Métodos de Predicción para Series Temporales de Intervalos e Histogramas

Tesis para la obtención del título de Doctor realizada por

Javier Arroyo Gallardo

Dirigida por el Profesor

Carlos Maté Jiménez

Departamento de Organización Industrial
Escuela Técnica Superior de Ingeniería (ICAI)
Universidad Pontificia Comillas

Madrid 2008

No me gusta el trabajo, a nadie le gusta,
pero me gusta lo que hay en el trabajo.
La posibilidad de descubrirte a ti mismo.
Tu propia realidad, para ti, no para los
demás, lo que nadie más puede saber.

Joseph Conrad, El corazón de las tinieblas

Agradecimientos

Antes de hablar,
tengo algo importante que decir.

Atribuida a Groucho Marx

En primer lugar, comenzaré la lista de agradecimientos por mi director, Carlos Maté, porque él es el principal culpable de que me embarcase en esta travesía que ahora llega a su fin. Ya no recuerdo cómo me convenció, pero supongo que si me hubiese hecho ver todo el esfuerzo que iba a suponerme completar el doctorado, yo habría salido corriendo. Así que no debió hacerlo. Pese a ello, no le guardo rencor, sino que le estoy muy agradecido por haber confiado en mí. También quiero darle las gracias por el tiempo y el empeño que ha mostrado en la tarea de dirigirme. Sé que ha sido mucho más de lo que le correspondía y también sé que, en algunos momentos, no se lo he puesto fácil. Por todo ello, su mérito es doble o triple. Ahora que ha terminado su labor, espero que pueda ver recompensados, al menos en parte, los esfuerzos realizados.

Mi doctorado ha tenido varias etapas y la primera de ellas se desarrolló en ICAI, dentro del Departamento de Organización Industrial. Guardo un grato recuerdo de aquella etapa y de la gente que conocí entonces. Después he seguido vinculado a través del proyecto de Predicción de Datos Simbólicos en el que he tenido la suerte de colaborar con Ángel y Antonio. Quiero agradecerles las molestias que se tomaron leyendo documentos de trabajo sobre mi tesis y ayudándome a robustecerlos. Ha sido un orgullo poder aprender de ellos. Gracias a los dos.

La mayor parte de esta tesis ha sido desarrollada durante mis años de profesor ayudante en la Universidad Complutense de Madrid. Por ello, quiero agradecer a los compañeros del Departamento de Ingeniería del Software e Inteligencia Artificial de dicha universidad el darme la oportunidad de trabajar con ellos. Es un placer y un lujo estar en este departamento, rodeado de personas de las que aprender y en un magnífico ambiente de trabajo. Mención especial para los ayudantes y becarios del departamento, buena gente todos ellos, con los que se ha hecho mucho más llevadera la lucha por sacar la tesis adelante.

También quiero agradecer al *Grupo de Agentes de Software: Ingeniería y Aplicaciones* el haberme acogido e integrado dentro del proyecto *Métodos y herramientas para modelado de sistemas multi-agente*, subvencionado por el MEC (TIN2005-08501-C03-01), dentro del cual he desarrollado mi labor. En especial, quiero agradecer al director del grupo, Juan Pavón, su apoyo. El respaldo que me ha dado a todos los niveles, incluido el económico, ha ido más allá de lo razonable. Sin ese respaldo, me hubiera sido muy difícil comenzar a cuajar una carrera en la universidad. Por ello, quiero expresarle mi agradecimiento más sincero. Espero que ahora, libre de las obligaciones de la tesis, pueda aportar mi granito de arena al formidable trabajo que realizan en su grupo.

Muchas personas me han ayudado en mayor o menor medida a sacar la tesis adelante. Gracias a todos ellos y en especial a los siguientes. A Rosa que se dejó engañar para echarme una mano con las predicciones de las casi infinitas series temporales de intervalos. A Marco y Pedro por prestarme su ya legendaria plantilla de LaTeXpara la tesis y por darme soporte cuando lo he necesitado. Y a Guille por llevar a cabo el diseño de la portada y por haberla retocado mil y una veces casi sin odiarme.

Marco, Pedro y Guille repiten mención porque, no sólo han colaborado en la tesis, sino que además son mis amigos. Seres míticos y entrañables los tres. De mayor quiero ser como ellos.

Turno para la familia que me soporta y me arropa incondicionalmente. Gracias a mi hermana Elena, a Jose y a las pequeñajas Paula e Irene, dos terremotos que contagian su alegria a todo el que se pone por delante.

A mis padres les debo todo. En realidad, ya se lo debía antes de empezar el doctorado, pero el 'todo'que les debo ahora es mucho mayor que el que les debía entonces. Son incontables las cosas que han hecho por mí durante estos años. Pequeños y grandes detalles con los que me han allanado enormemente el camino y que han permitido que sólo me tuviese que centrar en mi trabajo y en nada más. Sin ellos hubiera sido imposible. Sé que nunca podré saldar mi deuda, pero , al menos, poder hacer que se sientan orgullosos.

Por último, dar las gracias a Rocío (es genial estar contigo) que ha sido quien más ha padecido mis reiterados retrasos en la fecha del fin de la tesis, mi humor cambiante y mis encierros durante muchas tardes, especialmente durante el último año. Lo ha hecho con infinita comprensión y paciencia. Su cariño ha sido fundamental para no volverme más loco. Le debo una y espero devolvérsela en breve cuando se anime a continuar su doctorado.

