico, a un nivel de significación del 0.05%, hubiese sido 35.17, dando lugar a la región crítica $RC = \{X|X > 35.17\}$. Respecto de esta segunda región, el contraste sugeriría de forma adicional como posibles alternativas las familias F13 y de Clayton (CLA).

En cualquier caso, como era de esperar la cópula producto (I) resulta ser la que peor se aproxima a la verdadera distribución cópula asociada a los datos, lo cual refuerza la hipótesis de que efectivamente existe estructura de dependencia entre ellos. Por otra parte, las dos familias de cópulas que mejor parecen reflejar esta relación de dependencia entre X e Y (a través de U y V), con valores del estadístico de Pearson casi parejos, son la de Cola Derecha Pesada (C_{HRT}) y la de Gumbel (C_{GUMBEL}). Este resultado parece sensato teniendo en cuenta que, de antemano, esperábamos que la cópula a utilizar enfatizase la relación entre valores extremos de demanda y temperatura. De todas las candidatas planteadas son precisamente estas dos las especialmente orientadas a tal fin. La cópula C_{HRT} refuerza la relación entre las colas derechas de las distribuciones (máximos incrementos de demanda frente a máximas caídas de temperatura) y la de Gumbel es en sí misma una cópula perteneciente a la clase de valor extremo que, además de presentar colas pesadas, manifiesta cierta asimetría tendiendo a concentrar la probabilidad más en su cola derecha que en su cola izquierda. Decantarse por cualquiera de ellas es indiferente de acuerdo a los valores tan parecidos que presentan respecto del estadístico del contraste. Por respetar el resultado obtenido optaremos por la primera de ellas, o lo que es lo mismo, por la cópula representante de la familia HRT, de ecuación

$$C_{HRT}^*(u, v) = u + v - 1 + [(1-u)^{-0.52} + (1-v)^{-0.52} - 1]^{-0.52} \quad (7.24).$$

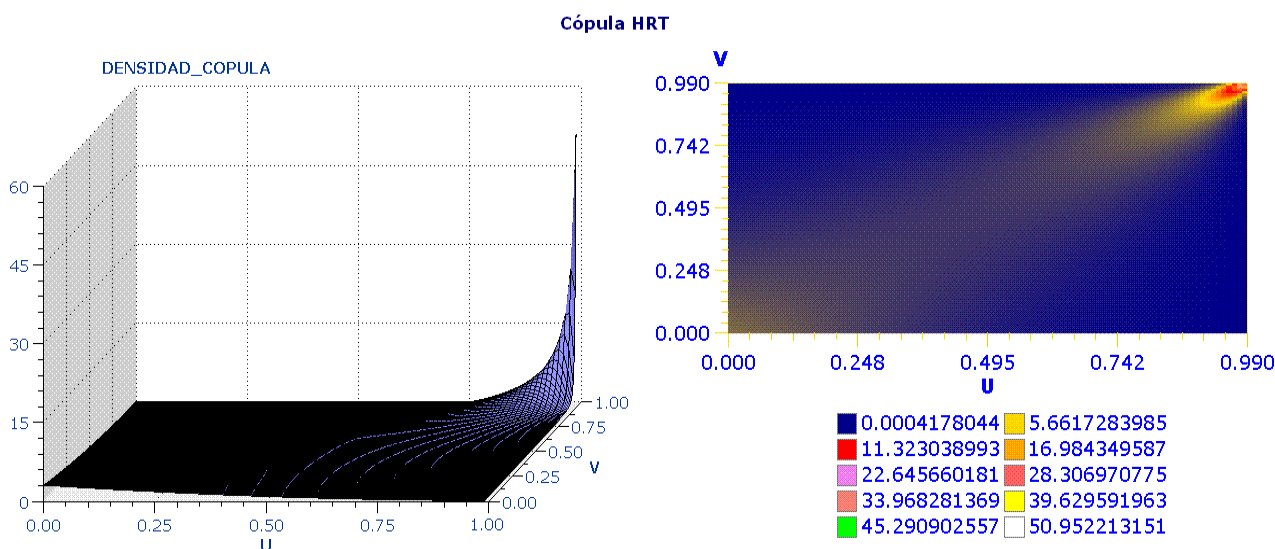


Ilustración 7.32: Densidad de la cópula de Cola Derecha Pesada ajustada a la muestra de entrenamiento en Febrero

Los gráficos anteriores son dos visiones posibles de la densidad cópula asociada a la expresión (7.24), una mediante un gráfico tridimensional y otra mediante uno de contorno. En ambos queda claramente puesta de manifiesto la concentración de probabilidad en la parte de la distribución conjunta en que se “cruzan” las dos colas derechas de las distribuciones marginales. La ecuación de esta densidad así como algunas otras propiedades y aplicaciones prácticas de esta cópula pueden ser encontradas en [VENTER]. En dicho artículo encontramos por ejemplo la relación entre el valor

de la cópula y el estadístico de Kendall, medida por $\tau = \frac{1}{2 * \theta + 1}$.

De aquí, se concluye que $\hat{\theta} = \frac{1-\hat{\tau}}{2*\hat{\tau}} = \frac{1-0.49}{0.98} = 0.52$, valor que entra dentro del rango de variación del parámetro de esta cópula ($\theta > 0$) y que es utilizado para calcular el representante de la familia (7.24).

De particular interés para nosotros va a ser la expresión de la cópula condicionada C_1 calculada a partir de la derivación respecto de U de la expresión (7.24) y que según se presenta en el artículo al que se hace referencia es:

$$C_{1,HRT}^*(u, v) = C_{HRT}^*(v|u) = 1 - [(1-u)^{\frac{-1}{0.52}} + (1-v)^{\frac{-1}{0.52}} - 1]^{-0.52-1} * (1-u)^{-1-\frac{1}{0.52}} \quad (7.25)$$

Esta expresión es invertible y por tanto se puede utilizar el método de la transformada inversa para generar valores de V a partir de un valor $U = u$ conocido (véase apartado 5). Efectivamente, si “ p ” es un valor generado de una variable aleatoria P uniforme en $(0,1)$, la igualdad

$$1 - [(1-u)^{\frac{-1}{0.52}} + (1-v)^{\frac{-1}{0.52}} - 1]^{-0.52-1} * (1-u)^{-1-\frac{1}{0.52}} = p \quad (7.26)$$

permite despejar “ v ” de manera sencilla y obtener la ecuación del valor generado para la variable V en función de U ,

$$\text{dada por } v = 1 - \left\{ 1 - (1-u)^{\frac{-1}{0.52}} + \left\{ (1-p) * (1-u)^{\left(1+\frac{1}{0.52}\right)} \right\}^{\frac{-1}{0.52+1}} \right\}^{-0.52} \quad (7.27).$$

La propuesta de valores aleatorios uniformes para “ p ” genera automáticamente valores de “ v ”.

En la ecuación anterior el valor de “ u ” no es más que el transformado por F_x de una bajada de “ x ” grados centígrados que suponemos del valor de la temperatura. Como nuestro interés se centra en conocer la respuesta incremental que se puede esperar para la demanda ante un descenso extremo de la temperatura, con vistas a poder evaluar la bondad de la simulación realizada mediante cópulas, vamos a seleccionar un día de Febrero de 2005 en el que el descenso de temperatura fuese máximo. Concretamente, el día 23 de Febrero de 2005 la temperatura máxima registrada en el aeropuerto de Barajas alcanzó un valor mínimo de 3.3°C, esto es, 10.4°C por debajo del valor mediano estimado para el mes (véase tabla (7.1)). Por tanto podemos proponer para “ x ” ese mismo valor, hallar su transformado $u = F_x(x)$ y generar valores de V . Obviamente, los valores que realmente nos van a interesar son las imágenes inversas mediante G_Y de cada “ v ” generado, pues serán estos valores simulados de Y los que correspondan a los incrementos porcentuales de demanda. A través de estos valores simulados, podremos obtener los respectivos valores de la propia variable *demanda* tras despejar en la ecuación (7.3) de la siguiente manera:

$$\text{Demanda_esperada}(t) = \text{Prediccion_demanda_en_condiciones_normales_de_temperatura}(t) * \left(1 + \frac{Y}{100}\right) \quad (7.28)$$

Después de simular 100 valores de Y , hemos procedido a su representación a través de un histograma de frecuencias, siendo el resultado el que se muestra a continuación:

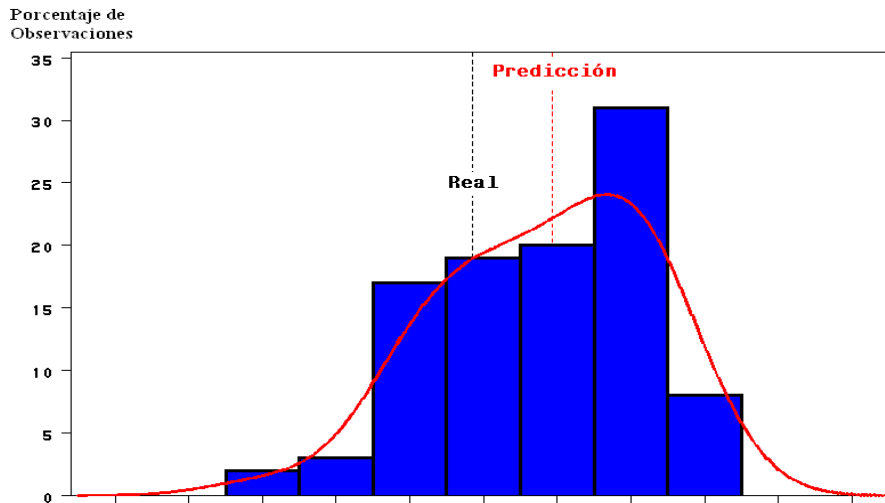


Ilustración 7.33: Distribución de las predicciones simuladas para la demanda cuando la temperatura máxima desciende 10°C respecto de su valor normal para Febrero

La mediana de estos valores, marcada por una fina línea punteada en rojo, podría ser tomada como el valor esperado para la demanda. El error porcentual de esta predicción con respecto al valor real que se presentó para la demanda es de un -2.82%, aproximación que se puede considerar francamente buena. Pero además, de manera adicional, el conocimiento de la distribución permite medir la probabilidad de que el consumo de gas sea aún mayor y evaluar posibles riesgos derivados de que se alcanzaran valores extremos. La forma asimétrica de la distribución, más pesada en su cola derecha, es una consecuencia directa del empleo de la cópula C_{HRT}^* .

7.2.5.3 Construcción de cópulas que optimizan el estadístico de Pearson

Vamos ahora a ajustar cópulas de carácter empírico, no paramétricas y que optimizan el valor del estadístico de Pearson. Para ello empezaremos resolviendo el problema de programación fraccional (4.7) que proporciona la relación de valores c'_{ij} que debe tomar una subcópula C' en cada uno de los vértices (i, j) de la partición para que el valor del estadístico de Pearson, calculado a partir de los propios c'_{ij} y del número de puntos muestrales N_{ij} que caen en cada uno de los rectángulos, sea lo menor posible. La siguiente tabla presenta la solución óptima c^*_{ij} a dicho problema. El valor del estadístico de Pearson que se obtiene a partir de esta solución es 0.357.

	j						
5	0.000	0.200	0.400	0.600	0.800	1.000	
4	0.000	0.200	0.389	0.578	0.766	0.800	
3	0.000	0.171	0.316	0.466	0.584	0.600	
2	0.000	0.171	0.275	0.335	0.400	0.400	
1	0.000	0.121	0.148	0.160	0.200	0.200	
0	0.000	0.000	0.000	0.000	0.000	0.000	
	0	1	2	3	4	5	i

Tabla 7.17: Valores que debe tomar la subcópula que optimiza el valor del estadístico de Pearson en cada uno de los vértices de la partición del cuadrado unidad

Se puede demostrar que la función C^* definida por esta relación de valores c^*_{ij} es una subcópula viendo que verifica cada una de las propiedades que caracterizan a estas funciones ((1.1) – (1.4)). La comprobación de las dos condiciones frontera es inmediata:

$$C'(\frac{i}{5}, 0) = C'(0, \frac{j}{5}) = 0 \quad \forall i, j \in \{1, 2, 3, 4, 5\} \quad \text{y} \quad C'(\frac{i}{5}, 1) = \frac{i}{5} \quad \text{y} \quad C'(0, \frac{j}{5}) = \frac{j}{5} \quad \forall i, j \in \{1, 2, 3, 4, 5\} \quad .$$

Para verificar que C^* es 2-creciente, basta ver que para cualquier rectángulo conformado por 4 valores adyacentes de dicha tabla (la evaluación de C^* sobre los vértices del rectángulo), la diferencia entre los que se sitúan en la diagonal principal y los que se ubican en la secundaria es siempre mayor o igual que cero. De ser así, como cualquier otro rectángulo se puede expresar siempre como un agregado de ellos, sus correspondientes volúmenes serán también mayores o iguales que cero. También se puede comprobar el lema 2.1.4 que encontramos en [NELSEN] según el cual, si efectivamente la función C^* es 2-creciente entonces debe ser no decreciente en cada una de sus variables. A este respecto, podemos observar que los valores de la tabla anterior crecen conforme nos movemos hacia arriba o hacia la derecha. Sin embargo, se trata de una propiedad necesaria (y no suficiente) que no garantiza la “2-crecencia” de C^* .

Los valores anteriores definen así una subcópula C^* que además presenta el menor valor posible del estadístico de Pearson para la partición dada, en comparación con cualquier otra posible subcópula definida sobre ella. Si construimos una función cópula que tome exactamente los mismos valores que la subcópula en los nodos de la partición, el resultado de la evaluación del estadístico de Pearson para dicha cópula será exactamente el mismo. De acuerdo a lo expuesto en los apartados 4.2.1, 4.2.2 y 4.2.3 de esta tesis, vamos a proponer una cópula generada por interpolación bilineal, otra por interpolación mediante polinomios de Bernstein y una tercera obtenida por interpolación mediante polinomios cúbicos de Hermite. Finalmente estableceremos la comparativa entre las curvas de distribución de demanda predichas para el 23 de Febrero de 2005 con cada una de las cópulas propuestas.

7.2.5.3.1 Construcción de una cópula por interpolación bilineal

Teniendo en cuenta el lema (4.1) y la ecuación (4.8) la construcción de la cópula $C^*_{BILINEAL}$ que se obtiene por interpolación bilineal de C^* es inmediata. El valor de la densidad cópula puede ser calculado directamente a partir de la tabla anterior ya que se define sobre cada rectángulo de la partición como la constante que resulta de realizar el cociente

entre el volumen del propio rectángulo y su área (véase ecuación (4.9)). Ésta última cantidad es común a todos los rectángulos y vale $\frac{1}{5} * \frac{1}{5} = \frac{1}{25}$. Así por ejemplo, teniendo en cuenta la tabla (7.17), el valor de la densidad en

cualquier punto del rectángulo superior derecho B sería $\frac{V_C^*(B)}{1/25} = 25 * (1 - 0.8 - 0.8 + 0.766) = 4.15$.

Los siguientes gráficos dan dos visiones distintas de la densidad cópula asociada a $C_{BILINEAL}^*$, función que por ser constante sobre cada uno de los rectángulos presenta un aspecto escalonado. En el gráfico de contorno resulta sencillo validar el resultado de la operación anterior.

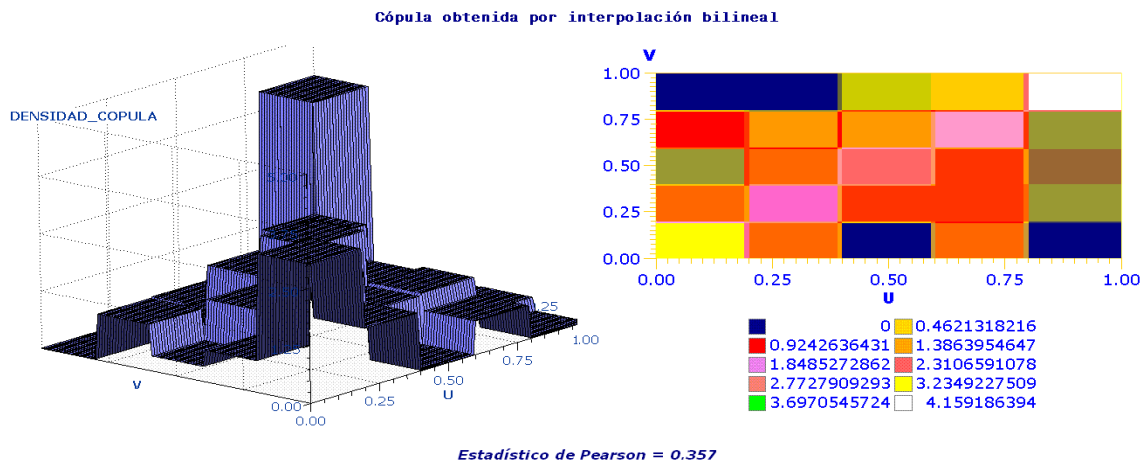


Ilustración 7.34: Densidad de la cópula obtenida por interpolación bilineal de la subcópula que optimiza el estadístico de Pearson ajustada a la muestra de entrenamiento en Febrero

Dado que esta cópula es una interpolación exacta de los valores que toma C^* en cada uno de los puntos de su dominio de definición, la evaluación del estadístico de Pearson vuelve a ser 0.357. Se puede observar como la mayor parte de la densidad se concentra en la región de la función en la que confluyen las colas derechas de las distribuciones de X e Y , lo cual es sinónimo de la fuerte correlación entre ellas y que ya recogía la cópula C_{HRT}^* . Aunque en menor medida, otra parte importante de la densidad se concentra en el rectángulo inferior izquierdo lo cual se puede entender si tenemos en cuenta que existen bastantes puntos en el mismo (véase ilustración (7.30)) que reflejan pequeños incrementos de demanda en respuesta a insignificantes descensos de la temperatura. Esto es una prueba de la buena calidad del proceso de construcción del histórico de demanda en condiciones normales de temperatura pues es de esperar que, si la temperatura se desvía pocos grados de su valor mensual esperado, el efecto sobre el comportamiento base de la demanda sea despreciable. La captura de esta segunda región en la que se concentra la probabilidad no es un hecho asociado exclusivamente al carácter empírico de la cópula interpoladora. Ya en el ajuste que hemos hecho con la cópula C_{HRT}^* se mostraban leves indicios de correlación entre las colas izquierdas de las distribuciones, circunstancia que no pasaba desapercibida en el gráfico térmico de la ilustración (7.32) en el cual se dejaba ver una tenue degradación del color cerca del origen. El último paso es nuevamente simular valores de Y a partir de la condicionada de $C_{BILINEAL}^*$, lo cual se hará a través de la expresión (5.16) según se detalla en el apartado 5.2.

7.2.5.3.2 Construcción de una cópula por interpolación cúbica

El segundo tipo de interpolación que vamos a llevar a cabo es a través de polinomios cúbicos. La idea es ajustar a cada uno de los rectángulos de la partición un polinomio de orden 3, o superficie cúbica en U y V , que interpole los valores c^*_{ij} que toma la subcópula óptima en cada uno de los vértices (i, j) de dichos rectángulos. Sin embargo, la construcción de una cópula cúbica que pase por dichos puntos no es siempre posible. De acuerdo al teorema (4.6) la construcción de una cópula mediante superficies cúbicas interpoladoras de Hermite está garantizada si C^* es una subcópula- V_{MIN} que de acuerdo a la definición (4.2) es una subcópula tal que el volumen a través de ella de cualquier rectángulo de la partición

$$\text{sea al menos } V_{MIN} = \frac{2}{3 * m * n} = \frac{2}{3 * 5 * 5} = \frac{2}{75} .$$

Nuestra subcópula óptima no verifica esta propiedad pues dado por ejemplo el rectángulo $B = [0, \frac{1}{5}] \times [\frac{4}{5}, 1]$ y de acuerdo a los valores presentados en la tabla (7.17), el volumen de dicho rectángulo sería

$$V_{c^*}(B) = 0.2 - 0.2 - 0 + 0 = 0 < \frac{2}{75} .$$

Por ello, un primer paso consiste en definir la subcópula- V_{MIN} C^{**} que optimice

nuevamente el valor de la expresión de Pearson o lo que es lo mismo, encontrar la solución óptima al problema de programación matemática (4.39). La siguiente tabla presenta la relación de valores c^{**}_{ij} que definen esta solución.

j								
5	0.000	0.200	0.400	0.600	0.800	1.000		
4	0.000	0.173	0.347	0.520	0.693	0.800		
3	0.000	0.147	0.277	0.414	0.520	0.600		
2	0.000	0.120	0.216	0.284	0.347	0.400		
1	0.000	0.084	0.111	0.137	0.173	0.200		
0	0.000	0.000	0.000	0.000	0.000	0.000		
	0	1	2	3	4	5	i	

Tabla 7.18: Valores que debe tomar la subcópula- V_{MIN} que optimiza el valor del estadístico de Pearson en cada uno de los vértices de la partición del cuadrado unidad

El valor del estadístico de Pearson que se obtiene a partir de esta solución y del número de puntos muestrales N_{ij} que caen en cada uno de los rectángulos es 19.00, valor que se conservará para la cópula cúbica C^{**}_{CUBICA} que construyamos y que, como era de esperar, es superior al obtenido para la subcópula óptima. Este valor no es directamente comparable con los presentados para dicho estadístico respecto de las familias uniparamétricas presentadas en la tabla (7.16) precisamente por la naturaleza no paramétrica de las cópulas interpoladoras que hace que el número de grados de libertad de ellas no sea el mismo. Aún así, el valor obtenido es muy similar al que ha resultado para las principales favoritas (próximo a 18 para C^*_{HRT} y C^*_{GUMBEL}), con lo cual, si simplemente interpretáramos los valores de la expresión de Pearson como una medida del error de ajuste entre lo observado en cada rectángulo (N_{ij}) y lo esperado a través de la cópula ($n * p_{ij}$), podríamos considerar que ambos son de magnitud parecida.

La definición de la superficie cúbica interpoladora se establece a partir de la expresión (4.19) la cual requiere del conocimiento del valor de $C^{**}_{CÚBICA}$ así como de sus derivadas de primer y segundo orden en cada uno de los vértices de la partición. Por su propio carácter interpolador, el valor de $C^{**}_{CÚBICA}$ en cada par (i, j) es el mismo que el de C^{**} , es decir, C^*_{ij} . Por tanto únicamente sería preciso determinar el valor de sus derivadas en cada (i, j) .

Por ser coherentes con el desarrollo presentado en el capítulo 4.2.3 de esta tesis, comenzaremos ajustando superficies de Ferguson, es decir, asumiendo que el valor de la derivada segunda de $C^{**}_{CÚBICA}$, a partir de ahora $C^{**}_{FERGUSON}$, en cada

nodo de la partición es 0, o sea,
$$\frac{\partial^2 C^{**}_{FERGUSON}(u_i, v_j)}{\partial u \partial v} = 0 \quad \forall i \in \{0, 1, 2, \dots, 5\} \quad \forall j \in \{0, 1, 2, \dots, 5\} .$$

Entonces, de acuerdo al teorema (4.4) cualquier relación de valores para las derivadas de primer orden

$$\frac{\partial C^{**}_{FERGUSON}(u_i, v_j)}{\partial u} \quad \text{y} \quad \frac{\partial C^{**}_{FERGUSON}(u_i, v_j)}{\partial v} \quad \forall i \in \{0, 1, 2, \dots, 5\} \quad \forall j \in \{0, 1, 2, \dots, 5\}$$

que sean solución al problema de programación no lineal entera “relajado” (4.26) va a generar una superficie interpoladora de Ferguson (ecuación (4.20)) que en sí misma es una función cópula.

De todo el espacio de posibles soluciones, vamos a tomar aquélla que minimiza la suma de las derivadas de primer orden. Si bien hubiera valido cualquier otra, la búsqueda de este objetivo pretende conseguir que las pendientes de los vectores tangentes a la superficie tengan menor pendiente y, en consecuencia, el resultado sea más suave.

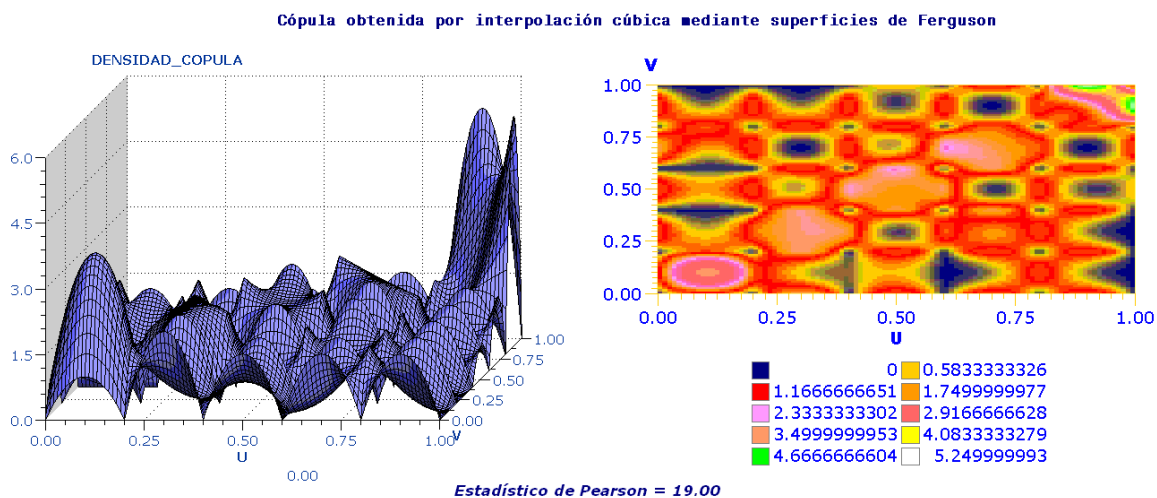


Ilustración 7.35: Densidad de la cópula obtenida por interpolación cúbica mediante superficies de Ferguson de la subcópula- V_{MIN} que optimiza el estadístico de Pearson ajustada a la muestra de entrenamiento en Febrero

El hecho de que la densidad cópula se obtenga a partir de la derivada segunda de $C^{**}_{FERGUSON}$ respecto de U y V , y de que hayamos forzado a que el valor de dicha derivada sea cero en todos los nodos de la partición, permite entender las drásticas caídas a cero que en dichos puntos reflejan los gráficos anteriores. Estos picos son por tanto consecuencia de haber ajustado una superficie de Ferguson. Con vistas a conseguir una curva más suave, vamos a imponer que el valor de la derivada segunda en los vértices de la partición de la nueva superficie a determinar $C^{**}_{CÚBICA}$, a partir de ahora

$C^{**}_{HERMITE}$, se mueva en consonancia con el nivel que presente esa misma superficie en un entorno de dichos vértices. Una aproximación a este nivel puede venir dada por el volumen a través de $C^{**}_{HERMITE}$ de los rectángulos que comparten ese vértice. Según se explica en el apartado 4.2.3.2.2, dado un vértice (u_i, v_j) un posible valor de referencia al que

aproximar $\frac{\partial^2 C^{**}_{HERMITE}(u_i, v_j)}{\partial u \partial v}$ sería $Ref_{V_B}(u_i, v_j) = \frac{V_{C^{**}(B)}}{\frac{1}{5} * \frac{1}{5}}$ siendo B el rectángulo de la partición cuyo vértice

inferior izquierdo es (u_i, v_j) es decir $B = [u_i, u_{i+1}] \times [v_j, v_{j+1}]$ (véase la ecuación (4.33)). En dicho apartado se proponía un segundo valor de referencia dado por la ecuación (4.34) (aproximación de Bernstein), pero éste ha sido finalmente desestimado pues la superficie obtenida tenía un aspecto más irregular.

Esta consideración conduce a que ahora, la búsqueda de suavidad implicará no sólo a las pendientes de los vectores tangentes a la superficie dadas por las derivadas primeras, sino también a las derivadas de segundo orden que intentaremos aproximar lo más posible a los valores de referencia citados. Esta aproximación se plantea a través de una función distancia que se convierte en un nuevo objetivo a minimizar en nuestro problema de programación fraccional dándole a éste la catalogación de problema multiobjetivo. El tratamiento de este problema se ha llevado a cabo a través del método de las ponderaciones, dando peso 0.5 a cada uno de los dos objetivos y resolviéndolo como si de un problema uniobjetivo se tratara. De acuerdo al teorema (4.5) cualquier relación de valores para las variables de decisión

$$\frac{\partial C^{**}_{HERMITE}(u_i, v_j)}{\partial u}, \frac{\partial C^{**}_{HERMITE}(u_i, v_j)}{\partial v} \text{ y } \frac{\partial^2 C^{**}_{HERMITE}(u_i, v_j)}{\partial u \partial v} \quad \forall i \in \{0,1,2,\dots,5\} \quad \forall j \in \{0,1,2,\dots,5\}$$

que sean solución al problema de programación no lineal entera multiobjetivo “relajado” (4.36), y en particular la óptima, va a generar una superficie interpoladora de Hermite (ecuación (4.19)) que en sí misma es una función cópula y que conserva el valor 19.00 del estadístico de Pearson. La optimalidad del problema no va a proporcionar un valor mejor para el estadístico de Pearson sino que pretende generar una cópula de aspecto más suave, condición que parece deseable a la hora de emplear ésta con fines predictivos. Los gráficos siguiente dan dos visiones distintas de la densidad cópula asociada a $C^{**}_{HERMITE}$.

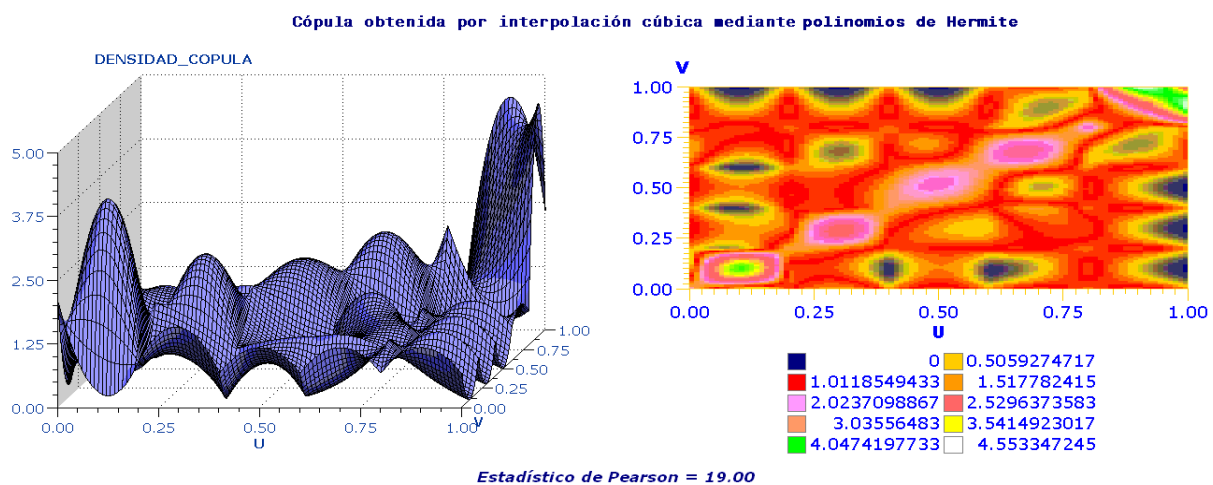


Ilustración 7.36: Densidad de la cópula obtenida por interpolación cúbica mediante polinomios de Hermite de la subcópula- V_{MIN} que optimiza el estadístico de Pearson ajustada a la muestra de entrenamiento en Febrero

Aun cuando el aspecto de esa densidad sigue presentando irregularidades, en el gráfico de contorno encontramos una primera diferencia que salta claramente a la vista con respecto a $C^{**}_{FERGUSON}$ (ilustración (7.35)) y es la ausencia de los pequeños círculos en color azul marino (representativo del valor 0) que existían en cada uno de los nodos de la partición. Nuevamente vemos dos indicios claros de correlación en el gráfico en respuesta a la concentración de puntos que se observa en la ilustración (7.30): una correlación más fuerte entre las colas derechas de las distribuciones y otra de intensidad ligeramente menor entre las colas izquierdas. La simulación de valores de esta cópula se puede llevar a cabo a través de la expresión (5.22) tal y como se explica en el apartado 5.3.

7.2.5.3.3 Construcción de una cópula por interpolación mediante polinomios de Bernstein

La última cópula que vamos a construir $C^*_{BERNSTEIN}$ utiliza polinomios de Bernstein (expresión (4.10)) como medio de aproximación al dominio de definición de la subcópula C^* aunque éstos no permiten hacer una interpolación exacta de ella. Esto supone obviamente que, como ya nos sucediera con C^{**}_{CUBICA} , tampoco va a ser posible alcanzar el valor óptimo del estadístico de Pearson obtenido para C^* , si bien a cambio la ventaja es que, por la propia definición de esta cópula, su densidad asociada presenta un aspecto muy suave, lejos del comportamiento escalonado y sinuoso que mostraban los ajustes anteriores.

Para definir la cópula $C^*_{BERNSTEIN}$ basta aplicar la ecuación (4.11) sobre las ternas de valores (i, j, c^*_{ij}) e igualmente a partir de ellos se puede calcular la densidad cópula correspondiente dada por la ecuación (4.12) y cuyo resultado se muestra a continuación:

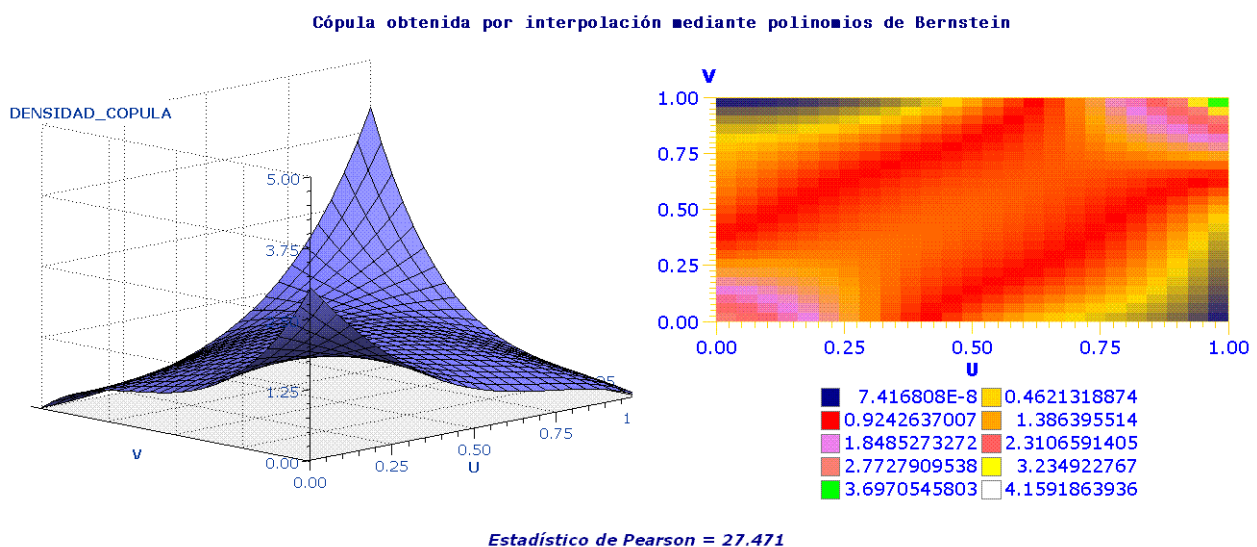


Ilustración 7.37: Densidad de la cópula obtenida por interpolación mediante polinomios de Bernstein de la subcópula que optimiza el estadístico de Pearson ajustada a la muestra de entrenamiento en Febrero

La suavidad a la que hacíamos referencia es la principal propiedad que observamos a la vista de uno y otro gráfico. Además, el carácter empírico de la cópula vuelve a quedar puesto de manifiesto. Como ocurriera con $C^*_{BILINEAL}$,

$C^{**}_{FERGUSON}$ y $C^{**}_{HERMITE}$ la densidad cópula enfatiza la correlación en aquellos rectángulos de la partición donde existe una mayor concentración de puntos. En coherencia con las cópulas C^*_{HRT} y C^*_{GUMBEL} , esta correlación es más fuerte entre las colas derechas de las distribuciones de X e Y y algo más débil entre los extremos izquierdos de ellas. El valor del estadístico de Pearson ha vuelto a crecer un poco respecto al de C^{**}_{CUBICA} y se distancia más del obtenido para las familias uniparamétricas. Sin embargo, hemos considerado que el aspecto suave de su densidad es una característica interesante y diferenciadora respecto al de las otras cópulas interpoladoras construidas, constituyendo una razón de suficiente peso, para plantearnos su empleo como posible alternativa para llevar a cabo la simulación de valores de la variable *demanda*. En este caso este proceso resulta algo más complicado pues, si bien la cópula condicionada responde a una expresión conocida (véase (5.23)), su inversión se antoja complicada. En su lugar, para realizar la simulación de un valor se puede proceder como se detalla en el apartado 5.4, generando un valor aleatorio “ p ” comprendido entre 0 y 1, igualando la expresión de la condicionada a dicho valor, y resolviendo en la variable V mediante métodos numéricos.

7.2.5.4 Comparativa de resultados entre las diferentes cópulas candidatas

A continuación, se presentan varios histogramas que reflejan la distribución de la demanda de gas en Madrid esperada para el día 23 de Febrero de 2005 en el supuesto de que dicho día se produjera una bajada de diez grados centígrados respecto del valor histórico mediano de las temperaturas máximas registradas en ese mes, que recordemos es 13.7°C (véase tabla (7.1)). Cada histograma es el resultado producido por una cópula distinta: C^*_{HRT} , C^*_{GUMBEL} , $C^*_{BILINEAL}$, $C^{**}_{BERNSTEIN}$ y C^{**}_{CUBICA} . Con la idea de poder comparar estas distribuciones todos los histogramas han sido anclados en un mismo origen y se ha considerado un número común de cajas de la misma amplitud. La línea punteada azul refleja la predicción que proporciona en condiciones normales de temperatura el modelo (7.13). La línea de color rojo marca la mediana de los valores simulados para la demanda supuesto un descenso de temperatura de 10°C. Finalmente la línea de color negro muestra el valor real de la demanda que se presentó dicho día. Obviamente, esta última junto con la primera son fijas en todos los gráficos. También se muestra el error con signo en el que se incurriría si se tomara como valor predicho para la demanda la mediana de la distribución.

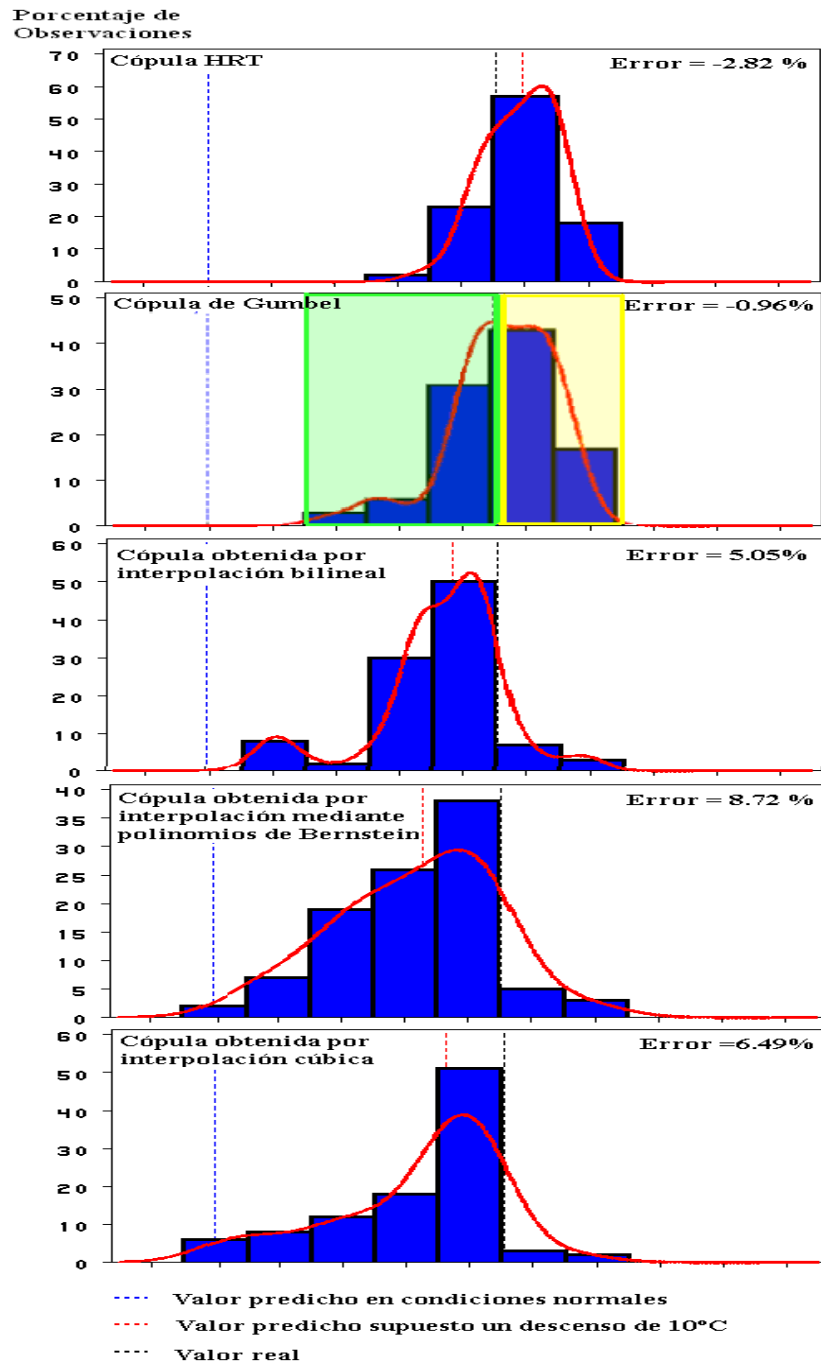


Ilustración 7.38: Distribución de las predicciones simuladas para la demanda mediante distintas familias de cópulas cuando la temperatura máxima desciende 10°C respecto de su valor normal para Febrero

A la vista de los histogramas, parece que son las cópulas C_{GUMBEL}^* y C_{HRT}^* las que proporcionan mejores resultados con unos márgenes de error sorprendentemente buenos. Las predicciones en mediana dadas por las cópulas interpoladoras quedan más distantes del valor real de la demanda. De todas maneras no debemos sacar conclusiones de forma precipitada. Puede ser que el valor mediano sugerido por las cópulas interpoladoras sea realmente más fiable que el de las otras dos pero que el dato registrado el 23 de Febrero haya sido un valor atípico y extremo como anuncian las colas de dichas distribuciones. Es decir, puede ser que el dato real de ese día sea un valor realmente alto que debería haber

sido capturado por el máximo de las distribuciones que ofrecen las cópulas C_{HRT}^* y C_{GUMBEL}^* y no así por su valor mediano respecto del cual estamos planteando la predicción. Por ello, de momento consideraremos importante el hecho de que el dato real (en negro) siempre esté comprendido dentro del rango de variación de todas las distribuciones presentadas. A raíz de este comentario, un posible análisis sería cuantificar en cuántas ocasiones el dato real queda incluido entre el mínimo y el máximo de la distribución de valores simulados por cada función cópula.

Para medir definitivamente cuál de las cópulas genera mejores predicciones en el mes de Febrero de 2005 vamos a hacer predicciones para todos aquellos días de ese mes en los que el descenso de la temperatura con respecto al valor establecido como normal fue al menos de 3°C. El número de datos según muestra la tabla (7.14) es 16, el mayor sujeto a estas circunstancias en comparación con el resto de los meses. De hecho ha sido ésta la razón que nos ha llevado a realizar el estudio sobre este mes. Así, para cada uno de esos días, vamos a hacer el mismo análisis que para el 23 de Febrero, es decir, sabiendo de antemano el número de grados que bajó la temperatura real, haremos las simulaciones de incrementos de demanda condicionando a ese valor numérico. Los criterios que vamos a establecer para medir la bondad de las cópulas candidatas son los siguientes:

1. **Valores reales en rango.-** Vamos a medir el número de veces en el que el dato real de consumo queda comprendido entre el mínimo y el máximo de los valores simulados. Premiaremos a las dos cópulas que presenten un mayor número de valores en rango.
2. **Fiabilidad de la distribución.-** Vamos a contar cuántos de los valores reales están cerca de los distintos percentiles de la distribución conformada por las simulaciones realizadas. Si el valor real cae entre el mínimo y el cuartil 1 de la distribución de la predicción lo contaremos dentro del grupo *BAJA* (parte baja de la distribución); si cae entre los cuartiles 1 y 3 lo incluiremos en el grupo *MEDIA* (parte media de la distribución); si cae entre el cuartil 3 y el percentil 95 lo asociaremos al grupo *ALTA* (parte alta de la distribución); si cae por encima del percentil 95, el valor corresponderá al grupo *EXTRM* (parte extrema de la distribución). Adicionalmente existen los grupos *BMIN* (bajo mínimo) y *SMAX* (sobre máximo) que harán referencia a aquellos valores reales que caen por debajo del valor mínimo de la distribución y por encima del valor máximo. La fiabilidad de la distribución dependerá de los tres factores siguientes:
 - Los valores reales deben estar concentrados cerca de la mediana, es decir, en la clase *MEDIA* el número de valores reales debe ser mayor o igual que en la suma de las clases *BMIN* y *BAJA* y también mayor o igual que en la suma de las clases *ALTA*, *EXTRM* y *SMAX*. La interpretación sería que cuando la temperatura baja un determinado número de grados, el valor real cae fundamentalmente en el rango de mayor densidad de la distribución (el rango esperado).
 - Interesa que los valores reales estén más concentrados en la cola derecha que en la izquierda, es decir, deben caer más veces entre los grupos *ALTA*, *EXTRM* y *SMAX* que entre *BMIN* y *BAJA*. De no ser

así, pudiera darse el caso de que las predicciones tendiesen a un máximo que realmente no se alcanza nunca y no parece conveniente alarmar a la compañía ante la posibilidad de un máximo que nunca llega a presentarse. O dicho de otra manera, parece fácil establecer un máximo si éste es inalcanzable. Por ello, debe existir cierta inclinación de los valores reales hacia la cola derecha de la distribución de las simulaciones (debe ser más pesada).

- Dentro de la cola derecha, el número de valores reales en el grupo *ALTA* debe ser mayor o igual que en el grupo *EXTRM* y esta misma relación debe guardar este último respecto del grupo *SMAX*, es decir, la frecuencia de valores en la cola derecha debe ser decreciente. De esta forma, el número de valores reales en cada clase de la cola derecha se moverá en consonancia con la densidad de las mismas.

Premiaremos a las dos cópulas que, de manera significativa, mejores resultados presenten respecto de cada uno de los criterios.

3. **Calidad de las predicciones.-** Vamos a calcular las medias de los errores de predicción cometidos. Se proponen 3 medidas:

- Media de los errores porcentuales absolutos de todos los días simulados (16 datos). Estos errores se miden, para cada día, comparando el dato real y el valor mediano de la distribución de la predicción que es el que se utilizará como referencia de demanda esperada para ese día.
- Media de los errores porcentuales absolutos de los días que caen en el grupo *MEDIA*. Estos errores se miden para cada día del grupo *MEDIA* comparando el dato real y el valor mediano de la distribución de la predicción. Supuesto que en un porcentaje alto de los casos el dato real estará cerca del valor mediano de la distribución, se trata de cuantificar el error entre ambos valores.
- Media de los errores porcentuales absolutos de los días que caen en los grupo *EXTRM* y *SMAX*. Estos errores se miden para cada día de los grupo *EXTRM* y *SMAX* comparando el dato real y el valor máximo de la distribución de la predicción. Se trata de cuantificar el error que se comete al dar como predicción el valor máximo de la distribución.

De igual manera que antes premiaremos a las dos cópulas que, de manera significativa, mejores resultados den para cada uno de estos criterios.

De acuerdo a lo expuesto, queda patente nuestro interés en medir la calidad de las predicciones en términos de sus distribuciones y no únicamente en función de su capacidad para acertar de forma puntual. Una de las principales ventajas que aporta el empleo de cópulas frente a otro tipo de metodologías más convencionales es la posibilidad de proporcionar resultados en términos de distribuciones que no tienen porqué ser necesariamente simétricas. Queremos destacar (véase por ejemplo la cópula de Gumbel en la ilustración (7.38)) el hecho de que, con independencia de la cópula que se utilice, los valores de las colas derechas (en amarillo) se encuentren más concentrados respecto de la mediana (que se propone como predicción) que los de las colas izquierdas (en verde), aun cuando por definición de la propia mediana (percentil 50), el número de ellos situados en una y otra cola será el mismo. Esta circunstancia resulta de gran relevancia para la compañía gasista pues le permitirá tener más controlada (acotada) la incertidumbre asociada al pico máximo de demanda que, no debemos de olvidar, es el verdadero objetivo de este estudio.

Habiendo realizado este apunte, presentamos a continuación el resultado obtenido para el mes de Febrero de 2005 respecto de cada uno de los 3 criterios establecidos.

Valores reales en rango

De acuerdo a este primer criterio existe una cópula claramente desfavorecida que es la ajustada mediante interpolación cúbica. En dicho caso, 5 de los 16 datos reales no se encuentran comprendidos entre el mínimo y el máximo valor predicho. Por tanto, de acuerdo a este primer criterio podríamos empezar descartando dicha cópula. Para las restantes cópulas, los valores son muy parecidos siendo tan sólo mejores en una unidad (un día) para las cópulas obtenidas por interpolación de C^* que para las pertenecientes a las familias de Gumbel y HRT. Tal vez lo más justo sería premiar con un punto a todas las cópulas salvo a C^{**}_{CUBICA} , aun cuando las que mejores resultados consiguen son las otras dos interpoladoras.

FAMILIA	EN_RANGO		Total
	0	1	
BER	1 6.25	15 93.75	16
BIL	1 6.25	15 93.75	16
CUB	5 31.25	11 68.75	16
GUM	2 12.50	14 87.50	16
HRT	2 12.50	14 87.50	16

Tabla 7.19: Evaluación del criterio “Conteo de valores en rango” para las cópulas candidatas a reflejar la relación entre el incremento de la demanda y el incremento de la temperatura en Febrero

Fiabilidad de la distribución

Evaluemos los tres factores que determinan la bondad de este criterio:

- Respecto del primero, las cópulas $C^*_{BERNSTEIN}$ y C^{**}_{CUBICA} son las que presentan más elementos (7) entre el cuartil 1 y el cuartil 3 (clase *MEDIA*) pero en ningún caso llegan a suponer al menor el 50% de los datos reales.
- Respecto del segundo, todas las cópulas verifican tener más valores a partir del cuartil 3 (grupos *ALTA*, *EXTRM* y *SMAX*) que por debajo del cuartil 1 (grupo *BAJA*), dándose el efecto asimétrico esperado. Las dos que más peso tienen en dicha cola son $C^*_{BERNSTEIN}$ y C^{**}_{CUBICA} pero tal vez fuera injusto penalizar a las distribuciones C^*_{HRT} y C^*_{GUMBEL} . Sólo en el caso de $C^*_{BILINEAL}$, el peso de la cola izquierda parece excesivo.
- Respecto del tercero, sólo la cópula $C^*_{BERNSTEIN}$ respeta el orden decreciente en cuanto al número de valores ubicados en las clases *ALTA*, *EXTRM* y *SMAX*.

Por tanto, premiaríamos con 2 puntos a $C^*_{BERNSTEIN}$ y con 1 a todas las demás salvo a $C^*_{BILINEAL}$ que no recibiría ningún punto.

FAMILIA	GRUPO					Total
	BAJA	MEDIA	ALTA	EXTRM	SMAX	
BER	0 0.00	7 43.75	6 37.50	2 12.50	1 6.25	16
BIL	6 37.50	2 12.50	2 12.50	5 31.25	1 6.25	16
CUB	0 0.00	7 43.75	3 18.75	1 6.25	5 31.25	16
GUM	3 18.75	6 37.50	2 12.50	3 18.75	2 12.50	16
HRT	4 25.00	5 31.25	2 12.50	3 18.75	2 12.50	16

Tabla 7.20: Evaluación del criterio “Fiabilidad de la distribución” para las cópulas candidatas a reflejar la relación entre el incremento de la demanda y el incremento de la temperatura en Febrero

Calidad de las predicciones

Respecto del error medio absoluto de predicción los resultados son los siguientes:

- Teniendo en cuenta todas las predicciones realizadas, los mejores resultados corresponden a las cópula C_{HRT}^* y C_{GUMBEL}^* .
- Si la medición se hace respecto de los días que caen en el grupo *MEDIA*, la media de error más baja se registra para las cópulas C_{HRT}^* y $C_{BERNSTEIN}^*$.
- En cuanto a los errores de predicción de los grupos *EXTRM* y *SMAX* (respecto del valor máximo de la distribución) el mejor resultado se obtiene para C_{HRT}^* y C_{GUMBEL}^* .

Así, como resultado de este criterio asignaríamos 3 puntos a la cópula C_{HRT}^* , 2 a C_{GUMBEL}^* y 1 a $C_{BERNSTEIN}^*$.

FAMILIA	ERROR_A_MEDIANA_TOTAL	ERROR_A_MEDIANA	ERROR_A_MAXIMO
BER	8.78	1.71	20.03
BIL	9.52	2.33	15.65
CUB	9.90	4.08	17.36
GUM	8.02	2.28	13.76
HRT	7.88	1.26	13.24

Tabla 7.21: Evaluación del criterio “Errores de predicción” para las cópulas candidatas a reflejar la relación entre el incremento de la demanda y el incremento de la temperatura en Febrero

El recuento de todos los puntos, daría como ganadora a la cópula C_{HRT}^* con 5 puntos, seguidas de $C_{BERNSTEIN}^*$ y C_{GUMBEL}^* con 4 y finalmente C_{CUBICA}^* y $C_{BILINEAL}^*$ con 1. Por tanto, en el mes de Febrero de 2005, la cópula que proporcionaría el mejor ajuste sería C_{HRT}^* , lo cual nos llevaría a sopesar la conveniencia de ésta para realizar predicciones en dicho mes del año. También podríamos plantearnos utilizar $C_{BERNSTEIN}^*$ atendiendo a que es la cópula que presenta mejores puntuaciones respecto del criterio “Valores en rango” y del primero de los criterios de “Fiabilidad de la distribución”, sin haber recibido un premio adicional por ninguno de ellos. Será la validación sobre el mes de Febrero de 2006 la que incline definitivamente la balanza a favor de una u otra cópula.

A continuación vamos a presentar estos mismos resultados para cada uno de los 9 meses invernales restantes comprendidos en los dos años del horizonte de predicción que hemos planteado. Dado que para un mismo mes y año es muy pequeño el número de días en los que la temperatura baja más de 3°C (véase por ejemplo el mes Noviembre de

2004 en la tabla (7.14)), se presentan los resultados por tipo de mes, es decir, en la categoría *Enero* estarían englobados Enero de 2005 y Enero de 2006. En cada caso, la propuesta de las cópulas candidatas (como en Febrero de 2005 eran C^*_{HRT} y C^*_{GUMBEL}) se plantea a partir de la evaluación del estadístico de Pearson sobre la nube de pares (x, y) asociados al mes en cuestión. La tabla siguiente presenta el orden en el que estas cópulas deberían ser tanteadas en función del valor del estadístico que para ellas se obtiene. Tan sólo se presentan las 3 primeras candidatas para cada uno de los meses.

ENERO			MARZO			NOVIEMBRE			DICIEMBRE		
COPULA	PEARSON	P_VALOR	COPULA	PEARSON	P_VALOR	COPULA	PEARSON	P_VALOR	COPULA	PEARSON	P_VALOR
HRT	16.3712	0.35782	NOR	16.5354	0.34739	GUM	18.3167	0.24638	HRT	9.3242	0.85996
GUM	17.0177	0.31781	GUM	18.6692	0.22912	PLA	18.3952	0.24246	GUM	10.9499	0.75614
NOR	18.0096	0.26216	F15	21.4659	0.12259	CLA	19.0401	0.21192	NOR	12.6324	0.63067

Tabla 7.22: Evaluación del estadístico de Pearson para una serie de cópulas ajustadas a cada uno de los meses invernales de la muestra de entrenamiento

La conclusión que podemos sacar a la vista de las tablas anteriores ((7.16) y (7.22)) es que durante los meses claramente invernales (Diciembre, Enero y Febrero), las cópulas C^*_{HRT} y C^*_{GUMBEL} (en este orden) deberían ser las primeras en ser utilizadas dado que ocupan siempre las dos primeras posiciones. Para los meses de Enero y Diciembre, la cópula gaussina se plantea como la tercera alternativa, mientras que para Febrero esta tercera posición está reservada a la familia de Plackett.

Para los meses que pudiéramos considerar de transición con el Otoño (Noviembre) y la Primavera (Marzo), la familia de Gumbel sigue ocupando los primeros puestos, pero el representante de la familia de Cola Derecha Pesada desaparece, permitiendo que aparezcan nuevamente la cópula Normal y la de Plackett.

Así, la familia de Gumbel perteneciente a la clase de valor extremo, aporta siempre con independencia del mes de estudio, un representante para analizar la relación de dependencia entre la variable de demanda y la climatológica. Bien es verdad que en aquellos casos en los cuales la familia de Cola Derecha Pesada aporta un representante, éste genera un p-valor algo más alto que permite no rechazar el contraste de bondad de ajuste con un mayor nivel de significatividad.

Finalmente, dado que es C^*_{NORMAL} la que en un mayor número de ocasiones ocupan el tercer puesto en litigio, será ésta la que participe juntamente con C^*_{HRT} , C^*_{GUMBEL} y las cópulas interpoladoras en la presentación final de resultados (tablas (7.23)-(7.27)). De acuerdo a ellos, las cópulas C^*_{HRT} , C^*_{GUMBEL} y C^{**}_{CUBICA} parecen ser las más apropiadas en líneas generales. Las dos primeras basan estos buenos resultados en la calidad de las predicciones que generan, mientras que la interpoladora, a parte de conseguir buenos resultados de predicción en Enero y Diciembre, justifica sus buenas puntuaciones en el criterio que hace referencia a la fiabilidad de la forma de la distribución.

Finalmente señalaremos que, respecto del debate que teníamos abierto entre la posibilidad de utilizar C^*_{HRT} o $C^*_{BERNSTEIN}$ para predecir en el mes de Febrero, la incorporación del año 2006 al conjunto de validación establece definitivamente como ganadora a la primera, siendo la principal razón de ello, la menor magnitud de los errores de predicción cometidos fuera de muestra (tercer criterio).