Resumen

Las series temporales clásicas, donde cada instante es descrito por un número real, sirven para representar una gran multitud de situaciones de la vida real, pero no son capaces de describir fielmente situaciones en las que en cada instante haya que reflejar cierta variabilidad. Los datos simbólicos de intervalo e histograma permiten representar dicha variabilidad a lo largo del tiempo, dando lugar a series temporales de intervalos e histogramas, respectivamente. En este trabajo se han abordado algunos aspectos relativos a estas series temporales como, por ejemplo, la medición del error, pero el principal objetivo de esta tesis es desarrollar métodos que permitan predecir estos nuevos tipos de series temporales de manera eficaz.

Las aproximaciones que se han propuesto para predecir series temporales de intervalos incluyen:

- Alisados exponenciales basados en la aritmética de intervalos
- Método de k-NN basado en la aritmética de intervalos
- Perceptrón multicapa basado en la aritmética de intervalos
- Predicción mediante las series temporales de sus componentes (mínimo, máximo, centro y radio) aplicando para ello métodos de predicción (univariantes o multivariantes) para series temporales clásicas.

Mientras que los métodos desarrollados para predecir series temporales de histogramas son:

- Alisados exponenciales basados en la aritmética de histogramas
- Alisados exponenciales basados en el concepto de baricentro
- Método de k-NN basado en el histograma basados en el concepto de baricentro

La capacidad predictiva de tods estos métodos ha sido probada con éxito en ejemplos reales de diferentes ámbitos como, por ejemplo, la meteorología o las finanzas.

Índice

Índice de figuras

Índice de Tablas

Capítulo 1

Introducción

No pretendo empezar con cuestiones precisas.
Creo que no se puede empezar con nada preciso.
Tienes que alcanzar dicha precisión
a medida que puedas, a medida que avances.

Bertrand Russell

Este primer capítulo ofrece una descripción general del problema sobre el que trata esta tesis: la predicción de series temporales de intervalos y de histogramas. Para ello, en primer lugar sitúa el problema a tratar dentro del área de la predicción de series temporales. A continuación, explica que la manera de abordar la predicción de las series temporales de intervalos y de histogramas que propone esta tesis, la cual se encuadra dentro del área de análisis de datos simbólicos. Por último, introduce brevemente el contenido del resto de capítulos de la tesis.

1.1. Motivación de la investigación

1.1.1. Introducción a la predicción de series temporales

Hoy en día, la predicción es una actividad crucial en áreas tan diversas como la economía, la meteorología, la actividad empresarial, las ciencias sociales, la ingeniería, las ciencias biológicas y ambientales, etc. De forma general, podemos entender que una predicción es el aviso o anuncio de algo que va a suceder. Resulta fácil comprender que el conocimiento anticipado de un hecho futuro permite prepararse de forma adecuada para hacerle frente. Por ejemplo, si se conoce anticipadamente la demanda de un producto para el mes próximo, se puede planificar mejor la producción del mismo para cubrir esa demanda sin quedarse corto, ni producir en exceso.

Esta tesis se sitúa en el ámbito de la predicción de magnitudes cuantitativas, como pueden ser la temperatura o el valor de un índice bursátil.

La predicción de una magnitud cuantitativa puede realizarse mediante dos aproximaciones: el conocimiento experto y la predicción de series temporales. En la primera de ellas, las predicciones son elaboradas por expertos que, basándose en su experiencia y conocimientos sobre la materia en cuestión, estiman el valor futuro de la magnitud a pronosticar. En la segunda aproximación, las predicciones son elaboradas mediante el análisis matemático de los datos históricos de la magnitud que nos interesa y, posiblemente, de otras magnitudes que se relacionan con ella. Sobre esta segunda aproximación se centrará esta tesis.

Una serie temporal es el resultado de la observación de los valores de una variable a lo largo del tiempo en intervalos regulares (diariamente, mensualmente, etc.). Las series temporales como disciplina estadística tienen unos 200 años de historia. En la introducción de Peña (2005) se realiza un completo recorrido por sus hitos más significativos, los cuales son repasados a continuación de forma más escueta.

1.1.1.1. Introducción histórica a las series temporales

Los orígenes de la disciplina se remontan a principios del siglo XIX cuando Laplace realizó un estudio sobre el efecto de las fases de la luna sobre las mareas y los movimientos del aire en la tierra. En esta etapa, resultan fundamentales las aportaciones de Fourier, que estudió las funciones periódicas; de Yule, que propuso en 1927 los procesos autorregresivos para explicar las manchas solares (Yule, 1927); y de Slutsky que estudió los procesos de media móvil para representar los ciclos económicos (Slutsky, 1937).

Otros muchos como Kolmogorov, Wiener, Cramer, Bartlett y Tukey realizaron contribuciones notables durante la primera mitad del siglo XX. La consolidación de la disciplina llegó en la segunda mitad del siglo XX con tres trabajos clave: el desarrollo por parte de Holt y Winters de métodos de predicción basados en alisados exponenciales (Holt, 1957; Winters, 1960); la propuesta de Box y Jenkins de una metodología unificada para la predicción de series estacionarias y no estacionarias (con y sin estacionalidad) que da lugar a los célebres modelos ARIMA (Box y Jenkins, 1970); y el desarrollo por parte de Kalman de un procedimiento para estimar las variables de estado y prever las observaciones futuras en sistemas lineales (Kalman, 1960).

Estas aportaciones hacen que la predicción entre en una fase de madurez y que surja el International Institute of Forecasters, IIF, promotor de foros y revistas de predicción como el Journal of Forecasting (en 1982), el International Journal of Forecasting (en 1985) y el International Symposium on Forecasting, conferencia anual de especialistas en el tema celebrada desde 1982. Otro hecho importante acontece en 2003 con la concesión del Premio Nobel de Economía a Granger y Engle que vino a reconocer el enorme impacto que tuvieron sus trabajos sobre la cointegración de series temporales (Engle y Granger, 1987) en la teoría económica.

En 2006, el IIF celebró su 25 aniversario y echó la vista atrás para hacer un repaso de los avances más significativos en el área durante ese periodo, los cuales quedan recogidos en el número especial de la revista (Hyndman y De Gooijer, 2006). Tal y como muestra dicho número especial, la predicción de series temporales es un área viva en continuo desarrollo y que aún tiene por delante un gran número de retos.

1.1.2. Descripción del problema

A lo largo de todos estos años, la investigación y la práctica sobre series temporales se ha centrado, principalmente, en series temporales en las que cada valor observado de la variable (o variables) de interés es representado mediante un número real (o entero). Este tipo de series, a las que llamaremos 'clásicas', sirven para representar una gran cantidad de situaciones que se dan en la realidad. Sin embargo, presentan limitaciones a la hora de representar algunas situaciones más complejas.

Una situación que no se puede representar fielmente mediante una serie temporal clásica se da cuando las observaciones tienen que recoger cierta incertidumbre. Una forma de representar dicha incertidumbre es utilizando conjuntos borrosos, lo que daría lugar a una serie temporal borrosa. Las primeras contribuciones en este área datan de la década de los 90 (Song y Chissom, 1993a,b; Song, Leland y Chissom, 1995). Este tipo de series temporales no serán abordadas en esta tesis, por lo que no se profundizará aquí sobre ellas.

Otra situación que no puede ser descrita de forma precisa por las series temporales clásicas se da cuando las observaciones deben reflejar variabilidad. Esto sucede cuando, para cada instante temporal, en lugar de contar con un valor observado, se cuenta con un conjunto de valores observados. Existen dos situaciones paradigmáticas donde esto sucede.

Caso paradigmático 1. Este caso se refiere a una situación en la que la variable de interés es medida con una determinada frecuencia (e.g. cada minuto), pero debe ser analizada a una frecuencia menor (e.g. diariamente). Si ante esa situación se opta por trabajar con la serie que se obtiene al muestrear la serie original con la frecuencia menor, se está ignorando la información contenida en los datos observados entre los instantes muestreados.

Este caso se presenta de manera habitual en las series temporales financieras que reflejan la cotización de una acción, de un índice bursátil o del cambio de unas divisas. En este tipo de series, los valores se generan con una frecuencia alta. Sin embargo cuando se analiza la evolución de estos valores normalmente se opta por utilizar únicamente los valores de cierre de sesión. Típicamente, se utilizan los cierres diarios, pero en algunos casos se emplean los cierres semanales o, incluso, mensuales. Al hacer esto, se prescinde de

los valores que se dan entre dos cierres consecutivos, por lo que se pierde la perspectiva sobre el comportamiento del bien financiero entre dichos valores de cierre, es decir, se están ignorando sus fluctuaciones intra-periodo.

Este tipo de situación no sólo aparece en el campo de las finanzas, sino también al manejar datos temporales recogidos de forma sistemática por sensores ya sea en procesos industriales, mediciones de niveles en hidrología, meteorología, medioambiente, etc.

Caso paradigmático 2. Este caso sucede cuando una variable es medida a lo largo del tiempo en los elementos de un conjunto, pero el interés no reside en conocer la evolución de cada uno de los elementos, sino del propio conjunto como un todo. Una forma de resumir los valores del conjunto en cada instante consiste en calcular su media. Sin embargo, la media, aunque es útil para describir un conjunto de datos, presenta ciertas limitaciones.

Una situación real donde esto ocurre se da cuando un instituto de estadística recoge de forma periódica los valores de una variable, como la renta, en todos los individuos de la población o en una muestra de la misma. Para representar de forma agregada la renta de la población en cada instante, se suele optar por utilizar la renta per capita, i.e. la media de las rentas de los individuos. Sin embargo, es bien conocido que la media es sensible a la presencia de valores extremos y que las medidas de tendencia central no informan sobre algunas características de la distribución subyacente como, por ejemplo, la dispersión.

Otro campo donde sucede esta situación es el del análisis de datos de panel (Baltagi, 2004). En los paneles se recoge información sobre, por ejemplo, los hábitos de consumo de un conjunto de familias a lo largo del tiempo. Un caso similar al de los datos de panel es el de los estudios de mercado, donde se suele medir de forma periódica la percepción que tienen los individuos sobre un determinado producto. En el área de control de calidad también pueden surgir este tipo de situaciones cuando se pretende representar de forma agregada los valores de un parámetro de calidad sobre todos los elementos que integran cada uno de los lotes de productos que se están fabricando.

El primero de los casos paradigmáticos describe una situación en la que los datos disponibles son agregados temporalmente para poder ser tratados. Mientras que el segundo de los casos ilustra una situación donde se emplea agregación contemporánea.

En ambos casos, los datos disponibles son resumidos para dar lugar a una serie temporal clásica donde cada instante temporal es descrito por un valor agregado, como la media o el valor de cierre. El problema reside en que este valor agregado puede suponer una simplificación excesiva, ya que puede acarrear la pérdida de información valiosa sobre el fenómeno a tratar.