Mes de Enero

Valores reales en rango

FAMILIA	EN_RANGO		Total
	0	1	
BER	3 23.08	10 76.92	13
BIL	3 23.08	10 76.92	13
CUB	4 30.77	9 69.23	13
GUM	3 23.08	10 76.92	13
HRT	3 23.08	10 76.92	13
NOR	3 23.08	10 76.92	13

Fiabilidad de la distribución

FAMILIA	GRUPO						Total
	BMIN	BAJA	MEDIA	ALTA	EXTRM	SMAX	
BER	2 15.38	1 7.69	2 15.38	2 15.38	5 38.46	1 7.69	13
BIL	2 15.38	1 7.69	2 15.38	3 23.08	4 30.77	1 7.69	13
CUB	2 15.38	0 0.00	3 23.08	0 0.00	6 46.15	2 15.38	13
GUM	2 15.38	1 7.69	3 23.08	5 38.46	1 7.69	1 7.69	13
HRT	2 15.38	1 7.69	3 23.08	5 38.46	1 7.69	1 7.69	13
NOR	2 15.38	1 7.69	2 15.38	4 30.77	3 23.08	1 7.69	13

Calidad de las predicciones

FAMILIA	ERROR_A_MEDIANA_TOTAL	ERROR_A_MEDIANA	ERROR_A_MAXIMO
BER	13.17	1.60	21.44
BIL	11.96	2.70	20.61
CUB	13.77	1.27	19.58
GUM	11.44	5.05	23.40
HRT	10.98	4.82	22.51
NOR	12.82	1.78	22.93

Tabla 7.23: Evaluación de los criterios que permiten determinar la cópula que mejor refleja la relación entre el incremento de la demanda y el incremento de la temperatura en Enero

Familia	Criterio 1	Criterio 2	Criterio 3	Total
BER	1	0+1+0	0+1+0	3
BIL	1	0+1+0	0+0+1	3
CUB	1	0+1+0	0+1+1	4
GUM	1	0+1+0	1+0+0	3
HRT	1	0+1+0	1+0+0	3
NOR	1	0+1+1	0+0+0	3

Mes de Febrero

Valores reales en rango

FAMILIA	EN_RANGO		Total
	0	1	
BER	1 4.00	24 96.00	25
BIL	2 8.00	23 92.00	25
CUB	5 20.00	20 80.00	25
GUM	2 8.00	23 92.00	25
HRT	2 8.00	23 92.00	25
NOR	1 4.00	24 96.00	25

Fiabilidad de la distribución

FAMILIA	GRUPO						Total
	BMIN	BAJA	MEDIA	ALTA	EXTRM	SMAX	
BER	0 0.00	1 4.00	14 56.00	7 28.00	2 8.00	1 4.00	25
BIL	1 4.00	8 32.00	6 24.00	4 16.00	5 20.00	1 4.00	25
CUB	0 0.00	1 4.00	13 52.00	4 16.00	2 8.00	5 20.00	25
GUM	0 0.00	6 24.00	11 44.00	3 12.00	3 12.00	2 8.00	25
HRT	0 0.00	7 28.00	10 40.00	3 12.00	3 12.00	2 8.00	25
NOR	0 0.00	3 12.00	14 56.00	4 16.00	3 12.00	1 4.00	25

Calidad de las predicciones

FAMILIA	ERROR_A_MEDIANA_TOTAL	ERROR_A_MEDIANA	ERROR_A_MAXIMO
BER	7.21	2.71	20.03
BIL	7.89	2.54	15.65
CUB	8.00	3.62	15.98
GUM	6.50	2.21	13.76
HRT	6.48	1.81	13.24
NOR	6.44	3.15	15.05

Tabla 7.24: Evaluación de los criterios que permiten determinar la cópula que mejor refleja la relación entre el incremento de la demanda y el incremento de la temperatura en Febrero

Familia	Criterio 1	Criterio 2	Criterio 3	Total
BER	1	1+1+1	0+0+0	4
BIL	1	0+0+0	0+0+0	1
CUB	0	1+1+0	0+0+0	2
GUM	1	0+1+0	1+1+1	5
HRT	1	0+1+0	1+1+1	5
NOR	1	1+1+1	1+0+0	5

Mes de Marzo

Valores reales en rango

FAMILIA	EN_RANGO		Total
	0	1	
BER	0 0.00	16 100.00	16
BIL	0 0.00	16 100.00	16
CUB	3 18.75	13 81.25	16
GUM	0 0.00	16 100.00	16
HRT	3 18.75	13 81.25	16
NOR	0 0.00	16 100.00	16

Fiabilidad de la distribución

FAMILIA	GRUPO					Total
	BAJA	MEDIA	ALTA	EXTRM	SMAX	
BER	2 12.50	6 37.50	2 12.50	6 37.50	0 0.00	16
BIL	1 6.25	6 37.50	4 25.00	5 31.25	0 0.00	16
CUB	0 0.00	7 43.75	1 6.25	5 31.25	3 18.75	16
GUM	2 12.50	4 25.00	4 25.00	6 37.50	0 0.00	16
HRT	3 18.75	3 18.75	3 18.75	4 25.00	3 18.75	16
NOR	2 12.50	4 25.00	5 31.25	5 31.25	0 0.00	16

Calidad de las predicciones

FAMILIA	ERROR_A_MEDIANA_TOTAL	ERROR_A_MEDIANA	ERROR_A_MAXIMO
BER	13.87	5.36	23.67
BIL	10.92	3.79	21.33
CUB	15.17	6.25	24.22
GUM	9.45	1.27	15.76
HRT	8.99	0.68	14.14
NOR	10.15	0.85	18.67

Tabla 7.25: Evaluación de los criterios que permiten determinar la cópula que mejor refleja la relación entre el incremento de la demanda y el incremento de la temperatura en Marzo

Familia	Criterio 1	Criterio 2	Criterio 3	Total
BER	1	0+1+0	0+0+0	2
BIL	1	0+1+0	0+0+0	2
CUB	0	0+1+0	0+0+0	1
GUM	1	0+1+0	1+0+1	4
HRT	0	0+1+0	1+1+1	4
NOR	1	0+1+0	0+1+0	3

Mes de Noviembre

Valores reales en rango

FAMILIA	EN_RANGO	Total
	1	
BER	10 100.00	10
BIL	10 100.00	10
CUB	10 100.00	10
GUM	10 100.00	10
HRT	10 100.00	10
NOR	10 100.00	10

Fiabilidad de la distribución

FAMILIA	GRUPO				Total
	BAJA	MEDIA	ALTA	EXTRM	
BER	1 10.00	5 50.00	4 40.00	0 0.00	10
BIL	1 10.00	6 60.00	3 30.00	0 0.00	10
CUB	1 10.00	6 60.00	2 20.00	1 10.00	10
GUM	1 10.00	7 70.00	2 20.00	0 0.00	10
HRT	1 10.00	7 70.00	2 20.00	0 0.00	10
NOR	1 10.00	6 60.00	3 30.00	0 0.00	10

Cantidad de las predicciones

FAMILIA	ERROR_A_MEDIANA_TOTAL	ERROR_A_MEDIANA	ERROR_A_MAXIMO
BER	5.99	3.03	.
BIL	5.57	3.22	.
CUB	5.62	2.97	15.73
GUM	3.82	2.38	.
HRT	3.59	2.23	.
NOR	4.83	2.49	.

Tabla 7.26: Evaluación de los criterios que permiten determinar la cópula que mejor refleja la relación entre el incremento de la demanda y el incremento de la temperatura en Noviembre

Familia	Criterio 1	Criterio 2	Criterio 3	Total
BER	1	1+1+0	0+0+0	3
BIL	1	1+1+0	0+0+0	3
CUB	1	1+1+1	0+0+0	4
GUM	1	1+0+0	1+1+0	4
HRT	1	1+0+0	1+1+0	4
NOR	1	1+1+0	0+1+0	4

Mes de Diciembre

Valores reales en rango

FAMILIA	EN_RANGO		Total
	0	1	
BER	2 50.00	2 50.00	4
BIL	2 50.00	2 50.00	4
CUB	2 50.00	2 50.00	4
GUM	2 50.00	2 50.00	4
HRT	2 50.00	2 50.00	4
NOR	2 50.00	2 50.00	4

Fiabilidad de la distribución

FAMILIA	GRUPO			Total
	BMIN	BAJA	MEDIA	
BER	2 50.00	0 0.00	2 50.00	4
BIL	2 50.00	2 50.00	0 0.00	4
CUB	2 50.00	0 0.00	2 50.00	4
GUM	2 50.00	1 25.00	1 25.00	4
HRT	2 50.00	1 25.00	1 25.00	4
NOR	2 50.00	1 25.00	1 25.00	4

Calidad de las predicciones

FAMILIA	ERROR_A_MEDIANA_TOTAL	ERROR_A_MEDIANA	ERROR_A_MAXIMO
BER	7.63	2.98	.
BIL	10.41	.	.
CUB	7.52	2.59	.
GUM	8.28	2.71	.
HRT	8.39	2.75	.
NOR	8.14	2.67	.

Tabla 7.27: Evaluación de los criterios que permiten determinar la cópula que mejor refleja la relación entre el incremento de la demanda y el incremento de la temperatura en Diciembre

Familia	Criterio 1	Criterio 2	Criterio 3	Total
BER	0	1+0+0	1+0+0	2
BIL	0	0+0+0	0+0+0	0
CUB	0	1+0+0	1+1+0	3
GUM	0	0+0+0	0+1+0	1
HRT	0	0+0+0	0+1+0	1
NOR	0	0+0+0	0+1+0	1

7.3 Predicción de la demanda de gas natural a corto plazo

7.3.1 Introducción al problema

El siguiente capítulo del bloque práctico de esta tesis versa sobre la predicción diaria de la demanda de gas a corto plazo. Una de las principales necesidades de disponer de una buena previsión a este horizonte es conseguir adaptar el ritmo uniforme y regular que imponen las suministradoras de gas a la gran heterogeneidad de demanda que a corto plazo se puede esperar por parte de los clientes finales en función de su perfil de consumo. No sólo deben tenerse en cuenta los recursos de gas disponibles sino los contratos firmados con comercializadoras y suministradoras que pueden implicar por ejemplo la movilización de flotas completas de buques metaneros (véase ilustración (7.39)) que traen el gas desde Argelia.

La dificultad para mantener la calidad del gas (presión) dentro del gasoducto justifica la necesidad de que la predicción no exceda en demasía el dato real de demanda: conforme pasa el tiempo, el gas pierde calidad, calienta menos y, en consecuencia se necesita más volumen para producir el mismo calor, con la consiguiente pérdida económica para la compañía. Si por el contrario el dato predicho se sitúa significativamente por debajo del real, los clientes industriales podrían ver reducidas sus tasas de producción y los domésticos plantearse el empleo de otra energía alternativa como forma de calefacción para su hogar. Si bien existen tanques de almacenamiento de gas y aun cuando las infraestructuras del sistema gasista español han sido dimensionadas buscando cierta flexibilidad en términos de regasificación y transporte, esta holgura es menor que la que existe en otros países como Francia o Italia. Por ello, son mayores las necesidades de ajustar modelos de predicción muy fiables, que proporcionen unas previsiones de demanda óptimas para garantizar el buen funcionamiento del transporte al menor coste posible.

La predicción de demanda a corto plazo es un problema clásico entre las compañías del sector energético. Existen multitud de técnicas contrastadas que proporcionan buenos resultados entre las que gozan de mayor popularidad los modelos ARIMA basados en la metodología Box-Jenkins. En este capítulo, nos basaremos en la idea propuesta para el tratamiento a medio plazo de las series de demanda de gas mediante funciones cópula, y propondremos un método de predicción alternativo a estos modelos clásicos.



Ilustración 7.39: Buque metanero descargando gas en la planta de Huelva

7.3.2 Planteamiento de la solución

Ya en el apartado 7.2.2 anunciábamos la conveniencia de emplear modelos ARIMA para abordar un problema de predicción de demanda de gas o electricidad a corto plazo frente a otras alternativas de modelización y argumentábamos que la razón fundamental es que estos modelos permiten identificar rasgos muy característicos de este tipo de series como son:

- La **tendencia** recogida por los parámetros asociados a los primeros retardos de la variable de consumo (diferencias regulares o parámetros autorregresivos de órdenes bajo).
- La **variabilidad** que puede cambiar (heterocedasticidad) o no (homocedasticidad) con el tiempo. En el primer caso suele ser necesario someter a la serie a algún tipo de transformación (mediante la función logarítmica o raíz cuadrada). Raramente la volatilidad de la serie es tal que resulta conveniente completar el proceso de ajuste con un tratamiento de la varianza del término error a través de un modelo **GARCH** (autorregresivo condicionalmente heterocedástico generalizado).
- La influencia del **calendario** tanto en el segmento doméstico (descenso de la demanda por desuso del gas en la residencia habitual durante periodos festivos) como en el industrial (caída en esos mismos periodos de la actividad de fábricas que emplean de forma regular el recurso energético). Parte de este efecto “calendario” será reflejado a través de los órdenes estacionales del modelo que permiten basar la predicción en el día de la semana anterior. De esta forma quedarán diferenciados los fines de semana del resto de días (por lo general, laborables). Otra parte más difícil de modelizar va asociada a situaciones anómalas que deben ser abordadas mediante un análisis de intervenciones traducido en la incorporación de variables binarias en el ARIMA:
 - El efecto de festividades aisladas puede ser identificado mediante intervenciones tipo pulso (AO = *Additive Outlier* o Atípico Aditivo).
 - El efecto de puentes también puede ser recogido mediante *outliers* tipo AO asociados al puente o, alternativamente, con *outliers* tipo TC (*Temporary Change* o Cambio Temporal) asociados a una festividad aislada que supone el impacto inicial en la serie y cuyo efecto decae de manera exponencial y amortiguada en los días siguientes (el del puente).
 - El efecto de un periodo vacacional se puede capturar mediante intervenciones tipo escalón (LS = *Level Shift* o Cambio de Nivel) que afecta a un grupo de días (Navidad, Agosto, S. Santa, etc.).

- En el sector residencial los **agentes climáticos** desempeñan un papel determinante pudiendo ser incluidos en la ecuación del ARIMA a través de modelos de función de transferencia. Ya comentamos en el apartado 7.2.3 que, después de haber contrastado la influencia de las distintas variables que nos proporcionó el Instituto Nacional de Meteorología asociadas a las estaciones más cercanas a la zona en que es medido el consumo de gas, fueron las variables de temperatura (máxima de Barajas y en menor medida, mínima de Cuatro Vientos) las que proporcionaron mejoras sustanciales en términos de predicción. Otras que en un principio se contemplaron pero que finalmente fueron descartadas por no resultar significativas y jugar en contra de la parsimonia buscada para el modelo son, por orden decreciente de importancia: la velocidad del viento, la humedad relativa del aire, la nubosidad (medida como la proporción de cielo cubierto) y las precipitaciones.

En resumen las predicciones a corto plazo de demanda doméstica de gas pueden obtenerse a partir de un modelo ARIMA debidamente complementado con un análisis de intervenciones para reflejar el efecto del calendario y la inclusión de modelos de función de transferencia asociados a las variables de temperatura. El criterio principal de decisión ante el abanico de técnicas existentes es obviamente la calidad de los resultados, pero también conviene citar otros como por ejemplo la facilidad con que pueden ser interpretados los parámetros que participan en el modelo o, por poner otro ejemplo, la estabilidad de los mismos, circunstancia que permite reducir el tiempo dedicado a su mantenimiento y distanciar los periodos dedicados a su revisión.

La metodología que vamos a proponer consiste en suplir el empleo de las funciones de transferencia por el de funciones cópula, para modelizar el efecto de la temperatura máxima de Barajas y la mínima de Cuatro Vientos. Partiremos de una predicción inicial dada por un modelo ARIMA y estudiaremos la relación de dependencia del proceso residual resultante del ajuste con estas variables de temperatura. Obsérvese que este planteamiento mantiene un claro paralelismo con el desarrollado en el tratamiento a medio plazo. Tenemos por un lado una variable asociada a la demanda medida en términos porcentuales que en este caso es el error de un modelo y que en el caso anterior era el incremento porcentual o por qué no decirlo, el error, entre la demanda real y la que estimábamos en condiciones normales de temperatura. Por otro lado, dispondremos de una variable de temperatura que en el modelo a medio plazo era la desviación con respecto a un valor estándar establecido para cada mes y que en este modelo será la diferencia en grados con respecto al día anterior, es decir, el número de grados centígrados que la temperatura ha bajado o subido de un día a otro. Nuevamente ambas variables eliminan la posible dependencia temporal cuya existencia nos obligaría, como también hemos mencionado en alguna ocasión, a tener ciertas precauciones si pretendemos utilizar funciones cópula.

Nuestro objetivo final es comparar los resultados que se obtienen a través de un modelo ARIMA en el que las variables de incremento de temperatura intervienen como modelos de función de transferencia y otro modelo ARIMA en el que el efecto que aportan estas funciones es incorporado a través de funciones cópula. Empezaremos ajustando este modelo ARIMA que sirve de punto de partida a ambas propuestas de solución.

7.3.2.1 Ajuste de una serie mediante un ARIMA complementado con modelos de función de transferencia

Para ajustar el modelo ARIMA de partida vamos a seguir cada una de las fases que componen la metodología Box-Jenkins y que fueron descritas en el apartado 6.1. Para empezar vamos a empezar distinguiendo entre un conjunto de entrenamiento con datos comprendidos entre el 1 de Noviembre de 1997 y el 31 de Octubre de 2005 y otro de validación cuyos datos se mueven entre el 1 de Noviembre de 2005 y el 31 de Marzo de 2006 que es el último invierno gasista del que disponemos y para el que parece tener sentido plantear la predicción fuera de muestra por ser el más reciente. Justificado así el hecho de que tomemos el 31 de Octubre como fecha de cierre del conjunto de entrenamiento diremos que, si bien se dispone de datos desde el 1 de Abril de 1997, hemos preferido tomar como fecha de inicio de dicho conjunto el día 1 de Noviembre con la idea de que en el ajuste sólo participen años completos, sin posibles descompensaciones entre meses. A partir de aquí, podemos empezar a aplicar la metodología citada.

El primer paso de la denominada **fase de identificación** consiste en determinar si es preciso o no hacer una transformación sobre la variable a predecir con vistas a reducir su variabilidad. En el ajuste a medio plazo en el que esta variable era la demanda en condiciones normales de temperatura, el criterio de información de Akaike (ecuación (2.6)) sugería el empleo de la transformación raíz cuadrada (véase tabla (7.10)). En este caso, en que el histórico es el mismo pero sin hacer el depurado del efecto climático dicho criterio propone nuevamente hacer esta misma transformación, circunstancia que obviamente no es de extrañar.

LAMBDA	AIC
1.0	19445.28
0.5	18806.32
0.0	19101.94

Tabla 7.28: Evaluación del AIC para varias propuestas de transformación para la variable demanda

El segundo paso es determinar la necesidad de diferenciar la variable a predecir para ésta sea estacionaria en media. Para ello vamos a especificar los órdenes regulares y estacionales del modelo tanto en lo que se refiere a la parte autorregresiva (AR) como a la de medias móviles (MA) a partir de los resultados parciales que vayan ofreciendo los gráficos de autocorrelación (correlogramas) obtenidos conforme vayan siendo ajustados los parámetros del modelo. Estos parámetros se irán recalculando a medida que vayan siendo incluidos nuevos términos en la fórmula. Así, la posible necesidad de diferenciar la serie se contemplará a partir de las estimaciones obtenidas para los coeficientes de la parte autorregresiva en caso de que éstas se muevan suficientemente cerca de de 1.

Para empezar observaremos el gráfico de autocorrelación simple asociada a la serie transformada así como el resultado de contrastar si dicho proceso es o no ruido blanco. Podríamos decir que el gráfico corresponde al residuo del “modelo sin estructura”.

$$X(t) = \epsilon(t) \quad (7.29)$$

siendo $X(t) = \sqrt{\text{Demanda}(t)}$.

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	9999.99	6	<.0001	0.998	0.994	0.993	0.992	0.991	0.992
12	9999.99	12	<.0001	0.992	0.990	0.988	0.987	0.986	0.985
18	9999.99	18	<.0001	0.986	0.987	0.985	0.982	0.981	0.981
24	9999.99	24	<.0001	0.980	0.981	0.982	0.980	0.978	0.976

Autocorrelation Plot of Residuals																								
Lag	Covariance	Correlation	-1	9	8	7	6	5	4	3	2	1	0	1	2	3	4	5	6	7	8	9	1	Std Error
0	53.672396	1.00000												*****										0
1	53.538469	0.99750												*****										0.018499
2	53.375303	0.99446												*****										0.031989
3	53.284665	0.99278												*****										0.041233
4	53.225498	0.99167												*****										0.048732
5	53.194122	0.99109												*****										0.055208
6	53.232348	0.99180												*****										0.060994
7	53.257796	0.99228												*****										0.066284
8	53.136766	0.99002												*****										0.071186
9	53.008486	0.98763												*****										0.075752
10	52.949354	0.98653												*****										0.080037
11	52.910147	0.98580												*****										0.084096
12	52.888666	0.98540												*****										0.087962
13	52.932354	0.98621												*****										0.091662
14	52.963025	0.98678												*****										0.095224
15	52.848069	0.98464												*****										0.098661
16	52.726279	0.98237												*****										0.101969
17	52.672290	0.98137												*****										0.105158
18	52.636441	0.98070												*****										0.108247
19	52.618500	0.98036												*****										0.111246
20	52.666279	0.98125												*****										0.114165
21	52.700351	0.98189												*****										0.117015
22	52.589365	0.97982												*****										0.119802
23	52.466421	0.97753												*****										0.122514
24	52.407859	0.97644												*****										0.125155

Ilustración 7.40: Test de ruido blanco y correlograma para la serie original de demanda de gas

A la vista de este primer gráfico se deduce la necesidad de ajustar un modelo AR(1) o incluso, si fuese preciso, una diferencia regular I(1). Por tanto, un primer modelo sería:

$$(1 - \phi_1 * B) X(t) = \epsilon(t) \quad (7.30)$$

Una vez ajustado este modelo, el resultado es el siguiente:

Conditional Least Squares Estimation					
Parameter	Estimate	Standard Error	t Value	Approx Pr > t	Lag
AR1,1	0.99785	0.0012610	791.32	<.0001	1

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	616.32	5	<.0001	0.124	-0.293	-0.132	-0.108	-0.278	0.051
12	2095.15	11	<.0001	0.589	0.037	-0.279	-0.084	-0.067	-0.259
18	3360.96	17	<.0001	0.054	0.585	0.035	-0.273	-0.075	-0.066
24	4785.92	23	<.0001	-0.253	0.050	0.583	0.056	-0.259	-0.083

Autocorrelation Plot of Residuals																								
Lag	Covariance	Correlation	-1	9	8	7	6	5	4	3	2	1	0	1	2	3	4	5	6	7	8	9	1	Std Error
0	0.249288	1.00000														*****								0
1	0.030998	0.12435													. **									0.018499
2	-0.073106	-.29326													***** .									0.018783
3	-0.032785	-.13152													*** .									0.020290
4	-0.027024	-.10840													** .									0.020580
5	-0.069389	-.27835													***** .									0.020774
6	0.012604	0.05056													. *									0.022013
7	0.146796	0.58886													. *****									0.022053
8	0.0092761	0.03721													. *									0.026901
9	-0.069535	-.27893													***** .									0.026919
10	-0.020905	-.08386													** .									0.027891
11	-0.016748	-.06718													* .									0.027977
12	-0.064465	-.25860													***** .									0.028032
13	0.013399	0.05375													. *									0.028837
14	0.145772	0.58475													. *****									0.028871
15	0.0087700	0.03518													. *									0.032674
16	-0.068086	-.27312													***** .									0.032687
17	-0.018818	-.07549													** .									0.033459
18	-0.016556	-.06641													* .									0.033517
19	-0.063157	-.25335													***** .									0.033562
20	0.012538	0.05030													. *									0.034210
21	0.145305	0.58288													. *****									0.034235
22	0.014032	0.05629													. *									0.037478
23	-0.064665	-.25940													***** .									0.037507
24	-0.020716	-.08310													** .									0.038116

Ilustración 7.41: Test de ruido blanco y correlograma para el modelo AR(1) ajustado a la serie de demanda de gas

Lo primero que observamos es que la estimación del parámetro ϕ_1 sale muy próxima a 1, lo cual es un claro indicio de la necesidad de llevar a cabo una diferencia regular ante la posibilidad de que la serie no sea estacionaria en media, y teniendo en cuenta que los modelos ARMA deben ser ajustados sobre procesos estacionarios. A pesar de que el valor de la desviación típica de la estimación es pequeño, puede considerarse suficiente para que el valor 1 esté incluido en un intervalo centrado en $\hat{\phi}_1$ y cuyo radio sea un par de veces dicho valor.

Por otra parte también puede apreciarse que alrededor de cada uno de los retardos estacionales reflejados en el gráfico, aparece el patrón del modelo MA que cabría ajustar a la vista de los tres primeros valores del correlograma, es decir, en los retardos 6 y 8 (a una unidad de distancia de 7) vemos una pequeña correlación asociada a la autocorrelación de orden 1 (MA(1)); en los retardos 5 y 9 (a dos unidades de distancia) vemos una correlación más fuerte asociada a la

autocorrelación de orden 2, que también es más significativa (MA(2)); finalmente en los retardos 4 y 10 (a tres unidades de distancia) vemos otra correlación bastante significativa en concordancia con la autocorrelación de orden 3. Este comportamiento peculiar está identificando que la componente estacional se incorpora de manera multiplicativa sobre la regular y que por tanto hemos de proponer un modelo multiplicativo en vez de aditivo. Sin embargo, antes de empezar incorporando la parte estacional del modelo, vamos a sustituir el AR(1) por un I(1). Consideremos así el modelo

$$(1 - B)X(t) = \epsilon(t) \quad (7.31)$$

El resultado asociado a este modelo es el siguiente:

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	689.32	6	<.0001	0.127	-0.311	-0.139	-0.115	-0.296	0.052
12	2332.18	12	<.0001	0.618	0.040	-0.299	-0.092	-0.072	-0.271
18	3737.91	18	<.0001	0.057	0.614	0.036	-0.292	-0.079	-0.070
24	5309.42	24	<.0001	-0.266	0.051	0.609	0.058	-0.280	-0.088

Autocorrelation Plot of Residuals																									
Lag	Covariance	Correlation	-1	9	8	7	6	5	4	3	2	1	0	1	2	3	4	5	6	7	8	9	1	Std Error	
0	0.239524	1.00000																						0	
1	0.030341	0.12667																							0.018503
2	-0.074410	-.31066																							0.018797
3	-0.033377	-.13935																							0.020480
4	-0.027571	-.11511																							0.020802
5	-0.070927	-.29612																							0.021019
6	0.012341	0.05152																							0.022401
7	0.148072	0.61819																							0.022442
8	0.0095126	0.03971																							0.027664
9	-0.071626	-.29903																							0.027684
10	-0.021959	-.09168																							0.028768
11	-0.017344	-.07241																							0.028868
12	-0.065019	-.27145																							0.028930
13	0.013660	0.05703																							0.029789
14	0.147124	0.61424																							0.029827
15	0.0086864	0.03627																							0.033882
16	-0.070024	-.29235																							0.033895
17	-0.018959	-.07915																							0.034747
18	-0.016862	-.07040																							0.034809
19	-0.063602	-.26554																							0.034858
20	0.012266	0.05121																							0.035544
21	0.145920	0.60921																							0.035569
22	0.013960	0.05828																							0.038978
23	-0.067055	-.27995																							0.039007
24	-0.021045	-.08786																							0.039689

Ilustración 7.42: Test de ruido blanco y correlograma para el modelo I(1) ajustado a la serie de demanda de gas

A la vista de cualquiera de los dos gráficos anteriores queda clara la necesidad de ajustar un término autorregresivo de orden 7 que, en función de su magnitud, también puede derivar en una nueva diferenciación de la serie. Las correlaciones asociadas a los retardos que son múltiplos de 7 son muy significativas y decrecen lentamente hacia 0.

Según hemos observado, este término deberá incorporarse de manera multiplicativa al modelo y por ello, pasaremos a distinguir entre la parte regular y la parte estacional del mismo. Sea entonces el modelo SARIMA(0,1,0)x(1,0,0)₇:

$$(1-B)(1-\Phi_7 * B^7)X(t) = \epsilon(t) \quad (7.32)$$

El resultado del ajuste es el siguiente:

Conditional Least Squares Estimation					
Parameter	Estimate	Standard Error	t Value	Approx Pr > t	Lag
AR1,1	0.61912	0.01455	42.55	<.0001	7

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	207.87	5	<.0001	0.192	-0.099	-0.105	-0.045	-0.080	-0.069
12	404.16	11	<.0001	-0.233	-0.074	-0.066	0.014	0.024	-0.046
18	485.67	17	<.0001	0.025	0.146	-0.017	-0.074	0.003	0.001
24	568.36	23	<.0001	-0.048	0.003	0.141	0.025	-0.066	-0.032

Autocorrelation Plot of Residuals																								
Lag	Covariance	Correlation	-1	9	8	7	6	5	4	3	2	1	0	1	2	3	4	5	6	7	8	9	1	Std Error
0	0.147900	1.00000																						0
1	0.028469	0.19249																						0.018503
2	-0.014631	-0.09892																						0.019176
3	-0.015551	-0.10514																						0.019350
4	-0.0067084	-0.04536																						0.019545
5	-0.011806	-0.07982																						0.019581
6	-0.010182	-0.06884																						0.019692
7	-0.034464	-0.23302																						0.019774
8	-0.011004	-0.07440																						0.020693
9	-0.0097081	-0.06564																						0.020784
10	0.0020666	0.01397																						0.020855
11	0.0035262	0.02384																						0.020858
12	-0.0067388	-0.04556																						0.020867
13	0.0036858	0.02492																						0.020901
14	0.021590	0.14598																						0.020912
15	-0.0025302	-0.01711																						0.021258
16	-0.010983	-0.07426																						0.021262
17	0.00037851	0.00256																						0.021351
18	0.00022174	0.00150																						0.021351
19	-0.0071016	-0.04802																						0.021351
20	0.00046434	0.00314																						0.021388
21	0.020789	0.14056																						0.021388
22	0.0037449	0.02532																						0.021702
23	-0.0097650	-0.06602																						0.021712
24	-0.0047041	-0.03181																						0.021781

Ilustración 7.43: Test de ruido blanco y correlograma para el modelo SARIMA(0,1,0)x(1,0,0)₇ ajustado a la serie de demanda de gas

En esta ocasión, el valor estimado para el parámetro ϕ_7 no es lo suficiente próximo a 1 como para traducirlo en una diferenciación. La siguiente autocorrelación residual que presenta mayor significatividad es la asociada al retardo 7 la cual sugiere la propuesta de un parámetro que cuantifique esa relación (MA(7)).

Tras incorporar este parámetro a la fórmula, tendremos un modelo SARIMA(0,1,0)x(1,0,1)₇ dado por:

$$(1-B)(1-\Phi_7*B^7)X(t)=(1-\Theta_7*B^7)\epsilon(t) \quad (7.33)$$

El resultado al que conduce el ajuste de este modelo es el siguiente:

Conditional Least Squares Estimation						Correlations of Parameter Estimates		
Parameter	Estimate	Standard Error	t Value	Approx Pr > t	Lag	Parameter	MA1,1	AR1,1
MA1,1	0.92917	0.0071955	129.13	<.0001	7	MA1,1	1.000	0.306
AR1,1	0.99938	0.0011483	870.29	<.0001	7	AR1,1	0.306	1.000

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	201.89	4	<.0001	0.180	-0.093	-0.138	-0.075	-0.057	-0.015
12	225.00	10	<.0001	0.012	-0.047	-0.064	-0.013	0.035	0.005
18	247.10	16	<.0001	0.000	0.006	-0.056	-0.048	0.019	0.040
24	249.97	22	<.0001	0.020	-0.014	-0.003	-0.001	-0.019	-0.002

Autocorrelation Plot of Residuals																								
Lag	Covariance	Correlation	-1	9	8	7	6	5	4	3	2	1	0	1	2	3	4	5	6	7	8	9	1	Std Error
0	0.099204	1.00000													*****									0
1	0.017809	0.17952													. ****									0.018503
2	-0.0092586	-.09333													** .									0.019090
3	-0.013679	-.13789													*** .									0.019245
4	-0.0074497	-.07509													** .									0.019580
5	-0.0056203	-.05665													* .									0.019679
6	-0.0014738	-.01486													. .									0.019735
7	0.0012161	0.01226													. .									0.019738
8	-0.0046438	-.04681													* .									0.019741
9	-0.0063580	-.06409													* .									0.019779
10	-0.0013363	-.01347													. .									0.019850
11	0.0034676	0.03495													. *									0.019853
12	0.00052890	0.00533													. .									0.019874
13	0.00003972	0.00040													. .									0.019875
14	0.00062588	0.00631													. .									0.019875
15	-0.0056040	-.05649													* .									0.019875
16	-0.0047769	-.04815													* .									0.019930
17	0.0019033	0.01919													. .									0.019970
18	0.0039701	0.04002													. *									0.019976
19	0.0019631	0.01979													. .									0.020004
20	-0.0014366	-.01448													. .									0.020010
21	-0.0002552	-.00257													. .									0.020014
22	-0.0000685	-.00069													. .									0.020014
23	-0.0018887	-.01904													. .									0.020014
24	-0.0002342	-.00236													. .									0.020020

Ilustración 7.44: Test de ruido blanco y correlograma para el modelo SARIMA(0,1,0)x(1,0,1)₇ ajustado a la serie de demanda de gas

Nuevamente el parámetro asociado a la parte autorregresiva sale muy próximo a la unidad. Parece evidente que también habremos de diferenciar nuestra variable en concepto de estacionalidad para evitar los riesgos de inestabilidad que pudieran derivarse de la no-estacionariedad semanal del proceso resultante.

Sin embargo, no queremos correr el riesgo de estar sobrediferenciando la serie. Por ello, si bien incluiremos una diferencia estacional en la fórmula, hemos decidido volver a proponer un AR(1) en vez de una diferencia para la parte regular para ver si efectivamente es necesaria la realización de ambas diferencias. Es decir, proponemos como modelo un SARIMA(1,0,0)x(0,1,1)₇ dado por:

$$(1 - \phi_1 * B)(1 - B^7)X(t) = (1 - \Theta_7 * B^7)\epsilon(t) \quad (7.34)$$

Para este modelo se obtiene el siguiente resultado:

Conditional Least Squares Estimation						Correlations of Parameter Estimates		
Parameter	Estimate	Standard Error	t Value	Approx Pr > t	Lag	Parameter	MA1,1	AR1,1
MA1,1	0.90164	0.0085364	105.62	<.0001	7	MA1,1	1.000	0.339
AR1,1	0.97833	0.0040867	239.39	<.0001	1	AR1,1	0.339	1.000

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	173.29	4	<.0001	0.179	-0.075	-0.125	-0.063	-0.042	-0.014
12	193.89	10	<.0001	0.011	-0.047	-0.048	-0.002	0.045	0.019
18	216.58	16	<.0001	0.003	0.007	-0.057	-0.033	0.030	0.049
24	220.09	22	<.0001	0.031	-0.011	-0.003	-0.001	-0.007	0.009

Autocorrelation Plot of Residuals																								
Lag	Covariance	Correlation	-1	9	8	7	6	5	4	3	2	1	0	1	2	3	4	5	6	7	8	9	1	Std Error
0	0.099926	1.00000													*****									0
1	0.017915	0.17928													.*****									0.018522
2	-0.0075055	-0.07511													** .									0.019108
3	-0.012485	-0.12494													** .									0.019209
4	-0.0063240	-0.06329													* .									0.019486
5	-0.0041946	-0.04198													* .									0.019556
6	-0.0014314	-0.01432													-. .									0.019587
7	0.0010507	0.01051													-. .									0.019590
8	-0.0047164	-0.04720													* .									0.019592
9	-0.0047758	-0.04779													* .									0.019631
10	-0.0002437	-0.00244													-. .									0.019671
11	0.0045453	0.04549													-. *									0.019671
12	0.0018504	0.01852													-. .									0.019707
13	0.00031288	0.00313													-. .									0.019713
14	0.00065701	0.00657													-. .									0.019714
15	-0.0056791	-0.05683													* .									0.019714
16	-0.0033214	-0.03324													* .									0.019770
17	0.0030418	0.03044													-. *									0.019790
18	0.0049142	0.04918													-. *									0.019806
19	0.0030529	0.03055													-. *									0.019847
20	-0.0011476	-0.01148													-. .									0.019864
21	-0.0002710	-0.00271													-. .									0.019866
22	-0.0001412	-0.00141													-. .									0.019866
23	-0.0006750	-0.00676													-. .									0.019866
24	0.00086118	0.00862													-. .									0.019867

Ilustración 7.45: Test de ruido blanco y correlograma para el modelo SARIMA(1,0,0)x(0,1,1)₇ ajustado a la serie de demanda de gas

Si bien el parámetro ϕ_1 sigue saliendo bastante próximo a la unidad, puede llegar a considerarse significativamente distinto de 1 si construimos un intervalo alrededor de él y de radio dos veces la desviación típica. En cualquier caso, la verdadera razón por la que hemos optado seguir con el SARIMA(1,0,0)x(0,1,1)₇ en vez del SARIMA(0,1,0)x(0,1,1)₇

es que la calidad de las predicciones fuera de muestra es mayor en el primer caso. Obsérvese también que la correlación existente entre los parámetros del modelo (ϕ_l y θ_l) se puede considerar pequeña.

El último paso a la vista del gráfico anterior sería incluir un coeficiente θ_l que mida la relación de dependencia entre el proceso residual y su retardo de orden 1 o lo que es lo mismo, entre el error de un día y el error del día anterior. La incorporación del término de medias móviles conduce al modelo final SARIMA(1,0,1)x(0,1,1)₇ de ecuación:

$$(1 - \phi_1 * B)(1 - B^7)X(t) = (1 - \theta_1 * B)(1 - \Theta_7 * B^7)\epsilon(t) \quad (7.35)$$

Una vez incorporado este término de medias móviles se obtiene el siguiente resultado:

Conditional Least Squares Estimation						Correlations of Parameter Estimates			
Parameter	Estimate	Standard Error	t Value	Approx Pr > t	Lag	Parameter	MA1,1	MA2,1	AR1,1
MA1,1	-0.21384	0.01859	-11.50	<.0001	1	MA1,1	1.000	0.089	0.227
MA2,1	0.88524	0.0092232	95.98	<.0001	7	MA2,1	0.089	1.000	0.356
AR1,1	0.96597	0.0052526	183.90	<.0001	1	AR1,1	0.227	0.356	1.000

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	36.07	3	<.0001	-0.004	-0.039	-0.098	-0.025	-0.023	-0.001
12	51.35	9	<.0001	0.014	-0.033	-0.029	0.005	0.052	0.018
18	68.80	15	<.0001	0.008	0.012	-0.045	-0.019	0.036	0.045
24	73.29	21	<.0001	0.033	-0.007	-0.005	0.009	-0.001	0.016

Autocorrelation Plot of Residuals																									
Lag	Covariance	Correlation	-1	9	8	7	6	5	4	3	2	1	0	1	2	3	4	5	6	7	8	9	1	Std Error	
0	0.096024	1.00000												*****										0	
1	-0.0003987	-0.00415												-											0.018522
2	-0.0037823	-0.03939												*											0.018522
3	-0.0094116	-0.09801												**											0.018551
4	-0.0024385	-0.02540												*											0.018728
5	-0.0022118	-0.02303												-											0.018739
6	-0.0001113	-0.00116												-											0.018749
7	0.0013881	0.01446												-											0.018749
8	-0.0031871	-0.03319												*											0.018753
9	-0.0027741	-0.02889												*											0.018773
10	0.00046683	0.00486												-											0.018788
11	0.0050111	0.05219												-*											0.018789
12	0.0017446	0.01817												-											0.018838
13	0.00072237	0.00752												-											0.018844
14	0.0011405	0.01188												-											0.018845
15	-0.0043232	-0.04502												*											0.018848
16	-0.0018575	-0.01934												-											0.018885
17	0.0034699	0.03614												-*											0.018892
18	0.0043383	0.04518												-*											0.018915
19	0.0032078	0.03341												-*											0.018952
20	-0.0007069	-0.00736												-											0.018973
21	-0.0004369	-0.00455												-											0.018973
22	0.00089230	0.00929												-											0.018974
23	-0.0000732	-0.00076												-											0.018975
24	0.0015219	0.01585												-											0.018975

Ilustración 7.46: Test de ruido blanco y correlograma para el modelo SARIMA(1,0,1)x(0,1,1)₇ ajustado a la serie de demanda de gas

El resultado final de esta fase de identificación es un modelo típico en el ajuste de series de demanda diaria de gas y electricidad. El modelo basa el comportamiento esperado de la demanda para un día concreto en el dato real del día anterior y el día del mismo tipo de la semana pasada, a la vez que se corrige en función de los errores que hubiera

cometido esos mismos días al hacer la estimación. Si bien en el gráfico no parece haber correlaciones demasiado significativas, el proceso residual no supera el test de ruido blanco. Esto puede resultar complicado de conseguir dado que por el elevado número de datos que compone la serie, el intervalo de confianza del residuo se estrecha bastante y hace fácil que las autocorrelaciones se “salgan” de dicha banda. De todas maneras, conviene incluir en el modelo algunos agentes explicativos que puedan terminar de eliminar la estructura observada en dicho gráfico manifestada por ejemplo a través de la autocorrelación de orden 3. A este respecto señalaremos que no es conveniente incluir un término de medias móviles θ_3 que elimine dicha correlación puesto que no existe razón para pensar que la predicción para la demanda de gas un día esté condicionada por el error que cometió el modelo hace tres días.

Las variables con las que trataremos de seguir explicando el proceso de error hacen referencia al calendario laboral. Como ya hemos comentado en varias ocasiones, la serie se encuentra fuertemente influenciada por el efecto de festividades aisladas, puentes y periodos vacacionales, en los que la actividad industrial manifiesta un sensible descenso. Para reflejar el efecto provocado por este tipo de fenómenos vamos a generar una relación de variables:

- Variables asociadas a cada una de las festividades aisladas: *VOIMAY* (asociada al 1 de Mayo), *V12OCT* (asociado al 12 de Octubre), etc. Son variables binarias que valen 1 el día en el que se produce la fiesta y 0 en el resto de días.
- Variable *PUENTE* asociadas a días de puente, que valen 1 si el día cae entre dos festivos (un lunes entre un domingo y un martes, un viernes entre un jueves y un sábado, etc.) y 0 el resto de días.
- Variables asociadas a cada uno de los 3 periodos vacacionales (Navidad, Semana Santa o Agosto) que valen 1 durante dichos periodos y 0 en el resto de días.

Las variables anteriores se denominan *variables de intervención*, siendo las pertenecientes a los dos primeros grupos de tipo “pulso” y las pertenecientes a la tercera clase de tipo “escalón”.

Conviene destacar también que el efecto de un mismo festivo (por ejemplo el 8 de Diciembre) puede depender del día de la semana en que se produzca. Si cae en lunes, dará lugar a un puente de 3 días (6,7 y 8), si cae en martes a un puente de 4 (5,6,7 y 8) y si cae en miércoles al famoso “*acueducto*” de 5 días de Diciembre (pues el día 6 es festivo). Por ello, incluiremos en el modelo otras variables binarias que identifiquen las festividades en función del día de la semana en el que caen. Así, un 1 de Noviembre que cae en jueves quedará intervenido mediante dos variables:

- Una variable *VOINOV* para la que se estimará un parámetro a partir del efecto que han producido los días 1 de Noviembre en años anteriores.
- Una variable *FESJUE* (jueves festivo) que se estima en función de todas las festividades recogidas en el histórico que hayan caído en jueves.

Esta forma de intervenir los festivos persigue también reflejar la influencia de este tipo de días en caso de no disponer de una muestra de datos suficientemente grande. En dicho caso, pudiera resultar por ejemplo que ninguno de los días 1

de Noviembre almacenados en el histórico hubiese caído entre lunes y viernes (en día laborable) y en consecuencia, la estimación asociada al parámetro que acompaña a la variable $V0INOV$ podría no ser significativamente negativa (o incluso positiva). Sin embargo, dispondremos de otra variable $FES(DIA_DE_LA_SEMANA)$ que por su propia definición, se entrena a partir de festivos que caen en días laborables, es decir, identifica días festivos que no caen ni en sábado ni en domingo, con lo cual su estimación sí será significativamente negativa y conseguirá, en conjunción con la anterior, hacer que el impacto sobre la serie sea de “caída”. Se trata por tanto, de que la suma de las estimaciones asociadas a los parámetros para cada par de variables que afecta a un festivo concreto sea siempre negativo, para que así quede reflejada sobre la serie la influencia de un día de tales características.

Dentro del entorno de las festividades aisladas consideraremos otras variables binarias que denominaremos $VISPFES$ y $POSFES$ que harán referencia respectivamente a los días que sean víspera de festivo o posterior al mismo, pues se ha observado que en ambos casos el nivel de la demanda suele ser algo más bajo de lo habitual. Parece razonable que la tarde anterior a una festividad pueda ser concedida a algunos trabajadores para que inicien su periodo de descanso y, del mismo modo, que la reincorporación a la actividad laboral después de un día festivo se produzca de forma gradual (arranque de maquinaria y hornos que utilizan gas, regreso de los usuarios domésticos de su segunda residencia, etc.).

La fórmula a la que respondería el modelo una vez incorporada las variables de intervención presentaría la siguiente estructura:

$$(1-B^7)X(t) = \frac{(1-\theta_1*B)(1-\Theta_7*B^7)}{(1-\phi_1*B)} \epsilon(t) + \sum_{i=0}^{Num_intervenciones} p_i * I_{\{t=i\}} \quad (7.36)$$

donde p_i son los coeficientes a estimar para cada una de las intervenciones e $I_{\{t=i\}}$ son funciones indicatrices que valen 1 exclusivamente el día o días a los que hace referencia la intervención. El resultado de la inclusión de estas nuevas variables es el que a continuación se presenta:

Conditional Least Squares Estimation							
Parameter	Estimate	Standard Error	t Value	Approx Pr > t	Lag	Variable	Shift
MA1,1	-0.24931	0.01850	-13.48	<.0001	1	RCONSUMO	0
MA2,1	0.92941	0.0072958	127.39	<.0001	7	RCONSUMO	0
AR1,1	0.98145	0.0038205	256.89	<.0001	1	RCONSUMO	0
NUM1	-0.77988	0.07955	-9.80	<.0001	0	V01JAN	0
NUM2	-0.59662	0.07600	-7.85	<.0001	0	V06jan	0
NUM3	-0.36530	0.05913	-6.18	<.0001	0	V19MAR	0
NUM4	-0.78476	0.10422	-7.53	<.0001	0	juesanto	0
NUM5	-1.03024	0.12405	-8.30	<.0001	0	viesanto	0
NUM6	-0.60131	0.08201	-7.33	<.0001	0	sabsanto	0
NUM7	-0.46368	0.08266	-5.61	<.0001	0	V01MAY	0
NUM8	-0.41520	0.07122	-5.83	<.0001	0	V02MAY	0
NUM9	-0.17459	0.05531	-3.16	0.0016	0	V15MAY	0
NUM10	-0.40569	0.07228	-5.61	<.0001	0	V15AUG	0
NUM11	-0.41058	0.07277	-5.64	<.0001	0	V12OCT	0
NUM12	-0.65666	0.05028	-13.06	<.0001	0	V01NOV	0
NUM13	-0.12709	0.05537	-2.30	0.0218	0	V09NOV	0
NUM14	-0.42413	0.07600	-5.58	<.0001	0	V06DEC	0
NUM15	-0.34574	0.07880	-4.39	<.0001	0	V08DEC	0
NUM16	-0.51275	0.07841	-6.54	<.0001	0	V24DEC	0
NUM17	-0.91464	0.08021	-11.40	<.0001	0	V25DEC	0
NUM18	-0.40942	0.07780	-5.26	<.0001	0	V31DEC	0
NUM19	-0.22638	0.03836	-5.90	<.0001	0	PUENTE	0
NUM20	-0.12895	0.02857	-4.51	<.0001	0	VISPFES	0
NUM21	-0.11342	0.02662	-4.26	<.0001	0	POSFES	0
NUM22	-0.31679	0.05671	-5.59	<.0001	0	FESLUN	0
NUM23	-0.32315	0.07183	-4.50	<.0001	0	FESMAR	0
NUM24	-0.25874	0.06417	-4.03	<.0001	0	FESMIE	0
NUM25	-0.30837	0.06467	-4.77	<.0001	0	FESJUE	0
NUM26	-0.40052	0.07234	-5.54	<.0001	0	FESVIE	0
NUM27	0.03391	0.06178	0.55	0.5832	0	NAVIDAD	0
NUM28	-0.05068	0.06178	-0.82	0.4121	0	AGOSTO	0
NUM29	-0.03979	0.06329	-0.63	0.5296	0	SEMANA_SANTA	0

Tabla 7.29: Estimaciones de los parámetros del modelo $SARIMA(1,0,1) \times (0,1,1)_7$ y de los asociados a las variables resultante de un análisis de intervención para la identificación del efecto “calendario”

De todas las variables contempladas únicamente las asociadas a los periodos vacacionales no resultan significativas y por ello finalmente han sido excluidas del ajuste. En el caso de la Navidad y la Semana Santa este hecho está justificado dado que existen multitud de variables de festividades, vísperas y puentes que ya recogen este efecto durante dichos periodos. Nos podríamos plantear la inclusión o no de la variable *AGOSTO* pero tampoco parece que vaya a tener demasiada importancia de cara a nuestro objetivo que es, recordemos, la calidad de las predicciones fundamentalmente durante el periodo invernal. Una vez eliminada la presencia de estas variables en el modelo, el resultado ha sido el siguiente:

Conditional Least Squares Estimation							
Parameter	Estimate	Standard Error	t Value	Approx Pr > t	Lag	Variable	Shift
MA1,1	-0.25002	0.01840	-13.59	<.0001	1	RCONSUMO	0
MA2,1	0.92914	0.0073052	127.19	<.0001	7	RCONSUMO	0
AR1,1	0.98154	0.0038099	257.63	<.0001	1	RCONSUMO	0
NUM1	-0.78231	0.07939	-9.85	<.0001	0	V01JAN	0
NUM2	-0.59998	0.07581	-7.91	<.0001	0	V06jan	0
NUM3	-0.36585	0.05907	-6.19	<.0001	0	V19MAR	0
NUM4	-0.78450	0.10416	-7.53	<.0001	0	juesanto	0
NUM5	-1.02820	0.12397	-8.29	<.0001	0	viesanto	0
NUM6	-0.59719	0.08168	-7.31	<.0001	0	sabsanto	0
NUM7	-0.46525	0.08260	-5.63	<.0001	0	V01MAY	0
NUM8	-0.41258	0.07107	-5.81	<.0001	0	V02MAY	0
NUM9	-0.17495	0.05526	-3.17	0.0016	0	V15MAY	0
NUM10	-0.40804	0.07209	-5.66	<.0001	0	V15AUG	0
NUM11	-0.41274	0.07258	-5.69	<.0001	0	V12OCT	0
NUM12	-0.65883	0.05016	-13.13	<.0001	0	V01NOV	0
NUM13	-0.12674	0.05532	-2.29	0.0220	0	V09NOV	0
NUM14	-0.42690	0.07584	-5.63	<.0001	0	V06DEC	0
NUM15	-0.34974	0.07854	-4.45	<.0001	0	V08DEC	0
NUM16	-0.50717	0.07765	-6.53	<.0001	0	V24DEC	0
NUM17	-0.91166	0.07938	-11.49	<.0001	0	V25DEC	0
NUM18	-0.40895	0.07776	-5.26	<.0001	0	V31DEC	0
NUM19	-0.22427	0.03826	-5.86	<.0001	0	PUENTE	0
NUM20	-0.13060	0.02848	-4.59	<.0001	0	VISPFES	0
NUM21	-0.11712	0.02598	-4.51	<.0001	0	POSFES	0
NUM22	-0.31667	0.05666	-5.59	<.0001	0	FESLUN	0
NUM23	-0.32298	0.07177	-4.50	<.0001	0	FESMAR	0
NUM24	-0.25923	0.06412	-4.04	<.0001	0	FESMIE	0
NUM25	-0.30836	0.06462	-4.77	<.0001	0	FESJUE	0
NUM26	-0.40032	0.07229	-5.54	<.0001	0	FESVIE	0

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	42.35	3	<.0001	0.009	0.018	-0.091	-0.054	-0.054	0.002
12	49.25	9	<.0001	-0.019	-0.029	0.006	-0.007	0.029	-0.015
18	58.79	15	<.0001	-0.014	0.024	0.017	-0.032	0.017	0.030
24	63.11	21	<.0001	0.001	-0.013	0.019	0.027	0.002	-0.014

Autocorrelation Plot of Residuals																				
Lag	Covariance	Correlation																		Std Error
0	0.066244	1.00000																		0
1	0.00062848	0.00949																		0.018522
2	0.0011866	0.01791																		0.018523
3	-0.0060163	-0.09082																		0.018529
4	-0.0035850	-0.05412																		0.018681
5	-0.0035732	-0.05394																		0.018735
6	0.00012533	0.00189																		0.018788
7	-0.0012819	-0.01935																		0.018788
8	-0.0019036	-0.02874																		0.018795
9	0.00040089	0.00605																		0.018810
10	-0.0004381	-0.00661																		0.018811
11	0.0019191	0.02897																		0.018812
12	-0.0010188	-0.01538																		0.018827
13	-0.0009050	-0.01366																		0.018831
14	0.0015598	0.02355																		0.018835
15	0.0011339	0.01712																		0.018845
16	-0.0020952	-0.03163																		0.018850
17	0.0011418	0.01724																		0.018868
18	0.0020119	0.03037																		0.018874
19	0.00007449	0.00112																		0.018891
20	-0.0008891	-0.01342																		0.018891
21	0.0012498	0.01887																		0.018894
22	0.0018086	0.02730																		0.018900
23	0.00010591	0.00160																		0.018914
24	-0.0009005	-0.01359																		0.018914

Ilustración 7.47: Test de ruido blanco, correlograma y estimaciones de los parámetros del modelo SARIMA(1,0,1)x(0,1,1)₇ con análisis de intervención ajustado a la serie de demanda de gas

Tomando como base este modelo, se han hecho predicciones en muestra y fuera de muestra siendo los resultados los que se presentan en la siguiente tabla:

Histórico de entrenamiento: (01/11/1997- 31/10/2005)

Modelo base: Sin tratamiento de la temperatura

Errores en muestra			Errores fuera de muestra		
MES	MEDIA	MEDIANA	MES	MEDIA	MEDIANA
TOTAL	5.81	4.64	TOTAL	5.49	4.30
1	5.10	4.03	1	5.53	4.30
2	5.45	4.32	2	3.97	3.48
3	7.47	5.97	3	6.85	7.45
11	6.03	4.88	11	5.80	4.47
12	4.98	4.17	12	5.17	3.30

Tabla 7.30: Errores de predicción del modelo SARIMA(1,0,1)x(0,1,1)₇ con análisis de intervención ajustado a la serie de demanda de gas

Vamos a ver cómo la inclusión de variables que recogen el efecto de las variables climáticas a través de modelos de función de transferencia mejora considerablemente la calidad de estas predicciones. Los modelos de función de transferencia se diseñan para relacionar la serie temporal “output” (*demanda de gas*) con otra u otras series “inputs” (las variables de *temperatura*). Estos modelos, según se puede consultar en el capítulo 13 de [AZTRI], responden a la siguiente expresión:

$$X(t) = \frac{\omega(B) * B^k}{\delta(B)} = \frac{(\omega_0 - \omega_1 * B - \omega_2 * B - \dots - \omega_n * B) * B^k}{(1 - \delta_1 * B - \delta_2 * B - \dots - \delta_n * B)} * Y(t) + N(t) \quad (7.37).$$

donde en nuestro caso $X(t)$ sería la demanda, $Y(t)$ la variable explicativa en cuestión, es decir, las temperaturas y $N(t)$ un proceso residual. A priori se considera la existencia de una relación de dependencia de causalidad unidireccional desde los “inputs” hacia el “output”. En cuanto a los polinomios respecto del operador de retardo B se refiere, observemos que:

- El orden del polinomio $\omega(B)$ determina el número de periodos (días) desde el cual se considera influye la variable $Y(t)$ en $X(t)$.
- El orden del polinomio $\delta(B)$ determina la forma en la que se transmite el efecto de $Y(t)$ a $X(t)$. Si por ejemplo este orden es 1, vendría a decir que el valor de la demanda de un día depende del valor de la temperatura de ese mismo día y que esa influencia decae de forma exponencial con mayor o menor velocidad en función de que el valor de δ_i esté más o menos próximo a 0.
- El operador de retardo B^k identifica el número de días que transcurren antes de que $Y(t)$ comience a tener efecto sobre la variable dependiente. Puede ser que un cambio de temperatura no afecte de manera instantánea en la demanda sino que comience a influir por ejemplo a partir del día siguiente. Este efecto que se define en el capítulo tercero de [GIDEDU] como **efecto de histéresis** o retardo se traduce en que los usuarios de gas

pueden no encender la calefacción inmediatamente después del momento en el que se produce el descenso de la temperatura, sino una vez que la misma se mantuvo baja por algunos días.

Hablaremos ahora de qué forma tendrán las variables $Y(t)$ que serán introducidas en la ecuación del ARIMA con su correspondiente modelo de función de transferencia ya que existen muchas maneras de definir las variables de temperaturas para que la información aportada por éstas refleje mejor su influencia sobre la demanda. La forma inmediata de hacerlo es incluir las propias variables en estado puro, es decir, como el número de grados centígrados que marcan los termómetros un día determinado. Otra manera de hacerlo es sometiendo a las variables a algún tipo de transformación que aumente su poder predictivo.

Por ejemplo, vamos a incluir la información de temperaturas como el incremento en grados que se produce de un día al siguiente para lo cual, vamos a definir las variables:

- INC_TMAX .- número de grados centígrados que sube o baja la temperatura máxima de un día al siguiente.
- INC_TMIN .- número de grados centígrados que sube o baja la temperatura mínima de un día al siguiente.

La idea es eliminar la dependencia temporal que llevan implícita las variables de temperatura puesto que nuestro objetivo es comparar el resultado de este ajuste con aquél que obtengamos a través de funciones cópula y éstas se desenvuelven bien cuando esta dependencia no existe. Previamente hemos comprobado que, de acuerdo a esta transformación, los resultados predictivos que se obtienen con un modelo ARIMA en el que las variables INC_TMAX e INC_TMIN son ajustadas con modelos de función de transferencia, son mejores que los que se obtienen cuando se utilizan las variables originales sin transformar, $TMAX$ y $TMIN$. También es verdad que existe una transformación aún mejor para dichas variables que conduce a resultados de mayor calidad y que más tarde comentaremos.

Otro hecho en el que queremos reincidir es que el efecto de la temperatura no es igual en meses diferentes. Así, vamos a considerar un par de variables (una asociada a la temperatura máxima, INC_TMAX_MES , y otra asociada a la temperatura mínima, INC_TMIN_MES) para cada uno de los meses no veraniegos del año y en los que por tanto se considera influye la temperatura. Estas variables se definen, en cada día del mes al que van asociadas, como el valor del incremento de temperatura de un día al siguiente, y como 0 en los días correspondientes el resto de los meses del año. En cuanto al modelo de función de transferencia ajustado a cada una de las variables, se contemplan sólo retardos de órdenes 1 y 2 de la propia variable ($n = 2$ en la ecuación (7.37)), de manera que el modelo evalúe la relación de dependencia entre la demanda de un día y el incremento climático respecto del día anterior, y cuantifique la influencia de este efecto en un horizonte máximo de dos días, pues no hemos considerado verosímil su repercusión más allá de este periodo. Tampoco se han tenido en cuenta estructuras más complejas de modelización en las que el efecto de las temperaturas comience a afectar pasados algunos días o en las que dicho efecto decaiga de acuerdo a un patrón exponencial o sinusoidal (órdenes “ k ” y “ n ” en la expresión (7.37)) Teniendo en cuenta estas consideraciones el modelo quedaría planteado como:

$$(1-B^7)X(t) = \frac{(1-\theta_1*B)(1-\theta_7*B^7)}{(1-\phi_1*B)} \epsilon(t) + \sum_{i=0}^{Num_intervenciones} p_i * I_{\{t=i\}} + \sum_{MES \in \{E, \dots, D\}} \sum_{jmax=0}^2 \omega_{MES, jmax}(B) * INC_TMAX_{MES} + \sum_{MES \in \{E, \dots, D\}} \sum_{jmin=0}^2 \omega_{MES, jmin}(B) * INC_TMIN_{MES} \quad (7.38)$$

Según puede verse en la ilustración (7.48), las variables propuestas para cada uno de los meses así como los retardos definidos a través de sus modelos de función de transferencia han resultado ser en su mayoría significativas, salvando algunas excepciones como son por ejemplo las variables de temperatura mínima correspondientes al mes de Mayo. Sí es verdad que el proceso residual sigue sin verificar todavía las hipótesis de ruido blanco de acuerdo a lo observado en la tabla de la derecha.

En cuanto a los errores de predicción, la tabla (7.31) nos permite comprobar como éstos han bajado notoriamente con la inclusión de las variables de temperatura y más aún conforme se han ido añadiendo términos en sus modelos de función de transferencia.