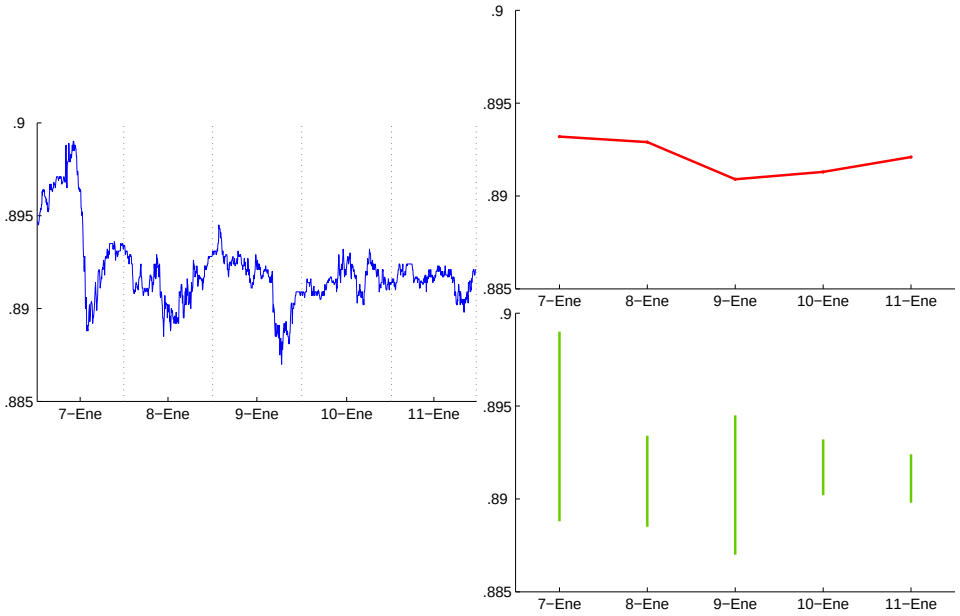


Figura 1.1: Serie intradiaria de la cotización €-$ (izqda.), serie temporal de los valores de cierre diarios (arriba dcha.) y serie temporal de intervalos que representan los valores mínimo y máximo diarios (abajo dcha.)

Para evitar estos inconvenientes, sería deseable contar con representaciones temporales que permitan describir de forma más completa los datos agregados. En esta tesis se proponen dos representaciones de ese tipo: las series temporales de intervalo y las series temporales de histogramas.

Un intervalo ofrece información sobre el rango de los valores observados, por lo que informa de la dispersión que se da en los datos. El intervalo puede representar el valor mínimo y máximo del conjunto o, alternativamente, otro intervalo como por ejemplo el recorrido intercuartílico, i.e. el intervalo limitado por la primera y la tercera cuartila. Por su parte, un histograma ofrece información no sólo sobre el rango de valores que toma el conjunto de datos, sino también de la distribución de los mismos a lo largo de dicho rango. La información que ofrece el histograma es, por tanto, notablemente más completa que la ofrecida por el intervalo, pero, a cambio, su estructura es también notablemente más compleja que la de éste. En algunos casos, puede bastar con un intervalo, mientras que en otros puede necesitarse un histograma.

La figura 1.1 muestra un ejemplo de las finanzas que sirve para ilustrar el *caso paradigmático 1*. La serie considerada es la de la cotización €-$ intradiaria entre el 7 y el 11 de enero de 2002. En la parte izquierda de la imagen, se muestra la serie completa; mientras que en la parte derecha, los valores intradiarios son resumidos de dos formas: tomando los valores de cie-

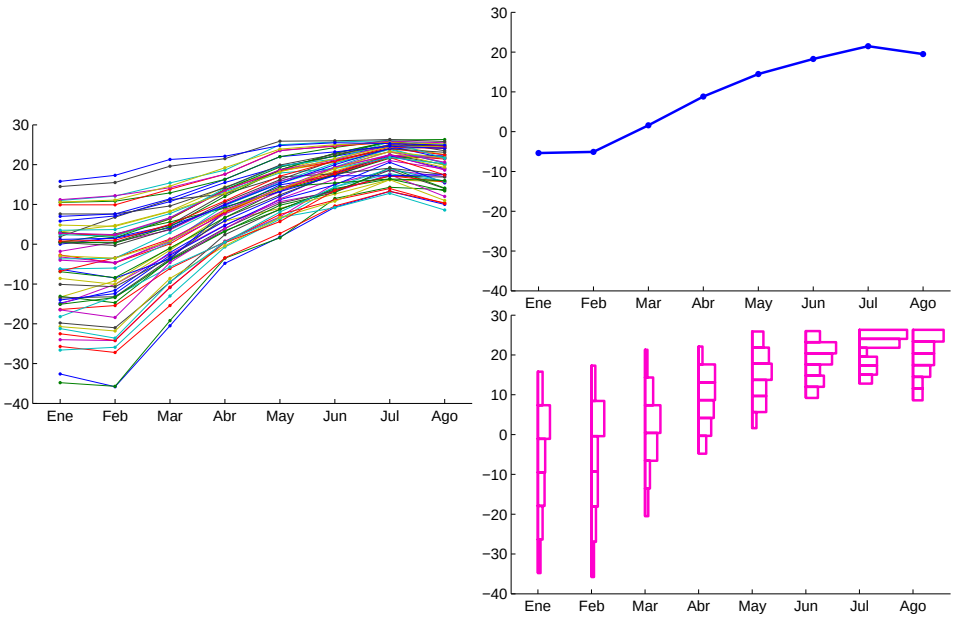


Figura 1.2: Series temporales de la temperatura media mensual en ºC en la red de estaciones meteorológicas de China (izqda.), serie temporal de la temperatura media en la red (arriba dcha.) y serie temporal de histogramas que representan la distribución de las temperaturas en la red (abajo dcha.)

rre diario y mediante el intervalo diario de los valores mínimo y máximo. Puede apreciarse muy claramente como, al tomar los valores de cierre, se pierde información sobre la volatilidad que se ha dado ese día. Dicha información es aportada por los intervalos, por lo que ambas representaciones son complementarias.

Por su parte, la figura 1.2 ilustra el *caso paradigmático 2* con un ejemplo del ámbito meteorológico. En dicha figura se muestra un conjunto de series temporales de la temperatura media mensual en ºC registrada en 60 estaciones distribuidas uniformemente a lo largo y ancho de China entre enero y agosto de 1952. Si la información de las 60 series temporales es agregada mediante la media, se obtiene una serie temporal clásica que representa la temperatura media en China y que se muestra en la parte superior derecha de la figura 1.2. Dicha serie enmascara la variabilidad de los datos originales, puesto que en los meses de invierno el rango de temperaturas es notablemente mayor que en verano. Por su parte, la serie temporal de histogramas que se muestra en la parte inferior derecha, no sólo informa del rango de la distribución, sino que también lo hace sobre la forma de ésta, la cual varía bastante entre los meses (véase, por ejemplo la diferencia que existe entre marzo y agosto). El gráfico de la STH es una herramienta exploratoria muy efectiva ya que facilita la interpretación visual de los datos. En ese sentido,

la interpretación del gráfico de las series temporales desagregadas, resulta sensiblemente más complicada.

Si el interés reside en predecir las n series temporales de manera individualizada, no tiene sentido agregar la información y, por tanto, no tiene sentido plantear una serie temporal de intervalos o de histogramas, ni tampoco una serie temporal de valores medios. Por otro lado, podría suceder que los elementos para los que se midiese la variable entre dos instantes temporales consecutivos no fuesen los mismos. Si sucede esto, no se pueden representar los datos originales como series temporales (tal y como se hace en el recuadro izquierdo de la figura 1.2) , ni tampoco obtener las predicciones de dichas series temporales de forma individualizada. En ese caso, la agregación es la única opción de manejar los datos y los intervalos y los histogramas son herramientas a considerar para este propósito.

1.1.3. Precedentes de las series temporales de intervalos y de histogramas

En este apartado se van a revisar los casos donde se ha detectado la necesidad de representar la variabilidad en una serie temporal y se ha recurrido para ello a una serie temporal de intervalos o de histogramas.

En meteorología es habitual representar las temperaturas mínima y máxima obtenidas o previstas para un determinado periodo de tiempo. La representación temporal de esta información daría lugar a una serie temporal de intervalos. Sin embargo, no se conoce ningún artículo en el que dichos datos sean reconocidos como tal o donde por su naturaleza sean analizados de manera especial. En este contexto, no es habitual utilizar la media, ni ningún otro valor agregado, sino que se emplean de forma habitual los intervalos para, de esa forma, representar la variabilidad térmica.

Por otro lado, las publicaciones y sitios web dedicados a las finanzas suelen mostrar de forma habitual los precios de apertura, de cierre, mínimo y máximo alcanzados por los diferentes bienes financieros durante el día o durante la semana. Estos cuatro valores dan lugar a dos intervalos: el que forman el mínimo y el máximo (del que ya se habló en el *caso paradigmático 1*) y el que forman la apertura y el cierre (que es un intervalo donde se determina si el bien incrementó o no su valor). Dichos intervalos ofrecen una muy buena síntesis de la variabilidad del bien financiero durante un determinado periodo, por ello son tan populares en finanzas. Además, tienen una representación gráfica propia: los gráficos de velas o *candlesticks*. Puede verse un ejemplo de dicha representación en la figura 1.3.

Los gráficos de velas fueron creados en el siglo XVIII por Munehisa Homna, comerciante japonés de arroz, para representar las variaciones diarias que alcanzaba el precio del arroz en el mercado. En finanzas, existe toda una teoría para interpretar estos gráficos en busca de señales de compra o de venta

Figura 1.3: Gráfico de velas diario del IBEX35 entre el 1 de agosto y el 22 de octubre de 2007 (Fuente: finanzas.com)

que sirvan de apoyo a la toma de decisiones, ver, por ejemplo, Morris (2006). Sin embargo, dicha teoría se encuentra alejada de las áreas de las series temporales y del aprendizaje estadístico. En la sección 3.5.2.2 se comentarán las aproximaciones para predecir series temporales de *candlesticks* basadas en el análisis estadístico o en la predicción de series temporales.

La existencia de los candlesticks evidencia que en finanzas resulta útil obtener representaciones temporales y predicciones que trasciendan a la serie temporal puntual y que informen sobre el rango de valores o la distribución que siguen los valores de la acción a lo largo del día.

Otro precedente interesante de representación de serie temporal de intervalos se muestra en Zellner y Tobias (2000). En dicho artículo se analizan las tasas de crecimiento anuales de 18 países industrializados con el fin de predecir la mediana de dichas tasas de dos formas, agregada y desagregada. Curiosamente, a la hora de representar gráficamente la serie temporal de las medianas, los autores optan por dibujar la mediana dentro de los intervalos que representan el rango intercuartílico de las 18 tasas de crecimiento anual. El gráfico resultante representa claramente una serie temporal de intervalos, tal y como se puede veren la figura 1.4. Dicha representación puede verse también como una serie temporal de histogramas incompleta. Si en ella apareciesen representados el mínimo y el máximo en cada instante temporal, se tendría perfectamente representada la distribución subyacente. La representación resultante sería un gráfico de cajas o *boxplot* (Tukey, 1977), que es, como se verá en el capítulo 5, un tipo particular de histograma.