Conditional Least Squares Estimation							
Parameter	Estimate	Standard Error	t Value	Approx Pr > t	Lag	Variable	Shift
NUM27	-0.04659	0.0040780	-11.42	<.0001	0	INC_THMAX_ENERO	0
NUM1,1	0.03726	0.0052527	7.09	<.0001	1	INC_THMAX_ENERO	0
NUM1,2	0.01647	0.0040570	4.06	<.0001	2	INC_THMAX_ENERO	0
NUM28	-0.02077	0.0039344	-5.28	<.0001	0	INC_THMIN_ENERO	0
NUM1,1	0.02067	0.0051333	4.03	<.0001	1	INC_THMIN_ENERO	0
NUM1,2	0.01282	0.0039264	3.27	0.0011	2	INC_THMIN_ENERO	0
NUM29	-0.06109	0.0042674	-14.32	<.0001	0	INC_THMAX_FEBRERO	0
NUM1,1	0.05350	0.0060311	8.87	<.0001	1	INC_THMAX_FEBRERO	0
NUM1,2	0.02352	0.0042695	5.51	<.0001	2	INC_THMAX_FEBRERO	0
NUM30	-0.02265	0.0048466	-4.67	<.0001	0	INC_THMIN_FEBRERO	0
NUM1,1	0.02240	0.0066005	3.39	0.0007	1	INC_THMIN_FEBRERO	0
NUM1,2	0.01302	0.0048386	2.69	0.0072	2	INC_THMIN_FEBRERO	0
NUM31	-0.06829	0.0040630	-16.81	<.0001	0	INC_THMAX_MARZO	0
NUM1,1	0.06491	0.0056867	11.41	<.0001	1	INC_THMAX_MARZO	0
NUM1,2	0.02867	0.0042565	6.74	<.0001	2	INC_THMAX_MARZO	0
NUM32	-0.0066004	0.0049720	-1.33	0.1844	0	INC_THMIN_MARZO	0
NUM1,1	0.01910	0.0066619	2.87	0.0042	1	INC_THMIN_MARZO	0
NUM1,2	0.01637	0.0047944	3.41	0.0006	2	INC_THMIN_MARZO	0
NUM33	-0.05560	0.0036285	-15.32	<.0001	0	INC_THMAX_ABRIL	0
NUM1,1	0.05077	0.0052039	9.76	<.0001	1	INC_THMAX_ABRIL	0
NUM1,2	0.02870	0.0041289	6.95	<.0001	2	INC_THMAX_ABRIL	0
NUM34	-0.01901	0.0048251	-3.94	<.0001	0	INC_THMIN_ABRIL	0
NUM1,1	0.02808	0.0068008	4.13	<.0001	1	INC_THMIN_ABRIL	0
NUM1,2	0.01907	0.0044012	4.33	<.0001	2	INC_THMIN_ABRIL	0
NUM35	-0.01963	0.0036825	-5.33	<.0001	0	INC_THMAX_MAYO	0
NUM1,1	0.02022	0.0049819	4.06	<.0001	1	INC_THMAX_MAYO	0
NUM1,2	0.0071498	0.0040379	1.77	0.0767	2	INC_THMAX_MAYO	0
NUM36	0.0029290	0.0048918	0.60	0.5494	0	INC_THMIN_MAYO	0
NUM1,1	0.0041932	0.0066017	0.64	0.5254	1	INC_THMIN_MAYO	0
NUM1,2	0.0011557	0.0044797	0.26	0.7964	2	INC_THMIN_MAYO	0
NUM37	-0.02477	0.0043580	-5.68	<.0001	0	INC_THMAX_OCTUBRE	0
NUM1,1	0.02122	0.0058324	3.64	0.0003	1	INC_THMAX_OCTUBRE	0
NUM1,2	0.01120	0.0045884	2.44	0.0147	2	INC_THMAX_OCTUBRE	0
NUM38	-0.01196	0.0049590	-2.41	0.0159	0	INC_THMIN_OCTUBRE	0
NUM1,1	0.01555	0.0064738	2.40	0.0164	1	INC_THMIN_OCTUBRE	0
NUM1,2	0.0068949	0.0048693	1.42	0.1569	2	INC_THMIN_OCTUBRE	0
NUM39	-0.04581	0.0043142	-10.62	<.0001	0	INC_THMAX_NOVIEMBRE	0
NUM1,1	0.03899	0.0053463	7.29	<.0001	1	INC_THMAX_NOVIEMBRE	0
NUM1,2	0.02878	0.0037928	7.59	<.0001	2	INC_THMAX_NOVIEMBRE	0
NUM40	-0.03112	0.0038376	-8.11	<.0001	0	INC_THMIN_NOVIEMBRE	0
NUM1,1	0.02532	0.0049217	5.15	<.0001	1	INC_THMIN_NOVIEMBRE	0
NUM1,2	0.01148	0.0037461	3.07	0.0022	2	INC_THMIN_NOVIEMBRE	0
NUM41	-0.04999	0.0046548	-10.74	<.0001	0	INC_THMAX_DICIEMBRE	0
NUM1,1	0.05049	0.0062232	8.11	<.0001	1	INC_THMAX_DICIEMBRE	0
NUM1,2	0.02822	0.0047315	5.96	<.0001	2	INC_THMAX_DICIEMBRE	0
NUM42	-0.01412	0.0043869	-3.22	0.0013	0	INC_THMIN_DICIEMBRE	0
NUM1,1	0.01366	0.0056897	2.40	0.0164	1	INC_THMIN_DICIEMBRE	0
NUM1,2	0.0027675	0.0044917	0.62	0.5379	2	INC_THMIN_DICIEMBRE	0

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	132.50	3	<.0001	0.052	0.146	-0.053	-0.079	-0.103	-0.040
12	150.17	9	<.0001	-0.021	-0.056	0.003	-0.023	0.034	-0.027
18	156.00	15	<.0001	-0.015	0.019	0.002	0.011	0.006	0.036
24	158.84	21	<.0001	-0.023	-0.008	0.008	0.017	0.005	0.002

Autocorrelation Plot of Residuals											
Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1								Std Error
0	0.042484	1.00000								*****	0
1	0.0022281	0.05244								. *	0.018528
2	0.0062189	0.14638								. ***	0.018579
3	-0.0022648	-0.05331								* .	0.018971
4	-0.0033561	-0.07900								** .	0.019022
5	-0.0043639	-0.10272								** .	0.019134
6	-0.0016973	-0.03995								* .	0.019323
7	-0.0009020	-0.02123								. .	0.019351
8	-0.0023945	-0.05636								* .	0.019359
9	0.00012667	0.00298								. .	0.019415
10	-0.0009749	-0.02295								. .	0.019415
11	0.0014369	0.03382								. *	0.019425
12	-0.0011518	-0.02711								* .	0.019445
13	-0.0006365	-0.01498								. .	0.019458
14	0.00078856	0.01856								. .	0.019462
15	0.00006454	0.00152								. .	0.019468
16	0.00045796	0.01078								. .	0.019468
17	0.00023998	0.00565								. .	0.019470
18	0.0015141	0.03564								. *	0.019471
19	-0.0009857	-0.02320								. .	0.019493
20	-0.0003507	-0.00825								. .	0.019503
21	0.00032909	0.00775								. .	0.019504
22	0.00070304	0.01655								. .	0.019505
23	0.00019814	0.00466								. .	0.019510
24	0.00008733	0.00206								. .	0.019510

Ilustración 7.48: Test de ruido blanco, correlograma y estimaciones de los parámetros del SARIMA(1,0,1)x(0,1,1) con análisis de intervención y modelos de función de transferencia ajustados a la serie de demanda de gas

Histórico de entrenamiento: (01/11/1997- 31/10/2005)

Modelos de función de transferencia con $n=0, m=0, k=0$

Errores en muestra			Errores fuera de muestra		
MES	MEDIA	MEDIANA	MES	MEDIA	MEDIANA
TOTAL	5.05	4.11	TOTAL	5.12	4.03
1	4.48	3.55	1	5.09	4.13
2	4.46	3.62	2	3.76	3.28
3	6.84	5.47	3	6.62	4.88
11	4.99	4.11	11	5.34	4.35
12	4.41	3.91	12	4.68	3.42

Modelos de función de transferencia con $n=1, m=0, k=0$

Errores en muestra			Errores fuera de muestra		
MES	MEDIA	MEDIANA	MES	MEDIA	MEDIANA
TOTAL	4.72	3.79	TOTAL	4.66	3.70
1	4.06	3.25	1	4.94	3.73
2	4.22	3.62	2	3.72	3.11
3	6.41	4.97	3	5.62	4.21
11	4.87	3.80	11	4.97	4.15
12	4.03	3.53	12	3.99	3.09

Modelos de función de transferencia con $n=2, m=0, k=0$

Errores en muestra			Errores fuera de muestra		
MES	MEDIA	MEDIANA	MES	MEDIA	MEDIANA
TOTAL	4.36	3.41	TOTAL	4.21	2.97
1	3.80	3.02	1	4.06	2.59
2	3.91	3.07	2	3.45	2.68
3	6.00	4.53	3	5.13	4.09
11	4.40	3.39	11	4.71	3.88
12	3.66	3.22	12	3.65	2.48

Tabla 7.31: Errores de predicción del modelo $SARIMA(1,0,1) \times (0,1,1)_7$ con análisis de intervención y modelos de función de transferencia ajustados a la serie de demanda de gas

Según hemos ido planteando este desarrollo podemos ver que estamos partiendo del error de predicción que comete una fórmula y reduciéndolo, en término medios, mediante la inclusión de nuevas variables que tratan de ir explicando parte de ese proceso residual desconocido. El objetivo final es que éste sea totalmente impredecible a partir de la información disponible y, en el mejor de los casos, que sea un proceso de ruido blanco. En concreto, el modelo (7.38) no es más que el resultado de tratar de predecir el error que comete el modelo (7.36) a través de la información aportada por variables de temperatura medidas en términos de incrementos.

7.3.2.2 Ajuste de una serie mediante un ARIMA complementado con un algoritmo basado en funciones cópula

Como camino alternativo, proponemos extraer el proceso de error vinculado a la ecuación (7.36) y buscar su relación de dependencia con las variables climáticas pero a través de funciones cópula en sustitución de las funciones de transferencia. Para ello vamos a aplicar el algoritmo iterativo de ajuste secuencial mediante funciones cópula descrito en el apartado 6.1. Así, de igual manera que hemos hecho con las funciones de transferencia, buscaremos la cópula que mejor explique el proceso residual del modelo (7.36) en función de las variables de temperatura asociadas a cada uno de los meses.

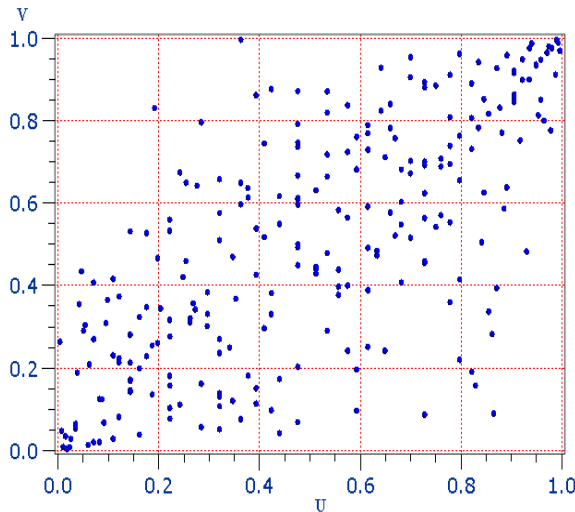
Tomemos por ejemplo el mes de Enero. Vamos a considerar un par de variables explicativas adicionales a INC_TMAX e INC_TMIN y que trataremos de relacionar con los procesos de error que vayan resultando de cada uno de los ajustes:

- INC_TMAX_2 .- diferencia de temperatura máxima en grados centígrados entre un día y el situado a dos unidades de tiempo, es decir, no el anterior sino el anterior al anterior.
- INC_TMIN_2 .- de definición análoga a la anterior pero respecto de la temperatura mínima.

El objetivo que perseguimos con éstas, es reflejar la información que en el modelo de función de transferencia introducen las variables retardadas (al especificar el orden $n = 1$ para el polinomio que aparece en el numerador de dicha función). En cuanto a las familias de cópulas candidatas a medir la relación de dependencia entre estas variables y los sucesivos procesos de error que vayan resultando, vamos a utilizar exactamente las mismas que las propuestas en el estudio de predicción a medio plazo (véase apartado 7.2.5.2).

Entonces, de acuerdo al algoritmo mencionado, empezaremos estableciendo $j = 1$, $i = 1$ y calculando $Mean^*_{HISTORICO} = 5.10\%$ la media del error sobre histórico de (7.36) en Enero (véase tabla (7.30)) y $Median^*_{HISTORICO} = 4.03\%$ la mediana del mismo. Este proceso de error calculado en función del valor real de la demanda ($demanda(t)$) y de la predicción hecha para ella a través de (7.36), $PI(t)$, sería la variable $ErrorI(t)$. Además $Mean^*_{FUTURO} = 5.53\%$ es el error medio fuera de muestra cometido para ese mes, es decir, el error medio del mes de Enero de 2006 mientras que $Median^*_{FUTURO} = 4.30\%$ es el error mediano.

Ya en el paso 2, tomamos la primera de las variables candidatas a explicar este primer proceso de error. Se trata entonces de buscar la cópula que mejor establezca la relación entre $ErrorI$ e INC_TMAX . Para ello, vamos a fijar una dimensión de partición del rectángulo unidad y vamos a ver cuál de las familias de cópulas de la relación de candidatas que proponemos proporciona un mejor valor del estadístico de Pearson. Como en ocasiones anteriores, hemos dividido el rectángulo unidad en un total de 25 rectángulos (dimensión 5 x 5). El siguiente gráfico muestra el espacio transformado que sirve de soporte a la función cópula buscada y en la tabla se presenta la relación de estadísticos de Pearson obtenidos para cada una de las candidatas planteadas:



COPULA	PARAMETRO_COPULA	PEARSON	P_VALOR
PLA	12.2339	22.614	0.092678
GUM	2.0446	22.958	0.085038
F15	1.5446	28.601	0.018095
F13	1.3631	35.426	0.002138
HRT	0.4786	35.605	0.002015
NOR	0.7191	39.807	0.000485
CLA	2.0892	62.858	0.000000
FRA	1.1474	93.253	0.000000
ALI	0.4811	95.820	0.000000
PRO	.	152.203	0.000000

La región crítica viene dada por $\{X|X > 24.99579014\}$
 Nivel de significación del contraste, $\alpha = 0.05$

Ilustración 7.49: Partición del espacio transformado de las variables INC_TMAX y $Error1$ (de predicción de la demanda) en Enero. Evaluación del estadístico de Pearson para una relación de cópulas

Seleccionamos por tanto la cópula de Plackett. De haber salido la cópula producto hubiéramos podido interpretar independencia entre las variables $Error1$ e INC_TMAX y haber descartado la posible relación entre ellas.

El siguiente paso del algoritmo (paso 3) nos lleva a simular para cada día “ t ” y, en función del valor de $INC_TMAX(t)$ asociado al mismo, valores de $Error1(t)$ a través de la cópula condicionada de Plackett. Para cada día “ t ” se propone como valor esperado de la variable $Error1$ la mediana de las simulaciones realizadas y, a partir de ella, proponemos una nueva predicción para dicho día $P'(t)$ a través de la fórmula (6.2). Esta predicción da lugar a un nuevo proceso de error, $Error'(t)$, cuya media y mediana sobre el histórico así como sobre el conjunto de validación es $Mean'_{HISTORICO} = 3.49\%$, $Median'_{HISTORICO} = 2.85\%$, $Mean'_{FUTURO} = 3.50\%$ y $Median'_{FUTURO} = 2.90\%$. Dado que $Mean'_{HISTORICO} < Mean^*_{HISTORICO}$ y que $Mean'_{FUTURO} < Mean^*_{FUTURO}$, establecemos $Error2(t) = Error'(t)$, $P2(t) = P'(t)$, $Mean^*_{HISTORICO} = Mean'_{HISTORICO}$, $Median^*_{HISTORICO} = Median'_{HISTORICO}$, $Mean^*_{FUTURO} = Mean'_{FUTURO}$, $Median^*_{FUTURO} = Median'_{FUTURO}$, $i = 2$, $j = 2$ y volvemos al paso 2.

Buscaremos ahora la cópula que mejor establezca la relación de dependencia entre el nuevo proceso de error $Error2(t)$ y la siguiente de las variables explicativas candidatas INC_TMAX_2 . Alternativamente se podría volver a intentar explicar el error nuevamente con INC_TMAX hasta que el error medio dejara de disminuir o bien hasta que la cópula que mejor se adaptase a la relación entre las variables fuese la cópula producto. De acuerdo a una partición de la misma dimensión que la anterior, el resultado es el siguiente:

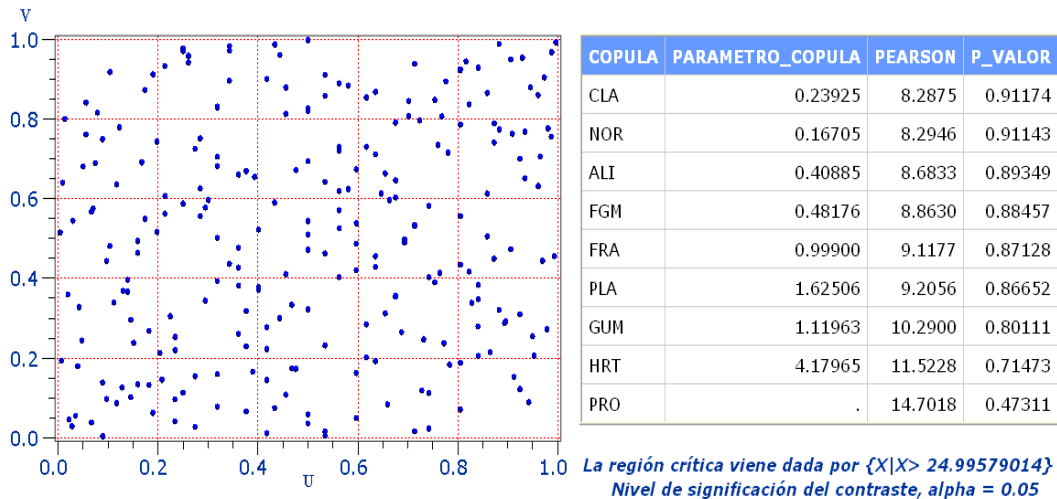


Ilustración 7.50: Partición del espacio transformado de las variables INC_TMAX_2 y $Error2$ de predicción de la demanda en Enero. Evaluación del estadístico de Pearson para una relación de cópulas

En este caso, la cópula seleccionada es la de Clayton. A través de su condicionada simularemos valores de $Error2(t)$ para cada valor de INC_TMAX_2 . Proponemos nuevamente como error esperado para cada día “ t ” la mediana de dichas simulaciones y utilizamos este valor esperado para proponer una nueva predicción $P'(t)$ a través de (6.2).

El nuevo proceso de error asociado a esta predicción $Error'(t)$, tiene por medias y medianas los valores: $Mean'_{HISTORICO} = 3.49\%$, $Median'_{HISTORICO} = 2.84\%$, $Mean'_{FUTURO} = 3.41\%$ y $Median'_{FUTURO} = 2.56\%$. Si bien no existen claras mejoras sobre histórico respecto de los últimos valores establecidos, sí se aprecia alguna respecto del conjunto de validación que hace que no sea descabellado incluir esta variable como explicativa en este proceso de ajuste. Así, establecemos $Error3(t) = Error'(t)$, $P3(t) = P'(t)$, $Mean^*_{HISTORICO} = Mean'_{HISTORICO}$, $Median^*_{HISTORICO} = Median'_{HISTORICO}$, $Mean^*_{FUTURO} = Mean'_{FUTURO}$, $Median^*_{FUTURO} = Median'_{FUTURO}$, $i = 3, j = 3$ y volvemos al paso 2.

La siguiente variable a considerar es INC_TMIN , debiendo analizar la relación de ésta con $Error3(t)$. La siguiente tabla propone la función cópula más apropiada a tal fin en función del valor del estadístico de Pearson.

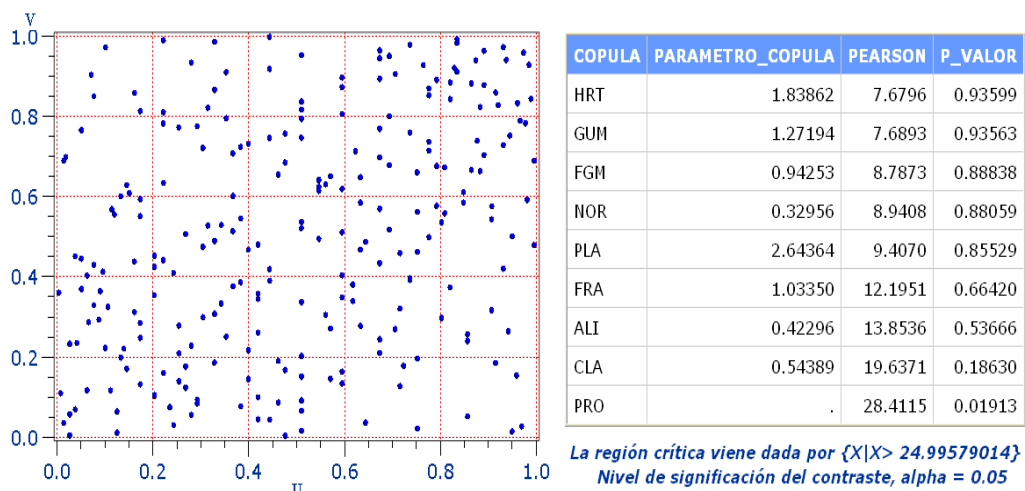


Ilustración 7.51: Partición del espacio transformado de las variables INC_TMIN y $Error3$ de predicción de la demanda en Enero. Evaluación del estadístico de Pearson para una relación de cópulas

La cópula de Cola Derecha Pesada (HRT) es la más apropiada para establecer la nueva relación entre variables. La nueva predicción generada $P'(t)$ y a través de ella, el nuevo proceso de error $Error'(t)$ proporciona las siguientes medidas de evaluación: $Mean'_{HISTORICO} = 3.33\%$, $Median'_{HISTORICO} = 2.64\%$, $Mean'_{FUTURO} = 2.98\%$ y $Median'_{FUTURO} = 1.89\%$. La mejora respecto de los últimos indicadores que teníamos vuelve a ser manifiesta y por ello establecemos $Error4(t) = Error'(t)$, $Mean^*_{HISTORICO} = Mean'_{HISTORICO}$, $Median^*_{HISTORICO} = Median'_{HISTORICO}$, $Mean^*_{FUTURO} = Mean'_{FUTURO}$, $Median^*_{FUTURO} = Median'_{FUTURO}$, $P4(t) = P'(t)$, $i = 4, j = 4$ y volvemos al paso 2. Podemos tomar ya la última de las variables candidatas disponibles INC_TMIN_2 y estudiar su relación con la variable $Error4(t)$. El espacio transformado que servirá de soporte a la nueva cópula así como los valores del estadístico de Pearson que se obtienen para las diferentes candidatas son los siguientes:

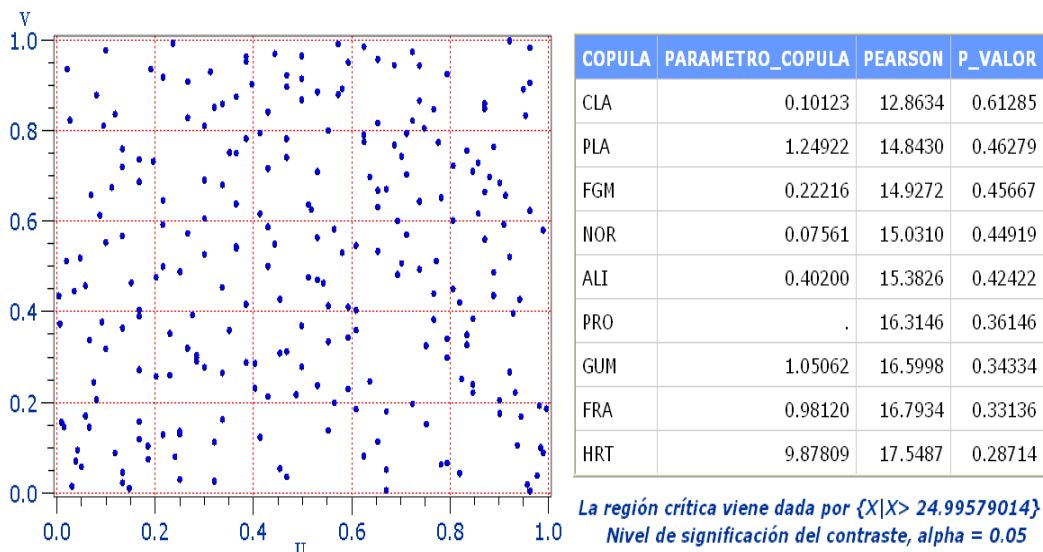


Ilustración 7.52: Partición del espacio transformado de las variables INC_TMIN_2 y $Error4$ de predicción de la demanda en Enero. Evaluación del estadístico de Pearson para una relación de cópulas

Nuevamente la cópula de Clayton es la que resulta más apropiada para representar la relación de dependencia entre las variables. Como nota curiosa, observaremos que por primera vez la cópula producto (sinónimo de independencia) empieza a tener más importancia que algunas otras alternativas como la de Gumbel o la de Frank. Éste puede ser es un indicio de que estamos llegando a un punto en el que encontraremos dificultades para seguir explicando el proceso de error. Los valores medios y medianos asociados al nuevo proceso $Error'(t)$ son: $Mean'_{HISTORICO} = 3.31\%$, $Median'_{HISTORICO} = 2.62\%$, $Mean'_{FUTURO} = 2.95\%$ y $Median'_{FUTURO} = 1.86\%$. La mejora respecto al paso anterior es muy pequeña tanto en términos históricos como en términos de validación. Esta circunstancia unida a la que acabamos de mencionar podría llevarnos a dar por finalizado el proceso iterativo de ajuste mediante funciones cópula. Sin embargo, también podríamos seguir adelante y, a falta de nuevas variables explicativas, volver a empezar a utilizar las mismas (una segunda ronda).

Establezcamos $Error5(t) = Error'(t)$, $P5(t) = P'(t)$, $Mean^*_{HISTORICO} = Mean'_{HISTORICO}$, $Median^*_{HISTORICO} = Median'_{HISTORICO}$, $Mean^*_{FUTURO} = Mean'_{FUTURO}$, $Median^*_{FUTURO} = Median'_{FUTURO}$, $i = 5, j = 5$ y volvamos al paso 2, considerando nuevamente como variable explicativa candidata INC_TMAX . Volveríamos a decantarnos por la cópula que respecto de la partición realizada para el espacio transformado proporcione un mejor valor del estadístico de Pearson.

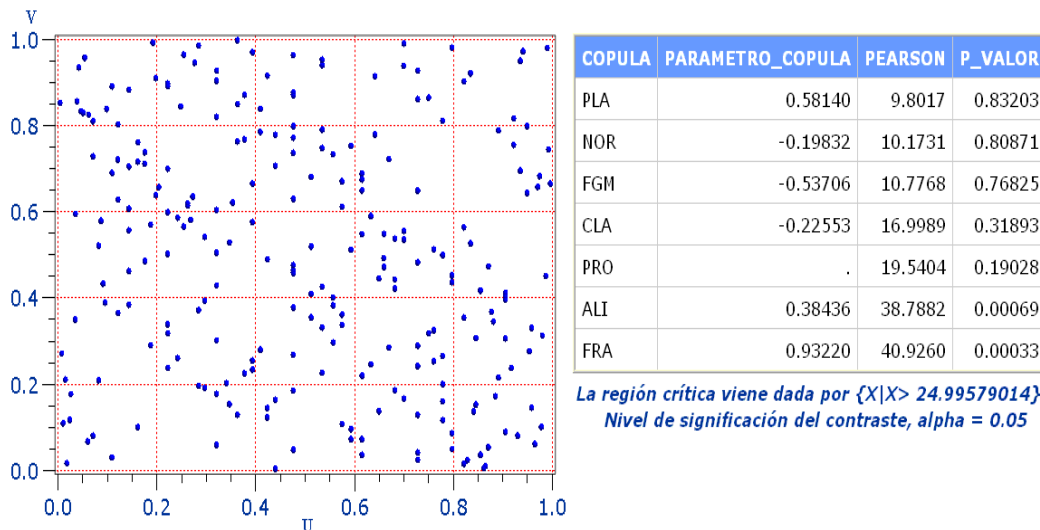


Ilustración 7.53: Partición del espacio transformado de las variables *INC_TMAX* y *Error5* de predicción de la demanda en Enero. Evaluación del estadístico de Pearson para una relación de cópulas

En esta ocasión, la cópula producto ha ganado una nueva posición y no todas las familias son apropiadas para reflejar dicha relación, dado que la estimación del parámetro que permite tomar un representante de la familia no se encuentra comprendido dentro del rango de variación permitido para dicho parámetro. En cuanto a los errores medios de predicción se refiere, los resultados que se obtienen son los siguientes: $Mean'_{HISTORICO} = 3.22\%$, $Median'_{HISTORICO} = 2.66\%$, $Mean'_{FUTURO} = 2.95\%$ y $Median'_{FUTURO} = 1.81\%$, que apenas mejora los últimos resultados conseguidos.

El criterio de parada que hemos decidido adoptar es que bien el error medio o bien el error mediano sobre el histórico deje de mejorar de una iteración a otra o que, alternativamente, la cópula que mejor explica la relación de dependencia entre el par de variables sea la cópula producto. En el caso que nos ocupa, vemos que aun cuando $Mean'_{HISTORICO} = 3.22\%$ mejora el valor de $Mean^*_{HISTORICO} = 3.31\%$, no sucede lo mismo respecto de los errores medianos dado que $Median'_{HISTORICO} = 2.66\%$ es mayor que $Median^*_{HISTORICO} = 2.62\%$. Por tanto, para el mes de Enero, la parada podría establecerse después de haber tanteado una sola vez cada una de las cuatro variables. Podríamos resumir entonces que para realizar una predicción en un día “t” del mes de Enero, el procedimiento consiste en lo siguiente:

1. Realizar una predicción $PI(t)$ a través de un modelo base (ecuación (7.36)) en el que no participan las variables de temperatura.
2. Iteración 1: Utilizar la cópula condicionada de la familia de Plackett cuyo valor del parámetro es 12.23 (ilustración (7.49)) para predecir el error porcentual $Error1(t)$ que se espera cometer en función del incremento esperado para la temperatura máxima desde el día anterior ($t-1$). Supuesto este error porcentual, proponer como nueva predicción $P2(t) = \frac{Error1(t) * PI(t)}{100} + PI(t)$.
3. Iteración 2: Utilizar la cópula condicionada de la familia de Clayton cuyo valor del parámetro es 0.24 (ilustración (7.50)) para predecir el error porcentual $Error2(t)$ que se espera cometer en función del

incremento esperado para la temperatura máxima desde hace dos días ($t-2$). Supuesto este error porcentual,

$$\text{proponer como nueva predicción } P3(t) = \frac{\text{Error}2(t) * P2(t)}{100} + P2(t) \text{ .}$$

4. Iteración 3: Utilizar la cópula condicionada de la familia HRT cuyo valor del parámetro es 1.84 (ilustración (7.51)) para predecir el error porcentual $\text{Error}3(t)$ que se espera cometer en función del incremento esperado para la temperatura mínima desde el día anterior ($t-1$). Supuesto este error porcentual, proponer como nueva

$$\text{predicción } P4(t) = \frac{\text{Error}3(t) * P3(t)}{100} + P3(t) \text{ .}$$

5. Iteración 4: Utilizar nuevamente la cópula condicionada de la familia de Clayton pero esta vez con valor del parámetro 0.10 (ilustración (7.52)) para predecir el error porcentual $\text{Error}4(t)$ que se espera cometer en función del incremento esperado para la temperatura mínima desde hace dos días ($t-2$). Supuesto este error porcentual,

$$\text{proponer como predicción final } P5(t) = \frac{\text{Error}4(t) * P4(t)}{100} + P4(t) \text{ .}$$

La siguiente tabla viene a resumir los errores porcentuales a los que se ha llegado para cada uno de los meses una vez completado el proceso de ajuste basado en las cuatro variables explicativas propuestas y después de probar un par de veces con cada una de ellas desde el $\text{Error}1$ asociado a la predicción del modelo (7.36), al $\text{Error}9$ resultante de intentar predecir el $\text{Error}8$ con la variable INC_TMIN_2 en su segunda participación.

De acuerdo al criterio de parada propuesto, el proceso se detiene después de un número diferente de iteraciones para cada mes. Se ha marcado con un rectángulo rojo la iteración desde la cual el error medio y el error mediano deja de mejorar, las cuales no tienen por qué coincidir. La mínima de ellas sirve para establecer el criterio de parada, pudiendo comprobarse que, en la mayoría de los casos, resulta ser una decisión acertada a tenor de lo que se observa en la evolución de los errores fuera de muestra, habiéndose marcado en azul la iteración tras la cual se detendría el algoritmo en función de lo establecido por el histórico. Así por ejemplo, el empeoramiento de resultados que se produce en el mes de Diciembre después de la primera iteración sobre el conjunto de entrenamiento, se refleja también posteriormente en el conjunto de validación, con lo que realizar la parada en ese preciso momento parece haber sido una decisión acertada. Además, si comparamos los errores finales en cada caso con los errores obtenidos para el modelo (7.38), recogidos en la tabla (7.31), se puede apreciar una mejora sustancial generalizada salvo en el mes de Diciembre en el que los resultados fuera de muestra parecen decantarse ligeramente a favor del modelo ARIMA con funciones de transferencia. Téngase en cuenta que en ambos casos hemos utilizado el mismo tipo de variables (incrementos de temperatura) e incluso podríamos decir que hemos empleado un número menor de ellas en el desarrollo mediante cópulas puesto que no hemos tenido en cuenta incrementos climáticos a 3 días que sí han sido contemplados en el modelo (7.38). Dado que, aun con todo y con eso, los resultados a los que se ha llegado pueden considerarse mejores en líneas generales, estaríamos en disposición de admitir la validez del método propuesto como alternativa de solución para un estudio de predicción a corto plazo de este tipo.

Errores en muestra (01/11/1997-31/10/2005)

MES	MEAN_E1	MEAN_E2	MEAN_E3	MEAN_E4	MEAN_E5	MEAN_E6	MEAN_E7	MEAN_E8	MEAN_E9
TOTAL	5.81	3.81	3.97	3.79	3.74	3.68	3.60	3.54	3.54
1	5.10	3.49	3.49	3.33	3.31	3.22	3.22	3.22	3.22
2	5.45	3.26	3.19	2.94	2.94	2.89	2.89	2.89	2.90
3	7.47	4.70	4.56	4.54	4.54	4.48	4.47	4.45	4.46
11	6.03	4.01	3.99	3.85	3.85	3.71	3.71	3.70	3.70
12	4.98	3.55	4.56	4.21	4.01	4.01	3.66	3.38	3.38

MES	P50_E1	P50_E2	P50_E3	P50_E4	P50_E5	P50_E6	P50_E7	P50_E8	P50_E9
TOTAL	4.64	3.05	3.29	3.12	3.01	2.85	2.86	2.82	2.80
1	4.03	2.85	2.84	2.64	2.62	2.66	2.66	2.67	2.66
2	4.32	2.70	2.60	2.28	2.33	2.39	2.41	2.41	2.37
3	5.97	3.61	3.61	3.48	3.42	3.21	3.15	3.18	3.28
11	4.88	3.45	3.35	3.29	3.22	3.08	3.13	3.13	3.12
12	4.17	3.00	4.00	3.57	3.09	3.12	3.03	2.78	2.76

Parar después de 4 iteraciones
 Parar después de 3 iteraciones
 Parar después de 6 iteraciones
 Parar después de 5 iteraciones
 Parar después de 1 iteración

Errores fuera de muestra (01/11/2005-31/03/2006)

MES	MEAN_E1	MEAN_E2	MEAN_E3	MEAN_E4	MEAN_E5	MEAN_E6	MEAN_E7	MEAN_E8	MEAN_E9
TOTAL	5.49	3.57	3.64	3.43	3.39	3.31	3.25	3.27	3.25
1	5.53	3.50	3.41	2.98	2.95	2.95	2.95	2.95	2.95
2	3.97	2.41	2.53	2.54	2.54	2.45	2.43	2.44	2.44
3	6.85	4.25	4.16	4.25	3.98	3.67	3.70	3.59	3.54
11	5.80	3.69	3.57	3.41	3.47	3.48	3.49	3.49	3.48
12	5.17	3.89	4.43	3.90	3.93	3.94	3.63	3.80	3.76

MES	P50_E1	P50_E2	P50_E3	P50_E4	P50_E5	P50_E6	P50_E7	P50_E8	P50_E9
TOTAL	4.30	2.64	2.63	2.58	2.33	2.23	2.21	2.26	2.24
1	4.30	2.90	2.56	1.89	1.86	1.81	1.83	1.83	1.82
2	3.48	1.98	1.87	1.90	1.78	1.71	1.69	1.70	1.69
3	7.45	2.76	2.63	2.90	3.03	2.24	2.18	2.12	1.97
11	4.47	3.06	2.83	2.75	2.72	3.22	3.28	3.36	3.31
12	3.30	2.67	3.84	3.41	2.66	2.69	3.39	2.85	2.95

E1 = Error1 = Error resultante del modelo ARMA que no utiliza funciones de transferencia
 E2 = Error2 = Error resultante de predecir Error1 con la variable INC_TMAX
 E3 = Error3 = Error resultante de predecir Error2 con la variable INC_TMAX_2
 E4 = Error4 = Error resultante de predecir Error3 con la variable INC_TMIN
 E5 = Error5 = Error resultante de predecir Error4 con la variable INC_TMIN_2
 E6 = Error6 = Error resultante de predecir Error5 con la variable INC_TMAX
 E7 = Error7 = Error resultante de predecir Error6 con la variable INC_TMAX_2
 E8 = Error8 = Error resultante de predecir Error7 con la variable INC_TMIN
 E9 = Error9 = Error resultante de predecir Error8 con la variable INC_TMIN_2

Tabla 7.32: Errores de predicción parciales conseguidos al utilizar iterativamente funciones cópula para explicar mediante las variables INC_TMAX/TMIN, los procesos de error que van resultando de sucesivos ajustes. El proceso de error inicial se obtiene tras ajustar un modelo SARIMA(1,0,1)x(0,1,1)₇ con análisis de intervención

Observación

Conviene observar que en este método de ajuste influye el orden en el que van siendo incluidas las variables explicativas para ir explicando los sucesivos términos de error. Así, en el caso anterior el orden establecido fue: INC_TMAX, INC_TMAX_2, INC_TMIN, INC_TMIN_2, INC_TMAX, INC_TMAX_2, INC_TMIN, INC_TMIN_2. Si en su lugar, hubiésemos seguido por ejemplo el orden INC_TMAX, INC_TMAX, INC_TMAX_2, INC_TMAX_2, INC_TMIN, INC_TMIN, INC_TMIN_2, INC_TMIN_2, el resultado hubiese sido el siguiente:

Errores en muestra (01/11/1997-31/10/2005)

MES	MEAN_E1	MEAN_E2	MEAN_E3	MEAN_E4	MEAN_E5	MEAN_E6	MEAN_E7	MEAN_E8	MEAN_E9
TOTAL	5.81	3.81	3.80	3.76	3.76	3.62	3.62	3.61	3.61
1	5.10	3.49	3.48	3.49	3.49	3.34	3.34	3.32	3.32
2	5.45	3.26	3.24	3.20	3.20	2.94	2.94	2.94	2.94
3	7.47	4.70	4.69	4.56	4.56	4.54	4.54	4.54	4.54
11	6.03	4.01	4.01	3.99	3.99	3.85	3.86	3.86	3.86
12	4.98	3.55	3.55	3.53	3.53	3.38	3.37	3.34	3.33

MES	P50_E1	P50_E2	P50_E3	P50_E4	P50_E5	P50_E6	P50_E7	P50_E8	P50_E9
TOTAL	4.64	3.05	3.07	3.04	3.07	2.93	2.97	2.94	2.93
1	4.03	2.82	2.78	2.84	2.87	2.68	2.65	2.59	2.60
2	4.32	2.70	2.63	2.59	2.62	2.37	2.31	2.37	2.38
3	5.97	3.61	3.70	3.61	3.58	3.48	3.47	3.42	3.43
11	4.88	3.45	3.37	3.35	3.35	3.31	3.21	3.20	3.21
12	4.17	3.00	3.03	3.05	3.05	2.77	2.73	2.79	2.71

Parar después de 2 iteraciones

Parar después de 3 iteraciones

Parar después de 1 iteración

Parar después de 5 iteraciones

Parar después de 1 iteración

Errores fuera de muestra (01/11/2005-31/03/2006)

MES	MEAN_E1	MEAN_E2	MEAN_E3	MEAN_E4	MEAN_E5	MEAN_E6	MEAN_E7	MEAN_E8	MEAN_E9
TOTAL	5.49	3.57	3.59	3.53	3.52	3.38	3.39	3.33	3.33
1	5.53	3.50	3.50	3.41	3.40	2.99	2.99	2.97	2.97
2	3.97	2.41	2.45	2.54	2.52	2.52	2.49	2.50	2.50
3	6.85	4.25	4.29	4.16	4.15	4.25	4.28	3.97	3.96
11	5.80	3.69	3.71	3.57	3.56	3.40	3.47	3.49	3.50
12	5.17	3.89	3.89	3.87	3.86	3.64	3.63	3.62	3.62

MES	P50_E1	P50_E2	P50_E3	P50_E4	P50_E5	P50_E6	P50_E7	P50_E8	P50_E9
TOTAL	4.30	2.64	2.63	2.57	2.53	2.58	2.55	2.42	2.40
1	4.30	2.90	2.88	2.64	2.55	1.91	1.82	1.75	1.78
2	3.48	1.98	1.82	1.96	1.96	1.80	1.67	1.68	1.68
3	7.45	2.76	3.06	2.63	2.66	2.91	2.98	3.03	3.07
11	4.47	3.06	3.08	2.83	2.84	2.75	2.75	2.74	2.74
12	3.30	2.67	2.60	2.76	2.69	3.47	3.21	3.14	3.11

E1 = Error1 = Error resultante del modelo ARMA que no utiliza funciones de transferencia
 E2 = Error2 = Error resultante de predecir Error1 con la variable INC_TMAX
 E3 = Error3 = Error resultante de predecir Error2 con la variable INC_TMAX_2
 E4 = Error4 = Error resultante de predecir Error3 con la variable INC_TMIN
 E5 = Error5 = Error resultante de predecir Error4 con la variable INC_TMIN_2
 E6 = Error6 = Error resultante de predecir Error5 con la variable INC_TMAX
 E7 = Error7 = Error resultante de predecir Error6 con la variable INC_TMAX_2
 E8 = Error8 = Error resultante de predecir Error7 con la variable INC_TMIN
 E9 = Error9 = Error resultante de predecir Error8 con la variable INC_TMIN_2

Tabla 7.33: Errores de predicción parciales conseguidos al utilizar iterativamente funciones cópula para explicar mediante las variables INC_TMAX/TMIN, los procesos de error que van resultando de sucesivos ajustes. El proceso de error inicial se obtiene tras ajustar un modelo SARIMA(1,0,1)x(0,1,1); con análisis de intervención

En este caso el resultado sale un poco peor que en el anterior. De entrada podemos ver que el criterio de parada suele actuar en un menor número de iteraciones (2-3-1-5-1 frente a 4-3-6-5-1), algo que a priori puede parecer que no es bueno porque el hecho de que se lleven a cabo más iteraciones es sinónimo de que en cada una de ellas el error de predicción ha ido bajando. Además se ha visto que siguiendo esta secuencia de ajuste alternativa, es mayor el número de veces en el que la mejor cópula para adecuar la relación de dependencia entre variables es la cópula producto o que, de ser otra, ésta no supera el contraste de bondad de ajuste.

Esta es una peculiaridad importante del algoritmo: “los resultados obtenidos están condicionados por el orden en el que entran las variables explicativas en el modelo”.

Hemos visto que para unas mismas variables explicativas, el procedimiento de ajuste basado en un ARIMA y complementado con funciones cópula proporciona mejores resultados que el que se complementa con modelos de función de transferencia. Sin embargo, de acuerdo a lo observado en la ilustración (7.48), podríamos pensar que este último no está suficientemente bien ajustado dado que el proceso residual no verifica las hipótesis de ruido blanco. Con vistas a poder comparar los resultados logrados con las cópulas con un modelo ARIMA de suficientes garantías, vamos a llevar a cabo una transformación específica de las variables de temperatura que aumenta el poder predictivo de éstas. Además de estas nuevas variables vamos a eliminar parte del histórico para ensanchar la banda de confianza que establece si las autocorrelaciones del término de error que deja el modelo son o no significativas y así facilitar que éste supere el test de ruido blanco.

En concreto, empezaremos restringiéndonos a los datos recogidos desde el 1 de Noviembre de 2001 al 31 de Octubre de 2005, es decir, a los últimos cuatro años de histórico. Para este periodo, de acuerdo a la relación que existe entre la demanda y cada una de las variables de temperatura (*TMAX* y *TMIN*) vamos a generar unas funciones *ad-hoc*, (*CURVATMAX* y *CURVATMIN* respectivamente), que recojan esta relación, la cual presentamos en los siguientes gráficos.

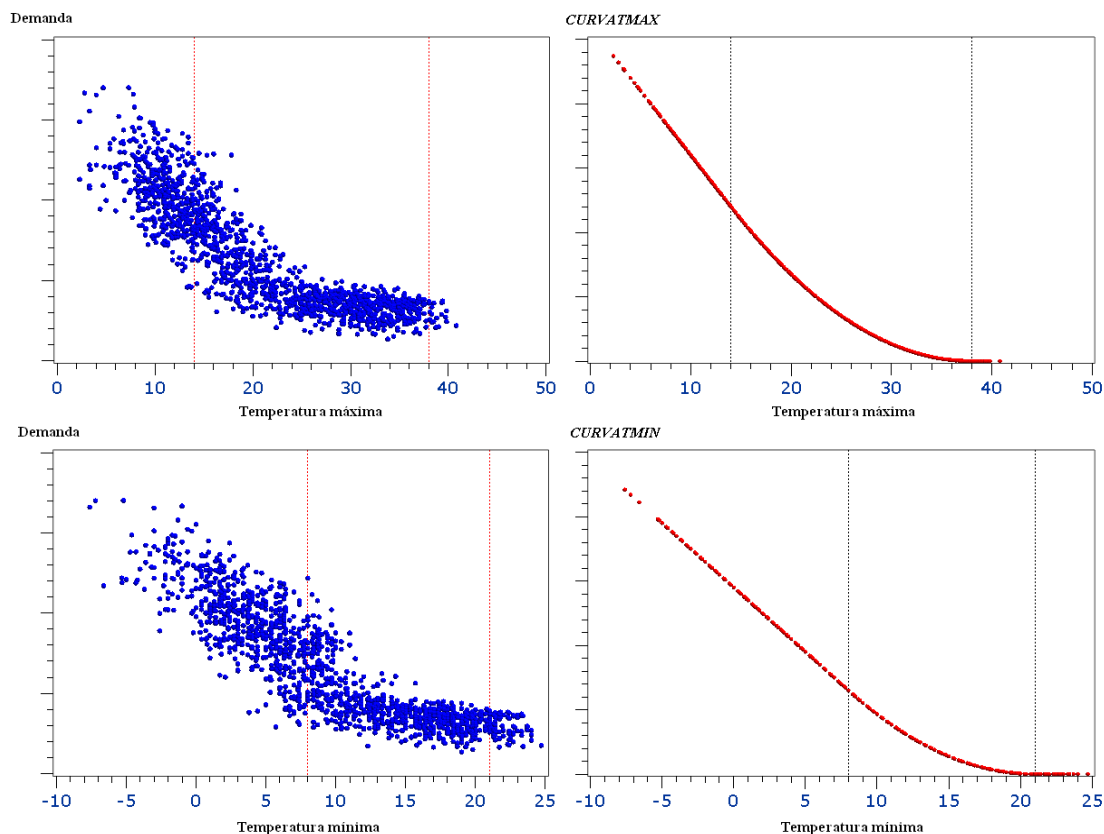


Ilustración 7.54: Transformaciones de las variables de temperatura que aumentan su capacidad para predecir la demanda cuando son utilizadas como modelos de función de transferencia en un modelo ARIMA

Estas funciones están definidas a trozos de acuerdo a unos umbrales a partir de los cuales la relación *demanda/temperatura* es diferente:

- Desde el mínimo valor registrado para la temperatura hasta el primer umbral T_1 la dependencia entre ambas variables es de tipo lineal. En el caso de la temperatura máxima, este límite lo hemos establecido en 14°C mientras que en el caso de la mínima lo hemos establecido en 8°C.
- Desde el umbral anterior hasta un segundo umbral T_2 , la relación entre las variables es de tipo cuadrático. En el caso de la temperatura máxima este valor es de 38°C y en el de la mínima es de 21°C.
- Desde el segundo umbral en adelante, la influencia de la temperatura sobre la demanda se mantiene constante produciéndose un efecto de saturación.

Respecto de estos dos umbrales T_1 y T_2 , la función se define de manera continua a través de la expresión:

$$F(T) = \left\{ \begin{array}{l} 0 \Leftrightarrow T > T_2 \\ \frac{(T-T_2)^2}{2*(T_2-T_1)} \Leftrightarrow T_1 < T \leq T_2 \\ \frac{T_1+T_2}{2} - T \Leftrightarrow T \leq T_1 \end{array} \right\} \quad (7.39)$$

De acuerdo a la observación que ya hemos hecho en anteriores ocasiones, el efecto de la temperatura no es el mismo en los distintos meses del año y por ello vamos a considerar un par de variables para cada uno de los meses no veraniegos, una asociada a la temperatura máxima, $CURVATMAX_MES$, y otra asociada a la temperatura mínima, $CURVATMIN_MES$. Estas variables valen cero fuera del mes de definición, mientras que en dicho mes, toman el valor asociado a la temperatura máxima y mínima del día a través de las respectivas $CURVATMAX$ y $CURVATMIN$. Nuevamente, los modelos de función de transferencia asociadas a estas variables sólo contemplan el valor retardado de las mismas para un máximo de dos días. Una vez incluidas estas variables en la ecuación (7.36) resulta el siguiente modelo:

$$(1-B^7)X(t) = \frac{(1-\theta_1*B)(1-\theta_7*B^7)}{(1-\phi_1*B)} \epsilon(t) + \sum_{i=0}^{Num_intervenciones} p_i * I_{\{t=i\}} + \sum_{MES \in \{E, \dots, D\}} \sum_{jmax=0}^2 \omega_{MES, jmax}(B) * CURVATMAX_{MES} + \sum_{MES \in \{E, \dots, D\}} \sum_{jmin=0}^2 \omega_{MES, jmin}(B) * CURVATMIN_{MES} \quad (7.40)$$

En las siguientes tablas presentamos las estimaciones obtenidas para esta nueva ecuación así como el correlograma asociado al proceso residual. En esta ocasión, no se puede rechazar la hipótesis que plantea que las autocorrelaciones de dicho proceso son iguales a 0 y, en consecuencia, podemos considerar que éste es ruido blanco, con lo que las predicciones ofrecidas por (7.40) ofrecerán más fiabilidad que las obtenidas por (7.38). Para demostrar que esta mejoría del proceso residual es una circunstancia asociada a la inclusión de las nuevas variables ($CURVATMAX$ y $CURVATMIN$ en lugar de INC_TMAX e INC_TMIN), y no al hecho de haber reducido el histórico de entrenamiento, se presentan en la ilustración (7.55) los procesos residuales asociados a uno y otro modelo. Es decir, presentamos también el resultado al que se hubiera llegado con las variables INC_TMAX e INC_TMIN recortando el histórico.

Conditional Least Squares Estimation							
Parameter	Estimate	Standard Error	t Value	Approx Pr > t	Lag	Variable	Shift
MA1,1	-0.13741	0.02844	-4.83	<.0001	1	RCONSUMO	0
MA2,1	0.85929	0.01513	56.79	<.0001	7	RCONSUMO	0
AR1,1	0.95168	0.0092582	102.79	<.0001	1	RCONSUMO	0
NUM1	-0.71317	0.10858	-6.57	<.0001	0	V01JAN	0
NUM2	-0.42139	0.09772	-4.31	<.0001	0	V06jan	0
NUM3	-0.36972	0.06071	-6.09	<.0001	0	V19MAR	0
NUM4	-0.81470	0.11333	-7.19	<.0001	0	juesanto	0
NUM5	-0.94717	0.14214	-6.66	<.0001	0	viesanto	0
NUM6	-0.69169	0.08441	-8.19	<.0001	0	sabsanto	0
NUM7	-0.26732	0.11782	-2.27	0.0234	0	V01MAY	0
NUM8	-0.45729	0.07557	-6.05	<.0001	0	V02MAY	0
NUM9	-0.19540	0.06141	-3.18	0.0015	0	V15MAY	0
NUM10	-0.45373	0.09280	-4.89	<.0001	0	V15AUG	0
NUM11	-0.16603	0.08315	-2.00	0.0460	0	V12OCT	0
NUM12	-0.44033	0.12360	-3.56	0.0004	0	V01NOV	0
NUM13	0.12631	0.05167	2.44	0.0146	0	V09NOV	0
NUM14	-0.40747	0.08692	-4.69	<.0001	0	V06DEC	0
NUM15	-0.41605	0.09131	-4.56	<.0001	0	V08DEC	0
NUM16	-0.44474	0.08071	-5.51	<.0001	0	V24DEC	0
NUM17	-0.77431	0.09011	-8.59	<.0001	0	V25DEC	0
NUM18	-0.32891	0.08671	-3.79	0.0002	0	V31DEC	0
NUM19	-0.31537	0.03632	-8.68	<.0001	0	PUENTE	0
NUM20	-0.08493	0.03202	-2.65	0.0081	0	VISPFES	0
NUM21	-0.17634	0.03009	-5.86	<.0001	0	POSFES	0
NUM22	-0.27552	0.07363	-3.74	0.0002	0	FESLUN	0
NUM23	-0.49616	0.08690	-5.71	<.0001	0	FESMAR	0
NUM24	-0.37030	0.07618	-4.86	<.0001	0	FESMIE	0
NUM25	-0.35920	0.07526	-4.77	<.0001	0	FESJUE	0
NUM26	-0.49635	0.09951	-4.99	<.0001	0	FESVIE	0
NUM27	0.10006	0.0059981	16.68	<.0001	0	CURVATMAX_ENERO	0
NUM1,1	-0.04998	0.0047702	-10.48	<.0001	1	CURVATMAX_ENERO	0
NUM1,2	-0.01877	0.0040650	-4.62	<.0001	2	CURVATMAX_ENERO	0
NUM28	0.10483	0.0057686	18.17	<.0001	0	CURVATMAX_FEBRERO	0
NUM1,1	-0.03986	0.0059016	-6.75	<.0001	1	CURVATMAX_FEBRERO	0
NUM1,2	-0.0081637	0.0037598	-2.17	0.0301	2	CURVATMAX_FEBRERO	0
NUM29	0.13187	0.0057737	22.84	<.0001	0	CURVATMAX_MARZO	0
NUM1,1	-0.05432	0.0050618	-10.73	<.0001	1	CURVATMAX_MARZO	0
NUM30	0.12130	0.0061400	19.76	<.0001	0	CURVATMAX_ABRIL	0
NUM1,1	-0.05804	0.0059568	-9.74	<.0001	1	CURVATMAX_ABRIL	0
NUM31	0.10400	0.0078250	13.29	<.0001	0	CURVATMAX_MAYO	0
NUM1,1	-0.07401	0.0089186	-8.30	<.0001	1	CURVATMAX_MAYO	0
NUM32	0.07232	0.0070710	10.23	<.0001	0	CURVATMAX_OCTUBRE	0
NUM1,1	-0.02872	0.0077929	-3.69	0.0002	1	CURVATMAX_OCTUBRE	0
NUM33	0.09474	0.0055101	17.19	<.0001	0	CURVATMAX_NOVIEMBRE	0
NUM1,1	-0.04558	0.0054124	-8.42	<.0001	1	CURVATMAX_NOVIEMBRE	0
NUM1,2	-0.02190	0.0048946	-4.48	<.0001	2	CURVATMAX_NOVIEMBRE	0
NUM34	0.09729	0.0059767	16.28	<.0001	0	CURVATMAX_DICIEMBRE	0
NUM1,1	-0.05059	0.0048421	-10.45	<.0001	1	CURVATMAX_DICIEMBRE	0
NUM1,2	-0.02420	0.0043281	-5.59	<.0001	2	CURVATMAX_DICIEMBRE	0
NUM35	0.03780	0.0064014	5.91	<.0001	0	CURVATMIN_ENERO	0
NUM36	0.03394	0.0063814	5.32	<.0001	0	CURVATMIN_FEBRERO	0
NUM1,1	-0.01414	0.0064392	-2.20	0.0283	1	CURVATMIN_FEBRERO	0
NUM37	0.04318	0.0070169	6.15	<.0001	0	CURVATMIN_NOVIEMBRE	0
NUM38	0.03026	0.0070080	4.32	<.0001	0	CURVATMIN_DICIEMBRE	0
NUM39	0.0073440	0.0064706	1.13	0.2566	0	CURVATMIN_MARZO	0
NUM40	0.01431	0.0066824	2.14	0.0325	0	CURVATMIN_ABRIL	0
NUM41	-0.04504	0.01005	-4.48	<.0001	0	CURVATMIN_MAYO	0
NUM42	0.02923	0.01052	2.78	0.0055	0	CURVATMIN_OCTUBRE	0

Tabla 7.34: Estimaciones de los parámetros del SARIMA(1,0,1)x(0,1,1), con análisis de intervención y modelos de función de transferencia asociadas a las variables CURVATMAX/TMIN

PROCESO RESIDUAL CON VARIABLES TMAXCON y TMINCON

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	5.47	3	0.1403	-0.003	-0.026	-0.055	0.004	0.004	0.000
12	10.19	9	0.3350	-0.017	-0.007	0.020	-0.020	-0.010	-0.045
18	16.94	15	0.3222	0.032	-0.030	0.043	0.017	-0.013	0.019
24	23.32	21	0.3270	-0.036	-0.011	0.014	0.022	0.045	0.013
30	30.97	27	0.2722	0.041	-0.032	0.015	0.029	0.027	-0.026
36	39.63	33	0.1984	0.056	-0.007	0.025	-0.005	0.006	0.044
42	49.07	39	0.1295	0.041	0.016	0.029	0.033	0.048	0.012
48	55.09	45	0.1441	-0.035	-0.004	0.004	0.044	-0.026	0.013

PROCESO RESIDUAL CON VARIABLES INC_TMAX e INC_TMIN

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	54.80	3	<.0001	0.058	0.127	-0.082	-0.071	-0.075	-0.025
12	61.33	9	<.0001	-0.028	-0.030	-0.035	-0.004	0.010	-0.038
18	62.93	15	<.0001	-0.011	-0.006	0.019	-0.005	0.014	0.019
24	69.32	21	<.0001	-0.052	-0.017	0.010	0.006	0.033	-0.008
30	75.29	27	<.0001	0.029	-0.041	-0.006	0.019	0.033	0.006
36	81.89	33	<.0001	0.047	0.009	0.029	0.018	0.028	0.014
42	99.99	39	<.0001	0.043	0.023	0.061	0.023	0.070	0.026
48	104.00	45	<.0001	-0.027	-0.013	-0.016	0.018	-0.015	-0.031

Autocorrelation Plot of Residuals												Autocorrelation Plot of Residuals												
Lag	Covariance	Correlation	-1	9	8	7	6	5	4	3	2	1	0	1	2	3	4	5	6	7	8	9	1	Std Error
0	0.034724	1.00000													*****									0
1	-0.0000888	-.00256													.									0.026243
2	-0.0009133	-.02630													*									0.026243
3	-0.0019101	-.05501													*									0.026262
4	0.00014688	0.00423													.									0.026341
5	0.00013240	0.00381													.									0.026341
6	0.00001429	0.00041													.									0.026342
7	-0.0005853	-.01686													.									0.026342
8	-0.0002585	-.00744													.									0.026349
9	0.00069684	0.02007													.									0.026350
10	-0.0006799	-.01958													.									0.026361
11	-0.0003347	-.00964													.									0.026371
12	-0.0015544	-.04476													*									0.026373
13	0.0011110	0.03200													*									0.026426
14	-0.0010549	-.03038													*									0.026452
15	0.0014869	0.04282													*									0.026476
16	0.00059307	0.01708													.									0.026524
17	-0.0004561	-.01313													.									0.026532
18	0.00065061	0.01874													.									0.026536
19	-0.0012587	-.03625													*									0.026545
20	-0.0003978	-.01145													.									0.026579
21	0.00049058	0.01413													.									0.026583
22	0.00076916	0.02215													.									0.026588
23	0.0015590	0.04490													*									0.026601
24	0.00045464	0.01309													.									0.026653

Ilustración 7.55: Test de ruido blanco y correlograma del SARIMA(1,0,1)x(0,1,1)₇ con análisis de intervención y modelos de función de transferencia asociados a las variables CURVATMAX/TMIN e INC_TMAX/TMIN

Esta notable mejoría en el ajuste del modelo a partir del nuevo tratamiento de la temperatura, se pone de manifiesto también en la calidad de los resultados en términos de error tanto sobre el histórico como fuera de muestra, tal y como se muestra en las tablas siguientes:

Histórico de entrenamiento: (01/11/2001- 31/10/2005)

Histórico de entrenamiento: (01/11/2001- 31/10/2005)

Modelos de función de transferencia: *CURVATMAX* y *CURVATMIN*

Modelos de función de transferencia: *INC_TMAX* e *INC_TMIN*

Errores en muestra			Errores fuera de muestra			Errores en muestra			Errores fuera de muestra		
MES	MEDIA	MEDIANA	MES	MEDIA	MEDIANA	MES	MEDIA	MEDIANA	MES	MEDIA	MEDIANA
TOTAL	3.24	2.57	TOTAL	3.08	2.35	TOTAL	4.55	3.38	TOTAL	4.52	3.67
1	3.01	2.22	1	2.43	1.59	1	4.30	3.63	1	4.53	3.71
2	2.46	2.14	2	2.60	1.90	2	3.58	3.12	2	3.94	3.19
3	4.25	3.53	3	3.79	3.15	3	5.85	4.82	3	5.23	5.11
11	3.65	2.62	11	3.61	2.85	11	4.77	3.02	11	4.91	3.85
12	2.82	2.54	12	2.97	2.12	12	4.18	3.33	12	3.95	3.15

Tabla 7.35: Errores de predicción del SARIMA(1,0,1)x(0,1,1)₇ con análisis de intervención y modelos de función de transferencia asociados a las variables CURVATMAX/TMIN e INC_TMAX/TMIN

Los errores medios y medianos de predicción son claramente favorables al modelo que utiliza las transformaciones dadas por (7.39) para las variables de temperatura. Los resultados asociados al modelo que utiliza las variables *INC_TMAX* e *INC_TMIN* salen incluso un poco peores que aquéllos que fueron obtenidos cuando contemplábamos todo el histórico (tabla (7.31)).

Así, terminamos decantándonos por el modelo (7.40) como alternativa a la solución propuesta mediante cópulas. Incluso podría proponerse un modelo para el que el término de error está aún más incorrelado con su pasado, es decir, que responde todavía mejor a la hipótesis de ruido blanco que sobre él se hace. Para ello, basta observar el correlograma de la izquierda presentado en la ilustración (7.55) y darse cuenta de que las autocorrelaciones de órdenes 2 y 3 presentan cierta significatividad. Basta incluir dos nuevos parámetros asociados a ellos en la parte regular de medias móviles θ_2 y θ_3 . El modelo resultante sería un SARIMA(1,0,3)x(0,1,1)₇ que responde a la expresión:

$$(1-B^7)X(t) = \frac{(1-\theta_1*B-\theta_2*B^2-\theta_3*B^3)(1-\Theta_7*B^7)}{(1-\Phi_1*B)} \epsilon(t) + \sum_{i=0}^{Num_intervenciones} p_i * I_{\{t=i\}} + \sum_{MES \in \{E, \dots, D\}} \sum_{jmax=0}^2 \omega_{MES, jmax}(B) * CURVATMAX_{MES} + \sum_{MES \in \{E, \dots, D\}} \sum_{jmin=0}^2 \omega_{MES, jmin}(B) * CURVATMIN_{MES} \quad (7.41)$$

Según se puede comprobar en la siguiente tabla, el nuevo modelo presenta estimaciones significativas para los nuevos parámetros y además no se establecen altas correlaciones entre los parámetros básicos del modelo, es decir, aquéllos que hacen referencia a la parte autorregresiva y a la parte de medias móviles. En cuanto al test de ruido blanco se refiere, existe aún menos evidencia empírica para rechazar que las autocorrelaciones del proceso residual valgan cero.

Conditional Least Squares Estimation							
Parameter	Estimate	Standard Error	t Value	Approx Pr > t	Lag	Variable	Shift
MA1,1	-0.11429	0.02905	-3.93	<.0001	1	RCONSUMO	0
MA1,2	0.04986	0.02969	1.68	0.0932	2	RCONSUMO	0
MA1,3	0.08076	0.02907	2.78	0.0055	3	RCONSUMO	0
MA2,1	0.87067	0.01490	58.43	<.0001	7	RCONSUMO	0
AR1,1	0.96518	0.0087897	109.81	<.0001	1	RCONSUMO	0

Correlations of Parameter Estimates					
Variable Parameter	RCONSUMO MA1,1	RCONSUMO MA1,2	RCONSUMO MA1,3	RCONSUMO MA2,1	RCONSUMO AR1,1
RCONSUMO MA1,1	1.000	0.215	0.063	0.125	0.326
RCONSUMO MA1,2	0.215	1.000	0.210	0.189	0.376
RCONSUMO MA1,3	0.063	0.210	1.000	0.111	0.308
RCONSUMO MA2,1	0.125	0.189	0.111	1.000	0.392
RCONSUMO AR1,1	0.326	0.376	0.308	0.392	1.000

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	0.44	1	0.5062	0.001	0.001	-0.001	-0.014	-0.007	-0.008
12	7.45	7	0.3836	-0.019	-0.016	0.006	-0.030	-0.023	-0.052
18	13.49	13	0.4108	0.026	-0.036	0.035	0.010	-0.025	0.015
24	19.42	19	0.4303	-0.044	-0.016	0.014	0.017	0.036	0.008
30	27.17	25	0.3476	0.040	-0.034	0.011	0.033	0.020	-0.029
36	35.43	31	0.2672	0.052	-0.014	0.021	-0.003	0.009	0.045
42	44.02	37	0.1988	0.039	0.015	0.032	0.028	0.044	0.014
48	49.24	43	0.2377	-0.032	-0.006	-0.003	0.037	-0.029	0.013

Autocorrelation Plot of Residuals																								
Lag	Covariance	Correlation	-1	9	8	7	6	5	4	3	2	1	0	1	2	3	4	5	6	7	8	9	1	Std Error
0	0.034585	1.00000												*****										0
1	0.00004128	0.00119												.										0.026243
2	0.00004986	0.00144												.										0.026243
3	-0.0000255	-0.00074												.										0.026243
4	-0.0004771	-0.01380												.										0.026243
5	-0.0002300	-0.00665												.										0.026248
6	-0.0002775	-0.00802												.										0.026249
7	-0.0006667	-0.01928												.										0.026251
8	-0.0005577	-0.01613												.										0.026261
9	0.00021659	0.00626												.										0.026268
10	-0.0010247	-0.02963												.										0.026269
11	-0.0007936	-0.02295												.										0.026292
12	-0.0017999	-0.05204												.										0.026306
13	0.00089686	0.02593												.	*									0.026376
14	-0.0012447	-0.03599												.	*									0.026394
15	0.0012114	0.03503												.	*									0.026428
16	0.00033337	0.00964												.										0.026460
17	-0.0008515	-0.02462												.										0.026462
18	0.00050953	0.01473												.										0.026478
19	-0.0015203	-0.04396												.	*									0.026483
20	-0.0005557	-0.01607												.										0.026534
21	0.00048422	0.01400												.										0.026540
22	0.00058475	0.01691												.										0.026545
23	0.0012408	0.03588												.	*									0.026553
24	0.00026929	0.00779												.										0.026586

Ilustración 7.56: Test de ruido blanco, correlograma, estimaciones de los parámetros y correlaciones entre ellos para el SARIMA(1,0,3)x(0,1,1)₇ con análisis de intervención y modelos de función de transferencia asociados a las variables CURVATMAX/TMIN

Lo único que podría atacarse a este modelo es el hecho de que puede parecer poco verosímil que la predicción a proporcionar para un día “t” tenga en cuenta el error que el modelo cometió hace 2 y 3 días. Teniendo en cuenta que los resultados en término de error sobre el histórico y fuera de muestra son muy similares a los que proporciona el modelo (7.40) podríamos quedarnos con el anterior (SARIMA(1,0,1)x(0,1,1)₇) de acuerdo a lo que establece el principio de parsimonia.

Histórico de entrenamiento: (01/11/2001- 31/10/2005)
 Modelos de función de transferencia: *CURVATMAX* y *CURVATMIN*
 SARIMA (1,0,1) x (0,1,1)₇

Histórico de entrenamiento: (01/11/2001- 31/10/2005)
 Modelos de función de transferencia: *CURVATMAX* y *CURVATMIN*
 SARIMA (1,0,3) x (0,1,1)₇

Errores en muestra			Errores fuera de muestra			Errores en muestra			Errores fuera de muestra		
MES	MEDIA	MEDIANA	MES	MEDIA	MEDIANA	MES	MEDIA	MEDIANA	MES	MEDIA	MEDIANA
TOTAL	3.24	2.57	TOTAL	3.08	2.35	TOTAL	3.22	2.52	TOTAL	3.04	2.37
1	3.01	2.22	1	2.43	1.59	1	3.01	2.14	1	2.36	1.65
2	2.46	2.14	2	2.60	1.90	2	2.45	2.00	2	2.55	1.91
3	4.25	3.53	3	3.79	3.15	3	4.22	3.54	3	3.77	3.06
11	3.65	2.62	11	3.61	2.85	11	3.61	2.38	11	3.58	2.89
12	2.82	2.54	12	2.97	2.12	12	2.79	2.51	12	2.94	2.21

Tabla 7.36: Errores de predicción de los modelos SARIMA(1,0,1)x(0,1,1)₇ y SARIMA(1,0,3)x(0,1,1)₇ con análisis de intervención y modelos de función de transferencia asociados a las variables CURVATMAX/TMIN

Los resultados proporcionados por cualquiera de estos dos modelos de predicción mejoran los que hemos obtenido en el ajuste que combina la predicción del modelo ARIMA con la corrección llevada a cabo mediante funciones cópula. Estos últimos, extraídos de la tabla (7.32) son los que se presentan en la siguiente tabla:

Histórico de entrenamiento: (01/11/1997- 31/10/2005)
 Variables explicativas: INC_TMAX (_2), INC_TMIN (_2)

Errores en muestra			Errores fuera de muestra		
MES	MEDIA	MEDIANA	MES	MEDIA	MEDIANA
TOTAL	3.61	2.92	TOTAL	3.33	2.22
1	3.31	2.62	1	2.95	1.86
2	2.94	2.28	2	2.54	1.90
3	4.47	3.15	3	3.70	2.18
11	3.71	3.08	11	3.48	3.22
12	3.55	3.00	12	3.89	2.67

Tabla 7.37: Errores de predicción finales conseguidos al utilizar iterativamente funciones cópula para explicar mediante las variables INC_TMAX/TMIN, los procesos de error que van resultando de sucesivos ajustes. El proceso de error inicial se obtiene tras ajustar un modelo SARIMA(1,0,1)x(0,1,1)₇ con análisis de intervención

La comparación de unos y otros resultados inclinan la balanza ligeramente a favor de los ARIMA que utilizan modelos de función de transferencia. Sin embargo, teniendo en cuenta que éstos emplean unas variables de temperatura que poseen un mayor poder de predicción que las que hemos usado para el ajuste mediante cópulas, y que a igualdad de factores explicativos los resultados obtenidas para estas últimas eran mejores, parece interesante plantearse la corrección del término de error del modelo ARIMA dado por (7.36) respecto de las variables climáticas

que resultan de aplicar la transformación (7.39). El único inconveniente que podemos encontrar es que en este caso, no está tan clara la independencia temporal de la variable de temperatura que vamos a utilizar para ir explicando los sucesivos términos de error que vayan surgiendo conforme vayamos aplicando el algoritmo, circunstancia que según se explica en [FERSCA] juega en contra del empleo de funciones cópula.

Hemos aplicado nuevamente el algoritmo descrito en el apartado 6.1 utilizando *CURVATMAX* y su retardo *CURVATMAX_2* y *CURVATMIN* y su retardo *CURVATMIN_2* como variables explicativas habiendo distinguido dos casos:

- En el primero de ellos hemos ajustado el modelo ARIMA (7.36) al histórico comprendido entre el 1 de Noviembre de 1997 y el 31 de Octubre de 2005, siendo el proceso residual resultante el que hemos ido aproximando de forma iterativa con las variables mencionadas.
- En el segundo de ellos hemos ajustado el modelo ARIMA (7.36) al histórico comprendido entre el 1 de Noviembre de 2001 y el 31 de Octubre de 2005, dado que ha sido éste el utilizado en los modelos de expresión (7.40) y (7.41) que hasta ahora han proporcionado las mejores predicciones. Un posible inconveniente es que, dado que el histórico es de menor dimensión, también lo será el número de datos de las sucesivas variables de error que tendremos para el ajuste de las funciones cópula.

Errores en muestra (01/11/1997-31/10/2005)

MES	MEAN_E1	MEAN_E2	MEAN_E3	MEAN_E4	MEAN_E5	MEAN_E6	MEAN_E7	MEAN_E8	MEAN_E9
TOTAL	5.81	5.31	4.60	4.79	4.65	4.28	4.14	4.04	3.91
1	5.10	4.50	3.93	3.77	3.64	3.47	3.38	3.36	3.35
2	5.45	4.75	4.20	4.02	3.89	3.56	3.37	3.33	3.31
3	7.47	7.06	6.25	6.19	6.10	5.60	5.35	5.36	5.36
11	6.03	5.60	4.79	6.39	6.13	5.47	5.31	4.86	4.25
12	4.98	4.61	3.82	3.61	3.49	3.31	3.29	3.29	3.26

MES	P50_E1	P50_E2	P50_E3	P50_E4	P50_E5	P50_E6	P50_E7	P50_E8	P50_E9	
TOTAL	4.64	4.20	3.64	3.64	3.50	3.22	3.11	3.09	2.97	
1	4.03	3.31	2.95	2.70	2.75	2.82	2.57	2.47	2.57	Parar después de 3 iteraciones
2	4.32	3.49	3.45	3.15	3.23	2.62	2.50	2.49	2.46	Parar después de 3 iteraciones
3	5.97	5.48	4.77	4.79	4.70	4.51	4.03	4.09	4.13	Parar después de 2 iteraciones
11	4.88	4.65	4.04	5.62	5.26	5.06	4.77	4.22	3.52	Parar después de 2 iteraciones
12	4.17	3.86	3.26	3.10	3.01	2.90	2.87	2.87	2.75	Parar después de 8 iteraciones

Errores fuera de muestra (01/11/2005-31/03/2006)

MES	MEAN_E1	MEAN_E2	MEAN_E3	MEAN_E4	MEAN_E5	MEAN_E6	MEAN_E7	MEAN_E8	MEAN_E9
TOTAL	5.49	5.04	4.19	4.44	4.27	3.95	3.76	3.66	3.59
1	5.53	4.77	3.96	3.77	3.78	3.48	3.05	3.08	3.05
2	3.97	3.91	3.06	3.33	3.09	3.07	2.75	2.82	2.74
3	6.85	6.22	5.67	5.61	5.64	4.91	4.76	4.74	4.73
11	5.80	5.73	4.11	5.68	5.21	4.75	4.64	4.05	3.87
12	5.17	4.47	4.03	3.77	3.52	3.48	3.52	3.52	3.49

MES	P50_E1	P50_E2	P50_E3	P50_E4	P50_E5	P50_E6	P50_E7	P50_E8	P50_E9
TOTAL	4.30	3.58	3.27	3.45	3.18	2.92	2.75	2.48	2.60
1	4.30	2.63	2.72	2.19	2.75	2.48	2.02	1.96	1.95
2	3.48	3.22	2.39	2.71	2.85	2.64	2.26	2.47	2.30
3	7.45	6.00	4.99	4.54	4.37	4.23	3.57	3.54	3.59
11	4.47	4.82	3.77	4.77	4.94	3.46	4.02	2.20	3.46
12	3.30	3.16	3.33	2.48	2.50	2.35	2.24	2.30	2.60

E1 = Error1 = Error resultante del modelo ARMA que no utiliza funciones de transferencia
 E2 = Error2 = Error resultante de predecir Error1 con la variable CURVATMAX
 E3 = Error3 = Error resultante de predecir Error2 con la variable CURVATMAX_2
 E4 = Error4 = Error resultante de predecir Error3 con la variable CURVATMIN
 E5 = Error5 = Error resultante de predecir Error4 con la variable CURVATMIN_2
 E6 = Error6 = Error resultante de predecir Error5 con la variable CURVATMAX
 E7 = Error7 = Error resultante de predecir Error6 con la variable CURVATMAX_2
 E8 = Error8 = Error resultante de predecir Error7 con la variable CURVATMIN
 E9 = Error9 = Error resultante de predecir Error8 con la variable CURVATMIN_2

Tabla 7.38: Errores de predicción parciales conseguidos al utilizar iterativamente funciones cópula para explicar mediante las variables CURVATMAX/TMIN, los procesos de error que van resultando de sucesivos ajustes. El proceso de error inicial se obtiene tras ajustar un modelo SARIMA(1,0,1)x(0,1,1)₇ con análisis de intervención

Modelo entrenado con datos del 1/11/1997 al 31/10/2005

Errores en muestra (01/11/2001-31/03/2005)

MES	MEAN_E1	MEAN_E2	MEAN_E3	MEAN_E4	MEAN_E5	MEAN_E6	MEAN_E7	MEAN_E8	MEAN_E9
TOTAL	5.93	5.45	4.84	4.69	4.45	4.14	4.06	4.02	4.01
1	5.65	5.13	4.63	4.44	3.95	3.79	3.73	3.66	3.63
2	5.11	4.41	3.81	3.59	3.53	3.22	3.14	3.07	3.09
3	7.53	7.13	6.37	6.26	6.19	5.66	5.45	5.43	5.43
11	6.46	6.05	5.36	5.17	4.83	4.47	4.44	4.35	4.33
12	4.86	4.52	4.00	3.92	3.69	3.53	3.50	3.52	3.51

MES	P50_E1	P50_E2	P50_E3	P50_E4	P50_E5	P50_E6	P50_E7	P50_E8	P50_E9	
TOTAL	4.54	4.41	3.85	3.75	3.41	3.13	3.04	2.98	3.00	
1	4.30	3.81	3.70	3.27	3.12	2.91	2.88	2.85	2.80	Parar después de 8 iteraciones
2	4.28	3.15	3.30	3.04	3.04	2.43	2.48	2.43	2.35	Parar después de 5 iteraciones
3	6.32	5.76	5.26	4.89	4.85	4.06	4.30	4.29	4.41	Parar después de 5 iteraciones
11	4.97	4.79	4.12	3.77	3.56	3.52	3.33	3.17	3.17	Parar después de 7 iteraciones
12	4.17	3.64	3.42	3.83	2.96	2.92	2.82	2.85	2.81	Parar después de 2 iteraciones

Errores fuera de muestra (01/11/2005-31/03/2006)

MES	MEAN_E1	MEAN_E2	MEAN_E3	MEAN_E4	MEAN_E5	MEAN_E6	MEAN_E7	MEAN_E8	MEAN_E9
TOTAL	5.89	5.54	4.63	4.52	4.40	4.20	3.98	3.97	3.95
1	5.58	5.05	4.21	4.14	4.13	4.00	3.73	3.77	3.69
2	4.10	4.09	3.06	3.24	3.10	3.31	2.94	3.00	2.96
3	7.16	6.69	5.75	5.62	5.70	5.15	4.74	4.73	4.74
11	7.06	6.89	5.68	5.45	5.27	4.99	4.90	4.77	4.81
12	5.40	4.88	4.33	4.07	3.70	3.49	3.50	3.49	3.47

MES	P50_E1	P50_E2	P50_E3	P50_E4	P50_E5	P50_E6	P50_E7	P50_E8	P50_E9
TOTAL	4.77	4.24	3.55	3.53	3.12	3.33	3.12	3.08	3.18
1	5.36	3.42	3.80	3.78	3.46	3.33	3.12	2.98	2.85
2	3.27	2.86	1.84	2.36	1.93	2.89	2.60	2.73	2.91
3	6.62	6.87	4.59	4.49	4.32	4.14	3.41	3.42	3.22
11	4.51	5.01	3.90	3.38	4.23	3.85	3.81	3.55	3.84
12	4.61	3.50	3.16	3.44	2.76	2.28	2.61	2.38	2.51

E1 = Error1 = Error resultante del modelo ARMA que no utiliza funciones de transferencia
 E2 = Error2 = Error resultante de predecir Error1 con la variable CURVATMAX
 E3 = Error3 = Error resultante de predecir Error2 con la variable CURVATMAX_2
 E4 = Error4 = Error resultante de predecir Error3 con la variable CURVATMIN
 E5 = Error5 = Error resultante de predecir Error4 con la variable CURVATMIN_2
 E6 = Error6 = Error resultante de predecir Error5 con la variable CURVATMAX
 E7 = Error7 = Error resultante de predecir Error6 con la variable CURVATMAX_2
 E8 = Error8 = Error resultante de predecir Error7 con la variable CURVATMIN
 E9 = Error9 = Error resultante de predecir Error8 con la variable CURVATMIN_2

Tabla 7.39: Errores de predicción parciales conseguidos al utilizar iterativamente funciones cópula para explicar mediante las variables CURVATMAX/TMIN, los procesos de error que van resultando de sucesivos ajustes. El proceso de error inicial se obtiene tras ajustar un modelo SARIMA(1,0,1)x(0,1,1), con análisis de intervención

Modelo entrenado con datos del 1/11/2001 al 31/10/2005

Los resultados de estos dos últimos conjuntos de tablas se pueden presentar de forma resumida como sigue:

Histórico de entrenamiento: (01/11/1997- 31/10/2005)

Variables explicativas: $CURVATMAX(2), CURVATMIN(2)$

Errores en muestra			Errores fuera de muestra		
MES	MEDIA	MEDIANA	MES	MEDIA	MEDIANA
TOTAL	4.43	3.39	TOTAL	4.09	3.05
1	3.77	2.70	1	3.77	2.19
2	4.02	3.15	2	3.33	2.71
3	6.25	4.77	3	5.67	4.99
11	4.79	4.04	11	4.11	3.77
12	3.29	2.87	12	3.52	2.30

Histórico de entrenamiento: (01/11/2001- 31/10/2005)

Variables explicativas: $CURVATMAX(2), CURVATMIN(2)$

Errores en muestra			Errores fuera de muestra		
MES	MEDIA	MEDIANA	MES	MEDIA	MEDIANA
TOTAL	4.19	3.21	TOTAL	4.26	3.29
1	3.63	2.80	1	3.69	2.85
2	3.22	2.43	2	3.31	2.89
3	5.66	4.06	3	5.15	4.14
11	4.35	3.17	11	4.77	3.55
12	4.00	3.42	12	4.33	3.16

Tabla 7.40: Errores de predicción parciales conseguidos al utilizar iterativamente funciones cópula para explicar mediante las variables $CURVATMAX/TMIN$, los procesos de error que van resultando de sucesivos ajustes. El proceso de error inicial se obtiene tras ajustar un modelo $SARIMA(1,0,1) \times (0,1,1)_7$ con análisis de intervención

Los resultados obtenidos son bastante parejos pero, en cualquier caso, peores que los que se obtenían con las variables de incremento de temperatura (véase tabla (7.37)). Es decir, las variables que tan buenos resultados proporcionan cuando son utilizadas como modelos de función de transferencia, no producen el mismo efecto si son utilizadas en el ajuste mediante cópulas. Esto, como hemos señalado, puede ser una consecuencia del hecho de que existe cierta estructura de dependencia temporal en estas variables.

Supongamos por ejemplo que un martes 14 de Noviembre se hiciera una predicción para el día 15, P_{15NOV} , a través del modelo (7.36) que no tiene en cuenta el efecto de la meteorología. Tal y como está planteado el modelo, dicha predicción estaría muy condicionada por el dato real de demanda del día 14 ($p = 1$), por el dato real de la semana pasada ($D = 1$) y por los posibles errores de predicción asociados a esos días ($q = 1$ y $Q = 1$). El día 15 en sí no presentaría ningún tipo de intervención adicional por no estar reconocido como una fiesta en el calendario laboral. Dado que el martes y el miércoles son días laborables, a igualdad de temperatura se podría esperar para ellos un mismo nivel de demanda. Así, si la temperatura máxima prevista para el día 15 fuera de 3°C y la temperatura registrada el día 14 estuviese comprendida ente 1°C y 5°C, la demanda no sufriría un cambio drástico de comportamiento y el modelo (7.36) incurriría en un bajo error de predicción. Sin embargo, para esa misma previsión de 3°C, un registro de 10°C el día 14 de Noviembre supondría, casi con total seguridad, que la demanda subiera (podría no ser así si el cambio de temperatura se produjera de un viernes a un sábado o de un laborable a un festivo). En este caso, dado que el modelo (7.36) no recoge el efecto de la temperatura, la predicción P_{15NOV} que proporcionaría el modelo quedaría muy por debajo del dato real y en consecuencia, el modelo daría lugar a un grave error de predicción. Vamos a ver que al intentar hacer la corrección mediante cópulas del error que cometería el modelo el día 15 en este último escenario, llegaríamos a resultados claramente diferenciados en función de que estableciéramos la relación de dicho error con el valor de la variable $CURVATMAX$ o INC_TMAX :

- Si utilizáramos $CURVATMAX$, simularíamos a través de la cópula condicionada, posibles errores que pudiera cometer el modelo (7.36) cuando la temperatura fuese de 3°C en Noviembre, errores que como hemos visto, presentan cierta dependencia temporal. Es decir, si la temperatura el día anterior fuese parecida, el error en que incurriría el modelo sería bajo, mientras que si el valor de temperatura fuese mucho mayor, el error sería

alto. En la muestra de aprendizaje que sirve a su vez de soporte a la función cópula a ajustar, tendríamos errores altos y errores bajos cometidos en Noviembre para un mismo valor de la temperatura, 3°C. Por ello, las simulaciones realizadas no tendrían una tendencia clara hacia errores porcentuales grandes o pequeños y la mediana de las mismas darían lugar a un error intermedio.

- Por otra parte, si utilizáramos *INC_TMAX*, la cópula condicionaría al valor de una caída de temperatura y las simulaciones se harían en función de ese valor. No se condicionaría al hecho de que la temperatura en Noviembre fuese de 3°C sino al hecho de que se hubiera producido una bajada de temperatura en Noviembre de 7°C (de 10°C el 14 de Noviembre a 7°C el día 15). La cópula se entrenaría a partir de caídas bruscas de temperatura en Noviembre ante las cuales no existe tanta dispersión en el término de error. Por lo general, los errores de predicción en estas circunstancias suelen ser muy altos.

Para corroborar estas hipótesis de una manera más formal, podemos utilizar el estadístico de **Durbin-Watson** el cual permite identificar la presencia de autocorrelación positiva o negativa de primer orden, es decir, la existencia de correlación entre el valor de la variable cuya dependencia temporal se cuestiona en el instante “*t*” (Y_t) y en el instante $t-1$ (Y_{t-1}) (la hipótesis nula plantea la ausencia de correlación significativa). Dicho estadístico tiene un rango de variación entre 0 y 4 y responde a la expresión:

$$D-W = \frac{\sum_{t=2}^n (Y_t - Y_{t-1})^2}{\sum_{t=1}^n Y_t^2} \quad (7.42)$$

En líneas generales, se puede considerar que valores del estadístico próximos a 2 indican escasa autocorrelación de la variable que se contrasta, valores entre 0 y 2 una autocorrelación positiva, y valores entre 2 y 4 una autocorrelación negativa. El cálculo de dicho estadístico sobre las distintas versiones de la variable *temperatura máxima*, conduce a los siguientes resultados:

Histórico utilizado: (01/11/1997-31/10/2005)

Evaluación del estadístico de Durbin-Watson

<i>TMAX</i>		<i>CURVATMAX</i>		<i>INC TMAX</i>	
MES	DURBIN_WATSON	MES	DURBIN_WATSON	MES	DURBIN_WATSON
TOTAL	0.0422	TOTAL	0.0387	TOTAL	2.3718
1	0.0611	1	0.0321	1	2.2724
2	0.0415	2	0.0420	2	2.5475
3	0.0295	3	0.0611	3	2.1709
11	0.0374	11	0.0428	11	2.4712
12	0.0560	12	0.0272	12	2.4153

Tabla 7.41: Contraste de dependencia temporal a través del estadístico de Durbin-Watson

La dependencia temporal de la variable original (*TMAX*) es más que evidente, pues la temperatura de un día será habitualmente parecida a la del día anterior. Del mismo modo, la transformación de dicha variable a través de la

expresión (7.39) (*CURVATMAX*) heredará esta propiedad, dado que dos valores de temperatura parecidos (de dos días consecutivos) serán transformados en valores también similares. Es por ello que los valores del estadístico no se mueven en entornos próximos a 2 para ninguna de estas variables. Sin embargo, para la variable *INC_TMAX* el estadístico de Durbin-Watson sí proporciona valores cercanos a 2 lo cual denota la ausencia de autocorrelación de orden 1. Esta circunstancia resulta obvia si tenemos en cuenta que dicha variable se obtiene al aplicar una diferencia de orden 1 sobre *TMAX* (a fin de cuentas un proceso autorregresivo de orden 1 en el que el parámetro ϕ_1 es igual a 1). De esta manera queda contrastada la dependencia temporal de la variable *CURVATMAX* y la independencia de *INC_TMAX* para la cual hemos obtenido mejores resultados en términos de error de predicción. Confirmamos así, la necesidad de eliminar la influencia temporal al utilizar las funciones cópula a las que venimos haciendo referencia o, alternativamente, emplear familias que sean capaces de reflejar esta dependencia (cópulas dependientes del tiempo).

8. Conclusiones

En esta tesis proponemos una metodología para realizar predicciones diarias de demanda de gas/electricidad a corto y medio plazo. Está orientada fundamentalmente al tratamiento de aquellos casos en los que un porcentaje considerable de este consumo es de carácter doméstico y se encuentra por tanto muy influenciado por la acción de agentes climáticos.

Una de las principales novedades que incorpora esta metodología es el empleo de una clase de funciones denominadas cópulas que posibilita, a partir de dos o más variables con sus respectivas marginales, la construcción de distribuciones multidimensionales que presenten dichas marginales (teorema de Sklar). Estas funciones permiten capturar la relación de dependencia existente entre las variables involucradas en el análisis (*demanda y temperatura*) y, a partir de ella, llevar a cabo la generación de valores de una de las variables condicionando al valor que pueden tomar las restantes. A través de la media o mediana de las simulaciones, somos capaces de proponer valores esperados para una distribución condicionada que, a fin de cuentas, viene a ser la esencia de un estudio de predicción.

El problema que se suele plantear es determinar aquella función cópula que genere la distribución conjunta más ajustada a la relación de dependencia existente entre todas las variables. El criterio de selección que hemos utilizado está basado en un contraste clásico de bondad de ajuste de distribuciones (contraste de la chi-cuadrado), cuantificado por el valor que toma el estadístico de Pearson \mathcal{P} sobre una partición dada del cuadrado unidad, el cual sirve de soporte a esta clase de funciones.

Hemos implementado un programa informático que, para un par de variables y una partición de unas dimensiones específicas, seleccione de manera automática una función cópula de acuerdo a este criterio. Partiendo de un conjunto de familias cuya expresión es conocida y que por sus propiedades se consideran candidatas interesantes, el programa propone como mejor alternativa aquella que, respecto de dicho estadístico, presente un menor valor.

Sin embargo, también hemos considerado importante la posibilidad de ofrecer una función cópula en aquellos casos en los que todos los valores calculados para los estadísticos de Pearson correspondientes a las distintas familias, conduzcan a rechazar el contraste planteado y resulte imposible disponer de una propuesta de ciertas garantías. En estos casos, el programa proporciona una cópula empírica C^* , no paramétrica, que garantiza un buen ajuste respecto del criterio de selección establecido. Dicha cópula se genera por extensión de la función subcópula C^{**} que, para la partición realizada, ofrece el menor (y por tanto mejor) valor posible de \mathcal{P} (\mathcal{P}^*). Para construir esta función subcópula hemos planteado un modelo de programación fraccional (por la naturaleza de su función objetivo), en el que las variables de decisión son los propios valores que debe tomar C^{**} en cada uno de los puntos de su dominio de definición y en el que el objetivo a minimizar es el propio estadístico. Este método de construcción de cópulas, que presentan un valor óptimo respecto del estadístico de Pearson, constituye la principal aportación teórica de esta tesis.

La subcópula producto es suficiente para demostrar la factibilidad del problema de programación matemática (teorema (4.1)) y además supone un buen punto de partida para la búsqueda de la solución óptima C^{**} , dado que, el resultado de la evaluación del estadístico sobre los puntos de su dominio, no presenta problemas de nulidad en el denominador.

La generación de la cópula C^* se lleva a cabo por interpolación de los valores c_{ij}^{**} que resultan de resolver el modelo de programación matemática, es decir, por interpolación del dominio de definición de la subcópula óptima. De esta manera, se pretende conservar el valor óptimo \mathcal{P}^* obtenido para ella.

Hemos expuesto tres formas distintas de interpolación, una bilineal (que permite obtener $C^*_{BILINEAL}$), otra mediante polinomios cúbicos de Hermite (que da lugar a $C^*_{HERMITE}$) y una tercera mediante polinomios de Bernstein (a partir de la cual se genera $C^*_{BERNSTEIN}$):

- La primera de ellas es la única que conserva el valor del estadístico de Pearson obtenido para C^* . Su construcción está siempre garantizada si bien el aspecto escalonado de su correspondiente densidad, resulta poco verosímil para reflejar una relación de dependencia, y por ello no consideramos apropiada su utilización con fines predictivos. Resulta conveniente en ocasiones hacer particiones muy finas para las cuales es necesario disponer de grandes volúmenes de datos.
- Por su parte, la construcción mediante interpolación cúbica no está siempre garantizada. Es necesario que la subcópula C^* a la que interpola pertenezca a una subclase de funciones que hemos definido como subcópulas- V_{MIN} (teorema (4.6)). Esta clase está constituida por subcópulas tales que el volumen a través de ellas de cualquiera de los rectángulos contenidos en su dominio de definición es mayor o igual que

$$V_{MIN} = \frac{2}{3 * m * n}, \text{ donde "m" y "n" representan los órdenes de la partición realizada sobre el cuadrado}$$

unidad. Como consecuencia de esta imposición, se puede asegurar que la subcópula- V_{MIN} generada, C^{**} , presenta el menor valor posible del estadístico de Pearson (\mathcal{P}^*) dentro de la clase de las subcópulas- V_{MIN} , pero no así dentro de la clase constituida por todas las subcópulas (\mathcal{P}).

A partir de C^{**} , hemos determinado unas condiciones suficientes (teorema (4.5)) que permiten asegurar que la superficie construida por interpolación mediante polinomios cúbicos de Hermite es una función cópula.

Las densidades asociadas a ambas cópulas, $C^*_{BILINEAL}$ y $C^*_{HERMITE}$, presentan cambios bruscos de comportamiento en su afán por interpolar exactamente los valores que toma C^* (alternativamente C^{**}).

La primera presenta un aspecto escalonado y la segunda uno muy peculiar ondulado, de continuos abombamientos. Esta circunstancia, si bien no parece aconsejable para reflejar relaciones “suaves” entre variables, sí puede resultar apropiada para capturar relaciones “multimodales” en las que existan núcleos dispersos, alrededor de los cuales se concentre la densidad multivariante.

- Finalmente, la densidad de la cópula $C^*_{BERNSTEIN}$ muestra mayor regularidad a costa de “aproximarse” a los valores de C^* pero, por las propiedades de los polinomios interpoladores (de Bernstein), no logra conseguir una interpolación exacta de los mismos. Así, si bien no se consigue conservar el valor óptimo \mathcal{P}^* , su aspecto más regular y uniforme sí parece más adecuado para reflejar relaciones de dependencia más convencionales.

En cualquier caso, haremos hincapié en que sugerimos el empleo de estas cópulas “interpoladoras” en aquellos casos en los cuales ninguna de las familias candidatas presente un valor del estadístico que le permita “superar” el contraste de bondad de ajuste. De hecho, éste no es el caso que se nos ha planteado al analizar la dependencia que existe entre la demanda de gas y las variables de temperatura en la serie temporal que hemos analizado.

Desde el punto de vista práctico, hemos incorporado este programa de selección de cópulas y de construcción de otras empíricas en aquellos casos en las que el contraste de bondad de ajuste sea rechazado para todas las candidatas, a una herramienta informática que hemos implementado y que automatiza el proceso de modelización a medio plazo de series de demanda energética de acuerdo a una metodología innovadora. El objetivo consiste en predecir el “pico” (valor máximo) diario de demanda que se puede esperar a un horizonte de dos años. Para ello, hemos propuesto 3 fases:

- **Fase I: Construcción de un histórico en condiciones normales de temperatura.-** Dado que a medio plazo el valor diario de las temperaturas es imposible de determinar, hemos decidido aislar del histórico, el efecto de las variables meteorológicas para, en función de él, realizar una predicción en condiciones de temperatura que establecemos como “normales”.
- **Fase II: Obtención de predicciones en condiciones normales de temperatura.-** De acuerdo al histórico obtenido, hemos utilizado técnicas de suavizado de curvas, como son las wavelets y los splines de regresión, para detectar el comportamiento básico de dicho histórico (ciclo anual). La extrapolación al futuro de esta componente de baja frecuencia permite “guiar” durante el horizonte de dos años que se plantea, a la predicción diaria realizada por un modelo lineal que tiene en cuenta únicamente la estacionalidad semanal de la serie.
- **Fase III: Simulación de escenarios climatológicos.-** A través de funciones cópula hemos analizado la relación de dependencia que vincula, en cada mes, al incremento existente entre el dato real de temperatura y el establecido mensualmente como normal, con el incremento porcentual que se produce entre el dato real de demanda y el dato establecido en condiciones climáticas de temperatura en la fase I. Concretamente hemos orientado el uso de funciones cópula, poco habituales en el ámbito de predicción de la demanda energética, a la simulación de las posibles desviaciones que puede experimentar el patrón de demanda establecido para el horizonte de dos años, ante diferentes escenarios climatológicos.

La herramienta que hemos programado da bastantes posibilidades a un analista familiarizado con todas estas técnicas expuestas, permitiéndole configurar a través de una relación de parámetros, el tipo de suavizado a utilizar (mediante wavelets o splines), el grado de suavidad del ajuste, la familia de cópulas para llevar a cabo la simulación, etc. La determinación del método de suavizado más conveniente (en nuestro caso los splines de regresión) está condicionado por la calidad de los resultados, evaluados en términos de error sobre datos recogidos en condiciones normales de temperatura, los cuales constituyen un grueso del histórico considerable. Sin embargo, respecto de la clase de cópulas a emplear, el criterio de selección se establece en función de sus posibilidades para predecir la distribución de la demanda bajo supuestos extremos del comportamiento de las variables de temperatura. Por ello, dado el escaso número de datos que permite hacer este contraste, la utilización de una u otra clase de cópulas puede antojarse en algunos casos un tanto subjetiva, siendo ésta la razón que nos llevado a volcar nuestros esfuerzos en que la herramienta pueda sugerir la familia más adecuada de forma automática.

Como punto de partida, se propone llevar a cabo la fase de simulación del comportamiento de la demanda bajo condiciones meteorológicas adversas a través de la distribución condicionada de aquella cópula, de expresión conocida, que presente un menor valor del estadístico de Pearson. En el caso en el que ninguna de ellas proporcione un buen

valor, y siempre y cuando la cópula que mejor se ajuste a los datos no sea la cópula producto (que refleja independencia), la herramienta lleva a cabo la simulación a partir de la distribución condicionada de alguna de las tres cópulas obtenidas por interpolación del dominio de la subcópula que optimiza el valor del estadístico de Pearson, C^* .

Hemos aplicado la metodología propuesta al análisis de la demanda de gas natural en Madrid, cuyo histórico de datos nos ha sido proporcionado por el **Departamento de Análisis y Simulación** de la compañía **Enagás**.

Esta serie presenta un porcentaje de consumo doméstico superior al 80%. Respecto de ella, las familias de cópulas para las cuales hemos obtenido un menor valor del estadístico de Pearson han sido la de Gumbel y la de Cola Derecha Pesada. Este resultado nos ha parecido bastante satisfactorio siendo nuestro objetivo predecir el valor máximo esperado para la demanda, dado que una y otra familia están especialmente orientadas a enfatizar la relación de dependencia entre sucesos extremos (entre una ola de frío y el “pico” que se puede esperar para la demanda).

Además, hemos comparado los resultados obtenidos para estas familias con los logrados mediante las cópulas generadas por interpolación de C^* . Si bien la balanza se inclina a favor de las primeras, hemos observado que en líneas generales no son muy diferentes a los que pueden proporcionar las segundas, las cuales, por su naturaleza empírica, también son capaces de amoldarse a las concentraciones de densidad que se producen en las regiones extremas de la distribución conjunta. Incluso, en las predicciones llevadas a cabo para algún mes concreto, los resultados obtenidos con estas cópulas “interpoladoras” han llegado a ser mejores, con lo cual estimamos que pueden constituir siempre una interesante alternativa de modelización.

En esta tesis proponemos también una metodología para realizar predicciones de demanda energética a corto plazo a través de un algoritmo (apartado 6.1) que, partiendo del proceso residual de un modelo ARIMA ajustado únicamente a partir del histórico de demanda, suple el empleo de funciones de transferencia por funciones cópula.

De igual manera que en el ajuste a medio plazo, se selecciona (de acuerdo al test de la chi-cuadrado) la función cópula que mejor refleja la relación de dependencia entre el error porcentual cometido por el ARIMA (a predecir) y las variables de incremento de temperatura de las que se dispone. En caso de que la cópula que mejor valor proporcione para el estadístico de Pearson sea la cópula producto, se considera que ambas variables son independientes y se rechaza la inclusión de la explicativa en el modelo. En caso contrario, la cópula es utilizada para corregir la predicción proporcionada por el ARIMA y dar lugar a un nuevo proceso residual que, de ser mejor que el anterior (de menor media y/o mediana en términos porcentuales), será relacionado con algún otro de los regresores disponibles. El proceso de ajuste termina cuando todos ellos han sido tanteados.

El algoritmo funciona bien si las variables implicadas son “ruidosas” y no presentan dependencia temporal. Respecto de la variable residual, esta independencia está siempre garantizada dado que, el punto de partida es el residuo de un modelo ARIMA y es precisamente al ajuste de la parte MA el que hace que el error obtenido sea incorrelado con su pasado. Por ello, todo se reduce a garantizar la independencia temporal de las variables explicativas. La razón que justifica esta necesidad, es que durante el proceso de ajuste se utilizan funciones cópula bidimensionales que únicamente analizan la relación existente entre el par de variables (X, Y) , no entre las integrantes de la terna (X, Y, T) donde T representa al tiempo.

Las primeras pruebas que hemos realizado perfilan un futuro bastante prometedor en cuanto a los resultados que podremos llegar a lograr, dado que, los que ya vamos obteniendo y que aquí hemos presentado, son directamente comparables a los que ofrecen los modelos de función de transferencia.

9. Futuros trabajos

Actualmente estamos incorporando nuevas funcionalidades a la herramienta de predicción a medio plazo. Pretendemos conseguir que participen más variables en la construcción del histórico de demanda en condiciones climatológicas normales y de momento hemos comenzado por hacer que se involucren de manera simultánea en dicho proceso las temperaturas máxima y mínima. Con la idea de que el usuario pueda simular escenarios en los que participen de manera conjunta ambas variables, estamos introduciendo funciones 3-cópulas que analicen la relación de dependencia de dichas variables con la demanda. Concretamente hemos incorporado la 3-cópula de Farlie-Gumbel-Morgenstern para la cual es sencillo calcular su expresión condicionada y realizar a partir de ella simulaciones, y estamos trabajando en la incorporación de algunas familias arquimedianas capaces de capturar otras relaciones interesantes como la de Gumbel (que como hemos visto acentúa la dependencia entre las colas derechas de las distribuciones), o la de Clayton (que acentúa más la relación de dependencia entre las colas izquierdas).

De manera adicional a la intención de mejorar esta herramienta, nuestra atención está principalmente centrada en el desarrollo de otra que permitirá el tratamiento a corto plazo de este tipo de series de acuerdo al algoritmo propuesto en el apartado 6.1. Conforme a algunos de los resultados presentados en esta tesis hemos observado que, partiendo de las predicciones realizadas por un modelo ARIMA que se construye únicamente a partir del histórico de demanda, la información aportada por ciertas variables exógenas (incrementos de temperatura) proporciona mejores predicciones haciendo uso del citado algoritmo que emplea funciones cópula, en lugar de los tradicionales modelos de función de transferencia. Este algoritmo relaciona de manera iterativa a través de una 2-cópula, pares de variables en las que una de ellas es siempre el proceso de error resultante del ajuste anterior. De igual modo que en el sistema de predicción a medio plazo, estamos incorporando funciones 3-cópulas que utilicen de forma conjunta la información aportada por el par de variables exógenas (temperatura máxima y mínima), en vez de utilizar éstas de manera independiente.

Nuestro interés actual en utilizar funciones cópula como parte de una metodología de predicción a corto plazo se fundamenta, no sólo en el hecho de que para un mismo grupo de variables haya proporcionado mejores resultados en términos de error que los modelos de función de transferencia que complementan a un modelo ARIMA, sino en la capacidad que tienen las cópulas de poder dar distribuciones de probabilidad asociadas a la predicción, algo que en condiciones climatológicas muy adversas puede ser especialmente interesante para el usuario de negocio. Desde este punto de vista, la desventaja que presenta el modelo ARIMA es que la estimación ofrecida por éste aparece centrada en un intervalo de confianza y no contempla posibles asimetrías en la predicción.

Por otra parte, teniendo en cuenta el alto poder de predicción que según hemos visto, parecen presentar algunas transformaciones de las variables de temperatura, pero cuya dependencia temporal las hace inapropiadas para su manipulación mediante cópulas, pretendemos estudiar aquéllas que permitan manejar variables en presencia de esta dependencia temporal o lo que es lo mismo, cópulas dependientes del tiempo en las cuales el parámetro asociado a la cópula puede venir dado en función de él (véase por ejemplo el caso de la cópula gaussiana comentado en el apartado 4.2.2 de [PATTON]). Funciones con estas propiedades podrían llevarnos incluso a plantear la posibilidad de aplicar el algoritmo que proponemos, para relacionar los procesos de error que iterativamente se van obteniendo con variables que reflejen la demanda del día o la semana anterior. De esta manera, las funciones cópula no serían utilizadas únicamente para buscar la relación del residuo de un modelo ARIMA con variables externas, sino que participarían

directamente en la identificación de características del histórico tales como la tendencia o la estacionalidad, lo cual podría conducir a que el proceso de modelización completo estuviese basado exclusivamente en el empleo de este tipo de funciones.

Incluso podría plantearse la posibilidad de llevar a cabo el tratamiento de festividades, puentes y periodos vacacionales (análisis de intervención) mediante funciones cópula. Si bien las variables que recogen este tipo de efecto son discretas e impiden por tanto el uso del teorema de Transformación Integral de Probabilidad (teorema (1.2)), existen algunos autores como **Denuit** y **Lambert** (2002) (véase [HEIREN]) que utilizan un método para hacer continuas estas variables, permitiendo así la aplicación de funciones cópula a marginales discretas. Podríamos por ejemplo definir las festividades a través de una variable que tomara 4 valores posibles en función de la intensidad de su efecto sobre la serie ($0 = \text{“no festivo”}$, $1 = \text{“festividad local o comunitaria”}$, $2 = \text{“puente asociado a festividad nacional”}$ y $3 = \text{“festividad nacional”}$), y utilizar el proceso que proponen los autores mencionados para hacer continuas dichas variables, de manera que puedan ser tratadas por nuestro algoritmo.

Esta línea de investigación focalizada en el tratamiento mediante cópulas de la dependencia dinámica asociada a la mayor parte de las series temporales, y en la que ya encontramos algunos trabajos (véase por ejemplo [PATTON], [PATTON2], [BOGASA] o [FERSCA]), será la que guíe nuestros próximos pasos en el futuro.

ANEXOS

ANEXO I: Algunas familias de cópulas

En este Anexo se presentan algunas de las familias de 2-cópulas más populares y se catalogan de acuerdo a algunos de los criterios expuestos en el apartado 1.5. La mayor parte de estas familias han sido extraídas de [NELSEN] y [JOE].

AI.1 Familias no paramétricas

AI.1.1 Familias de dependencia extrema

1. Cópula cota inferior de Fréchet-Hoeffding

$$W(u, v) = \max(u + v - 1, 0) \quad (\text{AI.1})$$

2. Cópula cota superior de Fréchet-Hoeffding

$$M(u, v) = \min(u, v) \quad (\text{AI.2})$$

3. Cópula producto

$$\Pi(u, v) = u * v \quad (\text{AI.3})$$

AI.1.2 Otras familias

1. Cópula empírica

$$C_n\left(\frac{i}{n}, \frac{j}{n}\right) = \frac{(n^\circ \text{ de pares } (x, y) \text{ en la muestra tales que } x \leq x_{(i)} \text{ e } y \leq y_{(j)})}{n} \quad (\text{AI.4})$$

AI.2 Familias uniparamétricas

AI.2.1 Familias elípticas

1. Cópula normal bivalente

$$C_{\theta}(u, v) = \Phi_{\theta}(\Phi^{-1}(u), \Phi^{-1}(v)) \quad (\text{AI.5})$$

siendo Φ_{θ} la función de distribución normal estándar bivalente ($N\left[\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \theta \\ \theta & 1 \end{bmatrix}\right]$ de correlación

$\theta = \text{corr}(\Phi^{-1}(u), \Phi^{-1}(v))$ y Φ^{-1} la inversa de Φ que es la función de distribución de una normal estándar univariante (0,1).

2. Cópula de Cauchy

Es un caso particular de la T-Cópula (véase (AI.40)) en el que el número de grados de libertad es 1 ($\delta = 1$). No se considera biparamétrica dado que el valor del segundo de los parámetros de los que depende viene fijado.

$$C_{\theta}(u, v) = t_{1,\theta}^2[t_1^{-1}(u), t_1^{-1}(v)] \quad \text{con } \delta > 2 \quad (\text{AI.6})$$

AI.2.2 Familias de valor extremo

1. Familia de Galambos (1975)

$$C_{\theta}(u, v) = u * v * \exp\left\{\left[(-\ln u)^{-\theta} + (-\ln v)^{-\theta}\right]^{\frac{-1}{\theta}}\right\} \quad \text{con } \theta \geq 0 \quad (\text{AI.7})$$

2. Familia de Gumbel (1960) o familia de Gumbel-Hougaard (también Arquimediana)

$$C_{\theta}(u, v) = \exp\left(-\left[(-\ln u)^{\theta} + (-\ln v)^{\theta}\right]^{\frac{1}{\theta}}\right) \quad \text{con } \theta \geq 1 \quad (\text{AI.8})$$

3. Familia de Hüsler y Reiss (1989)

$$C_{\theta}(u, v) = \exp\left\{-(-\ln u) * \Phi\left(\theta^{-1} + \frac{1}{2} * \theta * \log\left[\frac{-\ln u}{-\ln v}\right]\right) - (-\ln v) * \Phi\left(\theta^{-1} + \frac{1}{2} * \theta * \log\left[\frac{-\ln v}{-\ln u}\right]\right)\right\} \quad \text{con } \theta \geq 0 \quad (\text{AI.9})$$

siendo Φ^{-1} la inversa de Φ que es la función de distribución de una normal estándar univariante (0,1).

AI.2.3 Familias arquimedianas

1. Familia de Ali-Mikhail-Haq

$$C_{\theta}(u, v) = \frac{u * v}{1 - \theta * (1 - u) * (1 - v)} \quad \text{con } -1 \leq \theta \leq 1 \quad (\text{AI.10})$$

El generador de la cópula es $\phi_{\theta}(t) = \ln \frac{1 - \theta * (1 - t)}{t}$

2. Familia de Cook y Johnson o familia de Pareto o familia de Clayton

$$C_{\theta}(u, v) = \max([u^{-\theta} + v^{-\theta} - 1]^{-1/\theta}, 0) \quad \text{con } \theta \in [-1, \infty) \setminus \{0\} \quad (\text{AI.11})$$

El generador de la cópula es $\phi_{\theta}(t) = \frac{1}{\theta} * (t^{-\theta} - 1)$

3. Cópula de Cola Derecha Pesada

$$C_{\theta}(u, v) = u + v - 1 + [(1 - u)^{\frac{-1}{\theta}} + (1 - v)^{\frac{-1}{\theta}} - 1]^{-\theta} \quad \text{con } \theta > 0 \quad (\text{AI.12})$$

Esta cópula es realmente la cópula de supervivencia asociada a la familia de Clayton. Por definición, la cópula de supervivencia asociada a una cópula C es $\hat{C}(u, v) = u + v - 1 + C(1 - u, 1 - v)$.

4. Familia de Frank (1979)

$$C_{\theta}(u, v) = -\frac{1}{\theta} * \ln \left(1 + \frac{(e^{-\theta * u} - 1) * (e^{-\theta * v} - 1)}{e^{-\theta} - 1} \right) \quad \text{con } \theta \in \mathbb{R} \setminus \{0\} \quad (\text{AI.13})$$

El generador de la cópula es $\phi_{\theta}(t) = -\ln \frac{e^{-\theta * t} - 1}{e^{-\theta} - 1}$

5. Familia de Gumbel o familia de Gumbel-Barnett

$$C_{\theta}(u, v) = u * v * \exp(-\theta * \ln u * \ln v) \quad \text{con } 0 < \theta \leq 1 \quad (\text{AI.14})$$

y generador $\phi_{\theta}(t) = -\ln(1 - \theta * \ln t)$

6. Familia de Gumbel (1960) o familia de Gumbel-Hougaard (también de valor extremo)

$$C_{\theta}(u, v) = \exp(-[(-\ln u)^{\theta} + (-\ln v)^{\theta}]^{\frac{1}{\theta}}) \quad \text{con } \theta \geq 1 \quad (\text{AI.15})$$

El generador de la cópula es $\phi_{\theta}(t) = (-\ln t)^{\theta}$

7. Familia de Joe (1993)

$$C_{\theta}(u, v) = 1 - [(1 - u)^{\theta} + (1 - v)^{\theta} - (1 - u)^{\theta} * (1 - v)^{\theta}]^{\frac{1}{\theta}} \quad \text{con } \theta \geq 1 \quad (\text{AI.16})$$

El generador de la cópula es $\phi_{\theta}(t) = -\ln [1 - (1 - t)^{\theta}]$

8. Nelsen (1999)

$$C_{\theta}(u, v) = \max(1 - [(1-u)^{\theta} + (1-v)^{\theta}]^{1/\theta}, 0) \text{ con } \theta \geq 1 \quad (\text{AI.17})$$

El generador de la cópula es $\phi_{\theta}(t) = (1-t)^{\theta}$

9. Nelsen (1999)

$$C_{\theta}(u, v) = \max(\theta * u * v + (1-\theta) * (u+v-1), 0) \text{ con } 0 < \theta \leq 1 \quad (\text{AI.18})$$

El generador de la cópula es $\phi_{\theta}(t) = -\ln[\theta * t + (1-\theta)]$

10. Nelsen (1999)

$$C_{\theta}(u, v) = \max\left[\frac{\theta^2 * u * v - (1-u) * (1-v)}{\theta^2 - (\theta-1)^2 * (1-u) * (1-v)}, 0\right] \text{ con } \theta \geq 1 \quad (\text{AI.19})$$

El generador de la cópula es $\phi_{\theta}(t) = \frac{1-t}{1+(\theta-1)*t}$

11. Nelsen (1999)

$$C_{\theta}(u, v) = \frac{u * v}{[1 + (1-u^{\theta}) * (1-v^{\theta})]^{\frac{1}{\theta}}} \text{ con } 0 < \theta \leq 1 \quad (\text{AI.20})$$

El generador de la cópula es $\phi_{\theta}(t) = \ln(2 * t^{-\theta} - 1)$

12. Nelsen (1999)

$$C_{\theta}(u, v) = \max([u^{\theta} * v^{\theta} - 2 * (1-u^{\theta}) * (1-v^{\theta})]^{\frac{1}{\theta}}, 0) \text{ con } 0 < \theta \leq \frac{1}{2} \quad (\text{AI.21})$$

El generador de la cópula es $\phi_{\theta}(t) = \ln(2 - t^{\theta})$

13. Nelsen (1999)

$$C_{\theta}(u, v) = (1 + [(u^{-1} - 1)^{\theta} + (v^{-1} - 1)^{\theta}]^{\frac{1}{\theta}})^{-1} \text{ con } \theta \geq 1 \quad (\text{AI.22})$$

El generador de la cópula es $\phi_{\theta}(t) = \left(\frac{1}{t} - 1\right)^{\theta}$

14. Nelsen (1999)

$$C_{\theta}(u, v) = \exp(1 - [(1 - \ln u)^{\theta} + (1 - \ln v)^{\theta} - 1]^{\frac{1}{\theta}}) \text{ con } \theta > 0 \quad (\text{AI.23})$$

El generador de la cópula es $\phi_{\theta}(t) = (1 - \ln t)^{\theta} - 1$

15. Nelsen (1999)

$$C_{\theta}(u, v) = (1 + [(u^{-\frac{1}{\theta}} - 1)^{\theta} + (v^{-\frac{1}{\theta}} - 1)^{\theta}]^{\frac{1}{\theta}})^{-\theta} \text{ con } \theta \geq 1 \quad (\text{AI.24})$$

El generador de la cópula es $\phi_{\theta}(t) = (t^{-\frac{1}{\theta}} - 1)^{\theta}$

16. Nelsen (1999)

$$C_{\theta}(u, v) = \max(\{1 - [(1 - u^{\frac{1}{\theta}})^{\theta} + (1 - v^{\frac{1}{\theta}})^{\theta}]^{\frac{1}{\theta}}, 0\} \text{ con } \theta \geq 1 \quad (\text{AI.25})$$

El generador de la cópula es $\phi_{\theta}(t) = (1 - t^{\frac{1}{\theta}})^{\theta}$

17. Nelsen (1999)

$$C_{\theta}(u, v) = \frac{1}{2} * (S + \sqrt{(S^2 + 4 * \theta)}), S = u + v - 1 - \theta * (\frac{1}{u} + \frac{1}{v} - 1) \text{ con } \theta \geq 0 \quad (\text{AI.26})$$

El generador de la cópula es $\phi_{\theta}(t) = (\frac{\theta}{t} + 1) * (1 - t)$

18. Nelsen (1999)

$$C_{\theta}(u, v) = (1 + \frac{[(1 + u)^{-\theta} - 1] * [(1 + v)^{-\theta} - 1]}{2^{-\theta} - 1})^{\frac{-1}{\theta}} - 1 \text{ con } \theta \in \mathbb{R} \setminus \{0\} \quad (\text{AI.27})$$

El generador de la cópula es $\phi_{\theta}(t) = -\ln \frac{(1 + t)^{-\theta} - 1}{2^{-\theta} - 1}$

19. Nelsen (1999)

$$C_{\theta}(u, v) = \max(1 + \frac{\theta}{\ln[e^{\frac{\theta}{(u-1)}} + e^{\frac{\theta}{(v-1)}}]}, 0) \text{ con } \theta \geq 2 \quad (\text{AI.28})$$

El generador de la cópula es $\phi_{\theta}(t) = e^{\frac{\theta}{(t-1)}}$

20. Nelsen (1999)

$$C_{\theta}(u, v) = \frac{\theta}{\ln(e^{\frac{\theta}{u}} + e^{\frac{\theta}{v}} - e^{\theta})} \text{ con } \theta > 0 \quad (\text{AI.29})$$

El generador de la cópula es $\phi_{\theta}(t) = e^{\frac{\theta}{t}} - e^{\theta}$

21. Nelsen (1999)

$$C_{\theta}(u, v) = [\ln(\exp(u^{-\theta}) + \exp(v^{-\theta}) - e)]^{\frac{-1}{\theta}} \text{ con } \theta > 0 \quad (\text{AI.30})$$

El generador de la cópula es $\phi_{\theta}(t) = \exp(t^{-\theta}) - e$

22. Nelsen (1999)

$$C_{\theta}(u, v) = 1 - (1 - \{ \max([1 - (1 - u)^{\frac{1}{\theta}}]^{\frac{1}{\theta}} + [1 - (1 - v)^{\frac{1}{\theta}}]^{\frac{1}{\theta}} - 1, 0) \}^{\frac{1}{\theta}}) \text{ con } \theta \geq 1 \quad (\text{AI.31})$$

El generador de la cópula es $\phi_{\theta}(t) = 1 - [1 - (1 - t)^{\frac{1}{\theta}}]^{\frac{1}{\theta}}$

23. Nelsen (1999)

$$C_{\theta}(u, v) = \max\left(\left[1 - (1 - u^{\theta}) \sqrt{1 - (1 - v^{\theta})} - (1 - v^{\theta}) \sqrt{1 - (1 - u^{\theta})}\right]^{\frac{1}{\theta}}, 0\right) \text{ con } 0 < \theta \leq 1 \quad (\text{AI.32})$$

El generador de la cópula es $\phi_{\theta}(t) = \arcsen(1 - t^{\theta})$

AI.2.4 Otras familias uniparamétricas

1. Familia de Cuadras-Augé (1981)

$$C_{\theta}(u, v) = [\min\{u, v\}]^{\theta} * [u * v]^{1 - \theta} \text{ con } 0 \leq \theta \leq 1 \quad (\text{AI.33})$$

2. Familia de Farlie-Gumbel-Morgenstern (1956 Morgenstern, 1958 Gumbel, 1960 Farlie)

$$C_{\theta}(u, v) = u * v * [1 + \theta * (1 - u) * (1 - v)] \text{ con } -1 \leq \theta \leq 1 \quad (\text{AI.34})$$

2. Familia de Kimeldorf y Sampson (1975)

$$C_{\theta}(u, v) = (u^{-\theta} + v^{-\theta} - 1)^{\frac{-1}{\theta}} \text{ con } \theta \geq 0 \quad (\text{AI.35})$$

4. Familia de Mardia (1970)

$$C_{\theta}(u, v) = \frac{\theta^2 * (1 + \theta)}{2} * M(u, v) + (1 - \theta^2) * \Pi(u, v) + \frac{\theta^2 * (1 - \theta)}{2} * W(u, v) \text{ con } -1 \leq \theta \leq 1 \quad (\text{AI.36})$$

5. Familia de Plackett (1965)

$$C_{\theta}(u, v) = \frac{[1 + (\theta - 1) * (u + v)] - \sqrt{[1 + (\theta - 1) * (u + v)]^2 - 4 * u * v * \theta * (\theta - 1)}}{2 * (\theta - 1)} \text{ con } \theta \geq 0 \quad (\text{AI.37})$$

6. Familia de Sarmanov

$$C_{\theta}(u, v) = u * v + u * v * (1 - u) * (1 - v) * [3 * \theta + 5 * \theta^2 * (1 - 2 * u) * (1 - 2 * v)] \text{ con } 0 \leq \theta \leq 1 \quad (\text{AI.38})$$

7. $C_{\theta}(u, v) = \theta * \min\{u, v\} + (1 - \theta) * u * v \text{ con } 0 \leq \theta \leq 1 \quad (\text{AI.39})$

AI.3 Familias biparamétricas

AI.3.1 Familias elípticas

1. T-Cópula

$$C_{\theta, \delta}(u, v) = t_{\delta, \theta}^2[t_{\delta}^{-1}(u), t_{\delta}^{-1}(v)] \text{ con } \delta > 2 \quad (\text{AI.40})$$

siendo $t_{\delta, \theta}^2$ la función de distribución conjunta de una variable t-student con matriz de covarianzas igual a

$$\frac{\delta}{\delta-2} * R \text{ con } R = \begin{bmatrix} 1 & \theta \\ \theta & 1 \end{bmatrix}, \theta = \text{corr}(t_{\delta}^{-1}(u), t_{\delta}^{-1}(v)) \text{ y } \delta = \text{numero de grados de libertad} \text{ y siendo}$$

t_{δ} la función de distribución de una t-student con varianza igual a $\frac{\delta}{\delta-2}$.