En González-Rivera, Lee y Mishra (2008) podemos encontrar una ilustración con una serie temporal de histogramas. Dicha ilustración es reproducida

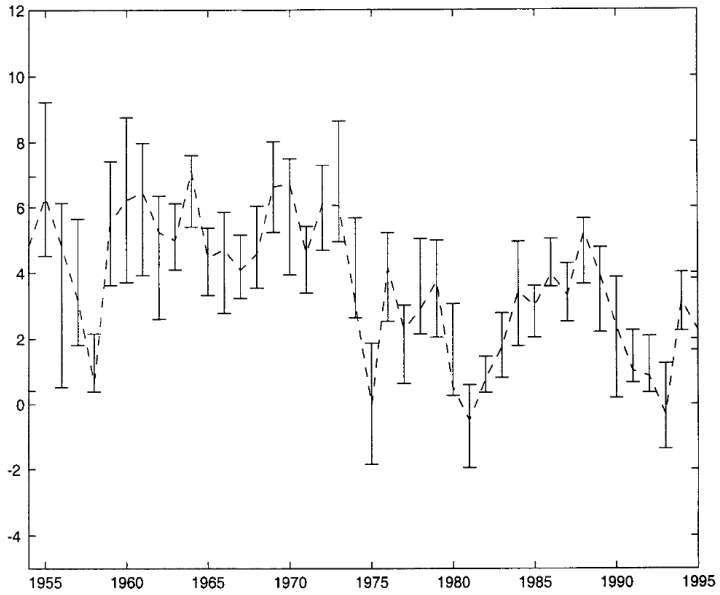


Figura 1.4: Medianas y rangos intercuartílicos de las tasas de crecimiento de producción real (%) de 18 países industrializados (Zellner y Tobias, 2000)

en la figura 1.5. En ella, los histogramas representan la distribución de los rendimientos en % de un conjunto de acciones en el instante temporal t o, lo que ellos llaman, el *ranking* de sección cruzada de los rendimientos en t. En dicha imagen, las líneas que unen cada histograma representan el cambio de posición de una acción dentro de la distribución (o *ranking*) de rendimientos. En el trabajo de González-Rivera et al. (2008), el interés reside en el estudio de los saltos que realiza una acción entre dos *rankings* de rendimientos consecutivos en el tiempo. Sin embargo, no se plantea la predicción de la serie temporal de histogramas que es uno de los objetivos de esta tesis.

La serie de Zellner y Tobias (2000) y la de González-Rivera et al. (2008) son ejemplos de series temporales de intervalos y de histograma obtenidas como síntesis o agregación de las observaciones de una variable en un conjunto de individuos. Por tanto, ambas corresponden al *caso paradigmático 2*. Por su parte, los candlesticks y los intervalos de temperaturas son ejemplos prácticos del *caso paradigmático 1*.

1.1.4. ¿Qué aportan las predicciones de las series temporales de intervalo y de histograma?

Esta tesis propone el uso de series temporales de intervalos y de histogramas para representar fenómenos donde las observaciones se vean afectadas de cierta variabilidad. Tal y como se mostró en el apartado 1.1.2, cuando surgen dichos fenómenos se suele recurrir a la agregación temporal (ver *caso*

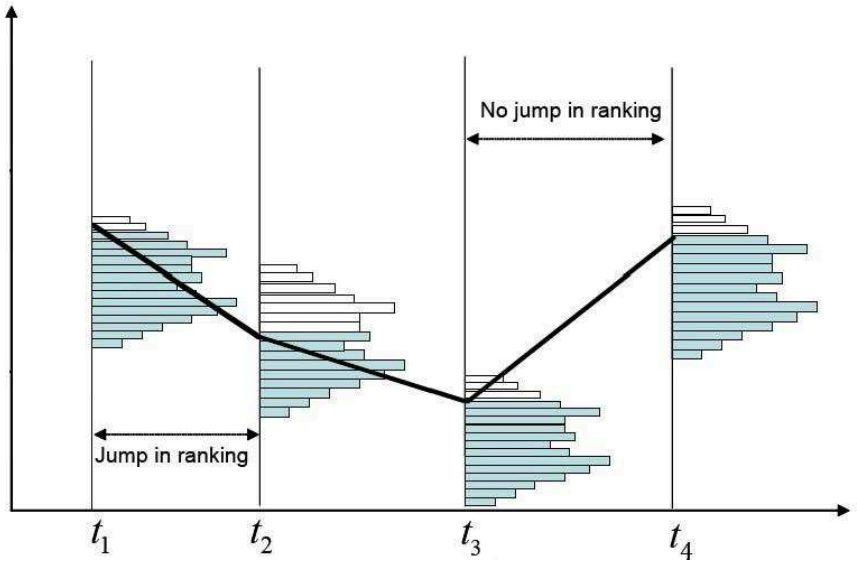


Figura 1.5: Serie temporal de histogramas que representan la distribución de los rendimientos en % de un conjunto de acciones a lo largo del tiempo (González-Rivera et al., 2008)

paradigmático 1) o contemporánea (ver *caso paradigmático 2*) de los datos. La predicción de una serie temporal de intervalos o de histogramas resulta muy útil si, en el fenómeno que se está analizando, la variabilidad desempeña un papel importante. En dichas situaciones, una predicción en forma de intervalo permite conocer anticipadamente la variabilidad futura en forma de rango, mientras que una predicción en forma de histograma informa además sobre cómo se distribuirán los valores dentro de dicho rango.

Si la variabilidad no es importante, un valor agregado como la media (para el *caso paradigmático 2*) o un valor muestreado en un determinado instante (para el *caso paradigmático 1*) puede ser suficiente. Sin embargo, si, como en las finanzas y en la meteorología, el interés reside precisamente en conocer la variabilidad del fenómeno y no tanto su tendencia central, entonces las series temporales de intervalos y de histogramas son de gran utilidad.

Además, puede darse el caso de que los datos que se están manejando deban ser tratados de forma agregada porque hacerlo de forma desagregada sea dificultoso o directamente infactible. A continuación, se muestran dos ejemplos que ilustran estas situaciones. El primero de ellos es del tipo descrito en el *caso paradigmático 1*, mientras que el segundo es del tipo descrito en el *caso paradigmático 2*.