AI.3.2 Familias de valor extremo

1. Joe (1997)

$$C_{\theta, \delta}(u, v) = \exp \left\{ - \left[(-\log u)^{\theta} + (-\log v)^{\theta} - \left((-\log u)^{-\theta * \delta} + (-\log v)^{-\theta * \delta} \right)^{\frac{-1}{\delta}} \right]^{\frac{1}{\theta}} \right\} \text{ con } \theta \geq 1, \delta > 0 \quad (\text{AI.41})$$

AI.3.3 Familias arquimedianas

1. Joe (1997)

$$C_{\theta, \delta}(u, v) = \left\{ \left[(u^{-\theta} - 1)^{\delta} + (v^{-\theta} - 1)^{\delta} \right]^{\frac{1}{\delta}} + 1 \right\}^{\frac{-1}{\theta}} \text{ con } \theta > 0, \delta \geq 1 \quad (\text{AI.42})$$

El generador de la cópula es $\phi_{\theta, \delta}(t) = (t^{-\theta} - 1)^{\delta}$

2. Joe (1997)

$$C_{\theta, \delta}(u, v) = \left\{ 1 + \delta^{-1} * \log \left(e^{\delta * (u^{-\theta} - 1)} + e^{\delta * (v^{-\theta} - 1)} - 1 \right) \right\}^{\frac{-1}{\theta}} \text{ con } \theta, \delta > 0 \quad (\text{AI.43})$$

El generador de la cópula es $\phi_{\theta, \delta}(t) = \exp(\delta * (t^{-\theta} - 1))$

3. Joe (1997)

$$C_{\theta, \delta}(u, v) = \exp \left\{ - \left[\delta^{-1} * \log \left(e^{\delta * (-\log u)} + e^{\delta * (-\log v)} - 1 \right) \right]^{\frac{1}{\theta}} \right\} \text{ con } \theta \geq 1, \delta > 0 \quad (\text{AI.44})$$

El generador de la cópula es $\phi_{\theta, \delta}(t) = \exp(\delta * (-\ln t)^{\theta}) - 1$

4. Joe (1997)

$$C_{\theta,\delta}(u,v) = 1 - (1 - \exp\{-[(-\log(1-(1-u)^\theta))^\delta + (-\log(1-(1-v)^\theta))^\delta]^\frac{1}{\delta}\})^\frac{1}{\theta} \text{ con } \theta \geq 1, \delta \geq 1 \quad (\text{AI.45})$$

El generador de la cópula es $\phi_{\theta,\delta}(t) = [-\log(1-(1-t)^\theta)]^\delta$

5. Joe (1997)

$$C_{\theta,\delta}(u,v) = 1 - (1 - [(1-(1-u)^\theta)^{-\delta} + (1-(1-v)^\theta)^{-\delta} - 1]^\frac{-1}{\delta})^\frac{1}{\theta} \text{ con } \theta \geq 1, \delta > 0 \quad (\text{AI.46})$$

El generador de la cópula es $\phi_{\theta,\delta}(t) = [1 - (1-t)^\theta]^{-\delta} - 1$

6. $C_{\theta,\delta}(u,v) = \max\{\{1 - [(1-u)^\theta + (1-v)^\theta]^\frac{1}{\delta}\}^\frac{1}{\theta}, 0\} \text{ con } 0 < \theta \leq 1, \delta \geq 1 \quad (\text{AI.47})$

El generador de la cópula es $\phi_{\theta,\delta}(t) = (1-t)^\theta$

7. Cópula arquimediana racional

$$C_{\theta,\delta}(u,v) = \max\left(\frac{u*v - \delta*(1-u)*(1-v)}{1 - \theta*(1-u)*(1-v)}, 0\right) \text{ con } 0 < \delta \leq 1 - |\theta| \quad (\text{AI.48})$$

AI.3.4 Otras familias biparamétricas

1. Familia de Fréchet (1958)

$$C_{\theta,\delta}(u,v) = \alpha*M(u,v) + (1-\alpha-\beta)*\Pi(u,v) + \beta*W(u,v) \text{ con } \alpha + \beta \leq 1 \quad (\text{AI.49})$$

2. Familia Farlie-Gumbel-Morgenstern iterada de Kotz y Johnson (1977)

$$C_{\theta,\delta}(u,v) = u*v + \theta*u*v*(1-u)*(1-v)*[1 + \delta*u*v] \text{ con } -1 \leq \theta \leq 1 \quad (\text{AI.50})$$

3. Familia Farlie-Gumbel-Morgenstern de Lin (1987)

$$C_{\theta,\delta}(u,v) = u*v + \theta*u*v*(1-u)*(1-v)*[1 + \delta*(1-u)*(1-v)] \text{ con } -1 \leq \theta \leq 1 \quad (\text{AI.51})$$

4. Familia de Marshall-Olkin o familia de Cuadras-Augé generalizada

$$C_{\theta,\delta}(u,v) = \min\{u^{1-\theta}, u*v^{1-\delta}\} \text{ con } 0 < \theta, \delta < 1 \quad (\text{AI.52})$$

5. Un miembro de la clase Archimax (combinación de las clases arquimediana y de valor extremo)

$$C_{\theta,\delta}(u,v) = (u^{-\theta} + v^{-\theta} - 1 - [(u^{-\theta} - 1)^{-\delta} + (v^{-\theta} - 1)^{-\delta}]^\frac{-1}{\delta})^\frac{-1}{\theta} \text{ con } \theta \geq 0, \delta > 0 \quad (\text{AI.53})$$

6. Familia de Joe (1997)

$$C_{\theta,\delta}(u,v) = \delta^{-1} * [1 - \{1 - [1 - (1-\delta)^\theta]^{-1} * [1 - (1-\delta*u)^\theta] * [1 - (1-\delta*v)^\theta]\}^\theta] \text{ con } \theta \geq 1, 0 \leq \delta \leq 1 \quad (\text{AI.54})$$

7. Familia de Joe (1997)

$$C_{\theta,\delta}(u,v) = \exp \{ - [(\delta - \log u)^\theta + (\delta - \log v)^\theta - \delta^\theta]^{\frac{1}{\theta}} + \delta \} \text{ con } \theta \geq 1, \delta > 0 \quad (\text{AI.55})$$

8. Familia de Joe (1997)

$$C_{\theta,\delta}(u,v) = u * v * [1 - \theta * (1 - u^{\frac{1}{\delta}}) * (1 - v^{\frac{1}{\delta}})]^{-\delta} \text{ con } 0 \leq \theta \leq 1, \delta > 0 \quad (\text{AI.56})$$

ANEXO II: Demostración del lema 4.2

“Sea una partición del cuadrado unidad en $n \times m$ rectángulos del mismo área y sea la solución óptima al problema de programación fraccional (4.7) dada por $c_{ij}^* = C^*\left(\frac{i}{n}, \frac{j}{m}\right) \forall i \in \{0,1,2,\dots,n\} \quad \forall j \in \{0,1,2,\dots,m\}$ que define una subcópula C^* de dominio los puntos de intersección entre los rectángulos que originan la partición.

Sea $C_{HERMITE}(u,v)$ la función que se obtiene por interpolación cúbica de C^* mediante polinomios de Hermite.

Si $\frac{\partial C_{HERMITE}(u_k,0)}{\partial u} = 0 \forall k \in \{0,1,2,\dots,n\}$ y $\frac{\partial C_{HERMITE}(0,v_l)}{\partial v} = 0 \forall l \in \{0,1,2,\dots,m\}$, entonces la función $C_{HERMITE}(u,v)$ cumple la primera de las condiciones para ser cópula: $C_{HERMITE}(u,0) = C_{HERMITE}(0,v) = 0$ para cualesquiera u y v ”.

demostración

Partiremos de la subcópula C^* la cual, por definición, verificará la propiedad (1.2) de manera que

$$C^*(u_i,0) = C^*(0,v_j) \quad \forall i \in \{0,1,2,\dots,n\}, j \in \{0,1,2,\dots,m\}$$

Vamos a hallar la imagen a través de $C_{HERMITE}$ del punto $(u,0)$. Asociado al punto $v = 0$ habrá que considerar como vértices del rectángulo al que pertenece $v_j = 0$ y $v_{j+1} = 1/m$. Teniendo en cuenta que

$$\frac{\partial C_{HERMITE}(u_k,0)}{\partial u} = 0 \quad \forall k \in \{0,1,2,\dots,n\} \quad \text{por hipótesis, y desarrollando la expresión (4.19) para } C_{HERMITE}(u,0) \text{ se}$$

obtendría lo siguiente:

$$\begin{aligned}
C_{HERMITE}(u,0)= & \left[\left(\frac{2*(u-u_i)^3}{(u_{i+1}-u_i)^3} - \frac{3*(u-u_i)^2}{(u_{i+1}-u_i)^2} + 1 \right) * C^*(u_i,0) + \left(\frac{-2*(u-u_i)^3}{(u_{i+1}-u_i)^3} + \frac{3*(u-u_i)^2}{(u_{i+1}-u_i)^2} \right) * C^*(u_{i+1},0) \right] * \\
& \left(\frac{2*(0-0)^3}{(1/m-0)^3} - \frac{3*(0-0)^2}{(1/m-0)^2} + 1 \right) + \\
& \left[\left(\frac{2*(u-u_i)^3}{(u_{i+1}-u_i)^3} - \frac{3*(u-u_i)^2}{(u_{i+1}-u_i)^2} + 1 \right) * C^*(u_i,1/m) + \left(\frac{-2*(u-u_i)^3}{(u_{i+1}-u_i)^3} + \frac{3*(u-u_i)^2}{(u_{i+1}-u_i)^2} \right) * C^*(u_{i+1},1/m) \right] * \\
& \left(\frac{-2*(0-0)^3}{(1/m-0)^3} + \frac{3*(0-0)^2}{(1/m-0)^2} \right) + \\
& \left[\left(\frac{2*(u-u_i)^3}{(u_{i+1}-u_i)^3} - \frac{3*(u-u_i)^2}{(u_{i+1}-u_i)^2} + 1 \right) * \frac{\partial C^*(u_i,0)}{\partial v} + \left(\frac{-2*(u-u_i)^3}{(u_{i+1}-u_i)^3} + \frac{3*(u-u_i)^2}{(u_{i+1}-u_i)^2} \right) * \frac{\partial C^*(u_{i+1},0)}{\partial v} + \right. \\
& \left. \left(\frac{(u-u_i)^3}{(u_{i+1}-u_i)^2} - \frac{2*(u-u_i)^2}{(u_{i+1}-u_i)} + (u-u_i) \right) * \frac{\partial^2 C^*(u_i,0)}{\partial u \partial v} + \left(\frac{(u-u_i)^3}{(u_{i+1}-u_i)^2} - \frac{(u-u_i)^2}{(u_{i+1}-u_i)} \right) * \frac{\partial^2 C^*(u_{i+1},0)}{\partial u \partial v} \right] * \\
& \left(\frac{(0-0)^3}{(1/m-0)^2} - \frac{2*(0-0)^2}{(1/m-0)} + (0-0) \right) + \\
& \left[\left(\frac{2*(u-u_i)^3}{(u_{i+1}-u_i)^3} - \frac{3*(u-u_i)^2}{(u_{i+1}-u_i)^2} + 1 \right) * \frac{\partial C^*(u_i,1/m)}{\partial v} + \left(\frac{-2*(u-u_i)^3}{(u_{i+1}-u_i)^3} + \frac{3*(u-u_i)^2}{(u_{i+1}-u_i)^2} \right) * \frac{\partial C^*(u_{i+1},1/m)}{\partial v} + \right. \\
& \left. \left(\frac{(u-u_i)^3}{(u_{i+1}-u_i)^2} - \frac{2*(u-u_i)^2}{(u_{i+1}-u_i)} + (u-u_i) \right) * \frac{\partial^2 C^*(u_i,1/m)}{\partial u \partial v} + \left(\frac{(u-u_i)^3}{(u_{i+1}-u_i)^2} - \frac{(u-u_i)^2}{(u_{i+1}-u_i)} \right) * \frac{\partial^2 C^*(u_{i+1},1/m)}{\partial u \partial v} \right] * \\
& \left(\frac{(0-0)^3}{(1/m-0)^2} - \frac{(0-0)^2}{(1/m-0)} \right) = \\
& \left[\left(\frac{2*(u-u_i)^3}{(u_{i+1}-u_i)^3} - \frac{3*(u-u_i)^2}{(u_{i+1}-u_i)^2} + 1 \right) * C^*(u_i,0) + \left(\frac{-2*(u-u_i)^3}{(u_{i+1}-u_i)^3} + \frac{3*(u-u_i)^2}{(u_{i+1}-u_i)^2} \right) * C^*(u_{i+1},0) \right] = 0 \\
& \quad \quad \quad \forall u \in [u_i, u_{i+1}]
\end{aligned}$$

(AII.1)

La igualdad a cero es consecuencia del hecho de que C^* sea una subcópula y por tanto, se cumple la primera de las propiedades de este tipo de funciones (véase (1.2)).

De igual manera, vamos a hallar la imagen a través de $C_{HERMITE}$ del punto $(0,v)$. Asociado al punto $u = 0$ habrá que considerar como vértices del rectángulo al que pertenece $u_i = 0$ y $u_{i+1} = 1/n$. Teniendo en cuenta que

$$\frac{\partial C_{HERMITE}(0, v_l)}{\partial v} = 0 \quad \forall l \in \{0, 1, 2, \dots, m\} \quad , \text{ y desarrollando la expresión (4.19) para } C_{HERMITE}(0,v) \text{ se tendría:}$$

$$\begin{aligned}
C_{HERMITE}(0,v)= & \left[\left(\frac{2*(0-0)^3}{(1/n-0)^3} - \frac{3*(0-0)^2}{(1/n-0)^2} + 1 \right) * C^*(0, v_j) \right] * \left(\frac{2*(v-v_j)^3}{(v_{j+1}-v_j)^3} - \frac{3*(v-v_j)^2}{(v_{j+1}-v_j)^2} + 1 \right) + \\
& \left[\left(\frac{2*(0-0)^3}{(1/n-0)^3} - \frac{3*(0-0)^2}{(1/n-0)^2} + 1 \right) * C^*(0, v_{j+1}) \right] * \left(\frac{-2*(v-v_j)^3}{(v_{j+1}-v_j)^3} + \frac{3*(v-v_j)^2}{(v_{j+1}-v_j)^2} \right) = 0 \\
& \quad \quad \quad \forall v \in [v_j, v_{j+1}]
\end{aligned}$$

(AII.2)

igualdad que también es consecuencia del hecho de que C^* sea subcópula y concluyéndose así la demostración del lema.

c.q.d

ANEXO III: Demostración del lema 4.3

“Sea una partición del cuadrado unidad en $n \times m$ rectángulos del mismo área y sea la solución óptima al problema de programación fraccional (4.7) dada por $c_{ij}^* = C^*\left(\frac{i}{n}, \frac{j}{m}\right) \forall i \in \{0,1,2,\dots,n\} \quad \forall j \in \{0,1,2,\dots,m\}$ que define una subcópula C^* de dominio los puntos de intersección entre los rectángulos que originan la partición.

Sea $C_{HERMITE}(u,v)$ la función que se obtiene por interpolación cúbica de C^* mediante polinomios de Hermite.

Si $\frac{\partial C_{HERMITE}(u_k,1)}{\partial u} = 1 \quad \forall k \in \{0,1,2,\dots,n\}$ y $\frac{\partial C_{HERMITE}(1,v_l)}{\partial v} = 1 \quad \forall l \in \{0,1,2,\dots,m\}$, entonces la función

$C_{HERMITE}(u,v)$ cumple la segunda de las condiciones para ser cópula: $C_{HERMITE}(u,1) = u$ y $C_{HERMITE}(1,v) = v$ para cualesquiera u y v , es decir, C tiene marginales uniformes”.

demostración

Empezaremos hallando la imagen a través de $C_{HERMITE}$ del punto $(u,1)$. Asociado al punto $v = 1$ habrá que considerar como vértices del rectángulo al que pertenece $v_j = (m-1)/m$ y $v_{j+1} = 1$. Teniendo en cuenta que

$$\frac{\partial C_{HERMITE}(u_k,1)}{\partial u} = 1 \quad \forall k \in \{0,1,2,\dots,n\} \quad \text{por hipótesis, y desarrollando la expresión (4.19) para } C_{HERMITE}(u,1) \text{ se}$$

obtendría lo siguiente:

$$\begin{aligned}
& C_{HERMITE}(u,1)= \\
& \left[\left(\frac{2*(u-u_i)^3}{(u_{i+1}-u_i)^3} - \frac{3*(u-u_i)^2}{(u_{i+1}-u_i)^2} + 1 \right) * C^*(u_i, \frac{m-1}{m}) + \left(\frac{-2*(u-u_i)^3}{(u_{i+1}-u_i)^3} + \frac{3*(u-u_i)^2}{(u_{i+1}-u_i)^2} \right) * C^*(u_{i+1}, \frac{m-1}{m}) + \right. \\
& \left. \left(\frac{(u-u_i)^3}{(u_{i+1}-u_i)^2} - \frac{2*(u-u_i)^2}{(u_{i+1}-u_i)} + (u-u_i) \right) * \frac{\partial C^*(u_i, \frac{m-1}{m})}{\partial u} + \left(\frac{(u-u_i)^3}{(u_{i+1}-u_i)^2} - \frac{(u-u_i)^2}{(u_{i+1}-u_i)} \right) * \frac{\partial C^*(u_{i+1}, \frac{m-1}{m})}{\partial u} \right] * \\
& \left(\frac{2*(1-\frac{m-1}{m})^3}{(1-\frac{m-1}{m})^3} - \frac{3*(1-\frac{m-1}{m})^2}{(1-\frac{m-1}{m})^2} + 1 \right) + \\
& \left[\left(\frac{2*(u-u_i)^3}{(u_{i+1}-u_i)^3} - \frac{3*(u-u_i)^2}{(u_{i+1}-u_i)^2} + 1 \right) * C^*(u_i,1) + \left(\frac{-2*(u-u_i)^3}{(u_{i+1}-u_i)^3} + \frac{3*(u-u_i)^2}{(u_{i+1}-u_i)^2} \right) * C^*(u_{i+1},1) + \right. \\
& \left. \left(\frac{(u-u_i)^3}{(u_{i+1}-u_i)^2} - \frac{2*(u-u_i)^2}{(u_{i+1}-u_i)} + (u-u_i) \right) * 1 + \left(\frac{(u-u_i)^3}{(u_{i+1}-u_i)^2} - \frac{(u-u_i)^2}{(u_{i+1}-u_i)} \right) * 1 \right] * \\
& \left(\frac{-2*(1-\frac{m-1}{m})^3}{(1-\frac{m-1}{m})^3} + \frac{3*(1-\frac{m-1}{m})^2}{(1-\frac{m-1}{m})^2} \right) + \\
& \left[\left(\frac{2*(u-u_i)^3}{(u_{i+1}-u_i)^3} - \frac{3*(u-u_i)^2}{(u_{i+1}-u_i)^2} + 1 \right) * \frac{\partial C^*(u_i, \frac{m-1}{m})}{\partial v} + \left(\frac{-2*(u-u_i)^3}{(u_{i+1}-u_i)^3} + \frac{3*(u-u_i)^2}{(u_{i+1}-u_i)^2} \right) * \frac{\partial C^*(u_{i+1}, \frac{m-1}{m})}{\partial v} + \right. \\
& \left. \left(\frac{(u-u_i)^3}{(u_{i+1}-u_i)^2} - \frac{2*(u-u_i)^2}{(u_{i+1}-u_i)} + (u-u_i) \right) * \frac{\partial^2 C^*(u_i, \frac{m-1}{m})}{\partial u \partial v} + \left(\frac{(u-u_i)^3}{(u_{i+1}-u_i)^2} - \frac{(u-u_i)^2}{(u_{i+1}-u_i)} \right) * \frac{\partial^2 C^*(u_{i+1}, \frac{m-1}{m})}{\partial u \partial v} \right] * \\
& \left(\frac{1-\frac{m-1}{m}}{(1-\frac{m-1}{m})^2} - \frac{2*(1-\frac{m-1}{m})^2}{(1-\frac{m-1}{m})} + (1-\frac{m-1}{m}) \right) + \\
& \left[\left(\frac{2*(u-u_i)^3}{(u_{i+1}-u_i)^3} - \frac{3*(u-u_i)^2}{(u_{i+1}-u_i)^2} + 1 \right) * \frac{\partial C^*(u_i,1)}{\partial v} + \left(\frac{-2*(u-u_i)^3}{(u_{i+1}-u_i)^3} + \frac{3*(u-u_i)^2}{(u_{i+1}-u_i)^2} \right) * \frac{\partial C^*(u_{i+1},1)}{\partial v} + \right. \\
& \left. \left(\frac{(u-u_i)^3}{(u_{i+1}-u_i)^2} - \frac{2*(u-u_i)^2}{(u_{i+1}-u_i)} + (u-u_i) \right) * \frac{\partial^2 C^*(u_i,1)}{\partial u \partial v} + \left(\frac{(u-u_i)^3}{(u_{i+1}-u_i)^2} - \frac{(u-u_i)^2}{(u_{i+1}-u_i)} \right) * \frac{\partial^2 C^*(u_{i+1},1)}{\partial u \partial v} \right] * \\
& \left(\frac{1-\frac{m-1}{m}}{(1-\frac{m-1}{m})^2} - \frac{(1-\frac{m-1}{m})^2}{(1-\frac{m-1}{m})} \right) = \\
& \left(\frac{1-\frac{m-1}{m}}{(1-\frac{m-1}{m})^2} - \frac{(1-\frac{m-1}{m})^2}{(1-\frac{m-1}{m})} \right) = \\
& \left[\left(\frac{2*(u-u_i)^3}{(u_{i+1}-u_i)^3} - \frac{3*(u-u_i)^2}{(u_{i+1}-u_i)^2} + 1 \right) * C^*(u_i,1) + \left(\frac{-2*(u-u_i)^3}{(u_{i+1}-u_i)^3} + \frac{3*(u-u_i)^2}{(u_{i+1}-u_i)^2} \right) * C^*(u_{i+1},1) + \right. \\
& \left. \left(\frac{(u-u_i)^3}{(u_{i+1}-u_i)^2} - \frac{2*(u-u_i)^2}{(u_{i+1}-u_i)} + (u-u_i) \right) * \frac{\partial C^*(u_i,1)}{\partial u} + \left(\frac{(u-u_i)^3}{(u_{i+1}-u_i)^2} - \frac{(u-u_i)^2}{(u_{i+1}-u_i)} \right) * \frac{\partial C^*(u_{i+1},1)}{\partial u} \right] \\
& \qquad \qquad \qquad \forall u \in [u_i, u_{i+1}]
\end{aligned}$$

Teniendo en cuenta que C^* es una subcópula y que por tanto cumple la segunda de las propiedades de este tipo de funciones (véase (1.3)), se tiene que $C^*(u_i,1) = u_i$ y que $C^*(u_{i+1},1) = u_{i+1}$. Como consecuencia,

$$\begin{aligned}
C_{HERMITE}(u,1) &= \\
& \left[\left(\frac{2*(u-u_i)^3}{(u_{i+1}-u_i)^3} - \frac{3*(u-u_i)^2}{(u_{i+1}-u_i)^2} + 1 \right) * u_i + \left(\frac{-2*(u-u_i)^3}{(u_{i+1}-u_i)^3} + \frac{3*(u-u_i)^2}{(u_{i+1}-u_i)^2} \right) * u_{i+1} + \right. \\
& \left. \left(\frac{(u-u_i)^3}{(u_{i+1}-u_i)^3} - \frac{2*(u-u_i)^2}{(u_{i+1}-u_i)^2} + (u-u_i) \right) * 1 + \left(\frac{(u-u_i)^3}{(u_{i+1}-u_i)^3} - \frac{(u-u_i)^2}{(u_{i+1}-u_i)^2} \right) * 1 \right] = \\
& \left[\frac{-2*(u-u_i)^3}{(u_{i+1}-u_i)^3} + \frac{3*(u-u_i)^2}{(u_{i+1}-u_i)^2} + \frac{2*(u-u_i)^3}{(u_{i+1}-u_i)^3} - \frac{3*(u-u_i)^2}{(u_{i+1}-u_i)^2} + u \right] = u \\
& \quad \forall u \in [u_i, u_{i+1}]
\end{aligned}$$

(AIII.1)

De igual manera, vamos a hallar la imagen a través de $C_{HERMITE}$ del punto $(1,v)$. Asociado al punto $u = 1$ habrá que considerar como vértices del rectángulo al que pertenece $u_i = (n-1)/n$ y $u_{i+1} = 1$. Teniendo en cuenta que

$$\frac{\partial C_{HERMITE}(1, v_l)}{\partial v} = 1 \quad \forall l \in \{0, 1, 2, \dots, m\} \quad , \text{ y desarrollando la expresión (4.19) de igual forma que antes para }$$

$C_{HERMITE}(u,1)$ se llegaría a:

$$\begin{aligned}
C_{HERMITE}(1,v) &= \\
& \left[\left(\frac{-2*(1-\frac{n-1}{n})^3}{(1-\frac{n-1}{n})^3} + \frac{3*(1-\frac{n-1}{n})^2}{(1-\frac{n-1}{n})^2} \right) * v_j \right] * \left(\frac{2*(v-v_j)^3}{(v_{j+1}-v_j)^3} - \frac{3*(v-v_j)^2}{(v_{j+1}-v_j)^2} + 1 \right) + \\
& \left[\left(\frac{-2*(1-\frac{n-1}{n})^3}{(1-\frac{n-1}{n})^3} + \frac{3*(1-\frac{n-1}{n})^2}{(1-\frac{n-1}{n})^2} \right) * v_{j+1} \right] * \left(\frac{-2*(v-v_j)^3}{(v_{j+1}-v_j)^3} + \frac{3*(v-v_j)^2}{(v_{j+1}-v_j)^2} \right) + \\
& \left[\left(\frac{-2*(1-\frac{n-1}{n})^3}{(1-\frac{n-1}{n})^3} + \frac{3*(1-\frac{n-1}{n})^2}{(1-\frac{n-1}{n})^2} \right) * 1 \right] * \left(\frac{(v-v_j)^3}{(v_{j+1}-v_j)^3} - \frac{2*(v-v_j)^2}{(v_{j+1}-v_j)^2} + (v-v_j) \right) + \\
& \left[\left(\frac{-2*(1-\frac{n-1}{n})^3}{(1-\frac{n-1}{n})^3} + \frac{3*(1-\frac{n-1}{n})^2}{(1-\frac{n-1}{n})^2} \right) * 1 \right] * \left(\frac{(v-v_j)^3}{(v_{j+1}-v_j)^3} - \frac{(v-v_j)^2}{(v_{j+1}-v_j)^2} \right) = \\
& \left[\left(\frac{-2*(1-\frac{n-1}{n})^3}{(1-\frac{n-1}{n})^3} + \frac{3*(1-\frac{n-1}{n})^2}{(1-\frac{n-1}{n})^2} \right) \right] * \left(\frac{-2*(v-v_j)^3}{(v_{j+1}-v_j)^3} + \frac{3*(v-v_j)^2}{(v_{j+1}-v_j)^2} + v_j \right) + \\
& \left[\left(\frac{-2*(1-\frac{n-1}{n})^3}{(1-\frac{n-1}{n})^3} + \frac{3*(1-\frac{n-1}{n})^2}{(1-\frac{n-1}{n})^2} \right) \right] * \left(\frac{2*(v-v_j)^3}{(v_{j+1}-v_j)^3} - \frac{3*(v-v_j)^2}{(v_{j+1}-v_j)^2} + v - v_j \right) = v \\
& \quad \forall v \in [v_j, v_{j+1}]
\end{aligned}$$

(AIII.2)

lo cual concluye la demostración del lema.

c.q.d

ANEXO IV: Demostración del lema 4.4

“Sea una partición del cuadrado unidad en $n \times m$ rectángulos del mismo área y sea la solución óptima al problema de programación fraccional (4.7) dada por $c_{ij}^* = C^* \left(\frac{i}{n}, \frac{j}{m} \right) \forall i \in \{0,1,2,\dots,n\} \quad \forall j \in \{0,1,2,\dots,m\}$ que define una subcópula C^* de dominio los puntos de intersección entre los rectángulos que originan la partición.

Sea $C_{HERMITE}(u,v)$ la función que se obtiene por interpolación cúbica de una subcópula C^* mediante polinomios de Hermite. Supongamos que las derivadas de segundo orden en el dominio de definición de la subcópula valen cero, es decir,

decir, $\frac{\partial^2 C_{HERMITE}(u_i, v_j)}{\partial u \partial v} = 0 \quad \forall i \in \{0,1,2,\dots,n\} \quad \forall j \in \{0,1,2,\dots,m\}$, en cuyo caso, $C_{HERMITE}(u,v)$ define una superficie interpoladora de Ferguson $C_{FERGUSON}(u,v)$.

Si $\frac{\partial C_{FERGUSON}(u_i, v_j)}{\partial u}$ y $\frac{\partial C_{FERGUSON}(u_i, v_j)}{\partial v} \quad \forall i \in \{0,1,2,\dots,n\} \quad \forall j \in \{0,1,2,\dots,m\}$

proporcionan junto con una relación de δ s y γ s una solución del problema de programación no lineal entera (4.21), entonces la función $C_{FERGUSON}(u,v)$ cumple la tercera de las condiciones para ser cópula:

$$V_{C_{FERGUSON}}([u_1, u_2] \times [v_1, v_2]) = C_{FERGUSON}(u_2, v_2) - C_{FERGUSON}(u_2, v_1) - C_{FERGUSON}(u_1, v_2) + C_{FERGUSON}(u_1, v_1) \geq 0 \quad , ,$$

$$\forall u_1, u_2, v_1, v_2. t. q. u_1 \leq u_2, v_1 \leq v_2$$

demostración

Vamos a ver que las condiciones que deben satisfacerse para que el volumen de un rectángulo arbitrario a través de $C_{FERGUSON}$ sea mayor o igual que cero, se traducen en la existencia de una solución al problema (4.21) dada por una relación de valores binarios asociados a las variables δ s y γ s y a las también variables

$$\frac{\partial C_{FERGUSON}(u_i, v_j)}{\partial u} \quad y \quad \frac{\partial C_{FERGUSON}(u_i, v_j)}{\partial v} \quad \forall i \in \{0,1,2,\dots,n\} \quad \forall j \in \{0,1,2,\dots,m\} \quad .$$

Realmente, basta establecer estas condiciones respecto de un rectángulo cualquiera (en rojo en el gráfico siguiente) totalmente contenido o coincidente con alguno uno de los $m \times n$ que surgen como consecuencia de la partición del cuadrado unidad dado que el volumen de cualquier otro rectángulo (en gris en dicho gráfico) se podrá expresar siempre como suma de los volúmenes de los del primer tipo.

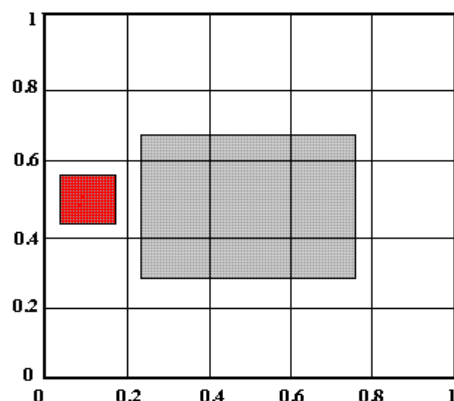


Ilustración AIV.1: Posibles rectángulos contenidos en el cuadrado unidad

Supongamos un rectángulo $B = [u_i^l, u_i^2] \times [v_j^l, v_j^2]$ totalmente contenido en uno de los rectángulos que conforman la rejilla, por ejemplo $B = [u_i, u_{i+1}] \times [v_j, v_{j+1}]$. Vamos a descomponer el volumen de dicho rectángulo a través de la función C_{FERGUSON} (aplicando (4.20)) en tantos bloques como pueden ser distinguidos en la matriz de la geometría de Hermite (4.18):

- Un primer bloque asociado al valor de la función en cada uno de los vértices de la superficie interpoladora.
- Un segundo bloque asociado al valor de las derivadas de primer orden respecto de U de la función en cada vértice.
- Un tercer bloque asociado al valor de las derivadas de primer orden respecto de V de la función en cada vértice.
- Un cuarto bloque (iría asociado a la función C_{HERMITE} pero no a C_{FERGUSON}) asociado al valor de las derivadas segundas de la función en cada vértice. No mostraremos este bloque dado que al hacer el desarrollo para C_{FERGUSON} saldría igual a 0.

Primer bloque: Correspondiente a los valores interpolados $C(u_i, v_j)$

$$\begin{aligned}
 V_{C_{HERMITE_BLOQUEI}}(B) = & \\
 & \left[\left(\frac{2*(u_i^2-u_i)^3}{(u_{i+1}-u_i)^3} - \frac{3*(u_i^2-u_i)^2}{(u_{i+1}-u_i)^2} + 1 \right) * C'(u_i, v_j) + \left(\frac{-2*(u_i^2-u_i)^3}{(u_{i+1}-u_i)^3} + \frac{3*(u_i^2-u_i)^2}{(u_{i+1}-u_i)^2} \right) * C'(u_{i+1}, v_j) \right] * \\
 & \left(\frac{2*(v_j^2-v_j)^3}{(v_{j+1}-v_j)^3} - \frac{3*(v_j^2-v_j)^2}{(v_{j+1}-v_j)^2} + 1 \right) + \\
 & \left[\left(\frac{2*(u_i^2-u_i)^3}{(u_{i+1}-u_i)^3} - \frac{3*(u_i^2-u_i)^2}{(u_{i+1}-u_i)^2} + 1 \right) * C'(u_i, v_{j+1}) + \left(\frac{-2*(u_i^2-u_i)^3}{(u_{i+1}-u_i)^3} + \frac{3*(u_i^2-u_i)^2}{(u_{i+1}-u_i)^2} \right) * C'(u_{i+1}, v_{j+1}) \right] * \\
 & \left(\frac{-2*(v_j^2-v_j)^3}{(v_{j+1}-v_j)^3} + \frac{3*(v_j^2-v_j)^2}{(v_{j+1}-v_j)^2} \right) - \\
 & \left[\left(\frac{2*(u_i^2-u_i)^3}{(u_{i+1}-u_i)^3} - \frac{3*(u_i^2-u_i)^2}{(u_{i+1}-u_i)^2} + 1 \right) * C'(u_i, v_j) + \left(\frac{-2*(u_i^2-u_i)^3}{(u_{i+1}-u_i)^3} + \frac{3*(u_i^2-u_i)^2}{(u_{i+1}-u_i)^2} \right) * C'(u_{i+1}, v_j) \right] * \\
 & \left(\frac{2*(v_j^1-v_j)^3}{(v_{j+1}-v_j)^3} - \frac{3*(v_j^1-v_j)^2}{(v_{j+1}-v_j)^2} + 1 \right) - \\
 & \left[\left(\frac{2*(u_i^2-u_i)^3}{(u_{i+1}-u_i)^3} - \frac{3*(u_i^2-u_i)^2}{(u_{i+1}-u_i)^2} + 1 \right) * C'(u_i, v_{j+1}) + \left(\frac{-2*(u_i^2-u_i)^3}{(u_{i+1}-u_i)^3} + \frac{3*(u_i^2-u_i)^2}{(u_{i+1}-u_i)^2} \right) * C'(u_{i+1}, v_{j+1}) \right] * \\
 & \left(\frac{-2*(v_j^1-v_j)^3}{(v_{j+1}-v_j)^3} + \frac{3*(v_j^1-v_j)^2}{(v_{j+1}-v_j)^2} \right) - \\
 & \left[\left(\frac{2*(u_i^1-u_i)^3}{(u_{i+1}-u_i)^3} - \frac{3*(u_i^1-u_i)^2}{(u_{i+1}-u_i)^2} + 1 \right) * C'(u_i, v_j) + \left(\frac{-2*(u_i^1-u_i)^3}{(u_{i+1}-u_i)^3} + \frac{3*(u_i^1-u_i)^2}{(u_{i+1}-u_i)^2} \right) * C'(u_{i+1}, v_j) \right] * \\
 & \left(\frac{2*(v_j^2-v_j)^3}{(v_{j+1}-v_j)^3} - \frac{3*(v_j^2-v_j)^2}{(v_{j+1}-v_j)^2} + 1 \right) - \\
 & \left[\left(\frac{2*(u_i^1-u_i)^3}{(u_{i+1}-u_i)^3} - \frac{3*(u_i^1-u_i)^2}{(u_{i+1}-u_i)^2} + 1 \right) * C'(u_i, v_{j+1}) + \left(\frac{-2*(u_i^1-u_i)^3}{(u_{i+1}-u_i)^3} + \frac{3*(u_i^1-u_i)^2}{(u_{i+1}-u_i)^2} \right) * C'(u_{i+1}, v_{j+1}) \right] * \\
 & \left(\frac{-2*(v_j^2-v_j)^3}{(v_{j+1}-v_j)^3} + \frac{3*(v_j^2-v_j)^2}{(v_{j+1}-v_j)^2} \right) + \\
 & \left[\left(\frac{2*(u_i^1-u_i)^3}{(u_{i+1}-u_i)^3} - \frac{3*(u_i^1-u_i)^2}{(u_{i+1}-u_i)^2} + 1 \right) * C'(u_i, v_j) + \left(\frac{-2*(u_i^1-u_i)^3}{(u_{i+1}-u_i)^3} + \frac{3*(u_i^1-u_i)^2}{(u_{i+1}-u_i)^2} \right) * C'(u_{i+1}, v_j) \right] * \\
 & \left(\frac{2*(v_j^1-v_j)^3}{(v_{j+1}-v_j)^3} - \frac{3*(v_j^1-v_j)^2}{(v_{j+1}-v_j)^2} + 1 \right) + \\
 & \left[\left(\frac{2*(u_i^1-u_i)^3}{(u_{i+1}-u_i)^3} - \frac{3*(u_i^1-u_i)^2}{(u_{i+1}-u_i)^2} + 1 \right) * C'(u_i, v_{j+1}) + \left(\frac{-2*(u_i^1-u_i)^3}{(u_{i+1}-u_i)^3} + \frac{3*(u_i^1-u_i)^2}{(u_{i+1}-u_i)^2} \right) * C'(u_{i+1}, v_{j+1}) \right] * \\
 & \left(\frac{-2*(v_j^1-v_j)^3}{(v_{j+1}-v_j)^3} + \frac{3*(v_j^1-v_j)^2}{(v_{j+1}-v_j)^2} \right) = \\
 V_{C'}(B) * & \left[\left(\frac{2*(u_i^2-u_i)^3}{(u_{i+1}-u_i)^3} - \frac{3*(u_i^2-u_i)^2}{(u_{i+1}-u_i)^2} \right) - \left(\frac{2*(u_i^1-u_i)^3}{(u_{i+1}-u_i)^3} - \frac{3*(u_i^1-u_i)^2}{(u_{i+1}-u_i)^2} \right) \right] * \\
 & \left[\left(\frac{2*(v_j^2-v_j)^3}{(v_{j+1}-v_j)^3} - \frac{3*(v_j^2-v_j)^2}{(v_{j+1}-v_j)^2} \right) - \left(\frac{2*(v_j^1-v_j)^3}{(v_{j+1}-v_j)^3} - \frac{3*(v_j^1-v_j)^2}{(v_{j+1}-v_j)^2} \right) \right]
 \end{aligned}$$

Por comodidad nos va a interesar descomponer el resultado final en dos sumandos iguales:

$$\begin{aligned}
 V_{C_{HERMITE_BLOQUEI}}(B) = & \\
 & \frac{V_{C'}(B)}{2} * \left[\left(\frac{2*(u_i^2-u_i)^3}{(u_{i+1}-u_i)^3} - \frac{3*(u_i^2-u_i)^2}{(u_{i+1}-u_i)^2} \right) - \left(\frac{2*(u_i^1-u_i)^3}{(u_{i+1}-u_i)^3} - \frac{3*(u_i^1-u_i)^2}{(u_{i+1}-u_i)^2} \right) \right] * \\
 & \left[\left(\frac{2*(v_j^2-v_j)^3}{(v_{j+1}-v_j)^3} - \frac{3*(v_j^2-v_j)^2}{(v_{j+1}-v_j)^2} \right) - \left(\frac{2*(v_j^1-v_j)^3}{(v_{j+1}-v_j)^3} - \frac{3*(v_j^1-v_j)^2}{(v_{j+1}-v_j)^2} \right) \right] + \\
 & \frac{V_{C'}(B)}{2} * \left[\left(\frac{2*(u_i^2-u_i)^3}{(u_{i+1}-u_i)^3} - \frac{3*(u_i^2-u_i)^2}{(u_{i+1}-u_i)^2} \right) - \left(\frac{2*(u_i^1-u_i)^3}{(u_{i+1}-u_i)^3} - \frac{3*(u_i^1-u_i)^2}{(u_{i+1}-u_i)^2} \right) \right] * \\
 & \left[\left(\frac{2*(v_j^2-v_j)^3}{(v_{j+1}-v_j)^3} - \frac{3*(v_j^2-v_j)^2}{(v_{j+1}-v_j)^2} \right) - \left(\frac{2*(v_j^1-v_j)^3}{(v_{j+1}-v_j)^3} - \frac{3*(v_j^1-v_j)^2}{(v_{j+1}-v_j)^2} \right) \right]
 \end{aligned}$$

(AIV.1)

Segundo bloque:Correspondiente a las derivadas primeras respecto de U de $C_{FERGUSON}$ en los vértices (u_i, v_j)

$$\begin{aligned}
V_{C_{HERMITE_BLOQUEZ}}(B) = & \\
& \left[\left(\frac{(u_i^2 - u_i)^3}{(u_{i+1} - u_i)^2} - \frac{2*(u_i^2 - u_i)^2}{(u_{i+1} - u_i)} + (u_i^2 - u_i) \right) * \frac{\partial C'(u_i, v_j)}{\partial u} + \left(\frac{(u_i^2 - u_i)^3}{(u_{i+1} - u_i)^2} - \frac{(u_i^2 - u_i)^2}{(u_{i+1} - u_i)} \right) * \frac{\partial C'(u_{i+1}, v_j)}{\partial u} \right] * \\
& \left(\frac{2*(v_j^2 - v_j)^3}{(v_{j+1} - v_j)^3} - \frac{3*(v_j^2 - v_j)^2}{(v_{j+1} - v_j)^2} + 1 \right) + \\
& \left[\left(\frac{(u_i^2 - u_i)^3}{(u_{i+1} - u_i)^2} - \frac{2*(u_i^2 - u_i)^2}{(u_{i+1} - u_i)} + (u - u_i) \right) * \frac{\partial C'(u_i, v_{j+1})}{\partial u} + \left(\frac{(u_i^2 - u_i)^3}{(u_{i+1} - u_i)^2} - \frac{(u_i^2 - u_i)^2}{(u_{i+1} - u_i)} \right) * \frac{\partial C'(u_{i+1}, v_{j+1})}{\partial u} \right] * \\
& \left(\frac{-2*(v_j^2 - v_j)^3}{(v_{j+1} - v_j)^3} + \frac{3*(v_j^2 - v_j)^2}{(v_{j+1} - v_j)^2} \right) - \\
& \left[\left(\frac{(u_i^2 - u_i)^3}{(u_{i+1} - u_i)^2} - \frac{2*(u_i^2 - u_i)^2}{(u_{i+1} - u_i)} + (u_i^2 - u_i) \right) * \frac{\partial C'(u_i, v_j)}{\partial u} + \left(\frac{(u_i^2 - u_i)^3}{(u_{i+1} - u_i)^2} - \frac{(u_i^2 - u_i)^2}{(u_{i+1} - u_i)} \right) * \frac{\partial C'(u_{i+1}, v_j)}{\partial u} \right] * \\
& \left(\frac{2*(v_j^1 - v_j)^3}{(v_{j+1} - v_j)^3} - \frac{3*(v_j^1 - v_j)^2}{(v_{j+1} - v_j)^2} + 1 \right) - \\
& \left[\left(\frac{(u_i^2 - u_i)^3}{(u_{i+1} - u_i)^2} - \frac{2*(u_i^2 - u_i)^2}{(u_{i+1} - u_i)} + (u - u_i) \right) * \frac{\partial C'(u_i, v_{j+1})}{\partial u} + \left(\frac{(u_i^2 - u_i)^3}{(u_{i+1} - u_i)^2} - \frac{(u_i^2 - u_i)^2}{(u_{i+1} - u_i)} \right) * \frac{\partial C'(u_{i+1}, v_{j+1})}{\partial u} \right] * \\
& \left(\frac{-2*(v_j^1 - v_j)^3}{(v_{j+1} - v_j)^3} + \frac{3*(v_j^1 - v_j)^2}{(v_{j+1} - v_j)^2} \right) - \\
& \left[\left(\frac{(u_i^1 - u_i)^3}{(u_{i+1} - u_i)^2} - \frac{2*(u_i^1 - u_i)^2}{(u_{i+1} - u_i)} + (u_i^1 - u_i) \right) * \frac{\partial C'(u_i, v_j)}{\partial u} + \left(\frac{(u_i^1 - u_i)^3}{(u_{i+1} - u_i)^2} - \frac{(u_i^1 - u_i)^2}{(u_{i+1} - u_i)} \right) * \frac{\partial C'(u_{i+1}, v_j)}{\partial u} \right] * \\
& \left(\frac{2*(v_j^2 - v_j)^3}{(v_{j+1} - v_j)^3} - \frac{3*(v_j^2 - v_j)^2}{(v_{j+1} - v_j)^2} + 1 \right) - \\
& \left[\left(\frac{(u_i^1 - u_i)^3}{(u_{i+1} - u_i)^2} - \frac{2*(u_i^1 - u_i)^2}{(u_{i+1} - u_i)} + (u - u_i) \right) * \frac{\partial C'(u_i, v_{j+1})}{\partial u} + \left(\frac{(u_i^1 - u_i)^3}{(u_{i+1} - u_i)^2} - \frac{(u_i^1 - u_i)^2}{(u_{i+1} - u_i)} \right) * \frac{\partial C'(u_{i+1}, v_{j+1})}{\partial u} \right] * \\
& \left(\frac{-2*(v_j^2 - v_j)^3}{(v_{j+1} - v_j)^3} + \frac{3*(v_j^2 - v_j)^2}{(v_{j+1} - v_j)^2} \right) + \\
& \left[\left(\frac{(u_i^1 - u_i)^3}{(u_{i+1} - u_i)^2} - \frac{2*(u_i^1 - u_i)^2}{(u_{i+1} - u_i)} + (u_i^1 - u_i) \right) * \frac{\partial C'(u_i, v_j)}{\partial u} + \left(\frac{(u_i^1 - u_i)^3}{(u_{i+1} - u_i)^2} - \frac{(u_i^1 - u_i)^2}{(u_{i+1} - u_i)} \right) * \frac{\partial C'(u_{i+1}, v_j)}{\partial u} \right] * \\
& \left(\frac{2*(v_j^1 - v_j)^3}{(v_{j+1} - v_j)^3} - \frac{3*(v_j^1 - v_j)^2}{(v_{j+1} - v_j)^2} + 1 \right) + \\
& \left[\left(\frac{(u_i^1 - u_i)^3}{(u_{i+1} - u_i)^2} - \frac{2*(u_i^1 - u_i)^2}{(u_{i+1} - u_i)} + (u - u_i) \right) * \frac{\partial C'(u_i, v_{j+1})}{\partial u} + \left(\frac{(u_i^1 - u_i)^3}{(u_{i+1} - u_i)^2} - \frac{(u_i^1 - u_i)^2}{(u_{i+1} - u_i)} \right) * \frac{\partial C'(u_{i+1}, v_{j+1})}{\partial u} \right] * \\
& \left(\frac{-2*(v_j^1 - v_j)^3}{(v_{j+1} - v_j)^3} + \frac{3*(v_j^1 - v_j)^2}{(v_{j+1} - v_j)^2} \right)
\end{aligned}$$

Desarrollando la expresión anterior se llegaría a

$$\begin{aligned}
V_{C_{HERMITE_BLOQUEZ}}(B) = & \left[\frac{\partial C'(u_i, v_j)}{\partial u} - \frac{\partial C'(u_i, v_{j+1})}{\partial u} \right] * \\
& \left[\left(\frac{(u_i^2 - u_i)^3}{(u_{i+1} - u_i)^2} - \frac{2*(u_i^2 - u_i)^2}{(u_{i+1} - u_i)} + (u_i^2 - u_i) \right) - \left(\frac{(u_i^1 - u_i)^3}{(u_{i+1} - u_i)^2} - \frac{2*(u_i^1 - u_i)^2}{(u_{i+1} - u_i)} + (u_i^1 - u_i) \right) \right] * \\
& \left[\left(\frac{2*(v_j^2 - v_j)^3}{(v_{j+1} - v_j)^3} - \frac{3*(v_j^2 - v_j)^2}{(v_{j+1} - v_j)^2} \right) - \left(\frac{2*(v_j^1 - v_j)^3}{(v_{j+1} - v_j)^3} - \frac{3*(v_j^1 - v_j)^2}{(v_{j+1} - v_j)^2} \right) \right] + \\
& \left[\frac{\partial C'(u_{i+1}, v_j)}{\partial u} - \frac{\partial C'(u_{i+1}, v_{j+1})}{\partial u} \right] * \\
& \left[\left(\frac{(u_i^2 - u_i)^3}{(u_{i+1} - u_i)^2} - \frac{(u_i^2 - u_i)^2}{(u_{i+1} - u_i)} \right) - \left(\frac{(u_i^1 - u_i)^3}{(u_{i+1} - u_i)^2} - \frac{(u_i^1 - u_i)^2}{(u_{i+1} - u_i)} \right) \right] * \\
& \left[\left(\frac{2*(v_j^2 - v_j)^3}{(v_{j+1} - v_j)^3} - \frac{3*(v_j^2 - v_j)^2}{(v_{j+1} - v_j)^2} \right) - \left(\frac{2*(v_j^1 - v_j)^3}{(v_{j+1} - v_j)^3} - \frac{3*(v_j^1 - v_j)^2}{(v_{j+1} - v_j)^2} \right) \right]
\end{aligned}$$

(AIV.2)

Las derivadas respecto de U y de V (vectores tangentes) van a ser las variables con las que vamos a poder jugar para tratar de conseguir que la función que estamos definiendo sea 2-creciente y por tanto una cópula, dado que los valores de la función vienen fijados de antemano (de hecho son los valores que estamos interpolando) y las derivadas de segundo orden estamos asumiendo que son iguales a 0 (superficie interpoladora de Ferguson). Vamos a asignar un nombre a las diferencias entre dichas derivadas (cuyo valor habremos de calcular):

$$\frac{\partial C'(u_i, v_j)}{\partial u} - \frac{\partial C'(u_i, v_{j+1})}{\partial u} = K'_{ij} \quad \text{y} \quad \frac{\partial C'(u_{i+1}, v_j)}{\partial u} - \frac{\partial C'(u_{i+1}, v_{j+1})}{\partial u} = L'_{ij}$$

En vez de dichos valores, podemos considerar sin restricción de generalidad, otros alternativos proporcionales que, por su forma, nos van a convenir más de cara a los cálculos posteriores. Sean:

$$\frac{\partial C'(u_i, v_j)}{\partial u} - \frac{\partial C'(u_i, v_{j+1})}{\partial u} = \frac{K_{ij} * V_C'(B)}{2*(u_{i+1} - u_i)} \quad \text{y} \quad \frac{\partial C'(u_{i+1}, v_j)}{\partial u} - \frac{\partial C'(u_{i+1}, v_{j+1})}{\partial u} = \frac{L_{ij} * V_C'(B)}{2*(u_{i+1} - u_i)} \quad , \text{ es decir,}$$

$$K_{ij} = \frac{2*(u_{i+1} - u_i)}{V_C'(B)} * \left[\frac{\partial C'(u_i, v_j)}{\partial u} - \frac{\partial C'(u_i, v_{j+1})}{\partial u} \right] \quad (\text{AIV.3})$$

$$L_{ij} = \frac{2*(u_{i+1} - u_i)}{V_C'(B)} * \left[\frac{\partial C'(u_{i+1}, v_j)}{\partial u} - \frac{\partial C'(u_{i+1}, v_{j+1})}{\partial u} \right] \quad (\text{AIV.4})$$

En función de ellos, podemos desarrollar la expresión anterior y escribir:

$$\begin{aligned} V_{C_{\text{HERMITE_BLOQUES}}}(B) &= \left(\frac{V_C'(B)}{2} \right) * \left[\left(\frac{2*(v_j^2 - v_j)}{(v_{j+1} - v_j)^3} - \frac{3*(v_j^2 - v_j)}{(v_{j+1} - v_j)^2} \right) - \left(\frac{2*(v_j^1 - v_j)}{(v_{j+1} - v_j)^3} - \frac{3*(v_j^1 - v_j)}{(v_{j+1} - v_j)^2} \right) \right] * \\ &\{ [K_{ij} * \left(\frac{(u_i^2 - u_i)^3}{(u_{i+1} - u_i)^3} - \frac{2*(u_i^2 - u_i)^2}{(u_{i+1} - u_i)^2} + \frac{(u_i^2 - u_i)}{(u_{i+1} - u_i)} \right) - \left(\frac{(u_i^1 - u_i)^3}{(u_{i+1} - u_i)^3} - \frac{2*(u_i^1 - u_i)^2}{(u_{i+1} - u_i)^2} + \frac{(u_i^1 - u_i)}{(u_{i+1} - u_i)} \right)] + \right. \\ &\left. L_{ij} * \left[\left(\frac{(u_i^2 - u_i)^3}{(u_{i+1} - u_i)^3} - \frac{(u_i^2 - u_i)^2}{(u_{i+1} - u_i)^2} \right) - \left(\frac{(u_i^1 - u_i)^3}{(u_{i+1} - u_i)^3} - \frac{(u_i^1 - u_i)^2}{(u_{i+1} - u_i)^2} \right) \right] \} \end{aligned}$$

(AIV.5)

Tercer bloque: Correspondiente a las derivadas primeras respecto de V de C_{FERGUSON} en los vértices (u_i, v_j)

El desarrollo se puede hacer de forma análoga al anterior. La única diferencia es que en este caso, las derivadas de primer orden son respecto de la variable V , pudiendo igualmente definir:

$$P_{ij} = \frac{2*(v_{j+1} - v_j)}{V_C'(B)} * \left[\frac{\partial C'(u_i, v_j)}{\partial v} - \frac{\partial C'(u_{i+1}, v_j)}{\partial v} \right] \quad (\text{AIV.6})$$

$$Q_{ij} = \frac{2*(v_{j+1} - v_j)}{V_C'(B)} * \left[\frac{\partial C'(u_i, v_{j+1})}{\partial v} - \frac{\partial C'(u_{i+1}, v_{j+1})}{\partial v} \right] \quad (\text{AIV.7})$$

En función de ellos, la expresión asociada al bloque 3 sería la siguiente:

$$\begin{aligned}
V_{C_{\text{HERMITE...BLOQUES}}}(B) &= \left(\frac{V_C'(B)}{2}\right) * \left[\left(\frac{2*(u_i^2 - u_i)^3}{(u_{i+1} - u_i)^3} - \frac{3*(u_i^2 - u_i)^2}{(u_{i+1} - u_i)^2} \right) - \left(\frac{2*(u_i^1 - u_i)^3}{(u_{i+1} - u_i)^3} - \frac{3*(u_i^1 - u_i)^2}{(u_{i+1} - u_i)^2} \right) \right] * \\
&\{ [P_{ij} * \left[\left(\frac{(v_j^2 - v_j)^3}{(v_{j+1} - v_j)^3} - \frac{2*(v_j^2 - v_j)^2}{(v_{j+1} - v_j)^2} + \frac{(v_j^2 - v_j)}{(v_{j+1} - v_j)} \right) - \left(\frac{(v_j^1 - v_j)^3}{(v_{j+1} - v_j)^3} - \frac{2*(v_j^1 - v_j)^2}{(v_{j+1} - v_j)^2} + \frac{(v_j^1 - v_j)}{(v_{j+1} - v_j)} \right) \right] + \right. \\
&Q_{ij} * \left. \left[\left(\frac{(v_j^2 - v_j)^3}{(v_{j+1} - v_j)^3} - \frac{(v_j^2 - v_j)^2}{(v_{j+1} - v_j)^2} \right) - \left(\frac{(v_j^1 - v_j)^3}{(v_{j+1} - v_j)^3} - \frac{(v_j^1 - v_j)^2}{(v_{j+1} - v_j)^2} \right) \right] \}
\end{aligned}$$

(AIV.8)

Recopilando las expresiones de los bloques 1,2 y 3 ((AIV.1), (AIV.5) y (AIV.8)) obtenemos la siguiente ecuación:

$$\begin{aligned}
V_{C_{\text{FERGUSON}}}(B) &= C(u_i^2, V_i^2) - C(u_i^2, V_i^1) - C(u_i^1, V_i^2) + C(u_i^1, V_i^1) = \\
&\left(\frac{V_C'(B)}{2}\right) * \left[\left(\frac{2*(v_j^2 - v_j)^3}{(v_{j+1} - v_j)^3} - \frac{3*(v_j^2 - v_j)^2}{(v_{j+1} - v_j)^2} \right) - \left(\frac{2*(v_j^1 - v_j)^3}{(v_{j+1} - v_j)^3} - \frac{3*(v_j^1 - v_j)^2}{(v_{j+1} - v_j)^2} \right) \right] * \\
&\{ [\left(\frac{2*(u_i^2 - u_i)^3}{(u_{i+1} - u_i)^3} - \frac{3*(u_i^2 - u_i)^2}{(u_{i+1} - u_i)^2} \right) - \left(\frac{2*(u_i^1 - u_i)^3}{(u_{i+1} - u_i)^3} - \frac{3*(u_i^1 - u_i)^2}{(u_{i+1} - u_i)^2} \right)] + \\
&K_{ij} * \left[\left(\frac{(u_i^2 - u_i)^3}{(u_{i+1} - u_i)^3} - \frac{2*(u_i^2 - u_i)^2}{(u_{i+1} - u_i)^2} + \frac{(u_i^2 - u_i)}{(u_{i+1} - u_i)} \right) - \left(\frac{(u_i^1 - u_i)^3}{(u_{i+1} - u_i)^3} - \frac{2*(u_i^1 - u_i)^2}{(u_{i+1} - u_i)^2} + \frac{(u_i^1 - u_i)}{(u_{i+1} - u_i)} \right) \right] + \\
&L_{ij} * \left[\left(\frac{(u_i^2 - u_i)^3}{(u_{i+1} - u_i)^3} - \frac{(u_i^2 - u_i)^2}{(u_{i+1} - u_i)^2} \right) - \left(\frac{(u_i^1 - u_i)^3}{(u_{i+1} - u_i)^3} - \frac{(u_i^1 - u_i)^2}{(u_{i+1} - u_i)^2} \right) \right] \} + \\
&\left(\frac{V_C'(B)}{2}\right) * \left[\left(\frac{2*(u_i^2 - u_i)^3}{(u_{i+1} - u_i)^3} - \frac{3*(u_i^2 - u_i)^2}{(u_{i+1} - u_i)^2} \right) - \left(\frac{2*(u_i^1 - u_i)^3}{(u_{i+1} - u_i)^3} - \frac{3*(u_i^1 - u_i)^2}{(u_{i+1} - u_i)^2} \right) \right] * \\
&\{ [\left(\frac{2*(v_j^2 - v_j)^3}{(v_{j+1} - v_j)^3} - \frac{3*(v_j^2 - v_j)^2}{(v_{j+1} - v_j)^2} \right) - \left(\frac{2*(v_j^1 - v_j)^3}{(v_{j+1} - v_j)^3} - \frac{3*(v_j^1 - v_j)^2}{(v_{j+1} - v_j)^2} \right)] + \\
&P_{ij} * \left[\left(\frac{(v_j^2 - v_j)^3}{(v_{j+1} - v_j)^3} - \frac{2*(v_j^2 - v_j)^2}{(v_{j+1} - v_j)^2} + \frac{(v_j^2 - v_j)}{(v_{j+1} - v_j)} \right) - \left(\frac{(v_j^1 - v_j)^3}{(v_{j+1} - v_j)^3} - \frac{2*(v_j^1 - v_j)^2}{(v_{j+1} - v_j)^2} + \frac{(v_j^1 - v_j)}{(v_{j+1} - v_j)} \right) \right] + \\
&Q_{ij} * \left[\left(\frac{(v_j^2 - v_j)^3}{(v_{j+1} - v_j)^3} - \frac{(v_j^2 - v_j)^2}{(v_{j+1} - v_j)^2} \right) - \left(\frac{(v_j^1 - v_j)^3}{(v_{j+1} - v_j)^3} - \frac{(v_j^1 - v_j)^2}{(v_{j+1} - v_j)^2} \right) \right] \}
\end{aligned}$$

(AIV.9)

Este es el volumen de cualquier rectángulo contenido en $[u_i, u_{i+1}] \times [v_j, v_{j+1}]$ a través de la función (4.20). Los elementos a determinar van a ser las variables K 's, L 's, P 's y Q 's, o lo que es lo mismo, las derivadas de primer orden (vectores tangente) en los vértices del rectángulo. Se trata de saber qué condiciones deben satisfacer para que la ecuación sea mayor o igual que 0 pues, en dicho caso, la función será 2-creciente, que es la tercera restricción que perseguimos para que finalmente sea una cópula.

Sean las funciones:

$$- f(x) = (2 + K_{ij} + L_{ij}) * x^3 + (-3 - 2 * K_{ij} - L_{ij}) * x^2 + K_{ij} * x \quad (\text{AIV.10})$$

con $x_1 = \frac{u_i^1 - u_i}{u_{i+1} - u_i} \leq \frac{u_i^2 - u_i}{u_{i+1} - u_i} = x_2$, que son valores comprendidos entre 0 y 1 porque

$$u_i^1 - u_i \geq 0 \text{ y } u_{i+1} \geq u_i \Rightarrow x_1 \geq 0 \text{ y } u_i^2 \leq u_{i+1} \Rightarrow u_i^2 - u_i \leq u_{i+1} - u_i \Rightarrow x_2 \leq 1. \text{ Además, } f(0) = 0 \text{ y } f(1) = -1$$

$$- \quad g(y) = (2 + P_{ij} + Q_{ij}) * y^3 + (-3 - 2 * P_{ij} - Q_{ij}) * y^2 + P_{ij} * y \quad (\text{AIV.11})$$

con $y_1 = \frac{v_j^1 - v_j}{v_{j+1} - v_j} \leq \frac{v_j^2 - v_1}{v_{j+1} - v_j} = y_2$, que son valores comprendidos entre 0 y 1 porque

$$v_j^1 - v_j \geq 0 \quad y \quad v_{j+1} \geq v_j \Rightarrow y_1 \geq 0 \quad y \quad v_i^2 \leq v_{j+1} \Rightarrow v_j^2 - v_j \leq v_{j+1} - v_j \Rightarrow y_2 \leq 1 \quad . \text{ Además, } g(0) = 0 \text{ y } g(1) = -1$$

$$- \quad h(z) = 2 * z^3 - 3 * z^2 \quad (\text{AIV.12})$$

Esta función es estrictamente decreciente entre 0 (siendo $h(0) = 0$) y 1 (siendo $h(1) = -1$). Por tanto, si $x_1 \leq x_2$ y $y_1 \leq y_2$, siendo valores comprendidos entre 0 y 1, $h(x_1) \geq h(x_2)$ y $h(y_1) \geq h(y_2)$ o lo que es lo mismo, $h(x_2) - h(x_1) \leq 0$ y $h(y_2) - h(y_1) \leq 0$

Observemos que podemos expresar el volumen del rectángulo B a través de $C_{FERGUSON}$ en función de “ f ”, “ g ” y “ h ” de la siguiente manera:

$$V_{C_{FERGUSON}}(B) = C(u_i^2, v_i^2) - C(u_i^2, v_i^1) - C(u_i^1, v_i^2) + C(u_i^1, v_i^1) = \\ \frac{V_{C'}(B)}{2} * \{h(x_2) - h(x_1)\} * \{f(x_2) - f(x_1)\} + \frac{V_{C'}(B)}{2} * \{h(y_2) - h(y_1)\} * \{g(y_2) - g(y_1)\}$$

(AIV.13)

Esta cantidad se compone de dos sumandos cada uno de ellos con 3 factores:

- El volumen a través de C' de B que es mayor o igual que 0 por ser C' una subcópula y B un rectángulo de su dominio de definición.
- La función “ h ” ((AIV.12)) cuya contribución al sumando es siempre negativa pues $h(x_2) - h(x_1) \leq 0$ y $h(y_2) - h(y_1) \leq 0$.
- El tercero de los factores viene representado por las funciones “ f ” y “ g ” respectivamente ((AIV.10) y (AIV.11)).

Por tanto, para que cada uno de los sumandos sea positivo y así también el resultado global (en cuyo caso la función sería 2-creciente) basta demostrar que las funciones “ f ” y “ g ” son estrictamente decrecientes en $[0,1]$ pues x_1, x_2, y_1 e y_2 están entre 0 y 1, y $x_1 \leq x_2$ y $y_1 \leq y_2$, con lo cual, si “ f ” y “ g ” son decrecientes, se tendrá que $f(x_1) \geq f(x_2)$ y $g(y_1) \geq g(y_2)$, y así, que $f(x_2) - f(x_1) \leq 0$ y $g(y_2) - g(y_1) \leq 0$.

Puesto que ambas funciones responden a la misma estructura, estudiemos para una cualquiera de ellas, por ejemplo “ f ”,

qué propiedad debería satisfacerse para que sea estrictamente decreciente en el intervalo $[0,1]$.

Para que una función sea decreciente en un intervalo, el signo de su derivada en el mismo debe ser menor o igual que 0.

Dicha derivada viene dada por: $f'(x) = 3*(2 + K_{ij} + L_{ij})*x^2 + 2*(-3 - 2*K_{ij} - L_{ij})*x + K_{ij}$ (AIV.14)

Analicemos las posibilidades que pueden presentarse:

1. Si $2 + K_{ij} + L_{ij} = 0$ la función anterior es una recta,

$$f'(x) = 2*(-3 - 2*K_{ij} - L_{ij})*x + K_{ij} = 2*(-1 - K_{ij})*x + K_{ij}$$

Para que esta recta sea siempre negativa entre 0 y 1, basta imponer que en los extremos, el valor de dicha recta sea negativo. Es decir, $f'(0) = K_{ij} \leq 0$ y $f'(1) = -2 - K_{ij} \leq 0 \Rightarrow K_{ij} \geq -2$

La condición a imponer sobre L_{ij} es consecuencia directa pues, como $L_{ij} = -2 - K_{ij} \Rightarrow -2 \leq L_{ij} \leq 0$.

Por tanto, si $2 + K_{ij} + L_{ij} = 0$, entonces f' será negativa si $(K_{ij}, L_{ij}) \in [-2, 0] \times [-2, 0]$

Definiremos así una primera **condición R0**: $2 + K_{ij} + L_{ij} = 0, -2 \leq K_{ij} \leq 0, -2 \leq L_{ij} \leq 0$ (AIV.15)

2. Supongamos ahora que $2 + K_{ij} + L_{ij} \neq 0$

Entonces f' define la ecuación de una parábola de vértice $x_{\text{vertice}} = \frac{(3 + 2*K_{ij} + L_{ij})}{(2 + K_{ij} + L_{ij})}$ (AIV.16).

En los extremos, el valor de dicha parábola es $f'(0) = K_{ij}$ y $f'(1) = L_{ij}$, que deben ser menores o iguales que 0, dado que debe tener signo negativo en todos los puntos del intervalo $[0,1]$.

Así, en este caso, una primera **condición P0** es que $K_{ij} \leq 0$ y $L_{ij} \leq 0$ (AIV.17).

Para que la parábola se mantenga menor o igual que 0 dentro del intervalo $[0,1]$ debe cumplir una de las siguientes 3 condiciones:

- **Condición P1.**- Que la “x” del vértice sea menor que 0.
- **Condición P2.**- Que la “x” del vértice sea mayor que 1.
- **Condición P3.**- Que la “y” del vértice (imagen de la “x” del vértice) sea menor que 0.

La única posibilidad que no debe darse es que la “x” del vértice se encuentre entre 0 y 1 y que su imagen sea positiva, porque en dicho caso, habría al menos un punto cuya imagen a través de la parábola positiva.

Analicemos cada una de estas condiciones:

La **condición P1** equivale a que $\frac{3+2*K_{ij}+L_{ij}}{3*(2+K_{ij}+L_{ij})} < 0$ (AIV.18),

la cual se dará cuando numerador y denominador tengan distinto signo lo cual implica que se dé alguna de las siguientes condiciones:

– **Condición P11.-** $3+2*K_{ij}+L_{ij} < 0$ y $2+K_{ij}+L_{ij} > 0$ (AIV.19)

o

– **Condición P12.-** $3+2*K_{ij}+L_{ij} > 0$ y $2+K_{ij}+L_{ij} < 0$ (AIV.20)

Esta región factible viene dada por restricciones lineales donde el operador disyuntivo se puede plantear a través de una variable binaria δ_1 .

La **condición P2** equivale a que $\frac{3+2*K_{ij}+L_{ij}}{3*(2+K_{ij}+L_{ij})} > 1$ (AIV.21)

Si $3*(2+K_{ij}+L_{ij}) > 0 \Rightarrow 3+2*K_{ij}+L_{ij} > 6+3*K_{ij}+3*L_{ij} \Rightarrow K_{ij}+2*L_{ij}+3 < 0$

Si $3*(2+K_{ij}+L_{ij}) < 0 \Rightarrow 3+2*K_{ij}+L_{ij} < 6+3*K_{ij}+3*L_{ij} \Rightarrow K_{ij}+2*L_{ij}+3 > 0$

Por tanto, que se de la condición P2 implica que se dé alguna de las siguientes condiciones:

– **Condición P21.-** $2+K_{ij}+L_{ij} > 0$ y $K_{ij}+2*L_{ij}+3 < 0$ (AIV.22)

o

– **Condición P22.-** $2+K_{ij}+L_{ij} < 0$ y $K_{ij}+2*L_{ij}+3 > 0$ (AIV.23)

Esta región factible viene también dada por restricciones lineales donde el operador disyuntivo se puede plantear a través de una variable binaria δ_2 .

La **condición P3** equivale a que $f(x_{\text{vertice}}) = f\left(\frac{3+2*K_{ij}+L_{ij}}{2+K_{ij}+L_{ij}}\right) < 0$ (AIV.24)

Basta sustituir para comprobar que $f(x_{\text{vertice}}) = \frac{-(3+2*K_{ij}+L_{ij})^2}{3*(2+K_{ij}+L_{ij})} + K_{ij}$. Para que esta cantidad sea menor que 0, debe satisfacerse alguna de las siguientes condiciones:

– **Condición P31.-** Si $2+K_{ij}+L_{ij} > 0$ (AIV.25)

el denominador es positivo, y se puede desarrollar la expresión sin cambiar el signo de la desigualdad.

$$-(3+2*K_{ij}+L_{ij})^2 \leq -3*(2+K_{ij}+L_{ij})*K_{ij} \Rightarrow K_{ij}^2+L_{ij}^2+6*K_{ij}+6*L_{ij}+K_{ij}*L_{ij}+9 \geq 0 \Rightarrow$$

$$(K_{ij}+L_{ij}+3)^2 - K_{ij}*L_{ij} \geq 0 \Rightarrow (K_{ij}+L_{ij}+3)^2 \geq K_{ij}*L_{ij}$$

Esta condición redundante con que $2+K_{ij}+L_{ij}>0, K_{ij} \leq 0, L_{ij} \leq 0$. Es decir, basta que se den éstas últimas para que se implique que $(K_{ij}+L_{ij}+3)^2 \geq K_{ij}*L_{ij}$ pues, efectivamente

$$K_{ij}+L_{ij}>-2 \Rightarrow K_{ij}*L_{ij}+L_{ij}^2 < -2*L_{ij} \Rightarrow K_{ij}*L_{ij}-1 < -2*L_{ij}-L_{ij}^2-1 = -(L_{ij}+1)^2 \Rightarrow$$

$$K_{ij}*L_{ij} < 1-(L_{ij}+1)^2 \leq 1$$

y como consecuencia directa de P31 se tiene que

$$K_{ij}+L_{ij}>-2 \Rightarrow K_{ij}+L_{ij}+3 > 1 \Rightarrow (K_{ij}+L_{ij}+3)^2 \geq 1$$

con lo cual

$$K_{ij}*L_{ij} < 1-(L_{ij}+1)^2 \leq 1 \leq (K_{ij}+L_{ij}+3)^2$$

Por ello, en este caso, la condición no lineal resulta redundante.

- **Condición P32.-** Si $2+K_{ij}+L_{ij} < 0$ (AIV.26)

el denominador es negativo, y desarrollando la expresión del mismo modo que antes se llega a que

$(K_{ij}+L_{ij}+3)^2 \leq K_{ij}*L_{ij}$ que, como veremos, se está refiriendo a la región interior de una **cónica**, concretamente una **elipse**.

Efectivamente, la ecuación general de una cónica es

$$a_{00}+2*a_{01}*x_1+2*a_{02}*x_2+a_{11}*x_1^2+2*a_{12}*x_1*x_2+a_{22}*x_2^2=0 \quad (\text{AIV.27})$$

que matricialmente se puede expresar de la siguiente manera:

$$\begin{bmatrix} 1 & x_1 & x_2 \end{bmatrix} * \begin{bmatrix} a_{00} & a_{01} & a_{02} \\ a_{01} & a_{11} & a_{12} \\ a_{02} & a_{12} & a_{22} \end{bmatrix} * \begin{bmatrix} 1 \\ x_1 \\ x_2 \end{bmatrix} = 0 \quad (\text{AIV.28})$$

siendo A la matriz de la cónica que en nuestro caso es

$$A = \begin{bmatrix} 9 & 3 & 3 \\ 3 & 1 & \frac{1}{2} \\ 3 & \frac{1}{2} & 1 \end{bmatrix} \quad \text{o bien} \quad \begin{bmatrix} 3 & 1 & 1 \\ 1 & \frac{1}{3} & \frac{1}{6} \\ 1 & \frac{1}{6} & \frac{1}{3} \end{bmatrix} \quad (\text{AIV.29})$$

Sería indiferente considerar cualquiera de las dos matrices pues bastaría dividir la ecuación de la cónica por 3 para que la segunda de las matrices planteadas fuera también válida.

Es decir, la ecuación de la cónica que define la condición P32 es

$$[1 \quad x_1 \quad x_2] * \begin{bmatrix} 3 & 1 & 1 \\ 1 & \frac{1}{3} & \frac{1}{6} \\ 1 & \frac{1}{6} & \frac{1}{3} \end{bmatrix} * \begin{bmatrix} 1 \\ x_1 \\ x_2 \end{bmatrix} = 0 \quad (\text{AIV.30})$$

La clasificación de esta cónica se puede hacer en función de sus **invariantes**, que son aquellas constantes que no varían a pesar de los cambios que efectuemos en el sistema de referencia respecto del cual se definen.

El **invariante cuadrático** es $J = \begin{vmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{vmatrix} = \begin{vmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{vmatrix} = \frac{3}{4} \neq 0$

que por ser distinto de cero denota que estamos ante una **cónica con centro único**.

El **invariante cúbico** es $K = \begin{vmatrix} a_{00} & a_{01} & a_{02} \\ a_{01} & a_{11} & a_{12} \\ a_{02} & a_{12} & a_{22} \end{vmatrix} = \begin{vmatrix} 9 & 3 & 3 \\ 3 & 1 & \frac{1}{2} \\ 3 & \frac{1}{2} & 1 \end{vmatrix} = \frac{-9}{4} \neq 0$

que por ser distinto de cero denota que es una **cónica no reducible** y que, junto al hecho de que J sea positivo, se traduce en que la cónica sea una **elipse**.

El **invariante lineal** es $I = a_{11} + a_{22} = 2$ cuyo signo es distinto al del invariante cúbico K , lo que se traduce en que la cónica que estamos manejando es una **elipse real**.

Además, el hecho de que en la matriz A , los términos a_{12} no sean cero sugiere que la elipse se encuentra **rotada**. Del mismo modo, el que los términos a_{01} y a_{02} no sean cero sugiere que además se encuentra **trasladada** del origen.

Vamos a considerar la forma canónica de la elipse, y vamos a ver como, aplicándole una rotación y una translación concreta obtenemos la matriz A .

La ecuación reducida de la elipse lleva por matriz $A' = \begin{bmatrix} a'_{00} & 0 & 0 \\ 0 & a'_{11} & 0 \\ 0 & 0 & a'_{22} \end{bmatrix}$ (AIV.31)

donde $a'_{00} = \frac{K}{J} = -3$.

La ecuación característica de la elipse es $t^2 - t * I + J = 0$ (AIV.32)

con raíces $1/2$ y $3/2$, que permiten obtener la ecuación en forma reducida:

$$-3 + \frac{3}{2} * p^2 + \frac{1}{2} * q^2 = 0 \Rightarrow \frac{p^2}{2} + \frac{q^2}{6} = 1 \quad (\text{AIV.33})$$

y, a través de ella, la ecuación canónica de la elipse: $(\frac{p}{\sqrt{2}})^2 + (\frac{q}{\sqrt{6}})^2 = 1$ (AIV.34)

cuya matriz es
$$A'' = \begin{bmatrix} -1 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{6} \end{bmatrix} \quad (\text{AIV.35})$$

que conduce finalmente a la ecuación reducida de la elipse en forma matricial, dada por

$$\begin{bmatrix} 1 & x_1 & x_2 \end{bmatrix} * \begin{bmatrix} -1 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{6} \end{bmatrix} * \begin{bmatrix} 1 \\ x_1 \\ x_2 \end{bmatrix} = 0 \quad (\text{AIV.36})$$

Vamos a ver qué transformaciones son necesarias hacer a la matriz canónica (AIV.35) para llegar a obtener aquella otra que representa a nuestra cónica en estudio y que viene dado por (AIV.29). Ya adelantábamos que dichas transformaciones iban a traducirse en una rotación y una traslación respecto del sistema de referencia por lo cual, vamos a considerar las matrices que permiten realizar este tipo de transformaciones.

Para rotar una matriz basta aplicar la matriz de una rotación
$$M = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos \alpha & \text{sen} \alpha \\ 0 & -\text{sen} \alpha & \cos \alpha \end{bmatrix} \quad (\text{AIV.37})$$

donde α determina el número de grados que debe ser rotado el sistema de referencia.

Para expresar $\begin{bmatrix} 1 & x_1 & x_2 \end{bmatrix}$ respecto de una rotación del sistema de referencia donde x'_1 y x'_2 definen las coordenadas en ese nuevo sistema, haremos

$$\begin{bmatrix} 1 & x_1 & x_2 \end{bmatrix} = \begin{bmatrix} 1 & x'_1 & x'_2 \end{bmatrix} * M \Rightarrow \begin{bmatrix} 1 & x'_1 & x'_2 \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_2 \end{bmatrix} * M^{-1} \quad (\text{AIV.38})$$

siendo
$$M^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos \alpha & -\text{sen} \alpha \\ 0 & \text{sen} \alpha & \cos \alpha \end{bmatrix} \quad (\text{AIV.39})$$

La ecuación de la cónica reducida respecto del nuevo sistema de referencia, será:

$$\begin{bmatrix} 1 & x'_1 & x'_2 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos \alpha & -\text{sen} \alpha \\ 0 & \text{sen} \alpha & \cos \alpha \end{bmatrix} \begin{bmatrix} -1 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{6} \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos \alpha & \text{sen} \alpha \\ 0 & -\text{sen} \alpha & \cos \alpha \end{bmatrix} \begin{bmatrix} 1 \\ x'_1 \\ x'_2 \end{bmatrix} = 0 \quad (\text{AIV.40})$$

siendo la matriz de esta cónica:

$$M * A'' * M^T = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos \alpha & -\operatorname{sen} \alpha \\ 0 & \operatorname{sen} \alpha & \cos \alpha \end{bmatrix} * \begin{bmatrix} -1 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{6} \end{bmatrix} * \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos \alpha & \operatorname{sen} \alpha \\ 0 & -\operatorname{sen} \alpha & \cos \alpha \end{bmatrix} =$$

$$\begin{bmatrix} -1 & 0 & 0 \\ 0 & \frac{1}{2} * \cos^2 \alpha + \frac{1}{6} * \operatorname{sen}^2 \alpha & \frac{1}{2} * \operatorname{sen} \alpha * \cos \alpha - \frac{1}{6} * \cos \alpha * \operatorname{sen} \alpha \\ 0 & \frac{1}{2} * \operatorname{sen} \alpha * \cos \alpha - \frac{1}{6} * \cos \alpha * \operatorname{sen} \alpha & \frac{1}{2} * \operatorname{sen}^2 \alpha + \frac{1}{6} * \cos^2 \alpha \end{bmatrix}$$

(AIV.41)

Dado que la matriz A que perseguimos tiene la forma $A = \begin{bmatrix} 3 & 1 & 1 \\ 1 & \frac{1}{3} & \frac{1}{6} \\ 1 & \frac{1}{6} & \frac{1}{3} \end{bmatrix}$

podemos igualar las 2 submatrices 2x2 situadas en la parte inferior derecha.

En principio, no podemos plantear la igualdad entre las coordenadas a_{01} y a_{02} pero esto no es más que una muestra de la necesidad de hacer posteriormente una translación del sistema de referencia. Así, igualando ambas submatrices resulta que

$$\frac{1}{2} * \cos \alpha * \operatorname{sen} \alpha - \frac{1}{6} * \cos \alpha * \operatorname{sen} \alpha = \frac{1}{6} \Rightarrow \frac{1}{6} \operatorname{sen}(2\alpha) = \frac{1}{6} \Rightarrow \operatorname{sen}(2\alpha) = 1$$

y por tanto, $2\alpha = \operatorname{arcsen}(1) \Rightarrow 2\alpha = 90^\circ \Rightarrow \alpha = 45^\circ$

En consecuencia, para hallar la matriz de la elipse rotada basta aplicar a la matriz de la base canónica A'' , la matriz de una rotación de 45° , como resultado del cual se obtiene:

$$A'_{\text{ROTADA}} = M * A'' * M^T = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{\sqrt{2}}{2} & \frac{-\sqrt{2}}{2} \\ 0 & \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{bmatrix} * \begin{bmatrix} -1 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{6} \end{bmatrix} * \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ 0 & \frac{-\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{bmatrix} = \begin{bmatrix} -1 & 0 & 0 \\ 0 & \frac{1}{3} & \frac{1}{6} \\ 0 & \frac{1}{6} & \frac{1}{3} \end{bmatrix}$$

(AIV.42)

La ecuación de la cónica respecto del nuevo sistema de referencia será

$$\begin{bmatrix} 1 & x'_1 & x'_2 \end{bmatrix} * \begin{bmatrix} -1 & 0 & 0 \\ 0 & \frac{1}{3} & \frac{1}{6} \\ 0 & \frac{1}{6} & \frac{1}{3} \end{bmatrix} * \begin{bmatrix} 1 \\ x'_1 \\ x'_2 \end{bmatrix} = 0 \quad (\text{AIV.43})$$

Finalmente para expresar ahora $\begin{bmatrix} 1 & x'_1 & x'_2 \end{bmatrix}$ respecto de una translación del sistema de referencia basta aplicar una matriz de translación, la cual responde a una estructura del tipo

$$C = \begin{bmatrix} 1 & C_1 & C_2 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (\text{AIV.44}).$$

donde C_1 y C_2 marcan las coordenadas respecto del nuevo sistema de referencia, es decir, las coordenadas de la translación.

De esta forma, si x''_1 y x''_2 definen las coordenadas respecto del nuevo sistema,

$$\begin{bmatrix} 1 & x'_1 & x'_2 \end{bmatrix} = \begin{bmatrix} 1 & x''_1 & x''_2 \end{bmatrix} * C \Rightarrow \begin{bmatrix} 1 & x''_1 & x''_2 \end{bmatrix} = \begin{bmatrix} 1 & x'_1 & x'_2 \end{bmatrix} * C^{-1} \quad (\text{AIV.45})$$

$$\text{siendo } C^{-1} = \begin{bmatrix} 1 & -C_1 & -C_2 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (\text{AIV.46})$$

La ecuación de la cónica respecto del nuevo sistema de referencia, será:

$$\begin{bmatrix} 1 & x''_1 & x''_2 \end{bmatrix} * \begin{bmatrix} 1 & -C_1 & -C_2 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} * \begin{bmatrix} -1 & 0 & 0 \\ 0 & \frac{1}{3} & \frac{1}{6} \\ 0 & \frac{1}{6} & \frac{1}{3} \end{bmatrix} * \begin{bmatrix} 1 & 0 & 0 \\ -C_1 & 1 & 0 \\ -C_2 & 0 & 1 \end{bmatrix} * \begin{bmatrix} 1 \\ x''_1 \\ x''_2 \end{bmatrix} = 0 \quad (\text{AIV.47})$$

La matriz de esta cónica sería la que estamos buscando. Por tanto, basta igualar el producto de las tres matrices centrales de la ecuación anterior a la matriz asociada a la elipse que buscamos (AIV.29) y de esta forma se podrán determinar los coeficientes de la translación.

$$\begin{bmatrix} 1 & -C_1 & -C_2 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} * \begin{bmatrix} -1 & 0 & 0 \\ 0 & \frac{1}{3} & \frac{1}{6} \\ 0 & \frac{1}{6} & \frac{1}{3} \end{bmatrix} * \begin{bmatrix} 1 & 0 & 0 \\ -C_1 & 1 & 0 \\ -C_2 & 0 & 1 \end{bmatrix} =$$

$$\begin{bmatrix} -1 + \frac{1}{3} * C_1^2 + \frac{1}{3} * C_1 * C_2 + \frac{1}{3} * C_2^2 & \frac{-1}{3} * C_1 - \frac{1}{6} * C_2 & \frac{-1}{6} * C_1 - \frac{1}{3} * C_2 \\ \frac{-1}{3} * C_1 - \frac{1}{6} * C_2 & \frac{1}{3} & \frac{1}{6} \\ \frac{-1}{6} * C_1 - \frac{1}{3} * C_2 & \frac{1}{6} & \frac{1}{3} \end{bmatrix} = \begin{bmatrix} 3 & 1 & 1 \\ 1 & \frac{1}{3} & \frac{1}{6} \\ 1 & \frac{1}{6} & \frac{1}{3} \end{bmatrix} \Rightarrow$$

$$\frac{-1}{3} * C_1 - \frac{1}{6} * C_2 = 1 \text{ y } \frac{-1}{6} * C_1 - \frac{1}{3} * C_2 = 1 \Rightarrow C_1 = -2 \text{ y } C_2 = -2$$

Por tanto, la región P32 hace referencia a una elipse que ha sido rotada 45°C y trasladada del origen (0,0) al punto (-2,-2)

Recopilando el conjunto de posibles condiciones (R0,P0,P1,P2 y P3) (dadas por (AIV.15) (AIV.17) (AIV.18) (AIV.21) y (AIV.24)) que deben cumplir los pares (K_{ij}, L_{ij}) , podemos hacer el planteamiento analítico de la región factible asociada a estos pares para que $C_{FERGUSON}$ sea 2-creciente en el rectángulo $[u_i, u_{i+1}] \times [v_j, v_{j+1}]$. Dicho planteamiento se puede expresar mediante restricciones conjuntivas enlazadas por variables binarias:

$$\begin{aligned}
K_{ij} + 2 * L_{ij} + 3 &\leq 3 * (1 - \delta_1) \\
-2 * K_{ij} - L_{ij} - 3 &\leq \left(\frac{24 + 15 * \sqrt{2}}{7}\right) * (1 - \delta_1) \\
2 * K_{ij} + L_{ij} + 3 &\leq 3 * (1 - \delta_2) \\
-K_{ij} - 2 * L_{ij} - 3 &\leq \left(\frac{24 + 15 * \sqrt{2}}{7}\right) * (1 - \delta_2) \\
-K_{ij} - L_{ij} - 2 &\leq 4 * (1 - \delta_3) \\
(K_{ij} + L_{ij} + 3)^2 - K_{ij} * L_{ij} &\leq 9 * (1 - \delta_4) \\
\delta_1 + \delta_2 + \delta_3 + \delta_4 &= 1 \\
K_{ij} &\leq 0 \\
L_{ij} &\leq 0 \\
\delta_1, \delta_2, \delta_3, \delta_4 &\in \{0, 1\}
\end{aligned}$$

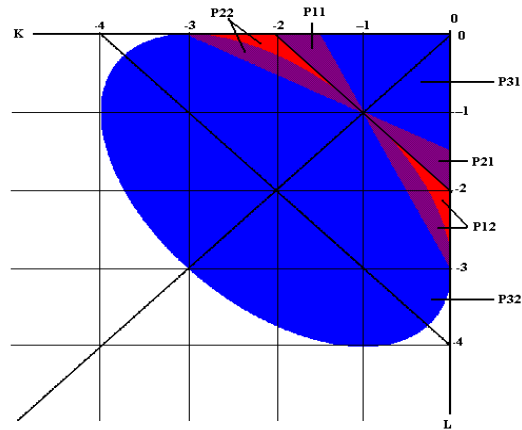


Ilustración AIV.2: Región de factibilidad del problema que permite construir una cópula por interpolación cúbica

En el gráfico anterior se pueden distinguir las zonas asociadas a cada una de las condiciones que hemos ido definiendo. Para empezar podemos ver que la región está definida sobre el tercer cuadrante del plano (condición P0). Además, en color rojo aparecen las regiones definidas por las restricciones lineales P1 y P2 y, en color azul, la región definida por la restricción no lineal P3 a la que pertenece la elipse centrada en el punto (-2,-2) y rotada 45°. En color morado se muestra la zona de intersección entre ambas regiones.

Para obtener cada una de las ecuaciones lo que se hace es asociar una variable binaria δ a cada una de las 4 regiones conexas que se van a definir. Éstas, no disjuntas, son las siguientes:

- Región 1: Determinada por las condiciones P11 y P22: $3 + 2 * K_{ij} + L_{ij} \leq 0$ y $2 + K_{ij} + L_{ij} \geq 0$ y $2 + K_{ij} + L_{ij} \leq 0$ y $K_{ij} + 2 * L_{ij} + 3 \geq 0$
- Región 2: Determinada por las condiciones P12 y P21: $3 + 2 * K_{ij} + L_{ij} \geq 0$ y $2 + K_{ij} + L_{ij} \leq 0$ y $2 + K_{ij} + L_{ij} \geq 0$ y $K_{ij} + 2 * L_{ij} + 3 \leq 0$
- Región 3: Determinada por las condición P31: $2 + K_{ij} + L_{ij} \geq 0$
- Región 4: Determinada por las condición P32: $(K_{ij} + L_{ij} + 3)^2 - K_{ij} * L_{ij} \leq 0$

La idea de asociar una variable binaria a cada región es que, como sólo una de ellas δ_i puede valer 1, pues $\delta_1 + \delta_2 + \delta_3 + \delta_4 = 1$ con $\delta_1, \delta_2, \delta_3, \delta_4 \in \{0, 1\}$, la posible solución factible esté localizada en una sola de las cuatro. Habrá que asegurar sin embargo que las restricciones asociadas a las restantes regiones cuando sus respectivos δ_j (con j distinto de i) valgan 0, no impongan restricciones adicionales sobre la región asociada a δ_i . Para ello, es necesario que

en dichos casos, las restricciones definen regiones que, al menos, contengan a la región seleccionada. Para hacer esto, se debe calcular el vector gradiente asociado a una restricción y mover ésta en dicha dirección para ver cuál es el último punto de la región factible al que corta.

Así por ejemplo, para la región 1 definida por las ecuaciones $3+2*K_{ij}+L_{ij}\leq 0$ y $2+K_{ij}+L_{ij}\geq 0$ y $2+K_{ij}+L_{ij}\leq 0$ y $K_{ij}+2*L_{ij}+3\geq 0$, o equivalentemente por $3+2*K_{ij}+L_{ij}\leq 0$ y $-K_{ij}-2*L_{ij}-3\leq 0$, los vectores gradientes asociados a las ecuaciones $f_1\equiv 3+2*K_{ij}+L_{ij}=0$ y $f_2\equiv -K_{ij}-2*L_{ij}-3=0$ son (2,1) y (-1,-2) respectivamente. Si movemos cada ecuación en la dirección de su vector gradiente, resulta que los últimos puntos en los que van a cortar a la región factible (dada por la unión de las 4 regiones) van a ser $C1=(0,0)$ y

$$C2=\left(\frac{-18-6*\sqrt{2}}{14}, \frac{-18-6*\sqrt{2}}{7}\right). \text{ Evaluando las ecuaciones en dichos puntos se obtienen las imágenes}$$

$$f_1(C1)=3 \text{ y } f_2(C2)=\frac{24+15*\sqrt{2}}{7} \text{ que van a servir de cotas a las restricciones.}$$

Procediendo de la misma manera para las variables P 's y Q 's que se obtienen a partir de las derivadas primeras respecto de V , y teniendo en cuenta la definición de estas variables dada por (AIV.3), (AIV.4), (AIV.6) y (AIV.7), la región factible puede plantearse en función de los vectores tangentes, pasando a ser éstos las variables de decisión. Si asociamos a dichas restricciones una función objetivo, lo que tendríamos sería un problema de programación no lineal (debido a la restricción no lineal P32) y además mixto, pues hay variables de decisión continuas (dadas por las derivadas de primer orden) y enteras (dadas por los δ 's asociados a cada región). La función objetivo podría consistir en maximizar o minimizar las derivadas, lo cual se traduciría en que los vectores tangentes a la superficie interpoladora en cada vértice de la partición, tuvieran mayor o menor pendiente y, en consecuencia, la superficie fuera más o menos suave.

Obsérvese que, si bien, se construye una región de factibilidad para cada par (K_{ij}, L_{ij}) (y análogamente para cada par (P_{ij}, Q_{ij}) que representan derivadas de primer orden respecto de V), lo importante es que los pares (K_{ij}, L_{ij}) de dos rectángulos consecutivos del plano, van a compartir derivadas de primer orden respecto de U .

Una región factible relaciona variables L_{ij} y K_{ij} . El valor de la variable L_{ij} de un rectángulo depende de la variable de

$$\text{decisión } \frac{\partial C'(u_{i+1}, v_j)}{\partial u} \text{ y depende también de } K_{ij} = \frac{2*(u_{i+1}-u_i)}{V_C'(B)} * \left[\frac{\partial C'(u_i, v_j)}{\partial u} - \frac{\partial C'(u_i, v_{j+1})}{\partial u} \right]$$

El rectángulo situado a su derecha, también llevará asociado a esa misma variable $\frac{\partial C'(u_{i+1}, v_j)}{\partial u}$, pero en esta ocasión a través del valor de la variable K_{ij} y a su vez dependerá también de otro L_{ij} distinto

$$L_{ij} = \frac{2*(u_{i+1}-u_i)}{V_C'(B)} * \left[\frac{\partial C'(u_{i+2}, v_j)}{\partial u} - \frac{\partial C'(u_{i+2}, v_{j+1})}{\partial u} \right]$$

Es decir, cada variable de decisión K_{ij} (y del mismo modo, cada L_{ij}, P_{ij} o Q_{ij}) recoge las derivadas correspondientes a dos vértices consecutivos de un rectángulo B que surge de realizar la partición del cuadrado unidad. Dado que un mismo vértice puede pertenecer a cuatro rectángulos distintos, existirán diferentes regiones factibles que se entrelazan unas con otras y por ello, las variables de decisión no se estiman de manera independiente para una región concreta, sino de

forma conjunta. Es precisamente esta combinación de regiones factibles la que integra el problema de programación matemática (4.21).

De acuerdo a la demostración constructiva que hemos desarrollado, la hipótesis de que las variables

$$\frac{\partial C_{FERGUSON}(u_i, v_j)}{\partial u} \text{ y } \frac{\partial C_{FERGUSON}(u_i, v_j)}{\partial v} \quad \forall i \in \{0, 1, 2, \dots, n\} \quad \forall j \in \{0, 1, 2, \dots, m\} \text{ , proporcionen junto con una}$$

relación de δ s y γ s una solución del problema de programación no lineal entera (4.21), implica directamente que el volumen de un rectángulo arbitrario totalmente contenido en otro de los que conforman la partición del cuadrado unidad tenga un volumen mayor o igual que cero a través de $C_{FERGUSON}$.

Así, el volumen de un rectángulo cualquiera trazado sobre el cuadrado unidad será también mayor o igual que cero pues se podrá expresar como suma de los volúmenes de rectángulos parcial o totalmente contenidos en los que definen la partición. Con ello quedaría probado el cumplimiento de la tercera de las propiedades de las cópulas (“ser 2-creciente”) y consecuentemente el lema.

c.q.d

ANEXO V: Demostración del lema 4.5

“Sea una partición del cuadrado unidad en $n \times m$ rectángulos del mismo área y sea la solución óptima al problema de programación fraccional (4.7) dada por $c_{ij}^* = C^* \left(\frac{i}{n}, \frac{j}{m} \right) \forall i \in \{0, 1, 2, \dots, n\} \quad \forall j \in \{0, 1, 2, \dots, m\}$ que define una subcópula C^* de dominio los puntos de intersección entre los rectángulos que originan la partición.

Sea $C_{HERMITE}(u, v)$ la función que se obtiene por interpolación cúbica de una subcópula C^* mediante polinomios de Hermite. Supongamos que las derivadas de segundo orden en el dominio de definición de la subcópula valen cero, es decir,

decir, $\frac{\partial^2 C_{HERMITE}(u_i, v_j)}{\partial u \partial v} = 0 \quad \forall i \in \{0, 1, 2, \dots, n\} \quad \forall j \in \{0, 1, 2, \dots, m\}$, en cuyo caso, $C_{HERMITE}(u, v)$ define una superficie interpoladora de Ferguson $C_{FERGUSON}(u, v)$.

Si $\frac{\partial C_{FERGUSON}(u_i, v_j)}{\partial u}$ y $\frac{\partial C_{FERGUSON}(u_i, v_j)}{\partial v} \quad \forall i \in \{0, 1, 2, \dots, n\} \quad \forall j \in \{0, 1, 2, \dots, m\}$

proporcionan una solución del problema de programación matemática (4.26), entonces la función $C_{FERGUSON}(u, v)$ cumple la tercera de las condiciones para ser cópula:

$$V_{C_{FERGUSON}}([u_1, u_2] \times [v_1, v_2]) = C_{FERGUSON}(u_2, v_2) - C_{FERGUSON}(u_2, v_1) - C_{FERGUSON}(u_1, v_2) + C_{FERGUSON}(u_1, v_1) \geq 0 \quad ,$$

$$\forall u_1, u_2, v_1, v_2 \text{ t.q. } u_1 \leq u_2, v_1 \leq v_2$$

demostración

Vamos a ver que dada una solución factible del problema (4.26) entonces también lo es del problema (4.21). En consecuencia, se podrá aplicar el lema (4.4) que conduce al cumplimiento de la tercera de las propiedades de las funciones cópula.

Sea $(K, L) \in [-3, 0] \times [-3, 0]$ y sean las 4 regiones factibles dadas por las condiciones R0, P0, P1, P2 y P3 que definen la región del problema (4.21) (véase ilustración (AIV.2) en el Anexo IV):

$$RF_0 = \{(K, L) \text{ t.q. } K \leq 0, L \leq 0, 2 + K + L = 0\} = \{(K, L) \text{ t.q. } (K, L) \text{ cumple } R0\}$$

$$RF_1 = \{(K, L) \text{ t.q. } K \leq 0, L \leq 0, \frac{3 + 2 * K + L}{3 * (2 + K + L)} < 0\} = \{(K, L) \text{ t.q. } (K, L) \text{ cumple } P0 \text{ y } P1\}$$

$$RF_2 = \{(K, L) \text{ t.q. } K \leq 0, L \leq 0, \frac{3 + 2 * K + L}{3 * (2 + K + L)} > 1\} = \{(K, L) \text{ t.q. } (K, L) \text{ cumple } P0 \text{ y } P2\}$$

$$RF_3 = \{(K, L) \text{ t.q. } K \leq 0, L \leq 0, \frac{-(3 + 2 * K + L)^2}{3 * (2 + K + L)} + K \leq 0\} = \{(K, L) \text{ t.q. } (K, L) \text{ cumple } P0 \text{ y } P3\}$$

Para ver la inclusión de (4.26) en (4.21), basta ver que (K, L) se encuentra en alguna de las 4 regiones. La demostración para el par (P, Q) resultaría análoga.

Como siempre se implicará que $K, L \leq 0$ (véase la propiedad P0 del Anexo IV), basta ver que se cumple alguna de las propiedades R0, P1, P2 o P3. Vamos a plantear 3 situaciones disjuntas para los valores que pueden tomar (K, L) en

función de su relación con la recta $K+L=-2$:

- Si $K+L=-2$, como $-2 \leq K \leq 0$ y $-2 \leq L \leq 0 \Rightarrow (K, L) \in RF_0$
- Si $K+L > -2 \Rightarrow -\frac{(3+2*K+L)^2}{3*(2+K+L)} < 0 \Rightarrow -\frac{(3+2*K+L)^2}{3*(2+K+L)} + K \leq 0 \Rightarrow (K, L) \in RF_3$
- Si $K+L < -2$, entonces vamos a distinguir otras tres situaciones disjunta respecto de la relación del par (K,L) con la región dada por $R = \{(K, L) | K+2*L+3 \geq 0 \text{ o } 2*K+L+3 \geq 0\}$

$$- \text{ Si } K+2*L+3 \geq 0 \Rightarrow 3*K+3*L+6 \geq 2*K+L+3 \Rightarrow 3*(K+L+2) \geq 2*K+L+3 \Rightarrow$$

$$3*(K+L+2) \geq 2*K+L+3 \Rightarrow \frac{(2*K+L+3)}{3*(K+L+2)} \geq 1 \Rightarrow (K, L) \in RF_2$$

$$- \text{ Si } 2*K+L+3 \geq 0 \Rightarrow \frac{(2*K+L+3)}{(K+L+2)} < 0 \Rightarrow (K, L) \in RF_1$$

- Si $(K, L) \notin R \Leftrightarrow [K+2*L+3 \leq 0 \text{ y } 2*K+L+3 \leq 0]$, condiciones que conducen a:

$$K+2*L+3 \leq 0 \Rightarrow K+L+3 \leq -L \quad (\text{AV.1})$$

$$2*K+L+3 \leq 0 \Rightarrow K+L+3 < -K \quad (\text{AV.2})$$

Además, por hipótesis, como $(K, L) \in [-3, 0] \times [-3, 0]$ se tiene que,

$$-3 \leq K \Rightarrow -3+L \leq K+L \Rightarrow L \leq K+L+3 \quad (\text{AV.3})$$

$$-3 \leq L \Rightarrow -3+K \leq K+L \Rightarrow K \leq K+L+3 \quad (\text{AV.4})$$

Como resultado de (AV.1)-(AV.4) resulta que

$$L \leq K+L+3 \leq -L \quad (\text{AV.5})$$

$$K \leq K+L+3 \leq -K \quad (\text{AV.6})$$

y, en consecuencia,

$$0 \leq |K+L+3| \leq K \quad (\text{AV.7})$$

$$0 \leq |K+L+3| \leq L \quad (\text{AV.8})$$

Por tanto,

$$0 \leq |K+L+3|^2 \leq K*L \Rightarrow (K+L+3)^2 \leq K*L \Rightarrow K^2+L^2+2*K*L+9+6*K+6*L \leq K*L \Rightarrow$$

$$K^2+L^2+K*L+9+6*K+6*L \leq 0 \quad (\text{AV.9})$$

Obsérvese ahora que la ecuación P3 se puede desarrollar y escribir como:

$$\frac{-(3+2*K+L)^2}{3*(2+K+L)} + K = \frac{-9-K^2-6*K-L^2-6*L-K*L}{3*(2+K+L)} \quad (\text{AV.10})$$

Como $2+K+L < 0$, el cociente anterior será menor o igual que 0
 $\Leftrightarrow -9-K^2-6*K-L^2-6*L-K*L > 0 \Leftrightarrow 9+K^2+6*K+L^2+6*L+K*L \leq 0$. Pero esta condición se cumple según acabamos de ver (AV.9). Luego la propiedad P3 se cumple y, por tanto, $(K, L) \in RF_3$

Por tanto los rectángulos del tipo $[-3,0] \times [-3,0]$ que definen las regiones factibles del problema (4.26) se encuentran contenidas en las regiones factibles dadas por P0, P1, P2 y P3 que caracterizan al problema (4.21).

c.q.d

ANEXO VI: Demostración del lema 4.6

“Sea una partición del cuadrado unidad en $n \times m$ rectángulos del mismo área y sea la solución óptima al problema de programación fraccional (4.7) dada por $c_{ij}^* = C^* \left(\frac{i}{n}, \frac{j}{m} \right) \forall i \in \{0, 1, 2, \dots, n\} \forall j \in \{0, 1, 2, \dots, m\}$ que define una subcópula C^* de dominio los puntos de intersección entre los rectángulos que originan la partición.

Sea $C_{HERMITE}(u, v)$ la función que se obtiene por interpolación cúbica de una subcópula C^* mediante polinomios de Hermite.

Si $\frac{\partial C_{HERMITE}(u_i, v_j)}{\partial u}$, $\frac{\partial C_{HERMITE}(u_i, v_j)}{\partial v}$ y $\frac{\partial^2 C_{HERMITE}(u_i, v_j)}{\partial u \partial v} \forall i \in \{0, 1, \dots, n\} \forall j \in \{0, 1, \dots, m\}$ proporcionan

junto con una relación de δ 's y γ 's (variables binarias) una solución al problema de programación multiobjetivo no lineal entera (4.35), entonces la función $C_{HERMITE}(u, v)$ cumple la tercera de las condiciones de las funciones cópula:

$$V_{C_{HERMITE}}([u_1, u_2] \times [v_1, v_2]) = C_{HERMITE}(u_2, v_2) - C_{HERMITE}(u_2, v_1) - C_{HERMITE}(u_1, v_2) + C_{HERMITE}(u_1, v_1) \geq 0 \quad ,, \\ \forall u_1, u_2, v_1, v_2 \text{ t.q. } u_1 \leq u_2, v_1 \leq v_2$$

demostración

De igual manera que se plantea en el Anexo IV, vamos a ver que las condiciones que deben satisfacerse para que el volumen de un rectángulo arbitrario a través de $C_{HERMITE}$ sea mayor o igual que cero se traducen en la existencia de una solución al problema (4.35) dada por una relación de valores para las variables

$$\frac{\partial C_{HERMITE}(u_i, v_j)}{\partial u}, \frac{\partial C_{HERMITE}(u_i, v_j)}{\partial v} \text{ y } \frac{\partial^2 C_{HERMITE}(u_i, v_j)}{\partial u \partial v} \forall i \in \{0, 1, \dots, n\} \forall j \in \{0, 1, \dots, m\} \text{ y una asignación}$$

de valores binarios asociados a las variables δ 's y γ 's.

Nuevamente, basta establecer estas condiciones respecto de un rectángulo cualquiera B totalmente contenido en uno de los $m \times n$ que surgen como consecuencia de la partición del cuadrado unidad puesto que el volumen de otro rectángulo genérico R se podría expresar siempre como suma de los volúmenes de rectángulos de tipo B .

Empezaremos aplicando la ecuación (4.19), escrita en función de las variables de decisión K_{ij} , L_{ij} , P_{ij} y Q_{ij} , a un rectángulo $B = [u_i^l, u_i^r] \times [v_j^l, v_j^r]$ totalmente contenido en uno de los rectángulos que conforman la rejilla, por ejemplo $B = [u_i, u_{i+1}] \times [v_j, v_{j+1}]$. Por comodidad, continuaremos imponiendo que las variables de decisión K_{ij} , L_{ij} , P_{ij} y Q_{ij} , se muevan en el rectángulo $[-3, 0] \times [-3, 0]$ (problema (4.26)) si bien no hemos de pasar por alto que estas variables tienen mayor margen de maniobra y pueden moverse dentro de la región no lineal presentada en la ilustración (AIV.2). Además, las variables de decisión asociadas a las derivadas segundas deberán ser mayores o iguales que cero dado que representan valores de la densidad cópula y ésta, como tal función de densidad, debe ser positiva.

De esta forma, el volumen del rectángulo B a través de (4.19) viene dado por:

$$\begin{aligned}
V_{C_{\text{HERMITE}}}(B) = & \\
\frac{V_C'(B)}{2} * [f(y_2) - f(y_1)] * [f(x_2) - f(x_1)] + K_{ij} * [g(x_2) - g(x_1)] + L_{ij} * [h(x_2) - h(x_1)] + & \\
\frac{V_C'(B)}{2} * [f(x_2) - f(x_1)] * [f(y_2) - f(y_1)] + P_{ij} * [g(y_2) - g(y_1)] + Q_{ij} * [h(y_2) - h(y_1)] + & \\
\frac{\partial^2 C'(u_i, v_j)}{\partial u \partial v} * [g(x_2) - g(x_1)] * [g(y_2) - g(y_1)] + & \\
\frac{\partial^2 C'(u_{i+1}, v_j)}{\partial u \partial v} * [h(x_2) - h(x_1)] * [g(y_2) - g(y_1)] + & \\
\frac{\partial^2 C'(u_i, v_{j+1})}{\partial u \partial v} * [g(x_2) - g(x_1)] * [h(y_2) - h(y_1)] + & \\
\frac{\partial^2 C'(u_{i+1}, v_{j+1})}{\partial u \partial v} * [h(x_2) - h(x_1)] * [h(y_2) - h(y_1)] &
\end{aligned}$$

siendo

$$\begin{aligned}
K_{ij} &= \frac{2 * (u_{i+1} - u_i)}{V_C'(B)} * \left[\frac{\partial C'(u_i, v_j)}{\partial u} - \frac{\partial C'(u_i, v_{j+1})}{\partial u} \right] \\
L_{ij} &= \frac{2 * (u_{i+1} - u_i)}{V_C'(B)} * \left[\frac{\partial C'(u_{i+1}, v_j)}{\partial u} - \frac{\partial C'(u_{i+1}, v_j)}{\partial u} \right] \\
P_{ij} &= \frac{2 * (v_{j+1} - v_j)}{V_C'(B)} * \left[\frac{\partial C'(u_i, v_j)}{\partial v} - \frac{\partial C'(u_{i+1}, v_j)}{\partial v} \right] \\
Q_{ij} &= \frac{2 * (v_{j+1} - v_j)}{V_C'(B)} * \left[\frac{\partial C'(u_i, v_{j+1})}{\partial v} - \frac{\partial C'(u_{i+1}, v_{j+1})}{\partial v} \right] \\
A_{ij} &= \frac{2 * (u_{i+1} - u_i) * (v_{j+1} - v_j)}{V_C'(B)} * \left[\frac{\partial^2 C'(u_i, v_j)}{\partial u \partial v} \right] \\
B_{i+1j} &= \frac{2 * (u_{i+1} - u_i) * (v_{j+1} - v_j)}{V_C'(B)} * \left[\frac{\partial^2 C'(u_i, v_{j+1})}{\partial u \partial v} \right] \\
C_{i+1j} &= \frac{2 * (u_{i+1} - u_i) * (v_{j+1} - v_j)}{V_C'(B)} * \left[\frac{\partial^2 C'(u_{i+1}, v_j)}{\partial u \partial v} \right] \\
D_{i+1j+1} &= \frac{2 * (u_{i+1} - u_i) * (v_{j+1} - v_j)}{V_C'(B)} * \left[\frac{\partial^2 C'(u_{i+1}, v_{j+1})}{\partial u \partial v} \right]
\end{aligned}$$

$$f(x) = 2 * x^3 - 3 * x^2; \quad g(x) = x^3 - 2 * x^2 + x; \quad h(x) = x^3 - x^2$$

$$x_k = \frac{u_i^k - u_i}{u_{i+1} - u_i}; \quad y_k = \frac{v_j^k - v_j}{v_{j+1} - v_j} \quad k \in \{1, 2\}$$

(AVI.1)

Vamos a tratar de presentar esta expresión de una forma más familiarizada con las que hemos utilizado en desarrollos anteriores (véase Anexo IV), para así poder aprovechar algunos de los resultados a los que hemos llegado para la superficie interpoladora de Ferguson (Anexo IV). Las funciones “ f ”, “ g ” y “ h ” presentadas en la expresión anterior no se corresponden con las definidas en el Anexo IV. Sí es cierto que la función que aquí definimos como “ f ” es la que en dicho anexo hemos definido como “ h ”. De cualquier modo, podemos proceder de dos formas paralelas para obtener una ecuación con ciertas similitudes a (AIV.9).

Expresión 1 para la búsqueda de condiciones que permitan que $C_{HERMITE}$ sea una función 2-creciente

$$\begin{aligned}
 V_{C_{HERMITE}}(B) = & \\
 & \frac{V_c'(B)}{2} * [f(x_2) - f(x_1)] * [(f(y_2) - f(y_1))] + \frac{V_c'(B)}{2} * [f(x_2) - f(x_1)] * [(f(y_2) - f(y_1))] + \\
 & \frac{V_c'(B)}{2} * [f(y_2) - f(y_1)] * [g(x_2) - g(x_1)] * K_{ij} + \frac{V_c'(B)}{2} * [g(y_2) - g(y_1)] * [g(x_2) - g(x_1)] * A_{ij} + \\
 & \frac{V_c'(B)}{2} * [f(y_2) - f(y_1)] * [h(x_2) - h(x_1)] * L_{ij} + \frac{V_c'(B)}{2} * [h(y_2) - h(y_1)] * [h(x_2) - h(x_1)] * D_{i+1, j+1} + \\
 & \frac{V_c'(B)}{2} * [f(x_2) - f(x_1)] * [g(y_2) - g(y_1)] * P_{ij} + \frac{V_c'(B)}{2} * [h(x_2) - h(x_1)] * [g(y_2) - g(y_1)] * C_{i+1, j} + \\
 & \frac{V_c'(B)}{2} * [f(x_2) - f(x_1)] * [h(y_2) - h(y_1)] * Q_{ij} + \frac{V_c'(B)}{2} * [g(x_2) - g(x_1)] * [h(y_2) - h(y_1)] * B_{i, j+1} = \\
 & \frac{V_c'(B)}{2} * [f(x_2) - f(x_1)] * [(f(y_2) - f(y_1)) + P_{ij} * (g(y_2) - g(y_1)) + Q_{ij} * (h(y_2) - h(y_1))] + \\
 & \frac{V_c'(B)}{2} * [g(x_2) - g(x_1)] * [K_{ij} * (f(y_2) - f(y_1)) + A_{ij} * (g(y_2) - g(y_1)) + B_{i, j+1} * (h(y_2) - h(y_1))] + \\
 & \frac{V_c'(B)}{2} * [h(x_2) - h(x_1)] * [L_{ij} * (f(y_2) - f(y_1)) + C_{i+1, j} * (g(y_2) - g(y_1)) + D_{i+1, j+1} * (h(y_2) - h(y_1))] + \\
 & \frac{V_c'(B)}{2} * [f(x_2) - f(x_1)] * [(f(y_2) - f(y_1))]
 \end{aligned}$$

(AVI.2)

Expresión 2 para la búsqueda de condiciones que permitan que $C_{HERMITE}$ sea una función 2-creciente

$$\begin{aligned}
 V_{C_{HERMITE}}(B) = & \\
 & \frac{V_c'(B)}{2} * [f(y_2) - f(y_1)] * [(f(x_2) - f(x_1)) + K_{ij} * (g(x_2) - g(x_1)) + L_{ij} * (h(x_2) - h(x_1))] + \\
 & \frac{V_c'(B)}{2} * [g(y_2) - g(y_1)] * [P_{ij} * (f(x_2) - f(x_1)) + A_{ij} * (g(x_2) - g(x_1)) + C_{i+1, j} * (h(x_2) - h(x_1))] + \\
 & \frac{V_c'(B)}{2} * [h(y_2) - h(y_1)] * [Q_{ij} * (f(x_2) - f(x_1)) + B_{i, j+1} * (g(x_2) - g(x_1)) + D_{i+1, j+1} * (h(x_2) - h(x_1))] + \\
 & \frac{V_c'(B)}{2} * [f(y_2) - f(y_1)] * [(f(x_2) - f(x_1))]
 \end{aligned}$$

(AVI.3)

Cualquiera de estas dos expresiones serviría para el propósito que nos proponemos. Observemos que el primer sumando de cada uno de ellos se corresponde con aquellos que componen la ecuación (AIV.9) y que habíamos demostrado que éstos eran mayores o iguales que cero cuando K_{ij} , L_{ij} , P_{ij} y Q_{ij} satisfacían las restricciones del problema (4.21) (o (4.26)). Del mismo modo, el cuarto sumando es también mayor o igual que cero pues es el producto de la mitad del volumen del rectángulo B por dos cantidades negativas dado que, como vimos, la función “ f ” (equivalentemente la función “ h ” en (AIV.12)) es decreciente. En consecuencia, nuestro análisis se va a centrar en los sumandos intermedios de dichas expresiones (el segundo y el tercero). Intuitivamente bastaría con que ambos fueran mayores o iguales que cero o bien, con que la cantidad que restaran (si alguno de los dos fuera negativo, o ambos) no impidiese que el resultado final fuera positivo (y así, que la función fuese 2-creciente).

Tomemos por ejemplo el segundo sumando de la expresión (AVI.2) y preguntémosnos bajo qué condiciones dicho

sumando es mayor o igual que cero.

$$\frac{V_c'(B)}{2} * [g(x_2) - g(x_1)] * [K_{ij} * (f(y_2) - f(y_1)) + A_{ij} * (g(y_2) - g(y_1)) + B_{ij+1} * (h(y_2) - h(y_1))] \quad (\text{AVI.4})$$

Esta expresión consta de tres factores de los cuales, sólo tenemos garantía de que sea positivo el primero de ellos pues representa el volumen de un rectángulo a través de una subcúpula. El valor del segundo podrá ser positivo o negativo en función de los valores de x_1 y x_2 . Será sobre el tercero sobre el que podamos ejercer algún tipo de influencia para conseguir nuestros propósitos a través de los valores de A_{ij} y B_{ij+1} , que están directamente relacionados con las derivadas de segundo orden (véase (AVI.1)).

Analicemos entonces el tercer factor, $[K_{ij} * (f(y_2) - f(y_1)) + A_{ij} * (g(y_2) - g(y_1)) + B_{ij+1} * (h(y_2) - h(y_1))]$ (AVI.5)

De antemano, sabemos que K_{ij} se puede mover entre -3 y 0 (tiene mayor maniobrabilidad, pero nos vale ese margen para esta demostración), $f(y_2) - f(y_1)$ es negativo (por ser “ f ” decreciente, (AIV.12)) y A_{ij} y B_{ij+1} son mayores o iguales que cero pues su definición involucra sólo a términos positivos (entre ellos el valor de la densidad cópula en los nodos de la rejilla que resultan de realizar la partición del cuadrado unidad). Así, el primero de los sumandos, $K_{ij} * (f(y_2) - f(y_1))$, es mayor o igual que cero. Las diferencias dadas por las imágenes de las funciones “ g ” y “ h ” son las que no vamos a tener controladas dado que dichas funciones no son monótonas como pone de manifiesto el gráfico siguiente:

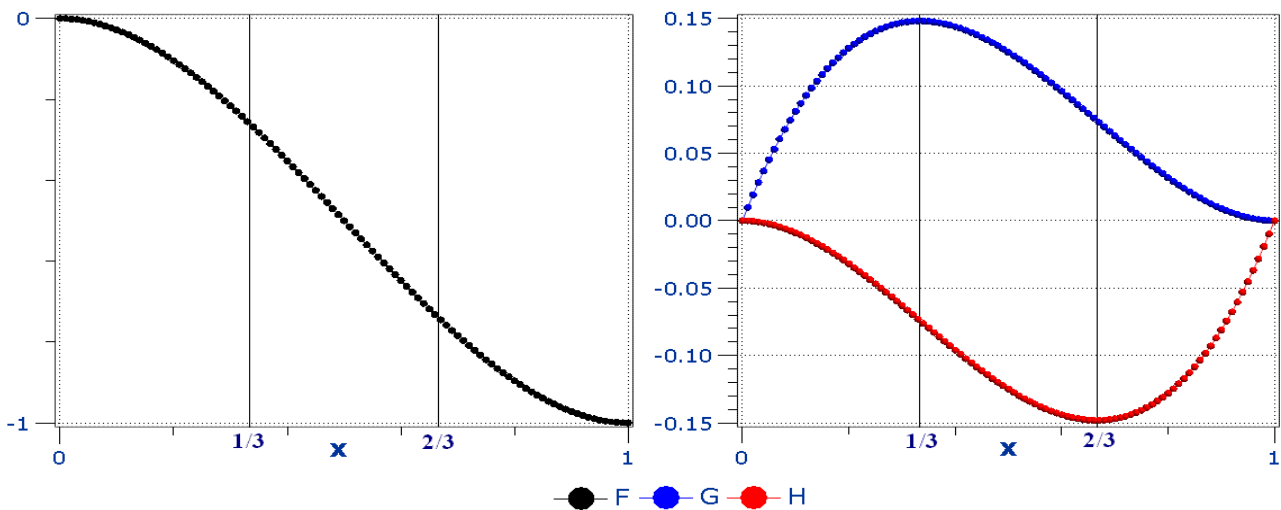


Ilustración AVI.1: Representación gráfica de las funciones f, g y h

En función de los valores de dichas funciones, podremos distinguir las siguientes situaciones:

- Obviamente, si $g(y_2) - g(y_1) \geq 0$ y $h(y_2) - h(y_1) \geq 0$ la expresión (AVI.5) es positiva pues sus tres sumandos lo son.

- Si $g(y_2)-g(y_1)<0$ y $h(y_2)-h(y_1)\geq 0$ los sumandos $K_{ij}*(f(y_2)-f(y_1))$ y $B_{ij+1}*(h(y_2)-h(y_1))$ son positivos y el sumando $A_{ij}*(g(y_2)-g(y_1))$ es negativo. Sin embargo, vamos a demostrar que $K_{ij}*(f(y_2)-f(y_1))+A_{ij}*(g(y_2)-g(y_1))$ es mayor o igual que cero. Para ello, vamos a

probar que $\frac{g(y_2)-g(y_1)}{f(y_2)-f(y_1)}<\frac{1}{3}$ (AVI.6).

Al ser $f(y_2)-f(y_1)<0$ y $g(y_2)-g(y_1)<0$, se tendrá que

$$\begin{aligned} \frac{g(y_2)-g(y_1)}{f(y_2)-f(y_1)}<\frac{1}{3} &\Leftrightarrow \frac{g(y_1)-g(y_2)}{f(y_1)-f(y_2)}<\frac{1}{3} \Leftrightarrow \frac{(y_1^3-2y_1^2+y_1)-(y_2^3-2y_2^2+y_2)}{(2y_1^3-3y_1^2)-(2y_2^3-3y_2^2)}<\frac{1}{3} \Leftrightarrow \\ 3y_1^3-6y_1^2+3y_1-3y_2^3+6y_2^2-3y_2 &< 2y_1^3-3y_1^2-2y_2^3+3y_2^2 \Leftrightarrow y_1^3-3y_1^2+3y_1 < y_2^3-3y_2^2+3y_2 \Leftrightarrow \\ m(y_1) &< m(y_2) \text{ siendo } m(y)=y^3-3y^2+3y \end{aligned}$$

Basta observar que la función m es creciente pues $m'(y)=3(y-1)^2\geq 0$ y así, $y_1<y_2\Rightarrow m(y_1)<m(y_2)$ con lo cual estaría probado (AVI.6). Además esta acotación se da con independencia del signo que presente $g(y_2)-g(y_1)$, sólo importa que " f " sea decreciente en $[0,1]$. Las implicaciones de esta acotación son las siguientes:

$$\begin{aligned} \frac{g(y_1)-g(y_2)}{f(y_1)-f(y_2)}<\frac{1}{3} &\Rightarrow 3*(g(y_1)-g(y_2))\leq f(y_1)-f(y_2)\Rightarrow \\ K_{ij}*(f(y_1)-f(y_2)) &\leq 3*K_{ij}*(g(y_1)-g(y_2))\Rightarrow K_{ij}*(f(y_2)-f(y_1))\geq 3*K_{ij}*(g(y_2)-g(y_1)) \end{aligned}$$

Dado que perseguimos que $K_{ij}*(f(y_2)-f(y_1))+A_{ij}*(g(y_2)-g(y_1))\geq 0$, y que A_{ij} es una de las variables de decisión con las que podremos jugar (de hecho estamos tratando de establecer su rango de valores posibles para garantizar que la función sea 2-creciente), basta hacer $0\leq A_{ij}\leq -3*K_{ij}\Rightarrow 0\geq -A_{ij}\geq 3*K_{ij}$ (AVI.7)

y como $g(y_2)-g(y_1)<0$, se puede concluir la implicación que buscábamos, es decir,

$$\begin{aligned} K_{ij}*(f(y_2)-f(y_1)) &\geq 3*K_{ij}*(g(y_2)-g(y_1))\geq -A_{ij}*(g(y_2)-g(y_1))\Rightarrow \\ K_{ij}*(f(y_2)-f(y_1))+A_{ij}*(g(y_2)-g(y_1)) &\geq 0 \end{aligned}$$

(AVI.8)

con lo que la expresión (AVI.5) es también positiva.