- Tal y como indican Engle y Russell (2009), las series temporales financieras intra-diarias presentan una serie de características que hace

que sea difícil tratarlas y predecirlas mediante métodos clásicos. Estas características son: espaciado temporal irregular, patrones de comportamiento diarios, el valor discreto de los precios y la existencia de dependencia compleja. Por esta razón, estas series se prestan especialmente bien a ser manejadas de forma agregada. Una forma de agregar esta información es mediante los intervalos de valores mínimo-máximo y apertura-cierre. Si lo que se quiere es conocer la distribución futura global de los datos, modelar la serie de alta frecuencia como una serie temporal de histogramas es una buena alternativa. Por último, hay que tener en cuenta que, en estos casos, predecir la serie intradiaria para el día siguiente es una tarea prácticamente imposible. Sin embargo, pronosticar el rango en el que se va a mover dicha serie o la distribución de sus valores es, como se verá en esta tesis, factible.

- Si se considera el caso de un instituto de estadística que está estudiando la evolución de una variable en el conjunto de la población, lo habitual es que dicho instituto trabaje con una muestra de la población de m individuos. En dicha muestra, muy posiblemente los individuos para los que se recoge la información en cada instante de tiempo no sean los mismos, por tanto no se pueden considerar m series temporales, una para cada individuo, y la agregación es la única forma de manejar dicha información. En esos casos, la media ofrece información interesante sobre la población, pero el intervalo complementa dicha información mediante el rango y el histograma sobre la distribución de valores sobre dicho rango. También puede suceder que, aunque sea posible predecir las m series temporales individuales, su número sea tan alto que no sea recomendable; especialmente, si el interés reside en conocer el comportamiento de la población y no de los individuos.

1.1.5. Métodos desarrollados al margen de la tesis

En el momento del inicio de esta tesis, el año 2004, no existía ningún método para predecir series temporales de intervalo, ni de histograma. Por ello, la tesis se planteó como la primera aportación dentro de dicho ámbito. Sin embargo, paralelamente al desarrollo de esta tesis, han ido apareciendo publicaciones que comienzan a abordar el estudio de series temporales de intervalo.

Teles y Brito (2005) plantean una primera aproximación a la predicción de estas series por medio de modelos ARMA. En esta aproximación, tanto la serie de los mínimos, como la serie de los máximos son modeladas cada una de ellas mediante una ecuación ARMA con los mismos parámetros, pero con distinta constante. Por su parte, Maia, de Carvalho y Ludermir (2006a) trabajan con una serie temporal de intervalos descomponiéndola en la serie temporal de los centros y la de los radios y ajustando para cada una de ellas

o un modelo ARMA o un modelo híbrido que combina el modelo ARMA y el perceptrón multicapa. Estos métodos serán explicados con mayor detalle en el capítulo 3.

Respecto a las series temporales de histograma, sólo el trabajo de Maté y González-Rivera (2007) plantea un método para predecir series temporales de histograma. En dicho trabajo, se predice la serie por medio de la técnica del análisis de componentes principales para datos de histograma desarrollado en Rodríguez, Diday y Winsberg (2000). La escasez de métodos propuestos para trabajar con series temporales de histogramas es debida a que el histograma es notablemente más complejo que el intervalo y que, por tanto, desarrollar métodos para analizar histogramas es más complicado que desarrollarlos para intervalos. Esta idea puede corroborarse consultando los métodos para uno y otro tipo de variable que aparecen en los principales libros de datos simbólicos (Bock y Diday, 2000; Billard y Diday, 2006b; Diday y Noirhomme, 2008). Sin embargo, pese a su complejidad, la información que aporta un histograma es mucho más rica que la que aporta un intervalo. Por ello, esta tesis propondrá métodos para pronosticar series temporales de histogramas.

En el siguiente apartado se mostrará en detalle el planteamiento que se ha seguido para abordar la predicción de series temporales de intervalos y de histogramas.

1.2. Planteamiento de la tesis

El objetivo principal de la tesis es el desarrollo de métodos de predicción para series temporales de intervalos y de histogramas y su aplicación en ejemplos reales para demostrar su efectividad. La tesis surge de la confluencia de dos disciplinas: el análisis de datos simbólicos (área que se ocupa del análisis estadístico de datos de intervalo y de histograma) y la predicción de series temporales.

El análisis de datos simbólicos (Bock y Diday, 2000) es una disciplina que considera que la realidad es, en algunos casos, demasiado compleja como para poder ser representada mediante variables clásicas. Por ello, para describir conceptos o grupos de elementos, propone el uso de variables simbólicas, ya que estas permiten representar la variabilidad que a menudo se observa en la realidad. Entre las variables simbólicas se encuentran las listas de valores, las variables de intervalo, las variables modales y las variables de histogramas. Los libros de Bock y Diday (2000), Billard y Diday (2006b), y Diday y Noirhomme (2008) compendian la mayor parte de los métodos para analizar este tipo de datos. Sin embargo,en estas obras no se propone ningún método de predicción de series temporales de intervalo o de histogramas. No obstante, para desarrollar los métodos de predicción que se proponen en esta tesis ha resultado crucial el conocimiento de los fundamentos en que se basan los métodos de análisis de datos simbólicos.

1.2.1. Métodos desarrollados en esta tesis

Los métodos de predicción que se van a proponer en esta tesis adaptan al contexto simbólico métodos de predicción ya existentes en el contexto de las series temporales clásicas. La complejidad que entraña trabajar con datos simbólicos es notablemente superior (especialmente en el caso de los histogramas) a la que entraña trabajar con datos clásicos. Por ello, los métodos que se van a adaptar al contexto simbólico son métodos que tienen un aparato matemático sencillo, pero, cuya capacidad predictiva ha sido probada en las series temporales clásicas. Entre los métodos que van a ser adaptados se encuentran los métodos de alisado (Gardner, 1985, 2006) y los métodos de k-NN o de los k vecinos más próximos (Yakowitz, 1987). En ambos métodos, las predicciones se generan, *grosso modo*, como un promedio ponderado de valores pasados de la serie. Por tanto, una parte muy importante de la adaptación de estos métodos al contexto simbólico consiste en determinar cómo se calculará el promedio de un conjunto de intervalos y el promedio de un conjunto de histogramas.