- Si $g(y_2)-g(y_1)\geq 0$ y $h(y_2)-h(y_1)<0$ los sumandos $K_{ij}*(f(y_2)-f(y_1))$ y $A_{ij}*(g(y_2)-g(y_1))$ son positivos y el sumando $B_{ij+1}*(h(y_2)-h(y_1))$ es negativo. Sin embargo, de igual manera que antes, vamos a demostrar que $K_{ij}*(f(y_2)-f(y_1))+B_{ij+1}*(h(y_2)-h(y_1))$ es mayor o igual que cero. Para ello,

vamos a probar que $\frac{h(y_2)-h(y_1)}{f(y_2)-f(y_1)}<\frac{1}{3}$ (AVI.9)

Al ser $f(y_2)-f(y_1)<0$ y $h(y_2)-h(y_1)<0$, se tendrá que

$$\frac{h(y_2)-h(y_1)}{f(y_2)-f(y_1)} < \frac{1}{3} \Leftrightarrow \frac{h(y_1)-h(y_2)}{f(y_1)-f(y_2)} < \frac{1}{3} \Leftrightarrow \frac{(y_1^3-y_1^2)-(y_2^3-y_2^2)}{(2y_1^3-3y_1^2)-(2y_2^3-3y_2^2)} < \frac{1}{3} \Leftrightarrow 3y_1^3-3y_1^2-3y_2^3+3y_2^2 < 2y_1^3-3y_1^2-2y_2^3+3y_2^2 \Leftrightarrow y_1^3 < y_2^3 \Leftrightarrow n(y_1) < n(y_2) \text{ siendo } n(y)=y^3$$

Nuevamente, la función “ n ” es creciente $n'(y)=3y^2 \geq 0$ y así, $y_1 < y_2 \Rightarrow n(y_1) < n(y_2)$ con lo cual estaría probado (AVI.9). Además esta acotación se da con independencia del signo que presente $h(y_2) - h(y_1)$, sólo importa que “ f ” sea decreciente en $[0,1]$. Las implicaciones de esta acotación son las siguientes:

$$\frac{h(y_1)-h(y_2)}{f(y_1)-f(y_2)} < \frac{1}{3} \Rightarrow 3*(h(y_1)-h(y_2)) \leq f(y_1)-f(y_2) \Rightarrow K_{ij}*(f(y_1)-f(y_2)) \leq 3*K_{ij}*(h(y_1)-h(y_2)) \Rightarrow K_{ij}*(f(y_2)-f(y_1)) \geq 3*K_{ij}*(h(y_2)-h(y_1))$$

Dado que perseguimos que $K_{ij}*(f(y_2)-f(y_1))+B_{i,j+1}*(h(y_2)-h(y_1)) \geq 0$, y que $B_{i,j+1}$ es otra de las variables de decisión, basta hacer $0 \leq B_{i,j+1} \leq -3*K_{ij} \Rightarrow 0 \geq -B_{i,j+1} \geq 3*K_{ij}$ (AVI.10)

y como $h(y_2)-h(y_1) < 0$, se puede concluir la implicación que buscábamos

$$K_{ij}*(f(y_2)-f(y_1)) \geq 3*K_{ij}*(h(y_2)-h(y_1)) \geq -B_{i,j+1}*(h(y_2)-h(y_1)) \Rightarrow K_{ij}*(f(y_2)-f(y_1))+B_{i,j+1}*(h(y_2)-h(y_1)) \geq 0$$

(AVI.11)

con lo que la expresión (AVI.5) es también positiva.

- Finalmente, si $g(y_2)-g(y_1) < 0$ y $h(y_2)-h(y_1) < 0$ el único sumando positivo es $K_{ij}*(f(y_2)-f(y_1))$ mientras que $A_{ij}*(g(y_2)-g(y_1))$ y $B_{i,j+1}*(h(y_2)-h(y_1))$ son negativos. Sin embargo, vamos a demostrar que

$K_{ij}*(f(y_2)-f(y_1))+A_{ij}*(g(y_2)-g(y_1))+B_{i,j+1}*(h(y_2)-h(y_1))$ es mayor o igual que cero. Para ello,

vamos a probar que $\frac{(g(y_2)-g(y_1))+h(y_2)-h(y_1))}{f(y_2)-f(y_1)} < \frac{1}{3}$ (AVI.12)

Al ser $f(y_2)-f(y_1) < 0, g(y_2)-g(y_1) < 0$ y $h(y_2)-h(y_1) < 0$, se tendrá que

$$\frac{(g(y_2)-g(y_1))+h(y_2)-h(y_1))}{f(y_2)-f(y_1)} < \frac{1}{3} \Leftrightarrow \frac{(g(y_1)-g(y_2))+h(y_1)-h(y_2))}{f(y_1)-f(y_2)} < \frac{1}{3} \Leftrightarrow \frac{(y_1^3-2y_1^2+y_1)-(y_2^3-2y_2^2+y_2)+(y_1^3-y_1^2)-(y_2^3-y_2^2)}{(2y_1^3-3y_1^2)-(2y_2^3-3y_2^2)} < \frac{1}{3} \Leftrightarrow 3*(2y_1^3-3y_1^2+y_1)-3*(2y_2^3-3y_2^2+y_2) < 2y_1^3-3y_1^2-(2y_2^3-3y_2^2) \Leftrightarrow 4y_1^3-6y_1^2+3y_1 < 4y_2^3-6y_2^2+3y_2 \Leftrightarrow r(y_1) < r(y_2) \text{ siendo } r(y)=4y^3-6y^2+3y$$

La función “ r ” es también creciente dado que $r'(y)=12y^2-12y+3=12(y-\frac{1}{2})^2 \geq 0$ y así,

$y_1 < y_2 \Rightarrow r(y_1) < r(y_2)$ con lo cual estaría probado (AVI.12). Además esta acotación se da con independencia del signo que presenten $g(y_2) - g(y_1)$ y $h(y_2) - h(y_1)$, sólo importa que “ f ” sea decreciente en $[0,1]$. Las implicaciones de esta acotación son las siguientes:

$$\frac{(g(y_1)-g(y_2))+h(y_1)-h(y_2)}{f(y_1)-f(y_2)} < \frac{1}{3} \Rightarrow 3*(g(y_1)-g(y_2))+3*(h(y_1)-h(y_2)) \leq f(y_1)-f(y_2) \Rightarrow$$

$$K_{ij}*f(y_1)-f(y_2) \leq 3*K_{ij}*(g(y_1)-g(y_2))+3*K_{ij}*(h(y_1)-h(y_2))$$

Por otro lado, respetando los rangos de variación propuestos para A_{ij} y B_{ij+1} dados por (AVI.7) y (AVI.10), y teniendo en cuenta que $g(y_2)-g(y_1) < 0$ y que $h(y_2)-h(y_1) < 0$ se tiene que

$$0 \leq A_{ij}*(g(y_1)-g(y_2)) \leq -3*K_{ij}*(g(y_1)-g(y_2)) \Rightarrow A_{ij}*(g(y_2)-g(y_1)) \geq 3*K_{ij}*(g(y_1)-g(y_2)) \quad (\text{AVI.13})$$

y procediendo de igual manera para B_{ij+1} ,

$$0 \leq B_{ij+1}*(h(y_1)-h(y_2)) \leq -3*K_{ij}*(h(y_1)-h(y_2)) \Rightarrow B_{ij+1}*(h(y_2)-h(y_1)) \geq 3*K_{ij}*(h(y_1)-h(y_2)) \quad (\text{AVI.14})$$

Así, sumando estas dos últimas expresiones, llegaríamos a que

$$A_{ij}*(g(y_2)-g(y_1))+B_{ij+1}*(h(y_2)-h(y_1)) \geq 3*K_{ij}*(g(y_1)-g(y_2))+3*K_{ij}*(h(y_1)-h(y_2))$$

y como $K_{ij}*(f(y_1)-f(y_2)) \leq 3*K_{ij}*(g(y_1)-g(y_2))+3*K_{ij}*(h(y_1)-h(y_2))$, se implica que $A_{ij}*(g(y_2)-g(y_1))+B_{ij+1}*(h(y_2)-h(y_1)) \geq K_{ij}*(f(y_1)-f(y_2)) = -K_{ij}*(f(y_2)-f(y_1))$ y por tanto, que $K_{ij}*(f(y_2)-f(y_1))+A_{ij}*(g(y_2)-g(y_1))+B_{ij+1}*(h(y_2)-h(y_1)) \geq 0$, que recordemos, es la expresión (AVI.5).

Así, la condiciones (AVI.7) y (AVI.10) son suficientes para que (AVI.5) sea positiva. Sin embargo, (AVI.5) es sólo uno de los factores del segundo sumando de (AVI.2) (véase (AVI.4)), el cual no es necesariamente positivo pues depende del signo de $g(x_2) - g(x_1)$ que, como hemos indicado, no es constante puesto que “g” no es una función monótona (véase ilustración (AVI.1)).

Antes de resolver este problema, notemos que puede procederse análogamente con el tercero de los sumandos de (AVI.2), dado por

$$\frac{V_c'(B)}{2} * [h(x_2)-h(x_1)] * [L_{ij}*(f(y_2)-f(y_1))+C_{i+1j}*(g(y_2)-g(y_1))+D_{i+1j+1}*(h(y_2)-h(y_1))] \quad (\text{AVI.15})$$

y demostrar que las condiciones $0 \leq C_{i+1j} \leq -3*L_{ij}$ (AVI.16)

y $0 \leq D_{i+1j+1} \leq -3*L_{ij}$ (AVI.17)

son suficientes para que el tercero de los factores de (AVI.15) sea positivo

$$[L_{ij}*(f(y_2)-f(y_1))+C_{i+1j}*(g(y_2)-g(y_1))+D_{i+1j+1}*(h(y_2)-h(y_1))] \quad (\text{AVI.18})$$

La imposición de las restricciones asociadas a las variables de decisión A , B , C y D permiten tener controlado el signo de casi todos los elementos que intervienen en (AVI.2), salvo $g(x_2) - g(x_1)$ y $h(x_2) - h(x_1)$. Estas restricciones vienen dadas por (AVI.7), (AVI.10), (AVI.16) y (AVI.17).

Veremos sin embargo que es necesario añadir algunas más para conseguir que la función $C_{HERMITE}$ sea 2-creciente pues, como decíamos, existen factores cuyo signo escapa a nuestro control. La intuición de cuáles son éstas otras condiciones la podríamos obtener haciendo un razonamiento análogo para (AVI.3) al realizado para (AVI.2).

Procediendo de la misma forma se puede comprobar la necesidad de imponer las siguientes restricciones para llegar a tener un control de todos los elementos que participan en (AVI.3) salvo $g(y_2) - g(y_1)$ y $h(y_2) - h(y_1)$:

$$0 \leq A_{ij} \leq -3 * P_{ij} \quad (\text{AVI.19})$$

$$0 \leq B_{i,j+1} \leq -3 * Q_{ij} \quad (\text{AVI.20})$$

$$0 \leq C_{i+1,j} \leq -3 * P_{ij} \quad (\text{AVI.21})$$

$$\text{y } 0 \leq D_{i+1,j+1} \leq -3 * Q_{ij} \quad (\text{AVI.22})$$

Son estas condiciones las que, en conjunción con las obtenidas en la búsqueda de que (AVI.5) y (AVI.18) sean factores positivos, nos van a permitir asegurar que la función C_{HERMITE} es 2-creciente. Para ello, recuperaremos la expresión (AVI.2) y estudiaremos los posibles signos de los factores $g(x_2) - g(x_1)$ y $h(x_2) - h(x_1)$.

$$\begin{aligned} & \frac{V_c'(B)}{2} * [f(x_2) - f(x_1)] * [(f(y_2) - f(y_1)) + P_{ij} * (g(y_2) - g(y_1)) + Q_{ij} * (h(y_2) - h(y_1))] + \\ & \frac{V_c'(B)}{2} * [g(x_2) - g(x_1)] * [K_{ij} * (f(y_2) - f(y_1)) + A_{ij} * (g(y_2) - g(y_1)) + B_{i,j+1} * (h(y_2) - h(y_1))] + \\ & \frac{V_c'(B)}{2} * [h(x_2) - h(x_1)] * [L_{ij} * (f(y_2) - f(y_1)) + C_{i+1,j} * (g(y_2) - g(y_1)) + D_{i+1,j+1} * (h(y_2) - h(y_1))] + \\ & \frac{V_c'(B)}{2} * [f(x_2) - f(x_1)] * [(f(y_2) - f(y_1))] \end{aligned}$$

siendo

$$0 \leq A_{ij} \leq -3 * K_{ij}; 0 \leq B_{i,j+1} \leq -3 * K_{ij}; 0 \leq C_{i+1,j} \leq -3 * L_{ij}; 0 \leq D_{i+1,j+1} \leq -3 * L_{ij};$$

Como consecuencia de estas restricciones,

$$\begin{aligned} & \frac{V_c'(B)}{2} * [f(x_2) - f(x_1)] * [(f(y_2) - f(y_1)) + P_{ij} * (g(y_2) - g(y_1)) + Q_{ij} * (h(y_2) - h(y_1))] \geq 0 \\ & [K_{ij} * (f(y_2) - f(y_1)) + A_{ij} * (g(y_2) - g(y_1)) + B_{i,j+1} * (h(y_2) - h(y_1))] \geq 0 \\ & [L_{ij} * (f(y_2) - f(y_1)) + C_{i+1,j} * (g(y_2) - g(y_1)) + D_{i+1,j+1} * (h(y_2) - h(y_1))] \geq 0 \\ & [f(x_2) - f(x_1)] * [(f(y_2) - f(y_1))] \geq 0 \end{aligned}$$

En función de los valores de $g(x_2) - g(x_1)$ y $h(x_2) - h(x_1)$, podremos distinguir las siguientes situaciones:

- Si $g(x_2) - g(x_1) \geq 0$ y $h(x_2) - h(x_1) \geq 0$ todos los sumandos de (AVI.2) son positivos y por tanto la función C_{HERMITE} sería 2-creciente
- Si $g(x_2) - g(x_1) < 0$ y $h(x_2) - h(x_1) \geq 0$, el único sumando de (AVI.2) que es negativo es (AVI.4). Sin embargo, vamos a demostrar que esta cantidad no es suficiente para contrarrestar el valor positivo acumulado por el resto de los sumandos, en particular, por el primero y el cuarto, que podemos agrupar en uno solo como:

$$\begin{aligned} & \frac{V_c'(B)}{2} * [f(x_2) - f(x_1)] * [(f(y_2) - f(y_1)) + P_{ij} * (g(y_2) - g(y_1)) + Q_{ij} * (h(y_2) - h(y_1))] + \\ & \frac{V_c'(B)}{2} * [f(x_2) - f(x_1)] * [(f(y_2) - f(y_1))] = \\ & \frac{V_c'(B)}{2} * [f(x_2) - f(x_1)] * [2 * (f(y_2) - f(y_1)) + P_{ij} * (g(y_2) - g(y_1)) + Q_{ij} * (h(y_2) - h(y_1))] \end{aligned}$$

(AVI.23)

Es decir, vamos a probar que

$$\begin{aligned} & [f(x_2) - f(x_1)] * [2 * (f(y_2) - f(y_1)) + P_{ij} * (g(y_2) - g(y_1)) + Q_{ij} * (h(y_2) - h(y_1))] + \\ & [g(x_2) - g(x_1)] * [K_{ij} * (f(y_2) - f(y_1)) + A_{ij} * (g(y_2) - g(y_1)) + B_{i,j+1} * (h(y_2) - h(y_1))] \geq 0 \end{aligned}$$

o equivalentemente, que

$$\begin{aligned} & [f(x_1) - f(x_2)] * [2 * (f(y_1) - f(y_2)) + P_{ij} * (g(y_1) - g(y_2)) + Q_{ij} * (h(y_1) - h(y_2))] \geq \\ & [g(x_1) - g(x_2)] * [K_{ij} * (f(y_2) - f(y_1)) + A_{ij} * (g(y_2) - g(y_1)) + B_{i,j+1} * (h(y_2) - h(y_1))] \end{aligned}$$

(AVI.24)

Para ello, empezaremos utilizaremos la acotación (AVI.6) ya probada que recordemos viene dada por

$$\frac{g(x_2) - g(x_1)}{f(x_2) - f(x_1)} < \frac{1}{3} .$$

Como $f(x_2) - f(x_1) < 0$ y $g(x_2) - g(x_1) < 0$, dicha expresión es equivalente a la siguiente:

$$\frac{g(x_1) - g(x_2)}{f(x_1) - f(x_2)} < \frac{1}{3} \Leftrightarrow f(x_1) - f(x_2) \geq 3 * (g(x_1) - g(x_2)) \quad \text{que es mayor que cero.}$$

Por otra parte, como $f(y_1) - f(y_2) \geq 0$ (por ser “ f ” decreciente), se verifica también la siguiente inecuación

$$\begin{aligned} & [2 * (f(y_1) - f(y_2)) + P_{ij} * (g(y_1) - g(y_2)) + Q_{ij} * (h(y_1) - h(y_2))] \geq \\ & [(f(y_1) - f(y_2)) + P_{ij} * (g(y_1) - g(y_2)) + Q_{ij} * (h(y_1) - h(y_2))] = \\ & - [(f(y_2) - f(y_1)) + P_{ij} * (g(y_2) - g(y_1)) + Q_{ij} * (h(y_2) - h(y_1))] \geq 0 \end{aligned}$$

(AVI.25)

Dado que todas estas cantidades son positivas podemos multiplicar sin cambiar el signo de la desigualdad

$$f(x_1) - f(x_2) \geq 3 * (g(x_1) - g(x_2)) \quad \text{y establecer que}$$

$$\begin{aligned} & [f(x_1) - f(x_2)] * [2 * (f(y_1) - f(y_2)) + P_{ij} * (g(y_1) - g(y_2)) + Q_{ij} * (h(y_1) - h(y_2))] \geq \\ & 3 * [g(x_1) - g(x_2)] * [2 * (f(y_1) - f(y_2)) + P_{ij} * (g(y_1) - g(y_2)) + Q_{ij} * (h(y_1) - h(y_2))] \end{aligned}$$

(AVI.26)

Vamos a probar que esta última expresión es mayor o igual que

$$[g(x_1) - g(x_2)] * [K_{ij} * (f(y_2) - f(y_1)) + A_{ij} * (g(y_2) - g(y_1)) + B_{i,j+1} * (h(y_2) - h(y_1))] \quad \text{(AVI.27)}$$

lo cual completaría la demostración que perseguimos.

Observemos que, como $g(x_2) - g(x_1) < 0$,

$$\begin{aligned}
& 3*[g(x_1)-g(x_2)]*[2*(f(y_1)-f(y_2))+P_{ij}*(g(y_1)-g(y_2))+Q_{ij}*(h(y_1)-h(y_2))] \geq \\
& [g(x_1)-g(x_2)]*[K_{ij}*(f(y_2)-f(y_1))+A_{ij}*(g(y_2)-g(y_1))+B_{i_{j+1}}*(h(y_2)-h(y_1))] \Leftrightarrow \\
& (K_{ij}+6)*(f(y_2)-f(y_1))+(A_{ij}+3*P_{ij})*(g(y_2)-g(y_1))+(B_{i_{j+1}}+3*Q_{ij})*(h(y_2)-h(y_1)) \leq 0
\end{aligned}$$

(AVI.28)

y que por tanto, bastaría probar esta última desigualdad. Por la estructura que ésta presenta, podemos intuir la conveniencia de que se cumplan restricciones adicionales para A_{ij} y $B_{i_{j+1}}$, concretamente las que ya fueran planteadas por las expresiones (AVI.19) $0 \leq A_{ij} \leq -3*P_{ij}$ y (AVI.20) $0 \leq B_{i_{j+1}} \leq -3*Q_{ij}$.

Estas restricciones si bien no habian sido deducidas del desarrollo de (AVI.2) ya mostramos que sí hubieran resultado de haber procedido de manera análoga para (AVI.3).

Imponer dichas condiciones implica conocer más signos en la expresión (AVI.28) y así, sabemos que $0 \leq 3 \leq K_{ij}+6 \leq 6$, $(f(y_2)-f(y_1)) < 0$ por ser “ f ” decreciente ente 0 y 1, $A_{ij}+3*P_{ij} \leq 0$ y $B_{i_{j+1}}+3*Q_{ij} \leq 0$, con lo cual los únicos signos que no controlamos son nuevamente los asociados a las funciones “ g ” y “ h ”.

Podemos nuevamente plantearnos cuatro casos:

- Si $g(y_2)-g(y_1) \geq 0$ y $h(y_2)-h(y_1) \geq 0$ es fácil verificar que (AVI.28) se cumple pues los tres sumandos de la expresión son negativos (cada uno de ellos es el producto de un factor positivo y otro negativo).
- El caso más conflictivo se podría presentar si $g(y_2)-g(y_1) < 0$ y $h(y_2)-h(y_1) < 0$ pues existen dos sumandos positivos $(A_{ij}+3*P_{ij})*(g(y_2)-g(y_1))$ y $(B_{i_{j+1}}+3*Q_{ij})*(h(y_2)-h(y_1))$, y sólo uno negativo $(K_{ij}+6)*(f(y_2)-f(y_1))$. Sin embargo, vamos a ver que la aportación conjunta de los positivos no va a ser superior a la de este último y que, por tanto, el resultado final será negativo.

Para ello, vamos a recordar la acotación que se establece entre las funciones “ f ”, “ g ” y “ h ”

$$\begin{aligned}
& \frac{(g(y_1)-g(y_2))+(h(y_1)-h(y_2))}{f(y_1)-f(y_2)} < \frac{1}{3} \Rightarrow f(y_1)-f(y_2) > 3*(g(y_1)-g(y_2))+3*(h(y_1)-h(y_2)) \Rightarrow \\
& -3*(f(y_1)-f(y_2)) \leq -9*(g(y_1)-g(y_2))-9*(h(y_1)-h(y_2))
\end{aligned}$$

(AVI.29)

y que $3 \leq K_{ij}+6 \Rightarrow (K_{ij}+6)*(f(y_2)-f(y_1)) \leq 3*(f(y_2)-f(y_1))$ pues “ f ” decrece en $[0,1]$. Entonces, podemos deducir que:

$$(K_{ij}+6)*(f(y_2)-f(y_1)) \leq -3*(f(y_1)-f(y_2)) \leq -9*(g(y_1)-g(y_2))-9*(h(y_1)-h(y_2)) \quad (\text{AVI.30})$$

Si esta expresión es inferior a

$$(A_{ij}+3*P_{ij})*(g(y_2)-g(y_1))+(B_{i_{j+1}}+3*Q_{ij})*(h(y_2)-h(y_1)) \quad (\text{AVI.31})$$

se implicaría directamente la desigualdad presentada en (AVI.28). Es decir, queremos comprobar que $-9(g(y_1)-g(y_2))-9(h(y_1)-h(y_2)) \leq (A_{ij}+3P_{ij})(g(y_2)-g(y_1))+(B_{ij+1}+3Q_{ij})(h(y_2)-h(y_1))$ o equivalentemente $0 \leq (A_{ij}+3P_{ij}+9)(g(y_1)-g(y_2))+(B_{ij+1}+3Q_{ij}+9)(h(y_1)-h(y_2))$ (AVI.32).

Pero esta última desigualdad es cierta pues ambos sumandos son positivos ya que cada uno de ellos es el producto de dos factores positivos.

Efectivamente, por hipótesis $g(y_1)-g(y_2)>0$ y $h(y_1)-h(y_2)>0$ y además, $P_{ij} \geq -3 \Rightarrow 3*P_{ij} \geq -9 \Rightarrow A_{ij}+3*P_{ij} \geq A_{ij}-9 \Rightarrow A_{ij}+3*P_{ij}+9 \geq A_{ij} \geq 0$ y de manera análoga, $Q_{ij} \geq -3 \Rightarrow B_{ij+1}+3*Q_{ij} \geq B_{ij+1}-9 \Rightarrow B_{ij+1}+3*Q_{ij}+9 \geq B_{ij+1} \geq 0$.

Demostrado este caso, no haría falta verificar los dos que quedan, puesto que en cada uno de ellos la contribución positiva es menor ya que sólo existe un sumando con ese signo (en vez de dos). Vamos a comprobar que esto es realmente así.

- Si $g(y_2)-g(y_1)<0$ y $h(y_2)-h(y_1) \geq 0$, los sumandos $(K_{ij}+6)*(f(y_2)-f(y_1))$ y $(B_{ij+1}+3*Q_{ij})*(h(y_2)-h(y_1))$ son negativos y, en cambio, $(A_{ij}+3*P_{ij})*(g(y_2)-g(y_1))$ es positivo. Sin embargo, vamos a ver que la aportación de este último no es superior a la del primero y que, por tanto, el resultado final es también negativo. Teniendo nuevamente en cuenta la acotación

$$\frac{g(y_1)-g(y_2)}{f(y_1)-f(y_2)} < \frac{1}{3} \Rightarrow 3*(g(y_1)-g(y_2)) \leq f(y_1)-f(y_2) \Rightarrow -9*(g(y_1)-g(y_2)) \geq -3*(f(y_1)-f(y_2))$$

(AVI.33)

y que $3 \leq K_{ij}+6 \Rightarrow (K_{ij}+6)(f(y_2)-f(y_1)) \leq 3(f(y_2)-f(y_1))$ por ser “f” decreciente entre 0 y 1, podemos escribir lo siguiente:

$$(K_{ij}+6)*(f(y_2)-f(y_1)) \leq 3(f(y_2)-f(y_1)) = -3(f(y_1)-f(y_2)) \leq -9(g(y_1)-g(y_2)) \quad (\text{AVI.34})$$

Vamos a probar que esta última cantidad es inferior a $(A_{ij}+3*P_{ij})*(g(y_1)-g(y_2))$ lo cual traería como consecuencia que

$$(K_{ij}+6)*(f(y_2)-f(y_1))+(A_{ij}+3*P_{ij})*(g(y_2)-g(y_1)) \leq 0 \quad (\text{AVI.35})$$

y completaría la demostración de que (AVI.28) es también negativo.

Efectivamente, probar $-9(g(y_1)-g(y_2)) \leq (A_{ij}+3*P_{ij})*(g(y_1)-g(y_2))$, equivale a probar $0 \leq (A_{ij}+3*P_{ij}+9)*(g(y_1)-g(y_2))$ lo cual es cierto pues ambos factores son positivos ya que $g(y_1)-g(y_2)>0$ por hipótesis y

$$P_{ij} \geq -3 \Rightarrow 3*P_{ij} \geq -9 \Rightarrow A_{ij}+3*P_{ij} \geq A_{ij}-9 \Rightarrow A_{ij}+3*P_{ij}+9 \geq A_{ij} \geq 0$$

- Finalmente, si $g(y_2)-g(y_1)\geq 0$ y $h(y_2)-h(y_1)<0$, los sumandos $(K_{ij}+6)*(f(y_2)-f(y_1))$ y $(A_{ij}+3*P_{ij})*(g(y_2)-g(y_1))$ son negativos y, en cambio, $(B_{ij+1}+3*Q_{ij})*(h(y_2)-h(y_1))$ es positivo. Sin embargo, vamos a ver que la aportación de este último no es superior a la del primero y que, por tanto, el resultado final es también negativo. Teniendo nuevamente en cuenta la acotación

$$\frac{h(y_1)-h(y_2)}{f(y_1)-f(y_2)} < \frac{1}{3} \Rightarrow 3*(h(y_1)-h(y_2)) \leq f(y_1)-f(y_2) \Rightarrow$$

$$-9*(h(y_1)-h(y_2)) \geq -3*(f(y_1)-f(y_2))$$

(AVI.36)

y que $3 \leq K_{ij}+6 \Rightarrow (K_{ij}+6)*(f(y_2)-f(y_1)) \leq 3*(f(y_2)-f(y_1))$ por ser “ f ” decreciente entre 0 y 1, podemos escribir lo siguiente:

$$(K_{ij}+6)*(f(y_2)-f(y_1)) \leq 3*(f(y_2)-f(y_1)) = -3*(f(y_1)-f(y_2)) \leq -9*(g(y_1)-g(y_2)) \quad (\text{AVI.37})$$

Vamos a probar que esta última cantidad es inferior a $(B_{ij+1}+3*Q_{ij})*(h(y_2)-h(y_1))$ lo cual traería como consecuencia que

$$(K_{ij}+6)*(f(y_2)-f(y_1)) + (B_{ij+1}+3*Q_{ij})*(h(y_2)-h(y_1)) \leq 0 \quad (\text{AVI.38})$$

y completaría la demostración de que (AVI.28) es también negativo.

Efectivamente, probar $-9*(h(y_1)-h(y_2)) \leq (B_{ij+1}+3*Q_{ij})*(h(y_1)-h(y_2))$, equivale a probar $0 \leq (B_{ij+1}+3*Q_{ij}+9)*(h(y_1)-h(y_2))$ que es cierto pues ambos factores son positivos ya que $h(y_1)-h(y_2) > 0$ por hipótesis y $Q_{ij} \geq -3 \Rightarrow B_{ij+1}+3*Q_{ij} \geq B_{ij+1}-9 \Rightarrow B_{ij+1}+3*Q_{ij}+9 \geq B_{ij+1} \geq 0$.

En definitiva, si $g(x_2)-g(x_1)<0$ y $h(x_2)-h(x_1)\geq 0$, la expresión (AVI.28) es cierta y, por tanto, el volumen a través de $C_{HERMITE}$ del rectángulo B es mayor o igual que cero, propiedad que convierte a $C_{HERMITE}$ en una función 2-creciente.

- Si $g(x_2)-g(x_1)\geq 0$ y $h(x_2)-h(x_1)<0$, el único sumando de (AVI.2) que es negativo es (AVI.15). Sin embargo, vamos a demostrar que esta cantidad no es suficiente para contrarrestar el valor positivo acumulado por el resto de los sumandos, en particular por el primero y el cuarto, que como se puede ver en la expresión (AVI.23) podemos agrupar así:

$$\frac{V_c'(B)}{2} * [f(x_2)-f(x_1)] * [2*(f(y_2)-f(y_1)) + P_{ij}*(g(y_2)-g(y_1)) + Q_{ij}*(h(y_2)-h(y_1))]$$

Es decir, vamos a probar que

$$[f(x_2)-f(x_1)] * [2*(f(y_2)-f(y_1)) + P_{ij}*(g(y_2)-g(y_1)) + Q_{ij}*(h(y_2)-h(y_1))] +$$

$$[h(x_2)-h(x_1)] * [L_{ij}*(f(y_2)-f(y_1)) + C_{i+1j}*(g(y_2)-g(y_1)) + D_{i+1j+1}*(h(y_2)-h(y_1))] \geq 0$$

o, equivalentemente, que

$$\begin{aligned} & [f(x_2)-f(x_1)]*[2*(f(y_2)-f(y_1))+P_{ij}*(g(y_2)-g(y_1))+Q_{ij}*(h(y_2)-h(y_1))] \geq \\ & [h(x_1)-h(x_2)]*[L_{ij}*(f(y_2)-f(y_1))+C_{i+1j}*(g(y_2)-g(y_1))+D_{i+1j+1}*(h(y_2)-h(y_1))] \geq 0 \end{aligned}$$

(AVI.39)

Para ello, empezaremos utilizando la acotación ya demostrada dada por (AVI.9) que recordemos era

$$\frac{h(x_2)-h(x_1)}{f(x_2)-f(x_1)} < \frac{1}{3} .$$

Como $f(x_2)-f(x_1) < 0$ y $h(x_2)-h(x_1) < 0$, dicha expresión es equivalente a la siguiente:

$$\frac{h(x_1)-h(x_2)}{f(x_1)-f(x_2)} < \frac{1}{3} \Leftrightarrow f(x_1)-f(x_2) \geq 3*(h(x_1)-h(x_2)) \quad \text{que es mayor que cero.}$$

Por otra parte, como $f(y_1)-f(y_2) \geq 0$ (por ser “f” decreciente), se verifica también la siguiente inecuación

$$\begin{aligned} & [2*(f(y_1)-f(y_2))+P_{ij}*(g(y_1)-g(y_2))+Q_{ij}*(h(y_1)-h(y_2))] \geq \\ & [(f(y_1)-f(y_2))+P_{ij}*(g(y_1)-g(y_2))+Q_{ij}*(h(y_1)-h(y_2))] = \\ & -[(f(y_2)-f(y_1))+P_{ij}*(g(y_2)-g(y_1))+Q_{ij}*(h(y_2)-h(y_1))] \geq 0 \end{aligned}$$

(AVI.40)

Dado que todas estas cantidades son positivas podemos multiplicar sin cambiar el signo de la desigualdad

$$f(x_1)-f(x_2) \geq 3*(h(x_1)-h(x_2)) \quad \text{y establecer que}$$

$$\begin{aligned} & [f(x_1)-f(x_2)]*[2*(f(y_1)-f(y_2))+P_{ij}*(g(y_1)-g(y_2))+Q_{ij}*(h(y_1)-h(y_2))] \geq \\ & 3*[h(x_1)-h(x_2)]*[2*(f(y_1)-f(y_2))+P_{ij}*(g(y_1)-g(y_2))+Q_{ij}*(h(y_1)-h(y_2))] \end{aligned}$$

(AVI.41)

Vamos a probar que esta última expresión es mayor o igual que

$$[h(x_1)-h(x_2)]*[L_{ij}*(f(y_2)-f(y_1))+C_{i+1j}*(g(y_2)-g(y_1))+D_{i+1j+1}*(h(y_2)-h(y_1))] \geq 0 \quad \text{(AVI.42)}$$

lo cual completaría la demostración que perseguimos.

Observemos que, como $h(x_2)-h(x_1) < 0$,

$$\begin{aligned} & 3*[h(x_1)-h(x_2)]*[2*(f(y_1)-f(y_2))+P_{ij}*(g(y_1)-g(y_2))+Q_{ij}*(h(y_1)-h(y_2))] \geq \\ & [h(x_1)-h(x_2)]*[L_{ij}*(f(y_2)-f(y_1))+C_{i+1j}*(g(y_2)-g(y_1))+D_{i+1j+1}*(h(y_2)-h(y_1))] \Leftrightarrow \\ & (L_{ij}+6)*(f(y_2)-f(y_1))+(C_{i+1j}+3*P_{ij})*(g(y_2)-g(y_1))+(D_{i+1j+1}+3*Q_{ij})*(h(y_2)-h(y_1)) \leq 0 \end{aligned}$$

(AVI.43)

y que por tanto, basta probar esta última desigualdad. Por la estructura que presenta ésta, podemos intuir la conveniencia de que se cumplan restricciones adicionales para C_{i+1j} y D_{i+1j+1} , concretamente las que ya fueran planteadas por las expresiones (AVI.21) $0 \leq C_{i+1j} \leq -3*P_{ij}$ y (AVI.22) $0 \leq D_{i+1j+1} \leq -3*Q_{ij}$.

Gracias a ellas, podemos conocer el signo de una gran parte de los factores que interviene en (AVI.43):

$$0 \leq 3 \leq L_{ij}+6 \leq 6, \quad (f(y_2)-f(y_1)) < 0 \quad \text{por ser “f” decreciente ente 0 y 1,} \quad C_{i+1j}+3*P_{ij} \leq 0 \quad \text{y}$$

$$D_{i+1j+1}+3*Q_{ij} \leq 0, \quad \text{con lo cual los únicos signos que no controlamos son nuevamente los asociados a$$

“g” y “h”, pudiendo presentarse las siguientes posibilidades:

- Si $g(y_2)-g(y_1)\geq 0$ y $h(y_2)-h(y_1)\geq 0$ es fácil verificar que (AVI.43) se cumple pues los tres sumandos de la expresión son negativos (cada uno de ellos es el producto de un factor positivo y otro negativo).
- Analicemos ahora el caso más desfavorable, en el cual $g(y_2)-g(y_1)<0$ y $h(y_2)-h(y_1)<0$. En dicho caso, existen dos sumandos positivos $(C_{i+1j}+3*P_{ij})*(g(y_2)-g(y_1))$ y $(D_{i+1j+1}+3*Q_{ij})*(h(y_2)-h(y_1))$, y sólo uno negativo $(L_{ij}+6)*(f(y_2)-f(y_1))$. Sin embargo, vamos a ver que la aportación conjunta de los positivos no va a ser superior a la de este último y que, por tanto, el resultado final será negativo.

Para ello, vamos a recordar la acotación que se establece entre las funciones “f”, “g” y “h”

$$\frac{(g(y_1)-g(y_2))+(h(y_1)-h(y_2))}{f(y_1)-f(y_2)} < \frac{1}{3} \Rightarrow f(y_1)-f(y_2) > 3*(g(y_1)-g(y_2))+3*(h(y_1)-h(y_2)) \Rightarrow -3*(f(y_1)-f(y_2)) \leq -9*(g(y_1)-g(y_2))-9*(h(y_1)-h(y_2))$$

(AVI.44)

y que $3 \leq L_{ij}+6 \Rightarrow (L_{ij}+6)*(f(y_2)-f(y_1)) \leq 3*(f(y_2)-f(y_1))$ pues “f” crece entre 0 y 1. Entonces, podemos deducir lo siguiente:

$$(K_{ij}+6)*(f(y_2)-f(y_1)) \leq -3*(f(y_1)-f(y_2)) \leq -9*(g(y_1)-g(y_2))-9*(h(y_1)-h(y_2)) \quad (\text{AVI.45})$$

Si esta expresión es inferior a

$$(C_{i+1j}+3*P_{ij})*(g(y_2)-g(y_1))+(D_{i+1j+1}+3*Q_{ij})*(h(y_2)-h(y_1)) \quad (\text{AVI.46})$$

se implicaría directamente la desigualdad presentada en (AVI.43). Es decir, queremos comprobar que

$$-9*(g(y_1)-g(y_2))-9*(h(y_1)-h(y_2)) \leq (C_{i+1j}+3*P_{ij})*(g(y_2)-g(y_1))+(D_{i+1j+1}+3*Q_{ij})*(h(y_2)-h(y_1))$$

o bien que $0 \leq (C_{i+1j}+3*P_{ij}+9)*(g(y_1)-g(y_2))+(D_{i+1j+1}+3*Q_{ij}+9)*(h(y_1)-h(y_2))$ (AVI.47).

Pero esta última desigualdad es cierta pues ambos sumandos son positivos ya que cada uno de ellos es el producto de dos factores positivos. Efectivamente, por hipótesis

$$g(y_1)-g(y_2) > 0, \quad h(y_1)-h(y_2) > 0 \quad \text{y además,}$$

$$P_{ij} \geq -3 \Rightarrow 3*P_{ij} \geq -9 \Rightarrow C_{i+1j}+3*P_{ij} \geq C_{i+1j}-9 \Rightarrow C_{i+1j}+3*P_{ij}+9 \geq C_{i+1j} \geq 0$$

De manera análoga,

$$Q_{ij} \geq -3 \Rightarrow D_{i+1j+1}+3*Q_{ij} \geq D_{i+1j+1}-9 \Rightarrow D_{i+1j+1}+3*Q_{ij}+9 \geq D_{i+1j+1} \geq 0.$$

No entraremos en la demostración de los otros dos casos que son una implicación directa de éste. Basta observar que en cada uno de ellos la contribución positiva a (AVI.2) sería menor ya que sólo existe un sumando con ese signo (en vez de dos).

En definitiva, si $g(x_2)-g(x_1)\geq 0$ y $h(x_2)-h(x_1)<0$, la expresión (AVI.43) es cierta y, por tanto, el

volumen a través de $C_{HERMITE}$ del rectángulo B es mayor o igual que cero, propiedad que convierte a $C_{HERMITE}$ en una función 2-creciente.

- El último caso, y también el que más en contra estaría del hecho de que (AVI.43) fuese positiva, se presentaría cuando $g(x_2) - g(x_1) < 0$ y $h(x_2) - h(x_1) < 0$, en cuyo caso existirían dos sumandos negativos en (AVI.2) que serían los dados por (AVI.4) y (AVI.15).

Sin embargo vamos a ver que tampoco en esta circunstancia, la aportación conjunta de ambos es suficiente para contrarrestar el valor positivo acumulado por los sumandos primero y cuarto que, como ya hemos hecho en casos anteriores (AVI.23), podemos agrupar de la siguiente manera:

$$\frac{V_C'(B)}{2} * [f(x_2) - f(x_1)] * [2 * (f(y_2) - f(y_1)) + P_{ij} * (g(y_2) - g(y_1)) + Q_{ij} * (h(y_2) - h(y_1))]$$

Es decir, vamos a probar que

$$\begin{aligned} & [f(x_2) - f(x_1)] * [2 * (f(y_2) - f(y_1)) + P_{ij} * (g(y_2) - g(y_1)) + Q_{ij} * (h(y_2) - h(y_1))] + \\ & [g(x_2) - g(x_1)] * [K_{ij} * (f(y_2) - f(y_1)) + A_{ij} * (g(y_2) - g(y_1)) + B_{ij+1} * (h(y_2) - h(y_1))] + \\ & [h(x_2) - h(x_1)] * [L_{ij} * (f(y_2) - f(y_1)) + C_{i+1j} * (g(y_2) - g(y_1)) + D_{i+1j+1} * (h(y_2) - h(y_1))] \geq 0 \end{aligned}$$

o, equivalentemente, que

$$\begin{aligned} & [f(x_1) - f(x_2)] * [2 * (f(y_1) - f(y_2)) + P_{ij} * (g(y_1) - g(y_2)) + Q_{ij} * (h(y_1) - h(y_2))] \geq \\ & [g(x_1) - g(x_2)] * [K_{ij} * (f(y_2) - f(y_1)) + A_{ij} * (g(y_2) - g(y_1)) + B_{ij+1} * (h(y_2) - h(y_1))] + \\ & [h(x_1) - h(x_2)] * [L_{ij} * (f(y_2) - f(y_1)) + C_{i+1j} * (g(y_2) - g(y_1)) + D_{i+1j+1} * (h(y_2) - h(y_1))] \geq 0 \end{aligned}$$

(AVI.48)

Para ello, empezaremos utilizaremos la acotación (AVI.12) que recordemos era

$$\frac{(g(x_2) - g(x_1)) + (h(x_2) - h(x_1))}{f(x_2) - f(x_1)} < \frac{1}{3}$$

Como $f(x_2) - f(x_1) < 0$ y $g(x_2) - g(x_1) < 0$, dicha expresión es equivalente a la siguiente:

$$\frac{(g(x_1) - g(x_2)) + (h(x_1) - h(x_2))}{f(x_1) - f(x_2)} < \frac{1}{3} \Leftrightarrow f(x_1) - f(x_2) \geq 3 * (g(x_1) - g(x_2)) + 3 * (h(x_1) - h(x_2)) \quad \text{que es}$$

mayor que cero.

Por otra parte, como $f(y_1) - f(y_2) \geq 0$ (por ser “ f ” decreciente), se verifica también la siguiente inecuación

$$\begin{aligned} & [2 * (f(y_1) - f(y_2)) + P_{ij} * (g(y_1) - g(y_2)) + Q_{ij} * (h(y_1) - h(y_2))] \geq \\ & [(f(y_1) - f(y_2)) + P_{ij} * (g(y_1) - g(y_2)) + Q_{ij} * (h(y_1) - h(y_2))] = \\ & - [(f(y_2) - f(y_1)) + P_{ij} * (g(y_2) - g(y_1)) + Q_{ij} * (h(y_2) - h(y_1))] \geq 0 \end{aligned}$$

(AVI.49)

Dado que todas estas cantidades son positivas podemos multiplicar sin cambiar el signo de la desigualdad

$$f(x_1) - f(x_2) \geq 3 * (g(x_1) - g(x_2)) + 3 * (h(x_1) - h(x_2)) \quad \text{y establecer que}$$

$$\begin{aligned}
& [f(x_1) - f(x_2)] * [2 * (f(y_1) - f(y_2)) + P_{ij} * (g(y_1) - g(y_2)) + Q_{ij} * (h(y_1) - h(y_2))] \geq \\
& 3 * [g(x_1) - g(x_2)] * [2 * (f(y_1) - f(y_2)) + P_{ij} * (g(y_1) - g(y_2)) + Q_{ij} * (h(y_1) - h(y_2))] + \\
& 3 * [h(x_1) - h(x_2)] * [2 * (f(y_1) - f(y_2)) + P_{ij} * (g(y_1) - g(y_2)) + Q_{ij} * (h(y_1) - h(y_2))]
\end{aligned}$$

(AVI.50)

Vamos a probar que esta última expresión es mayor o igual que

$$\begin{aligned}
& [g(x_1) - g(x_2)] * [K_{ij} * (f(y_2) - f(y_1)) + A_{ij} * (g(y_2) - g(y_1)) + B_{i,j+1} * (h(y_2) - h(y_1))] + \\
& [h(x_1) - h(x_2)] * [L_{ij} * (f(y_2) - f(y_1)) + C_{i+1,j} * (g(y_2) - g(y_1)) + D_{i+1,j+1} * (h(y_2) - h(y_1))]
\end{aligned}$$

(AVI.51)

lo cual completaría la demostración que perseguimos.

Observemos que, como $g(x_2) - g(x_1) < 0$ y $h(x_2) - h(x_1) < 0$

$$\begin{aligned}
& 3 * [g(x_1) - g(x_2)] * [2 * (f(y_1) - f(y_2)) + P_{ij} * (g(y_1) - g(y_2)) + Q_{ij} * (h(y_1) - h(y_2))] + \\
& 3 * [h(x_1) - h(x_2)] * [2 * (f(y_1) - f(y_2)) + P_{ij} * (g(y_1) - g(y_2)) + Q_{ij} * (h(y_1) - h(y_2))] \geq \\
& [g(x_1) - g(x_2)] * [K_{ij} * (f(y_2) - f(y_1)) + A_{ij} * (g(y_2) - g(y_1)) + B_{i,j+1} * (h(y_2) - h(y_1))] + \\
& [h(x_1) - h(x_2)] * [L_{ij} * (f(y_2) - f(y_1)) + C_{i+1,j} * (g(y_2) - g(y_1)) + D_{i+1,j+1} * (h(y_2) - h(y_1))] \Leftrightarrow \\
& (K_{ij} + 6) * (f(y_2) - f(y_1)) + (A_{ij} + 3 * P_{ij}) * (g(y_2) - g(y_1)) + (B_{i,j+1} + 3 * Q_{ij}) * (h(y_2) - h(y_1)) + \\
& (L_{ij} + 6) * (f(y_2) - f(y_1)) + (C_{i+1,j} + 3 * P_{ij}) * (g(y_2) - g(y_1)) + (D_{i+1,j+1} + 3 * Q_{ij}) * (h(y_2) - h(y_1)) \leq 0
\end{aligned}$$

(AVI.52)

y que por tanto, basta probar esta última desigualdad.

Atendiendo a las demostraciones de los casos anteriores, basta darse cuenta de que las condiciones (AVI.19)-(AVI.22) son suficientes para que se dé la desigualdad (AVI.52) no siendo necesario repetir los cuatro casos que podrían plantearse para las funciones “g” y “h”.

Dado que las restricciones adicionales (AVI.7), (AVI.10), (AVI.16), (AVI.17) y (AVI.19)-(AVI.22) se dan por hipótesis, puesto que éstas son una traducción directa de que las variables

$$\frac{\partial C_{HERMITE}(u_i, v_j)}{\partial u}, \frac{\partial C_{HERMITE}(u_i, v_j)}{\partial v} \text{ y } \frac{\partial^2 C_{HERMITE}(u_i, v_j)}{\partial u \partial v} \quad \forall i \in \{0, 1, \dots, n\} \quad \forall j \in \{0, 1, \dots, m\} \text{ proporcionen}$$

junto con una relación de δ s y γ s una solución del problema de programación multiobjetivo no lineal entera (4.35), se implicará, de acuerdo a la demostración que hemos desarrollado, que el volumen de un rectángulo arbitrario totalmente contenido en otro de los que conforman la partición del cuadrado unidad tenga un volumen mayor o igual que cero a través de $C_{HERMITE}$.

De igual manera que deducíamos en el Anexo IV, el volumen de otro rectángulo cualquiera trazado sobre el cuadrado unidad será también mayor o igual que cero pues se podrá expresar como suma de los volúmenes de rectángulos parcial o totalmente contenidos en los que definen la partición. Así, quedaría probado el cumplimiento de la tercera de las propiedades de las cópulas (“ser 2-creciente”) y consecuentemente el lema.

c.q.d

ANEXO VII: Demostración del teorema 4.6

“Sea una partición del cuadrado unidad en $n \times m$ rectángulos del mismo área y sea la solución óptima al problema de programación fraccional (4.7) dada por $c_{ij}^* = C^*\left(\frac{i}{n}, \frac{j}{m}\right) \forall i \in \{0, 1, 2, \dots, n\} \quad \forall j \in \{0, 1, 2, \dots, m\}$ que define una subcópula C^* de dominio $\Delta = \{\text{los puntos de intersección entre los rectángulos que originan la partición}\}$.

La condición necesaria y suficiente para poder definir una cópula $C_{FERGUSON}$ por interpolación del dominio de C^* a través de las condiciones suficientes establecidas en los lemas (4.2), (4.3) y (4.5), es que C^* sea una 2-subcópula- V_{MIN}

siendo $V_{MIN} = \frac{2}{3 * m * n}$ ”.

demostración

Vamos a probar las implicaciones en cada uno de los sentidos.

\Rightarrow) Supongamos que tenemos una cópula $C_{FERGUSON}$ cuya definición responde a (4.20) y tal que el valor de sus derivadas primeras respecto de U y V responde a las condiciones establecidas en los lemas (4.2), (4.3) y (4.5), las cuales permiten la definición de dicha cópula.

Es decir, tenemos una relación de valores (variables del problema (4.26)) para las derivadas respecto de U y V ,

$$\frac{\partial C_{FERGUSON}(u_i, v_j)}{\partial u}, \quad \frac{\partial C_{FERGUSON}(u_i, v_j)}{\partial v} \quad \forall i \in \{0, 1, 2, \dots, n\} \quad \forall j \in \{0, 1, 2, \dots, m\} \quad \text{tales que:}$$

- $\frac{\partial C_{FERGUSON}(u_k, 0)}{\partial u} = 0 \quad \forall k \in \{0, 1, 2, \dots, n\} \quad \text{y} \quad \frac{\partial C_{FERGUSON}(0, v_l)}{\partial v} = 0 \quad \forall l \in \{0, 1, 2, \dots, m\} \quad (\text{lema (4.2)})$
- $\frac{\partial C_{FERGUSON}(u_k, 1)}{\partial u} = 1 \quad \forall k \in \{0, 1, 2, \dots, n\} \quad \text{y} \quad \frac{\partial C_{FERGUSON}(1, v_l)}{\partial v} = 1 \quad \forall l \in \{0, 1, 2, \dots, m\} \quad (\text{lema (4.3)})$
- Proporcionan una solución factible al problema (4.26) (lema (4.5))

Sea V_{MIN} el volumen mínimo que deben tener cada uno de los $n \times m$ rectángulos (que resultan de hacer la partición del cuadrado unidad) a través de la función C^* (y por tanto a través de $C_{FERGUSON}$) por obtenerse ésta por interpolación de C^* .

Vamos a ver que necesariamente $V_{MIN} \geq \frac{2}{3 * m * n}$ y por tanto que C^* es 2-subcópula- V_{MIN} .

Consideremos de forma genérica la rejilla que divide el cuadrado unidad en $n \times m$ rectángulos:

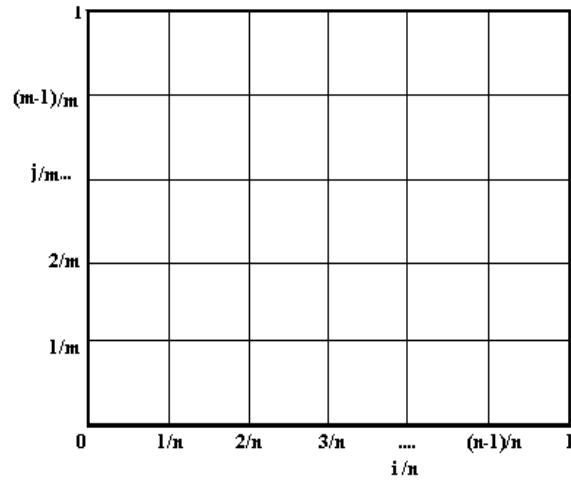


Ilustración AVII.1: Partición $n \times m$ del cuadrado unidad

Vamos a analizar en primer lugar las variables de decisión asociadas a hileras horizontales, la primera de las cuales se mueve en la ordenada $v = 0$. Estas variables responden a la nomenclatura de P 's y Q 's y son las que relacionan las primeras derivadas de la cópula respecto de V (véanse (4.24) y (4.25)):

$$P_{ij} = \frac{2*(v_{j+1} - v_j)}{V_{C^*}(B)} * \left[\frac{\partial C^*(u_i, v_j)}{\partial v} - \frac{\partial C^*(u_{i+1}, v_j)}{\partial v} \right]$$

$$Q_{ij} = \frac{2*(v_{j+1} - v_j)}{V_{C^*}(B)} * \left[\frac{\partial C^*(u_i, v_{j+1})}{\partial v} - \frac{\partial C^*(u_{i+1}, v_{j+1})}{\partial v} \right]$$

Basaremos el análisis en variables P 's, siendo indiferente el haberlo hecho para variables Q 's puesto que unas y otras estarían referidas al suelo (v_j) y al techo (v_{j+1}) de los rectángulos de cada hilera horizontal, circunstancia irrelevante para este estudio. Si fuéramos absolutamente puristas con la notación, deberíamos utilizar variables P 's para la ordenada 0, variables Q 's para la ordenada 1 y cualquiera de las dos para ordenadas comprendidas entre 1 y $m-1$.

Entonces, como $P_{0,0}$ pertenece a la región factible asociada al problema (4.26), debe ser menor o igual que 0 y, mayor o igual que -3. Además, también por hipótesis, suponemos que deben darse las condiciones impuestas por los lemas (4.2) y (4.3).

Teniendo en cuenta todo esto, podemos hacer el desarrollo siguiente:

$$P_{0,0} = \frac{2*\frac{1}{m}}{V_{MIN}} * \left[\frac{\partial C^*(0,0)}{\partial v} - \frac{\partial C^*(u_1,0)}{\partial v} \right] \leq 0 \Rightarrow \frac{\partial C^*(u_1,0)}{\partial v} \geq 0 \quad (\text{AVII.1})$$

$$P_{0,0} = \frac{2*\frac{1}{m}}{V_{MIN}} * \left[\frac{\partial C^*(0,0)}{\partial v} - \frac{\partial C^*(u_1,0)}{\partial v} \right] \geq -3 \Rightarrow \frac{\partial C^*(u_1,0)}{\partial v} \leq \frac{3*V_{MIN}*m}{2} \quad (\text{AVII.2})$$

Del mismo modo, Como $P_{n-1,0}$ debe pertenecer a la región factible,

$$P_{n-1,0} = \frac{2 * \frac{1}{m}}{V_{MIN}} * \left[\frac{\partial C^*(u_{n-1},0)}{\partial v} - \frac{\partial C^*(1,0)}{\partial v} \right] \leq 0 \Rightarrow \left[\frac{\partial C^*(u_{n-1},0)}{\partial v} - 1 \right] \leq 0 \Rightarrow \left[\frac{\partial C^*(u_{n-1},0)}{\partial v} \right] \leq 1 \quad (\text{AVII.3})$$

$$P_{n-1,0} = \frac{2 * \frac{1}{m}}{V_{MIN}} * \left[\frac{\partial C^*(u_{n-1},0)}{\partial v} - \frac{\partial C^*(1,0)}{\partial v} \right] \geq -3 \Rightarrow \left[\frac{\partial C^*(u_{n-1},0)}{\partial v} - 1 \right] \geq \frac{-3 * V_{MIN} * m}{2} \Rightarrow \left[\frac{-\partial C^*(u_{n-1},0)}{\partial v} \right] \leq -1 + \frac{3 * V_{MIN} * m}{2}$$

(AVII.4)

En función de las ecuaciones (AVII.2) y (AVII.4) se tiene que:

$$\frac{\partial C^*(u_1,0)}{\partial v} - \frac{\partial C^*(u_{n-1},0)}{\partial v} \leq \frac{3 * V_{MIN} * m}{2} - 1 + \frac{3 * V_{MIN} * m}{2} = -1 + 3 * V_{MIN} * m \quad (\text{AVII.5})$$

Por otra parte, teniendo en cuenta la cota inferior para el resto de las P 's de las variables de ordenada 0, se tiene que:

$$P_{1,0} = \frac{2 * \frac{1}{m}}{V_{MIN}} * \left[\frac{\partial C^*(u_1,0)}{\partial v} - \frac{\partial C^*(u_2,0)}{\partial v} \right] \geq -3 \Rightarrow \left[\frac{\partial C^*(u_1,0)}{\partial v} - \frac{\partial C^*(u_2,0)}{\partial v} \right] \geq \frac{-3 * V_{MIN} * m}{2} \quad (\text{AVII.6})$$

$$P_{2,0} = \frac{2 * \frac{1}{m}}{V_{MIN}} * \left[\frac{\partial C^*(u_2,0)}{\partial v} - \frac{\partial C^*(u_3,0)}{\partial v} \right] \geq -3 \Rightarrow \left[\frac{\partial C^*(u_2,0)}{\partial v} - \frac{\partial C^*(u_3,0)}{\partial v} \right] \geq \frac{-3 * V_{MIN} * m}{2} \quad (\text{AVII.7})$$

...

$$P_{n-2,0} = \frac{2 * \frac{1}{m}}{V_{MIN}} * \left[\frac{\partial C^*(u_{n-2},0)}{\partial v} - \frac{\partial C^*(u_{n-1},0)}{\partial v} \right] \geq -3 \Rightarrow \left[\frac{\partial C^*(u_{n-2},0)}{\partial v} - \frac{\partial C^*(u_{n-1},0)}{\partial v} \right] \geq \frac{-3 * V_{MIN} * m}{2} \quad (\text{AVII.8})$$

La suma de estas $n-2$ expresiones da lugar a

$$\left[\frac{\partial C^*(u_{1,0})}{\partial v} - \frac{\partial C^*(u_{n-1},0)}{\partial v} \right] \geq \frac{-3 * V_{MIN} * m * (n-2)}{2} \quad (\text{AVII.9})$$

Así, juntando las expresiones (AVII.5) y (AVII.9) se llega a que

$$\frac{-3 * V_{MIN} * m * (n-2)}{2} \leq \left[\frac{\partial C^*(u_1,0)}{\partial v} - \frac{\partial C^*(u_{n-1},0)}{\partial v} \right] \leq -1 + 3 * V_{MIN} * m \quad (\text{AVII.10})$$

Atendiendo a los extremos de la expresión anterior, podemos establecer finalmente que,

$$\frac{-3 * V_{MIN} * m * (n-2)}{2} \leq -1 + 3 * V_{MIN} * m \Rightarrow V_{MIN} \geq \frac{2}{3 * m * n} \quad (\text{AVII.11}).$$

A este mismo resultado se podría haber llegado para otra hilera horizontal cualquiera. Efectivamente, si consideramos las variables de decisión asociadas a una ordenada arbitraria "j", se tiene que:

Como $P_{0,j}$ debe pertenecer a la región factible, debe ser menor o igual que 0 y, mayor o igual que -3

$$P_{0,j} = \frac{2 * \frac{1}{m}}{V_{MIN}} * \left[\frac{\partial C^*(0,v_j)}{\partial v} - \frac{\partial C^*(u_1,v_j)}{\partial v} \right] \leq 0 \Rightarrow \frac{\partial C^*(u_1,v_j)}{\partial v} \geq 0 \quad \forall j \in \{0,1,\dots,m\} \quad (\text{AVII.12})$$

$$P_{0,j} = \frac{2 * \frac{1}{m}}{V_{MIN}} * \left[\frac{\partial C^*(0, v_j)}{\partial v} - \frac{\partial C^*(u_1, v_j)}{\partial v} \right] \geq -3 \Rightarrow \frac{\partial C^*(u_1, v_j)}{\partial v} \leq \frac{3 * V_{MIN} * m}{2} \quad \forall j \in \{0, 1, \dots, m\} \quad (\text{AVII.13})$$

Del mismo modo, Como $P_{n-1,j}$ debe pertenecer a la región factible,

$$P_{n-1,j} = \frac{2 * \frac{1}{m}}{V_{MIN}} * \left[\frac{\partial C^*(u_{n-1}, v_j)}{\partial v} - \frac{\partial C^*(1, v_j)}{\partial v} \right] \leq 0 \Rightarrow \left[\frac{\partial C^*(u_{n-1}, v_j)}{\partial v} - 1 \right] \leq 0 \Rightarrow \left[\frac{\partial C^*(u_{n-1}, v_j)}{\partial v} \right] \leq 1 \quad \forall j \in \{0, 1, \dots, m\}$$

(AVII.14)

$$P_{n-1,j} = \frac{2 * \frac{1}{m}}{V_{MIN}} * \left[\frac{\partial C^*(u_{n-1}, v_j)}{\partial v} - \frac{\partial C^*(1, v_j)}{\partial v} \right] \geq -3 \Rightarrow \left[\frac{\partial C^*(u_{n-1}, v_j)}{\partial v} - 1 \right] \geq \frac{-3 * V_{MIN} * m}{2} \Rightarrow \left[\frac{-\partial C^*(u_{n-1}, v_j)}{\partial v} \right] \leq -1 + \frac{3 * V_{MIN} * m}{2} \quad \forall j \in \{0, 1, \dots, m\}$$

(AVII.15)

En función de (AVII.13) y (AVII.15) se tiene que

$$\frac{\partial C^*(u_1, v_j)}{\partial v} - \frac{\partial C^*(u_{n-1}, v_j)}{\partial v} \leq \frac{3 * V_{MIN} * m}{2} - 1 + \frac{3 * V_{MIN} * m}{2} = -1 + 3 * V_{MIN} * m \quad \forall j \in \{0, 1, \dots, m\} \quad (\text{AVII.16})$$

Por otra parte, teniendo en cuenta la cota inferior para el resto de las P 's asociadas a la ordenada "j", se tiene que:

$$P_{1,j} = \frac{2 * \frac{1}{m}}{V_{MIN}} * \left[\frac{\partial C^*(u_1, v_j)}{\partial v} - \frac{\partial C^*(u_2, v_j)}{\partial v} \right] \geq -3 \Rightarrow \left[\frac{\partial C^*(u_1, v_j)}{\partial v} - \frac{\partial C^*(u_2, v_j)}{\partial v} \right] \geq \frac{-3 * V_{MIN} * m}{2} \quad \forall j \in \{0, 1, \dots, m\} \quad (\text{AVII.17})$$

$$P_{2,j} = \frac{2 * \frac{1}{m}}{V_{MIN}} * \left[\frac{\partial C^*(u_2, v_j)}{\partial v} - \frac{\partial C^*(u_3, v_j)}{\partial v} \right] \geq -3 \Rightarrow \left[\frac{\partial C^*(u_2, v_j)}{\partial v} - \frac{\partial C^*(u_3, v_j)}{\partial v} \right] \geq \frac{-3 * V_{MIN} * m}{2} \quad \forall j \in \{0, 1, \dots, m\} \quad (\text{AVII.18})$$

...

$$P_{n-2,j} = \frac{2 * \frac{1}{m}}{V_{MIN}} * \left[\frac{\partial C^*(u_{n-2}, v_j)}{\partial v} - \frac{\partial C^*(u_{n-1}, v_j)}{\partial v} \right] \geq -3 \Rightarrow \left[\frac{\partial C^*(u_{n-2}, v_j)}{\partial v} - \frac{\partial C^*(u_{n-1}, v_j)}{\partial v} \right] \geq \frac{-3 * V_{MIN} * m}{2} \quad \forall j \in \{0, 1, \dots, m\}$$

(AVII.19)

La suma de estas $n-2$ expresiones da lugar a

$$\left[\frac{\partial C^*(u_1, v_j)}{\partial v} - \frac{\partial C^*(u_{n-1}, v_j)}{\partial v} \right] \geq \frac{-3 * V_{MIN} * m * (n-2)}{2} \quad \forall j \in \{0, 1, \dots, m\} \quad (\text{AVII.20})$$

Así, teniendo en cuenta (AVII.16) y (AVII.20) se llega nuevamente a que

$$\frac{-3 * V_{MIN} * m * (n-2)}{2} \leq \left[\frac{\partial C^*(u_1, v_j)}{\partial v} - \frac{\partial C^*(u_{n-1}, v_j)}{\partial v} \right] \leq -1 + 3 * V_{MIN} * m \quad \forall j \in \{0, 1, \dots, m\} \quad (\text{AVII.21})$$

y en consecuencia a que $\frac{-3*V_{MIN}*m*(n-2)}{2} \leq -1 + 3*V_{MIN}*m \Rightarrow V_{MIN} \geq \frac{2}{3*m*n}$ (AVII.22)

El mismo razonamiento aplicado a las hileras verticales nos va a conducir a la acotación de V_{MIN} en función de la dimensión “n”. Dichas hileras van asociadas a variables K 's y L 's (véase (4.22) y (4.23)).

$$K_{ij} = \frac{2*(u_{i+1}-u_i)}{V_{C^*}(B)} * \left[\frac{\partial C^*(u_i, v_j)}{\partial u} - \frac{\partial C^*(u_i, v_{j+1})}{\partial u} \right] \quad \text{y} \quad L_{ij} = \frac{2*(u_{i+1}-u_i)}{V_{C^*}(B)} * \left[\frac{\partial C^*(u_{i+1}, v_j)}{\partial u} - \frac{\partial C^*(u_{i+1}, v_{j+1})}{\partial u} \right]$$

Dado que el utilizar unas u otras resulta nuevamente irrelevante, emplearemos por ejemplo las variables K 's.

Supuesta una hilera vertical cualquiera asociada a la abscisa “i”, entonces $K_{i,0}$ debe pertenecer a la región factible asociada al problema (4.26) y por tanto, debe ser menor o igual que 0 y, mayor o igual que -3 . Esto, junto con el hecho de que se verifican las condiciones impuestas por los lemas (4.2) y (4.3) sobre las derivadas primeras, permite hacer el siguiente desarrollo:

$$K_{i,0} = \frac{2*\frac{1}{n}}{V_{MIN}} * \left[\frac{\partial C^*(u_i, 0)}{\partial u} - \frac{\partial C^*(u_i, v_1)}{\partial u} \right] \leq 0 \Rightarrow \frac{\partial C^*(u_i, v_1)}{\partial u} \geq 0 \quad \forall i \in \{0, 1, \dots, n\} \quad (\text{AVII.23})$$

$$K_{i,0} = \frac{2*\frac{1}{n}}{V_{MIN}} * \left[\frac{\partial C^*(u_i, 0)}{\partial u} - \frac{\partial C^*(u_i, v_1)}{\partial u} \right] \geq -3 \Rightarrow \frac{\partial C^*(u_i, v_1)}{\partial u} \leq \frac{3*V_{MIN}*n}{2} \quad \forall i \in \{0, 1, \dots, n\} \quad (\text{AVII.24})$$

Del mismo modo, Como $K_{i,m-1}$ debe pertenecer a la región factible,

$$K_{i,m-1} = \frac{2*\frac{1}{n}}{V_{MIN}} * \left[\frac{\partial C^*(u_i, v_{m-1})}{\partial u} - \frac{\partial C^*(u_{i,1})}{\partial u} \right] \leq 0 \Rightarrow \left[\frac{\partial C^*(u_i, v_{m-1})}{\partial u} - 1 \right] \leq 0 \Rightarrow$$

$$\left[\frac{\partial C^*(u_i, v_{m-1})}{\partial u} \right] \leq 1 \quad \forall i \in \{0, 1, \dots, n\}$$

(AVII25)

$$K_{i,m-1} = \frac{2*\frac{1}{n}}{V_{MIN}} * \left[\frac{\partial C^*(u_i, v_{m-1})}{\partial u} - \frac{\partial C^*(u_{i,1})}{\partial u} \right] \geq -3 \Rightarrow \left[\frac{\partial C^*(u_i, v_{m-1})}{\partial u} - 1 \right] \geq \frac{-3*V_{MIN}*n}{2} \Rightarrow$$

$$\left[\frac{-\partial C^*(u_i, v_{m-1})}{\partial u} \right] \leq -1 + \frac{3*V_{MIN}*n}{2} \quad \forall i \in \{0, 1, \dots, n\}$$

(AVII.26)

En función de (AVII.24) y (AVII.26) se tiene que:

$$\frac{\partial C^*(u_i, v_1)}{\partial u} - \frac{\partial C^*(u_i, v_{m-1})}{\partial u} \leq \frac{3*V_{MIN}*n}{2} - 1 + \frac{3*V_{MIN}*n}{2} = -1 + 3*V_{MIN}*n \quad \forall i \in \{0, 1, \dots, n\} \quad (\text{AVII.27})$$

Por otra parte, teniendo en cuenta la cota inferior para el resto de las K 's asociadas a la abscisa “i”, se tiene que:

$$K_{i,1} = \frac{2*\frac{1}{n}}{V_{MIN}} * \left[\frac{\partial C^*(u_i, v_1)}{\partial u} - \frac{\partial C^*(u_i, v_2)}{\partial u} \right] \geq -3 \Rightarrow \left[\frac{\partial C^*(u_i, v_1)}{\partial u} - \frac{\partial C^*(u_i, v_2)}{\partial u} \right] \geq \frac{-3*V_{MIN}*n}{2} \quad \forall i \in \{0, 1, \dots, n\} \quad (\text{AVII.28})$$

$$K_{i,2} = \frac{2 * \frac{1}{n}}{V_{MIN}} * \left[\frac{\partial C^*(u_i, v_2)}{\partial u} - \frac{\partial C^*(u_i, v_3)}{\partial u} \right] \geq -3 \Rightarrow \left[\frac{\partial C^*(u_i, v_2)}{\partial u} - \frac{\partial C^*(u_i, v_3)}{\partial u} \right] \geq \frac{-3 * V_{MIN} * n}{2} \quad \forall i \in \{0, 1, \dots, n\} \quad (\text{AVII.29})$$

...

$$K_{i,m-2} = \frac{2 * \frac{1}{n}}{V_{MIN}} * \left[\frac{\partial C^*(u_i, v_{m-2})}{\partial u} - \frac{\partial C^*(u_i, v_{m-1})}{\partial u} \right] \geq -3 \Rightarrow \left[\frac{\partial C^*(u_i, v_{m-2})}{\partial u} - \frac{\partial C^*(u_i, v_{m-1})}{\partial u} \right] \geq \frac{-3 * V_{MIN} * n}{2} \quad \forall i \in \{0, 1, \dots, n\}$$

(AVII.30)

La suma de estas $m-2$ expresiones da lugar a

$$\left[\frac{\partial C^*(u_i, v_1)}{\partial u} - \frac{\partial C^*(u_i, v_{m-1})}{\partial u} \right] \geq \frac{-3 * V_{MIN} * n * (m-2)}{2} \quad \forall i \in \{0, 1, \dots, n\} \quad (\text{AVII.31})$$

A partir de las expresiones (AVII.27) y (AVII.31) se obtiene la relación

$$\frac{-3 * V_{MIN} * n * (m-2)}{2} \leq \left[\frac{\partial C^*(u_i, v_1)}{\partial u} - \frac{\partial C^*(u_i, v_{m-1})}{\partial u} \right] \leq -1 + 3 * V_{MIN} * n \quad \forall i \in \{0, 1, \dots, n\} \quad (\text{AVII.32})$$

y en consecuencia a que
$$\frac{-3 * V_{MIN} * n * (m-2)}{2} \leq -1 + 3 * V_{MIN} * n \Rightarrow V_{MIN} \geq \frac{2}{3 * m * n} \quad (\text{AVII.33})$$

Así, por cualquiera de las vías ((AVII.11), (AVII.22) y (AVII.33)) se llega a que la condición necesaria para poder encontrar solución al problema (4.26) es que $V_{MIN} \geq \frac{2}{3 * m * n}$ o lo que es lo mismo, que C^* sea 2-subcópula- V_{MIN} con

$$V_{MIN} = \frac{2}{3 * m * n} .$$

⇔) Supongamos ahora que C^* es 2-subcópula- V_{MIN} con $V_{MIN} = \frac{2}{3 * m * n} .$

Sea entonces C^* la solución óptima al problema de programación fraccional dada por

$$c_{ij}^* = C^*\left(\frac{i}{n}, \frac{j}{m}\right) \quad \forall i \in \{0, 1, 2, \dots, n\} \quad \forall j \in \{0, 1, 2, \dots, m\} .$$

Veamos que se puede definir $C_{FERGUSON}$ dada por (4.20)

cuyo valor para las derivadas de primer orden respecto de U y V viene dado por

$$\frac{\partial C_{FERGUSON}}{\partial u}\left(\frac{i}{n}, \frac{j}{m}\right) = \frac{j}{m}, \quad \frac{\partial C_{FERGUSON}}{\partial v}\left(\frac{i}{n}, \frac{j}{m}\right) = \frac{i}{n} \quad \forall i \in \{0, 1, 2, \dots, n\} \quad \forall j \in \{0, 1, 2, \dots, m\} \quad (\text{AVII.34})$$

Es decir, vamos a ver que con esta definición se van a dar las condiciones para que se cumplan los lemas (4.2), (4.3) y (4.5). Efectivamente,

-
$$\frac{\partial C_{FERGUSON}}{\partial u}\left(\frac{i}{n}, 0\right) = 0 \quad \forall i \in \{0, 1, 2, \dots, n\} \quad \text{y} \quad \frac{\partial C_{FERGUSON}}{\partial v}\left(0, \frac{j}{m}\right) = 0 \quad \forall j \in \{0, 1, 2, \dots, m\}$$
 (se cumple el lema (4.2))

$$- \frac{\partial C_{FERGUSON}}{\partial u} \left(\frac{i}{n}, \frac{m}{m} = 1 \right) = \frac{m}{m} = 1 \quad \forall i \in \{0, 1, 2, \dots, n\} \quad y$$

$$\frac{\partial C_{FERGUSON}}{\partial v} \left(\frac{n}{n} = 1, \frac{j}{m} \right) = \frac{n}{n} = 1 \quad \forall l \in \{0, 1, 2, \dots, m\} \quad (\text{se cumple también el lema (4.3)})$$

- Proporcionan una solución factible al problema (4.26) dado que, teniendo en cuenta que $u_{i+1} - u_i = \frac{1}{n}$,

$$\text{y que } \left[\frac{\partial C_{FERGUSON}}{\partial u} \left(\frac{i}{n}, \frac{j}{m} \right) - \frac{\partial C_{FERGUSON}}{\partial u} \left(\frac{i}{n}, \frac{j+1}{m} \right) \right] = \frac{j}{m} - \frac{j+1}{m} = \frac{-1}{m} \quad (\text{AVII.35})$$

$$\text{se tiene que } \frac{2*(u_{i+1} - u_i)}{V_{MIN}} * \left[\frac{\partial C_{FERGUSON}}{\partial u} \left(\frac{i}{n}, \frac{j}{m} \right) - \frac{\partial C_{FERGUSON}}{\partial u} \left(\frac{i}{n}, \frac{j+1}{m} \right) \right] = \frac{2*\frac{1}{n}}{V_{MIN}} * \frac{-1}{m} \leq 0 \quad (\text{AVII.36})$$

Por otro lado, como consecuencia del hecho de ser 2-subcópula- V_{MIN} , y considerando (AVII.35) resulta que

$$\begin{aligned} V_{MIN} &\geq \frac{2}{3*m*n} \Rightarrow \frac{3*n*V_{MIN}}{2} \geq \frac{1}{m} \Rightarrow \frac{-1}{m} \geq \frac{-3*n*V_{MIN}}{2} \Rightarrow \\ &\left[\frac{\partial C_{FERGUSON}}{\partial u} \left(\frac{i}{n}, \frac{j}{m} \right) - \frac{\partial C_{FERGUSON}}{\partial u} \left(\frac{i}{n}, \frac{j+1}{m} \right) \right] \geq \frac{-3*n*V_{MIN}}{2} \Rightarrow \\ &\frac{2*\frac{1}{n} * \left[\frac{\partial C_{FERGUSON}}{\partial u} \left(\frac{i}{n}, \frac{j}{m} \right) - \frac{\partial C_{FERGUSON}}{\partial u} \left(\frac{i}{n}, \frac{j+1}{m} \right) \right]}{V_{MIN}} \geq -3 \Rightarrow \\ &\frac{2*(u_{i+1} - u_i) * \left[\frac{\partial C_{FERGUSON}}{\partial u} \left(\frac{i}{n}, \frac{j}{m} \right) - \frac{\partial C_{FERGUSON}}{\partial u} \left(\frac{i}{n}, \frac{j+1}{m} \right) \right]}{V_{MIN}} \geq -3 \end{aligned}$$

(AVII.37)

La conjunción de (AVII.36) y (AVII.37) da lugar a la primera de las restricciones asociadas a las derivadas respecto de U del problema (4.26).

Para la segunda, basta observar que, igualmente,

$$\left[\frac{\partial C_{FERGUSON}}{\partial u} \left(\frac{i+1}{n}, \frac{j}{m} \right) - \frac{\partial C_{FERGUSON}}{\partial u} \left(\frac{i+1}{n}, \frac{j+1}{m} \right) \right] = \frac{j}{m} - \frac{j+1}{m} = \frac{-1}{m} \quad (\text{AVII.38})$$

con lo cual, los desarrollos que conducen a las acotaciones (AVII.36) y (AVII.37) son exactamente los mismos.

Respecto de las restricciones asociadas a las derivadas en V ,

$$\left[\frac{\partial C_{FERGUSON}}{\partial v} \left(\frac{i}{n}, \frac{j}{m} \right) - \frac{\partial C_{FERGUSON}}{\partial v} \left(\frac{i+1}{n}, \frac{j}{m} \right) \right] = \frac{i}{n} - \frac{i+1}{n} = \frac{-1}{n} \quad (\text{AVII.39})$$

$$\text{y } \left[\frac{\partial C_{FERGUSON}}{\partial v} \left(\frac{i}{n}, \frac{j+1}{m} \right) - \frac{\partial C_{FERGUSON}}{\partial v} \left(\frac{i+1}{n}, \frac{j+1}{m} \right) \right] = \frac{i}{n} - \frac{i+1}{n} = \frac{-1}{n} \quad (\text{AVII.40})$$

donde para cualquiera de ellas se puede hacer un razonamiento análogo al realizado para las derivadas en U .

Teniendo en cuenta que $v_{j+1} - v_j = \frac{1}{m}$ y procediendo por ejemplo a partir de (AVII.39) se tiene que:

$$\frac{2*(v_{j+1} - v_j)}{V_{MIN}} * \left[\frac{\partial C_{FERGUSON}}{\partial u} \left(\frac{i}{n}, \frac{j}{m} \right) - \frac{\partial C_{FERGUSON}}{\partial u} \left(\frac{i+1}{n}, \frac{j}{m} \right) \right] = \frac{2*\frac{1}{m}}{V_{MIN}} * \frac{-1}{n} \leq 0 \quad (\text{AVII.41}).$$

La conjunción de esta última expresión con la hipótesis de que C^* es 2-subcópula- V_{MIN} permite obtener la otra acotación,

$$\begin{aligned} V_{MIN} \geq \frac{2}{3*m*n} &\Rightarrow \frac{3*m*V_{MIN}}{2} \geq \frac{1}{n} \Rightarrow \frac{-1}{n} \geq \frac{-3*m*V_{MIN}}{2} \Rightarrow \\ \left[\frac{\partial C_{FERGUSON}}{\partial u} \left(\frac{i}{n}, \frac{j}{m} \right) - \frac{\partial C_{FERGUSON}}{\partial u} \left(\frac{i+1}{n}, \frac{j}{m} \right) \right] &\geq \frac{-3*m*V_{MIN}}{2} \Rightarrow \\ \frac{2*\frac{1}{m} * \left[\frac{\partial C_{FERGUSON}}{\partial u} \left(\frac{i}{n}, \frac{j}{m} \right) - \frac{\partial C_{FERGUSON}}{\partial u} \left(\frac{i+1}{n}, \frac{j}{m} \right) \right]}{V_{MIN}} &\geq -3 \Rightarrow \\ \frac{2*(v_{j+1} - v_j) * \left[\frac{\partial C_{FERGUSON}}{\partial u} \left(\frac{i}{n}, \frac{j}{m} \right) - \frac{\partial C_{FERGUSON}}{\partial u} \left(\frac{i+1}{n}, \frac{j}{m} \right) \right]}{V_{MIN}} &\geq -3 \end{aligned}$$

(AVII.42).

De esta forma, los vectores tangentes en el dominio de definición de la subcópula definidos por (AVII.34) constituyen una solución factible del problema de programación matemática (4.26) y por tanto se cumplen las condiciones del lema (4.5).

Así, hemos encontrado una relación de valores para las derivadas primeras (AVII.34) que satisfacen las condiciones suficientes establecidas en los lemas (4.2), (4.3) y (4.5). Esto permite asegurar que la función $C_{FERGUSON}$ construida por interpolación cúbica de C^* es una cópula, quedando completada así la demostración de este teorema.

c.q.d

ANEXO VIII: Contrastes de bondad de ajuste a una distribución dada

Los test de bondad de ajuste nos van a permitir contrastar si una muestra de “ n ” valores, $\{x_1, \dots, x_n\}$ procede de una distribución conocida F_0 . En otras palabras, si F es la verdadera función de distribución de la muestra, el test que se plantea es:

$$\begin{aligned} H_0: F &= F_0 \\ H_1: F &\neq F_0 \end{aligned}$$

Para llevarlo a cabo, se trocea el rango de variación de la distribución a contrastar en una serie de clases y se comprueba si el número de valores muestrales observados en cada una de las clases (O_i) se parece al número de ellos que cabría esperar (E_i) bajo el supuesto de que se distribuyeran según F_0 .

Para distribuciones unidimensionales estas categorías serían intervalos, mientras que por ejemplo para distribuciones bidimensionales (2-cóputas), serían rectángulos.

Formalmente, según se explica por ejemplo en [NEWBOLD], supuesta una partición en K clases, la hipótesis nula

vendría a especificar las probabilidades p_1, p_2, \dots, p_K (con $\sum_{i=1}^K p_i = 1$) de que una observación x_i (o x_{ij} en el caso bidimensional) se encuentre en cada categoría.

Dado que hay “ n ” observaciones muestrales, el número esperado de datos en cada categoría bajo H_0 sería $E_i = n * p_i$. Este valor representa la probabilidad de que una observación muestral se encuentre en la clase “ i ” siendo tales observaciones las que van a permitir contrastar la hipótesis. Si el número de valores muestrales observados en cada categoría es muy próximo al que se espera, entonces existirá evidencia a favor de la hipótesis nula y se podrá decir que los datos se ajustan a la distribución de probabilidad supuesta F_0 .

Una forma obvia de medir esta proximidad es comparando la diferencia entre ambas cantidades en concepto de distancias, es decir, en valor absoluto o bien en términos cuadráticos para que no se produzca un efecto de compensación entre unas y otras que pudiera conducir a una conclusión errónea.

En términos cuadráticos, el **contraste de la Chi-cuadrado** establece esta comparativa a través del estadístico de **Pearson**, el cual responde a la expresión:

$$P = \frac{\sum_{i=1}^k (O_i - E_i)^2}{E_i} \quad (\text{AVIII.1})$$

Efectivamente, este estadístico cuantifica la diferencia entre el número de valores muestrales en una clase “ i ” (O_i) y el valor esperado en ella (E_i). Por ello, valores bajos del estadístico de Pearson conducirán a no rechazar la hipótesis nula.

Se puede demostrar que $P \sim \chi^2_{K-1, \alpha}$ y así, cuando $P > \chi^2_{K-1, \alpha}$ dicha hipótesis es rechazada.

Este contraste necesita muestras de gran tamaño y su potencia crece conforme crece el número de clases. Existe un convenio empírico que aconsejan el empleo de este contraste cuando el tamaño muestral es mayor que 30, el número de intervalos considerados (clases) es mayor que 5 y la frecuencia teórica de cada clase (E_i) es mayor o igual que 5 (de no ser así, se recomienda reagrupar las clases).

Alternativamente, cuando el tamaño de la muestra no es muy grande ($n \sim 30$), se utiliza el **contraste de Kolmogorov-Smirnov** que mide la distancia a la que hacíamos referencia en términos absolutos. En este caso, el estadístico del contraste representa la máxima diferencia entre la función de distribución empírica de la muestra $F_n(x)$ y la distribución teórica a contrastar, es decir,

$$D_n = \sup_{x \in \mathbb{R}} |F_n(x) - F_0(x)| \quad (\text{AVIII.2})$$

Kolmogorov y Smirnov estudiaron la distribución asintótica de este estadístico. Si H_0 es cierta ($F = F_0$), según el teorema de **Glivenko-Cantelli** se tiene que

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \xrightarrow{c.s.} 0 \quad (\text{AVIII.3})$$

Por tanto, valores bajos de D_n conducen a “admitir” que los datos se ajustan a la distribución de probabilidad supuesta F_0 o más estrictamente hablando, a no rechazar la hipótesis H_0 .

ANEXO IX: Contraste de Kruskal-Wallis

El test de **Kruskal-Wallis** permite realizar un análisis de la varianza desde la perspectiva no paramétrica. Según puede consultarse por ejemplo en el capítulo 15 de [NEWBOLD], el contraste plantea si “ k ” muestras $\{X_{ij}\}_{\{j=1,2,\dots,n_i\}}$ $\forall i \in \{1,2,\dots,k\}$ corresponden a distribuciones con una misma media, es decir,

$$\begin{aligned} H_0: \mu_1 = \mu_2 = \dots = \mu_k \\ H_1: \mu_i \neq \mu_j \text{ para algunos } i, j \in \{1,2,\dots,k\} \end{aligned} \quad (\text{AIX.1})$$

o equivalentemente, si μ define la media global de las “ k ” poblaciones combinadas y F_i la diferencia entre dicho valor y la media de la población del i -ésimo grupo ($F_i = \mu_i - \mu \quad \forall i \in \{1,2,\dots,k\}$),

$$H_0: F_1 = F_2 = \dots = F_k = 0 \quad (\text{AIX.2})$$

Bajo el enfoque no paramétrico, este problema se transforma en realizar el test respecto de los rangos $\{R_{ij}\}$ asociados a las observaciones muestrales $\{X_{ij}\}$. La ventaja respecto del test ANOVA es que únicamente presupone la aleatoriedad en la extracción de las muestras sin hacer referencia a las hipótesis de homocedasticidad y normalidad que éste precisa. Por ello, acostumbra a utilizarse cuando el analista intuye que la población que está analizando tiene una distribución significativamente diferente de la Normal.

Los pasos a seguir para realizar este contraste son:

1. Ordenar las $\sum_{i=1}^k n_i = n$ observaciones $\{X_{ij}\}$ de menor a mayor asignando a cada una de ellas su rango $\{R_{ij}\}$ (posición). En caso de existir datos repetidos, se puede proceder de la siguiente manera:

Supongamos que tenemos 4 datos repetidos que ocupan las posiciones 2, 3, 4 y 5, es decir,

18 (1), 22.3 (2), 22.3 (3), 22.3 (4), 22.3 (5), 25 (6),...

Entonces, se puede tomar la media aritmética de los rangos $\frac{2+3+4+5}{4} = 3.5$ y asignar a los 4 datos este rango. El dato sexto y los sucesivos (salvo nueva repetición), conservarían su rango.

2. Para cada una de las “ k ” muestras, calcular R_i como la suma de los rangos $\{R_{ij}\}$ de sus observaciones (también en orden creciente). La hipótesis H_0 será falsa si estas cantidades son muy diferentes.

3. Calcular el estadístico del contraste de las medias poblacionales

$$K - W = \frac{12}{n * (n + 1)} \sum_{i=1}^k \frac{R_i^2}{n_i} - 3 * (n + 1) \quad (\text{AIX.3})$$

que, bajo H_0 , se distribuye según una Chi-cuadrado con $k-1$ grados de libertad

4. Comparar el valor del estadístico con su valor teórico $\chi_{k-1, \alpha}^2$ y rechazar H_0 si $K - W > \chi_{k-1, \alpha}^2$.

ANEXO X: Análisis mediante wavelets

Las **wavelets** son unas herramientas matemáticas utilizadas habitualmente en el ámbito de las series temporales (eliminación de ruido, análisis de señales geológicas y sísmicas, etc.) y en el de procesamiento de imágenes (lectura de huellas digitales, reconocimiento de firmas, decodificación de señales de sonido, etc.). Su traducción exacta sería “onda pequeña” (diminutivo del término inglés *wave* que significa “onda” u “ola”), aunque normalmente las veremos referenciadas en nuestra literatura a través de las palabras “*ondícula*”, “*ondeleta*” u “*ondita*”. De acuerdo a su nombre, las wavelets son ondas que crecen y decaen en un periodo de tiempo limitado en comparación con las funciones seno y coseno que mantienen su oscilación a lo largo de toda la recta real. La mayor parte de las explicaciones que a continuación se presentan pueden ser encontradas en los capítulos primero y cuarto de [PERWAL].

Analíticamente, una wavelet es una función integrable y oscilatoria ψ cuya media es cero, o lo que es lo mismo, una función que satisface las dos propiedades siguientes:

1. Su integral es cero, es decir,
$$\int_{-\infty}^{\infty} \psi(u) du = 0 \quad (\text{AX.1})$$

Esta primera condición asegura que cualquier actividad de la wavelet en el eje real positivo se verá compensada por otra de signo contrario dándole así el aspecto de una **onda**.

2. Es cuadrado-integrable, es decir,
$$\int_{-\infty}^{\infty} \psi^2(u) du = 1 \quad (\text{AX.2})$$

Esta segunda propiedad implica que para cierto $0 < \epsilon < 1$, se podrá determinar un intervalo $[-T, T]$ tal que

$$\int_{-T}^T \psi^2(u) du < 1 - \epsilon \quad \text{y así,} \quad \int_{-T}^T \psi^2(u) du \approx 1 \quad \text{y} \quad \int_{(-\infty, -T)}^{(T, \infty)} \psi^2(u) du \approx 0 \quad \text{cuando} \quad \epsilon \rightarrow 0.$$

Así, la actividad de la wavelet está restringido a un intervalo de tiempo **pequeño** en comparación con la recta real.

En el contexto de las series temporales, estas funciones miden cómo varía el valor medio de una serie entre dos periodos de longitud λ . Valores de λ suficientemente grandes permiten detectar comportamientos suaves o de baja frecuencia (en los históricos de demanda de gas que se presenta en las aplicaciones prácticas de esta tesis irían referidos a periodos de 365 días que definen el ciclo anual) mientras que valores pequeños de dicho parámetro identifican cambios rápidos (de una semana a otra) y, en el caso extremo, una señal de error (de un día a otro).

A continuación se presenta la idea intuitiva que justifica esta cualidad de las wavelets para medir variaciones del nivel en que se mueve una función.

El valor medio de una función $x(\cdot)$ sobre un intervalo $[a, b]$ viene dado por
$$\alpha(a, b) = \frac{1}{(b-a)} \int_a^b x(u) du \quad (\text{AX.3}).$$

En consecuencia, fijado un instante de tiempo “ t ”, el valor medio de la función en el intervalo de longitud λ centrado

en dicho instante vendrá dado por
$$A(\lambda, t) \equiv \alpha \left(t - \frac{\lambda}{2}, t + \frac{\lambda}{2} \right) = \frac{1}{\lambda} \int_{t-\frac{\lambda}{2}}^{t+\frac{\lambda}{2}} x(u) du \quad (\text{AX.4}).$$

La variación de este valor medio entre dos intervalos consecutivos $(t-\lambda, t)$ y $(t, t+\lambda)$ vendrá dada por

$$D(\lambda, t) = A(\lambda, t + \frac{\lambda}{2}) - A(\lambda, t - \frac{\lambda}{2}) = \frac{1}{\lambda} \int_t^{t+\lambda} x(u) du - \frac{1}{\lambda} \int_{t-\lambda}^t x(u) du = \int_{-\infty}^{\infty} \psi_{\lambda,t}(u) x(u) du \quad (\text{AX.5})$$

donde
$$\psi_{\lambda,t}(u) = \begin{cases} \frac{-1}{\lambda} & \text{si } t-\lambda < u \leq t \\ \frac{1}{\lambda} & \text{si } t < u \leq t+\lambda \end{cases} \quad (\text{AX.6})$$

que para $\lambda = 1$ y $t = 0$, es proporcional a la primera wavelet que encontramos en la literatura, debida a **Alfred Haar**

(principios del siglo XX), y que responde a la expresión
$$\psi^{HAAR}(u) = \begin{cases} \frac{-1}{\sqrt{2}} & \text{si } -1 < u \leq 0 \\ \frac{1}{\sqrt{2}} & \text{si } 0 < u \leq 1 \end{cases} \quad (\text{AX.7}).$$

.Dado que $\psi^{HAAR}(t) = \frac{1}{\sqrt{2}} * \psi_{1,0}(t) \Rightarrow \psi^{HAAR}(t) \propto \psi_{1,0}(t)$, de la misma manera se implica que

$$W^{HAAR}(1,0) = \int_{-\infty}^{\infty} \psi^{HAAR}(u) x(u) du \propto \int_{-\infty}^{\infty} \psi_{1,0}(u) x(u) du = D(1,0)$$

La extrapolación de la función de Haar a un instante arbitrario “ t ” y una escala cualquiera λ proporciona la wavelet

$$\psi_{\lambda,t}^{HAAR}(u) = \begin{cases} \frac{-1}{\sqrt{2}\lambda} & \text{si } t-\lambda < u \leq t \\ \frac{1}{\sqrt{2}\lambda} & \text{si } t < u \leq t+\lambda \end{cases} \quad (\text{AX.8})$$

que permite, de igual forma, definir la familia de funciones $W^{HAAR}(\lambda, t) = \int_{-\infty}^{\infty} \psi_{\lambda,t}^{HAAR}(u) x(u) du \propto D(\lambda, t)$.

Se define λ como el **parámetro de escalamiento o dilatamiento** y “ t ” como el de **translación o desplazamiento**.

La colección de variables $\{W^{HAAR}(\lambda, t) : \lambda > 0, -\infty < t < \infty\}$ recibe el nombre de **transformada wavelet continua (CWT) de Haar** de la función $x(\cdot)$. Esta transformada es proporcional a la diferencia entre dos medias adyacentes de escala λ , la primera comenzando en el instante “ t ” (sobre el intervalo $(t, t+\lambda)$) y la segunda terminando en dicho instante (sobre $(t-\lambda, t)$) quedando así justificado el hecho de que estas funciones permitan medir la variación del valor medio de una señal entre dos periodos de tiempo consecutivos.

A través de otro tipo de funciones wavelets se puede hacer que las medias calculadas sobre dichos intervalos se encuentren ponderadas en función de su proximidad a un punto interior de los mismos. Por ejemplo las CWT asociadas a wavelets *gaussianas*, las cuales se obtienen a partir de la derivación de la función de densidad gaussiana, pueden ser proporcionales a la diferencia entre dos medias ponderadas asociadas a periodos consecutivos. Las transformadas continuas correspondientes a otras wavelets de esta familia o de alguna otra como la popularmente conocida como “*sombrero mexicano*” (segunda derivada de la función de densidad gaussiana) recogen la variación entre una media ponderada asociada a un periodo de tiempo y la media de otras dos calculadas sobre los dos intervalos circundantes a él. El siguiente gráfico ilustra la forma que presentan las funciones wavelets citadas.

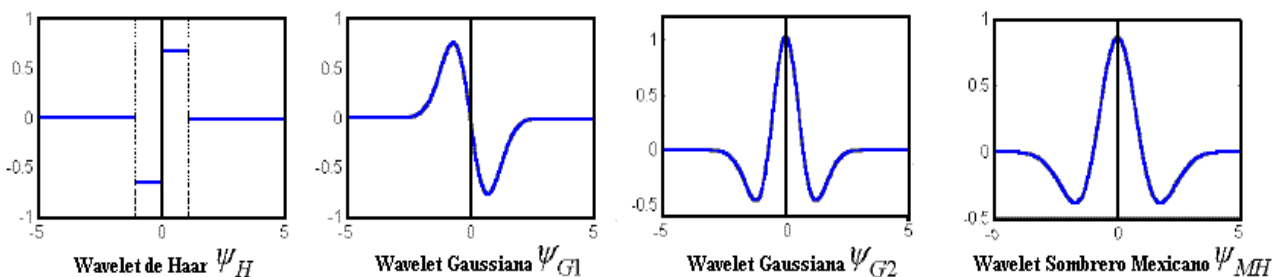


Ilustración AX.1: Algunas funciones wavelets

Además de las presentadas en dicho gráfico existen otras muchas clases de wavelets como las *daublets*, las *symmlets* o las *coiflets*, por citar algunos de los ejemplos más conocidos. Asociada a cada una de estas familias existen unos representantes o “*padres*” denominados “*wavelet padre*” y “*wavelet madre*” a partir de los cuales se pueden obtener todos los miembros restantes mediante operaciones de escalado (o dilatación) y translación (o desplazamiento).

De igual manera que sucede con las funciones cópula, uno de los problemas principales consiste en determinar qué familia de wavelets puede ser más apropiada para el tratamiento de una determinada señal en función de las características de ésta. De hecho, una parte importante de la investigación actual en este campo esta focalizada en la construcción de funciones wavelets (“*cabezas de familia*”) que proporcionen una descripción informativa y eficiente de la señal que se desea analizar en función de las propiedades de ésta. Esto se traduce normalmente en la búsqueda de aquellas restricciones a imponer sobre las funciones wavelets (adicionales a (AX.1) y (AX.2)) para poder realizar con ellas un análisis útil. Una de las propiedades que goza de mayor popularidad, garantiza la reconstrucción de una señal $x(\cdot)$ a partir de la CWT asociada a una wavelet determinada ψ , si su transformada de Fourier, dada por

$$\psi(f) = \int_{-\infty}^{\infty} \psi(u) e^{-i*2*\pi*f*u} du \quad , \text{ verifica que } 0 < C_{\psi} = \int_0^{\infty} \frac{|\psi(u)|^2}{f} df < \infty \quad (\text{propiedad de admisibilidad}) \text{ y la señal}$$

$$x(\cdot) \text{ satisface que } \int_{-\infty}^{\infty} x^2(t) dt < \infty \quad .$$

La reconstrucción se puede conseguir a través de la expresión

$$x(t) = \frac{1}{C} \int_0^{\infty} \left[\int_{-\infty}^{\infty} W(\lambda, t) * \frac{1}{\sqrt{\lambda}} * \psi\left(\frac{t-u}{\lambda}\right) du \right] * \frac{1}{\lambda^2} d\lambda$$

y la consecuencia directa es que la CWT contiene la misma información que la función $x(\cdot)$ pero presentada de forma diferente, descompuesta como suma de funciones que se mueven a diferentes velocidades o frecuencias. Así, de la misma manera que el análisis de Fourier consiste en descomponer una señal como una combinación de funciones senos que se mueven a diferentes frecuencias, el análisis mediante wavelets permite la descomposición en versiones escaladas y trasladadas de una wavelet madre.

Si bien la CWT, como su propio nombre indica, permite el tratamiento de señales continuas, existe una colección de variables equivalente para la descomposición de funciones discretas definidas sobre un conjunto finito de instantes “ t ”, ψ 's, ϕ , como son las series temporales y que van a ser las que verdaderamente nos van a interesar. Así, a través de la **transformada wavelet discreta** (DWT) podemos expresar una serie temporal como combinación lineal de funciones wavelets ψ 's, ϕ que se obtienen por dilatación o translación de sus “padres” y que miden la variación de la señal $x(\cdot)$ a distintas escalas. Las wavelets más relacionadas con su “padre” (ϕ 's) van asociada a las partes más suaves ($S = suave$ o *del inglés smooth*) o de baja frecuencia de la serie y en contraposición, aquellas que heredan las características de la “wavelet madre” (ψ 's) resultan apropiadas para la representación de las partes más detalladas ($d = detalles$ o *del inglés details*) o componentes de más alta frecuencia. Es decir, los “padres” de una clase generan una base de $L^2(\mathbb{R})$ dada por una relación de funciones wavelets en función de las cuales es posible representar una señal.

$$x(t) = \sum_k S_{J,k} \phi_{J,k}(t) + \sum_k d_{J,k} \psi_{J,k}(t) + \sum_k d_{J-1,k} \psi_{J-1,k}(t) + \dots + \sum_k d_{1,k} \psi_{1,k}(t) \quad (\text{AX.9})$$

Los coeficientes de esta combinación van asociados a un instante concreto “ t ” y a una escala de orden 2^j (**escalas “dyadic”**). Por ejemplo, el coeficiente $d_{j,k}$ informa sobre el comportamiento de la función $x(\cdot)$ cerca del instante

$\frac{k}{2^j}$ en la escala $\frac{1}{2^j}$. De esta forma, las wavelets consiguen adaptarse a características locales de la señal $x(\cdot)$.

En forma matricial la expresión anterior se puede escribir como $X = \sum_{j=1}^J D_j + S_j$ (AX.10)

que define un **análisis multirresolución** (MRA) de la serie temporal X y que originalmente fue investigada por **Stéphane Mallat** e **Yves Meyer** (1989). A través de esta ecuación, la serie queda representada como suma de $J + 1$ vectores denominados “*crustales*” que vienen a ser componentes de la serie X que se mueven a diferentes escalas o niveles de resolución (de ahí el calificativo de “multirresolución”). En nuestro caso, los niveles de resolución más altos (S_6, D_6 , etc.) son los que emplearemos para la detección del comportamiento cíclico anual del histórico de demanda construido bajo unas supuestas condiciones “normales” de temperatura.

Otra aplicación del MRA podría ser llevar a cabo un blanqueado de la serie eliminando el ruido mediante la sustracción de la componente de más alta frecuencia (D_1) que vendría a representar las variaciones de consumo que se producen de un día al siguiente. Sobre nuestro histórico de datos original estas variaciones estarían relacionadas con, aparte obviamente del efecto del calendario (impacto de un día festivo respecto al anterior) con los cambios de temperatura que puedan provocar el que la demanda suba o baje de un día al siguiente.

ANEXO XI: Análisis mediante splines de regresión

Los **splines** son unas herramientas matemáticas que resultan muy útiles en muchas ramas de la ciencia como son el diseño gráfico, el procesamiento de señales, la búsqueda de soluciones numéricas a ecuaciones diferenciales, etc. Dentro del ámbito de las series temporales suelen ser utilizadas para la estimación de valores desconocidos no recogidos en el histórico (*missings*) o bien, como es nuestro caso, para realizar predicciones de un comportamiento básico (de baja frecuencia) subyacente a los datos. La gran popularidad que han ganado los splines se debe a la simplicidad con que pueden ser construidos y, sobre todo, al reducido coste computacional que conlleva su uso. La mayor parte de las explicaciones que a continuación se presentan pueden ser encontradas en el capítulo segundo de [GREEN].

Los splines son básicamente curvas definidas a trozos mediante polinomios que vienen a representar una aproximación de una función “*f*” cuyo valor se desconoce salvo en una relación de puntos de su dominio. El grado de estos polinomios suele ser bajo con la idea de que la aproximación a “*f*” no presente fuertes oscilaciones tan poco deseables y verosímiles. De hecho, como veremos, el proceso de construcción del spline se basa en buscar aquella curva “*g*” que minimice el error cuadrático respecto de “*f*” a la que vez que se penalice las posibles fluctuaciones que puedan producirse, con vistas a reducir al máximo el error resultante del ajuste (penaliza el sobreajuste). El spline que se utiliza más frecuentemente es el de grado 3 (spline cúbico) cuya definición analítica es la siguiente:

Sean t_1, t_2, \dots, t_n una serie de puntos dados en un intervalo $[a, b]$ (denominados nodos) y tales que $a < t_1 < t_2 < \dots < t_n < b$. Un **spline cúbico** es una función “*g*” definida en $[a, b]$ que satisface las dos condiciones siguientes:

1. En cada uno de los intervalos $(a, t_1), (t_1, t_2), \dots, (t_n, b)$, “*g*” es un polinomio cúbico
2. En cada t_i , “*g*” así como sus dos primeras derivadas (g' y g'') son continuas. Por tanto las tres funciones son continuas en todo el intervalo $[a, b]$

Cuando se verifica de manera adicional que la derivada segunda y tercera de “*g*” en “*a*”, y “*b*” toma el valor 0 (condiciones frontera naturales) el spline recibe la denominación de **spline cúbico natural**.

Para escribir la ecuación del spline, basta dar su definición en cada uno de los trozos en que ha sido dividido su dominio de definición. Por ejemplo, un spline definido sobre el intervalo $[a, b]$ de acuerdo a la partición anterior sería

$$g(t) = d_i * (t - t_i)^3 + c_i * (t - t_i)^2 + b_i * (t - t_i) + a_i \quad \forall t_i \leq t \leq t_{i+1}, i = 0, 1, \dots, n, t_0 = a, t_{n+1} = b \quad (\text{AXI.1})$$

La razón que justifica el empleo de este tipo de funciones como aproximación de una curva dada es la siguiente. Supongamos que tenemos una serie temporal $Y(t)$, es decir, la respuesta de una variable Y en una relación de instantes de tiempo “*t*”. Podemos plantear el ajuste de Y en función de “*t*” a través de la ecuación $Y(t) = g(t) + \epsilon(t)$ (AXI.2)

En el caso en el que $g(t)$ sea una recta nos encontraríamos con el típico modelo de regresión lineal. Sin embargo, si el ajuste de una recta resulta inapropiado, una alternativa clásicamente considerada consiste en proponer para “*g*” un

polinomio de orden bajo y proceder a la estimación de sus coeficientes a través del **método de los mínimos cuadrados**. Al utilizar esta técnica debe ponerse especial cuidado dado que siempre puede encontrarse un polinomio de orden $n-1$ que pase por los “ n ” pares $(t, Y(t))$ de manera que el error cuadrático medio es exactamente cero y por tanto mínimo. Para ello, bastaría considerar la poligonal que pasa por todos ellos, es decir, unir mediante líneas rectas todos los pares. Pero no es necesario tomar una función tan poco “estética” como la poligonal, de la cual conocemos su falta de diferenciabilidad en los puntos de unión. También es posible fijar unas condiciones de suavizado sobre éstos y construir otra curva distinta con derivada continua en todos los pares y que pase por todos ellos. Es decir puede disponerse de una curva suave “ g ” que interpola todos los puntos y cuyo error de ajuste es 0 y por tanto mínimo. El sobreajuste de una curva a unos datos consiste precisamente en esto, reducir inconscientemente el error sobre la muestra sin tener en cuenta que la función resultante no será buena para sacar conclusiones: “la interpolación abusiva de los valores muestrales se mueve en contra de la extrapolación del resultado a otras muestras distintas”. Por ello, una posible alternativa a la hora de estimar la curva es penalizar el sobreajuste, es decir, castigar la presencia de bruscas oscilaciones que buscan pasar lo más cerca posible de los pares muestrales. A este respecto existe el método de los **mínimos cuadrados penalizados** que estima aquella función “ g ” que minimiza la suma de cuadrados penalizados dada por la expresión

$$S(g) = \sum_{i=1}^n (Y_i - g(t_i))^2 + \alpha * \int_a^b \{g''(x)\}^2 dx \quad (\text{AXI.3}).$$

El primer término de la expresión anterior es el error cuadrático medio, cantidad que trata de minimizar el método de mínimos cuadrados. El segundo es una evaluación del grado de fluctuación de la curva “ g ”. La motivación de este término como medida de la oscilación de una curva, es que dos funciones g_1 y g_2 que difieran únicamente en una constante o en una función lineal, tengan la misma medida de oscilación, algo que parece deseable dentro del contexto de la regresión. Si bien podrían tomarse otros funcionales de g'' , el empleo de la integral del cuadrado de la función presenta ventajas importantes desde el punto de vista computacional.

El parámetro α cuantifica la importancia que se da a uno u otro término. Valores grandes de α conducen a curvas especialmente suaves pues el objetivo se centra en minimizar la medida de fluctuación. Por el contrario, valores pequeños quitan importancia a dicha medida y centran la atención en la minimización del error cuadrático medio, dando lugar a curvas más variables (más ajustadas a los datos).

Resulta que si $S[a,b]$ representa el conjunto de todas las funciones suaves en el intervalo $[a,b]$, el miembro de este conjunto que interpola los pares $(t_i, y(t_i))$ y minimiza el valor de $\int_a^b \{g''(x)\}^2 dx$ es un spline cúbico natural con nodos t_i . Es esta propiedad la que sustenta el empleo de los splines como métodos de aproximación de curvas.

Además, existe un teorema que demuestra que este spline puede ser siempre construido y para ello basta con resolver un sistema no singular de ecuaciones lineales para las derivadas de “ g ” en los nodos t_i . La complejidad lineal del problema ($O(n)$ operaciones aritméticas) se traduce en una sencillez computacional especialmente deseable para su implementación y aplicación.

Una de las principales utilidades que se saca al empleo de los splines, es la posibilidad de aproximar una función “ f ” a través de una combinación lineal de ellos (de una base de splines), planteando un modelo de regresión en el que participan como regresores para explicar la variable dependiente “ f ”.

A modo ilustrativo vamos a mostrar cómo podría llevarse a cabo la construcción de los elementos de la base en función de la cual quedaría representada nuestro histórico de demanda. La idea es mostrar la facilidad con que ésta pueda ser construida, propiedad que como decíamos hace especialmente atractivo el empleo de este tipo de funciones.

Dado que nuestra intención es que la curva suavizadora de nuestro histórico (el *Ciclo*) se comporte de forma periódica (anual), no debemos considerar una clase de splines cualquiera, sino que resulta conveniente considerar un espacio constituido por splines periódicos orientados a reflejar esta propiedad. Sea entonces,

$$S_p = \{s \in C^1[a, b] \mid t, q \quad s(t) = s(t+365), s|_{[t_{i-1}+365*k, t_{i+1}+365*k]} \in \mathbb{R}_3[t] \quad \forall i \in \{1, 2, \dots, n\}, k \in \mathbb{N}\} \quad .$$

Dado que S_p es hilbertiano por ser un subespacio de $L^2(a, b)$, una base del mismo está compuesta por funciones $\phi_i, \psi_i \in S_p$ que verifican las siguientes propiedades:

$$\phi_i(t_j) = \begin{cases} 0 & \text{si } j \neq i \\ 1 & \text{si } j = i \end{cases} ; \quad \phi_i'(t_j) = 0 \quad (\text{AXI.4})$$

$$\psi_i(t_j) = 0 ; \quad \psi_i'(t_j) = \begin{cases} 0 & \text{si } j \neq i \\ 1 & \text{si } j = i \end{cases} \quad (\text{AXI.5})$$

Bastaría determinar el valor que deben tomar los coeficientes de la ecuación (AXI.1), dada por

$$\phi(t), \psi(t) = d_i * (t - t_i)^3 + c_i * (t - t_i)^2 + b_i * (t - t_i) + a_i \quad \forall t_i \leq t \leq t_{i+1}$$

para garantizar la satisfacción de dichas propiedades.

En el primer caso, dos de los coeficientes se estiman de forma inmediata:

$$\phi_i(t_i) = a_i = 1 \quad (\text{AXI.6})$$

$$\phi_i'(t_i) = 3 * d_i * (t_i - t_i)^2 + 2 * c_i * (t_i - t_i) + b_i = 0 \Rightarrow b_i = 0 \quad (\text{AXI.7})$$

Los otros dos, se calculan a partir del valor de la función ϕ y su derivada en t_{i+1} :

$$\phi_i(t_{i+1}) = d_i * (t_{i+1} - t_i)^3 + c_i * (t_{i+1} - t_i)^2 + 1 = 0 \quad (\text{AXI.8})$$

$$\phi_i'(t_{i+1}) = 3 * d_i * (t_{i+1} - t_i)^2 + 2 * c_i * (t_{i+1} - t_i) = 0 \quad (\text{AXI.9})$$

La resolución del sistema proporciona los valores buscados:

$$c_i = \frac{-3}{(t_{i+1} - t_i)^2} \quad (\text{AXI.10})$$

$$d_i = \frac{2}{(t_{i+1} - t_i)^3} \quad (\text{AXI.11})$$

En consecuencia,

$$\phi_i(t) = \frac{2}{(t_{i+1} - t_i)^3} * (t - t_i)^3 - \frac{3}{(t_{i+1} - t_i)^2} * (t - t_i)^2 + 1 \quad \forall t_i \leq t \leq t_{i+1} \quad (\text{AXI.12})$$

En el segundo caso, dos de los coeficientes también se estiman de forma inmediata:

$$\psi_i(t_i) = a_i = 0 \quad (\text{AXI.13})$$

$$\psi_i'(t_i) = 3*d_i*(t_i - t_i)^2 + 2*c_i*(t_i - t_i) + b_i = 1 \Rightarrow b_i = 1 \quad (\text{AXI.14})$$

Los otros dos, se calculan igualmente a partir del valor de la función ψ y de su derivada en t_{i+1} :

$$\psi_i(t_{i+1}) = d_i*(t_{i+1} - t_i)^3 + c_i*(t_{i+1} - t_i)^2 + (t_{i+1} - t_i) = 0 \quad (\text{AXI.15})$$

$$\psi_i'(t_{i+1}) = 3*d_i*(t_{i+1} - t_i)^2 + 2*c_i*(t_{i+1} - t_i) + 1 = 0 \quad (\text{AXI.16})$$

La resolución del sistema proporciona los valores buscados:

$$c_i = \frac{-2}{t_{i+1} - t_i} \quad (\text{AXI.17})$$

$$d_i = \frac{1}{(t_{i+1} - t_i)^2} \quad (\text{AXI.18})$$

En consecuencia,

$$\psi_i(t) = \frac{1}{(t_{i+1} - t_i)^2} * (t - t_i)^3 - \frac{2}{(t_{i+1} - t_i)} * (t - t_i)^2 + (t - t_i) \quad \forall t_i \leq t \leq t_{i+1} \quad (\text{AXI.19})$$

Debe tenerse en cuenta que la variable “ t ” no crece de forma constante con el tiempo a lo largo de toda la recta real sino que lo hace exclusivamente dentro de cada intervalo $[1,365]$ definido sobre ella, es decir, los intervalos $[t_i, t_{i+1}]$ están contenidos siempre en el segmento $[1,365]$. Así, por ejemplo si el primer día que forma parte de nuestro histórico es el 1 de Abril de 1997, se le asociará a dicho día el valor $t = 1$, al día 2 de Abril de 1997 el valor 2, al día 31 de Marzo de 1998 el valor 365 y al día 1 de Abril de 1998 nuevamente el valor $t = 1$. Procediendo sucesivamente de esta forma, las expresiones (AXI.12) y (AXI.19) definen funciones periódicas que, se puede comprobar, verifican las propiedades (AXI.4) y (AXI.5) y definen por tanto una base.

En el caso en el que el número de trozos fuera 12, los valores a los que harían referencia los intervalos serían

$$t_i = \frac{12*i}{365} \quad \text{y cada función estaría centrada en el día 1 de cada mes abarcando el periodo de 60 días comprendido por}$$

los meses que empiezan y terminan dicho día.

Finalmente presentamos en un gráfico la forma que tienen algunos de los elementos de la base de S_p en función de los cuales se expresará la serie temporal que define el histórico de demanda. Cada par de funciones

$(\phi_i, \psi_i) \quad \forall i \in \{1, 2, \dots, 12\}$ y con él cada intervalo $[t_i, t_{i+1}] \quad \forall i \in \{1, 2, \dots, 12\}$, va referido a un mes concreto del año (Enero, Febrero, ..., Diciembre).

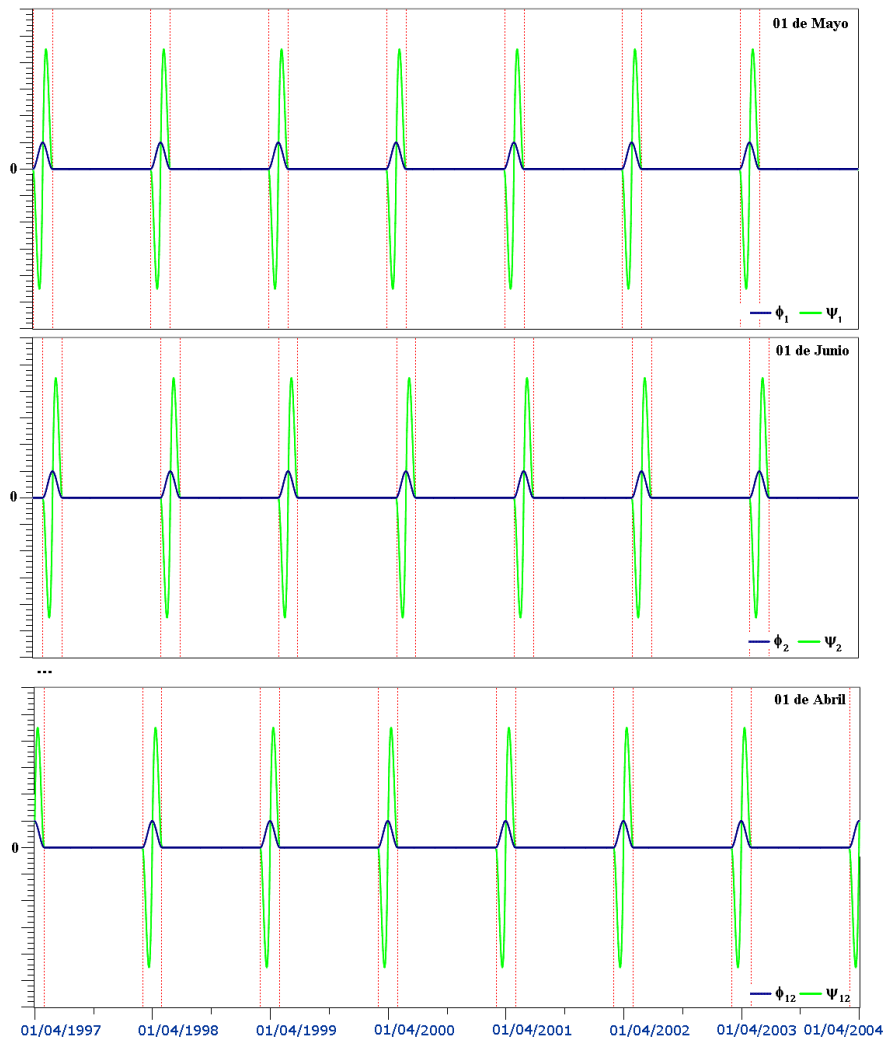


Ilustración AXI.1: Base de funciones splines utilizada para suavizar el histórico de demanda

Bibliografía

[ARMGAL], Margaret Armstrong y Alain Galli (2002), "Sequential non gaussian simulations using the FGM Copula", Cerna Working Paper.

[AZTRI], Antonio Aznar y Francisco Javier Trivez (1993), "Métodos de predicción en economía, Vol. II", 1ª ed., Barcelona, Ariel Economía.

[BOGASA], Eric Bouyé, Nicolas Gaussel y Mark Salmon (2001), "Investigating dynamic dependence using copulae", Financial Econometrics Research Center.

[DOSCH], Jadran Dobric y Friedrich Schmid (2005), "Testing goodness of fit for parametric families of copulas-an application to financial data", Communications in Statistics: Simulation and Computation, Vol. 34, N° 4, pp. 1053 – 1068.

[ENAGAS], Enagás, <http://www.enagas.es>.

[FERSCA], Jean-David Fermanian y Olivier Scaillet (2005), "Some statistical pitfalls in copula modelling for financial applications", Technical report.

[FIREAST], R.Fildes, A.Randall y P.Stubbs (1997), "One day ahead demand forecast in the utility industries: Two case studies", Journal of the Operational Research Society, Vol. 48, N° 1, pp. 15-24(10).

[FOVAFEHU], James D. Foley, Andries van Dam, Steven K. Feiner y John F. Hughes (1995), "Computer graphics. Principles and practice", 2ª ed., Addison-Wesley Publishing Company Inc.

[FRENERO], Gregory A. Fredricks, Roger B. Nelsen, José Antonio Rodríguez-Lallena (2005), "Copulas with fractal supports", Insurance: Mathematics and Economics 37, 42–48.

[GIDEDU], S.Gil, J. Deferrari y L.Duperron (2002), "Modelo generalizado de predicción de consumos de gas natural a mediano y corto plazo", Tercer Congreso Latinoamericano y del Caribe de gas y electricidad, Santa Cruz de la Sierra, Bolivia.

[GREEN], P.J. Green (1994), "Non parametric regression and generalized linear models" , London, Chapman and Hall.

[GUTIER], Ester Gutiérrez Moya (2003), "La demanda residencial de energía eléctrica en la comunidad autónoma de Andalucía: un análisis cuantitativo", Tesis doctoral Universidad de Sevilla: Facultad de Ciencias Económicas y Empresariales.

[HEIREN], Andréas Heinen y Erik Rengifo (2003), "Multivariate autoregressive modelling of time series count data using copulas", CORE Discussion Paper 2003/25, Université Catholique de Louvain.

[HILIE], Frederick S. Hillier y Gerald J. Lieberman (2001), "Investigación de operaciones", 7ª ed., México, McGraw-Hill.

[HUSASC], Matthew Hurd, Mark Salmonn y Christoph Schleicher (2005), "Using copulas to construct foreign exchange distributions with an application to the sterling exchange rate index", CEPR Discussion Papers 5114.

[JOE], Harry Joe (1997), "Multivariate Models and Dependence Concepts", London, Chapman and Hall.

[KETTLER], Paul C. Kettler y Fred Espen Benth (2006), "Dynamic copula models for the spark spread", Pure Mathematics, N° 14.

[LIU], Wei Liu (2005), "Distributions of currencies portfolio Return: A copula methodology".

[LYXUE], Ly Fie Sugianto y Xue-Bing Lu, "Demand forecasting in the deregulated market: a bibliography survey", School of Business Systems, Monash University 3800 Victoria, Australia.

[MATTEIS], Roberto De Matteis (2001), "Fitting copulas to data", Diploma Thesis, Institute of Mathematics of the University of Zurich.

[MEHRA], M.Mehra (2002), "Demand forecasting for electricity".

[MOLIFE], Rashed Molife (2003), "Using copulas as a measure of dependence between competing causes of mortality", London.

[MUHAJI], Muhammad Riaz Khan y Ajith Abraham (2003), "Short term load forecasting models in Czech Republic using soft computing paradigms", Journal-ref: International Journal of Knowledge-Based Intelligent Engineering Systems, IOS Press Netherlands, Vol. 7, N° 4, pp. 172-179.

[NELQUERO], R.B. Nelsen, J.J. Quesada-Molina y J.A.Rodríguez-Lallena (1997), "Bivariate copulas with cubic sections", Journal of nonparametric statistics, Vol. 7, pp. 205-220.

[NELSEN], R.B. Nelsen (1999), "An introduction to copulas", New York, Inc, Springer-Verlag.

[NEWBOLD], Paul Newold (1998), "Estadística para los negocios y la economía", 4ª ed., Prentice Hall.

[PARVAL], Leandro Pardo y Teófilo Valdés (1987), "Simulación. Aplicaciones prácticas en la empresa", Madrid, Diaz de Santos S. A.

[PATTON], Andrew J. Patton (2001), "Modelling time-varying exchange rate dependence using the conditional copula", Universidad de California de San Diego, Discussion Paper N° 01-09.

[PATTON2], Andrew J. Patton (2001), "Estimation of copula models for times series of possibly different lengths", Universidad de California de San Diego, Economics Working Paper Series, Número 2001-17.

[PEMEX], PEMEX: Gas y petroquímica básica, <http://www.gas.pemex.com>.

[PERWAL], Donal B. Percival y Andrew T. Walden (2000), "Wavelet methods for time series analysis", Cambridge, Cambridge University Press.

[PERWHI], Tony Perchard y Clive Whitehand (2000), "Short term gas demand forecasting", en Ellul I, ed. Proceedings of the 32nd Annual Meeting on Pipeline Simulation Interest Group. Savannah.

[QUEGAR], Vicente Quesada Paloma y Alfonso García Pérez (1985), "Curso básico de cálculo de probabilidades", Madrid, ICE.

[ROMANO], Claudio Romano (2002), "Calibrating and simulating copula functions: an application to the italian stock market", Working paper N° 12, CIDEM.

[SANSAT], Alessio Sancetta y Stephen Satchell (2004), "The Bernstein copula and its applications to modeling and approximations of multivariate distributions", *Econometric Theory* 20: 535–562.

[SASDOC], SAS Institute, <http://v8doc.sas.com/sashtml>.

[SEGRS], Johan Segers (2004), "Extreme-Value Copulas".

[SIXTO], Sixto Ríos Insua (1993), "Investigación operativa: optimización", 2ª ed., Madrid, Centro de Estudios Ramos Areces, S.A.

[VACHUNTA], Vanaja Iyer, Chun Che Fung y Tamas Gedeon (2003), "A fuzzy neural approach to electricity load and spot price forecasting in a deregulated electricity", *TENCON 2003: Conference on Convergent Technologies for Asia-Pacific Region*, Vol. 4, pp. 1479-1482.

[VANLAM], F. Vandenhende y P. Lambert (2002), "Local dependence estimation using non-parametric archimedian copulas".

[VENTER], Gary G. Venter (2001), "Tails of copulas", en *Proceedings ASTIN Washington, USA*, pp. 68–113.