Para predecir con las series temporales de intervalos se proponen dos aproximaciones:

- Predecir la serie mediante métodos de predicción clásicos: consiste en tratar la serie temporal de intervalos como un par de series temporales clásicas que pueden ser la serie de los mínimos y la de los máximos; o, alternativamente, la de los centros y la de los radios. Para predecir cada una de estas series se puede utilizar la técnica de predicción clásica que mejor se ajuste a cada serie o se pueden modelar las posibles interrelaciones entre las dos series consideradas mediante modelos multivariantes (Lütkepohl, 2005), como el modelo de vectores autorregresivos (VAR) o el modelo vectorial de corrección del error (VECM).
- Predecir la serie mediante métodos de predicción que consideren el intervalo como tal, se han desarrollado los siguientes, todos ellos basados en la aritmética de intervalos (Moore, 1966):
  - métodos de alisado exponencial
  - algoritmo de los k vecinos más cercanos (k-NN)
  - perceptrón multicapa: trabajo liderado por Antonio Muñoz (Muñoz San Roque, Maté, Arroyo y Sarabia, 2007)

En cuanto a la predicción de series temporales de histograma se han propuesto las siguientes técnicas:

- Dos adaptaciones distintas de los métodos de alisado exponencial
  - Utilizando la aritmética de histogramas (Colombo y Jaarsma, 1980) para realizar los promedios.

  - Sustituyendo el cálculo del promedio, que permiten alisar la serie, por el cálculo del baricentro de un conjunto de histogramas (Irpino y Verde, 2006b).

- Una adaptación del algoritmo de k-NN que utiliza también el concepto de baricentro para generar las predicciones y las distancias de Mallows y de Wasserstein para encontrar a los vecinos más próximos.

Con estos métodos, la tesis abre un camino que hasta la fecha estaba prácticamente inexplorado pero que ofrece muchas posibilidades tanto a nivel de investigación, como a nivel de aplicación.

1.3. Organización de la exposición

Este primer capítulo ha servido como introducción a la tesis, describiendo el problema que se va a tratar y la forma en que se va a abordar. En el capítulo 2 se dará un repaso a los métodos propuestos en la literatura para analizar datos simbólicos de intervalo y de histograma. En dicho capítulo, se analizará la forma en que estos métodos trabajan con intervalos y con histogramas, ya que algunos de ellos servirán como base para los métodos que se propondrán más adelante.

En el capítulo 3 se definirá el área en la que se centra la tesis: la predicción de series temporales donde la variable observada es una variable simbólica de intervalo o de histograma. En dicho capítulo se relacionará éste área con las series temporales clásicas y con otras propuestas existentes que trabajan con series más complejas que las clásicas o que pretender obtener predicciones que vayan más allá de la mera estimación puntual. Las aproximaciones que se abordarán son:

- Dentro del área de las series temporales clásicas: las predicciones de intervalo (Chatfield, 2001a) y densidad (Tay y Wallis, 2000), las series temporales multivariantes y la agregación de series temporales.

- Al margen de las series temporales clásicas: la predicción de series temporales de *candlesticks* en el ámbito financiero (Lee et al., 2006), las series temporales simbólicas basadas en alfabetos de símbolos (Daw et al., 2003) y las series temporales simbólicas basadas en variables simbólicas, sobre las que se centra la tesis.

Puede verse que las dos últimas aproximaciones tienen el mismo nombre, *series temporales simbólicas*. Sin embargo, ambas proceden de áreas distintas: el análisis simbólico y el análisis de datos simbólicos, respectivamente. El primero trabaja con datos expresados mediante un alfabeto de símbolos y el segundo con datos expresados mediante intervalos, histogramas, listas

de valores (categóricos o no), etc. Pese a algunas similitudes entre las dos disciplinas, las diferencias son notables.

La tesis se centra en las series temporales de intervalos y de histogramas. El capítulo 3 revisará las métodos que se han propuesto dentro de este enfoque y al margen de esta tesis.

El capítulo 4 se ocupa de las series temporales de intervalos. Propone una definición de las mismas, explica cómo se pueden obtener, define medidas de error para estas series y propone distintas aproximaciones para predecirlas: el uso de métodos clásicos sobre las series de los componentes del intervalo; y la propuesta de métodos que trabajan con el intervalo como tal, como son los alisados exponenciales, el k-NN y el perceptrón multicapa. El capítulo también incluye la comparación de las distintas aproximaciones sobre series termporales de intervalos reales.

El capítulo 5 está dedicado a las series temporales de histogramas. En él, se aborda la definición de estas series, la propuesta de medidas de error paraellas y el desarrollo de métodos de predicción, entre los que se encuentran: los alisados exponenciales basados en aritmética de histogramas y en el concepto de baricentro de un conjunto de histogramas, y una adaptación del algoritmo de k-NN basado también en el concepto de baricentro de un conjunto de histogramas. Los métodos de predicción propuestos se prueban sobre series provenientes de diferentes ámbitos.

Los resultados que obtienen los métodos de predicción propuestos para series temporales de intervalos y de histogramas son, en líneas generales, muy satisfactorios y vienen a subrayar la utilidad de los métodos propuestos. El capítulo 6 concluye y presenta las líneas de trabajo futuro.

Además, la tesis incluye dos apéndices. En el apéndice A se revisan los fundamentos de los métodos de predicción clásicos que son adaptados en la tesis al contexto simbólico. Por otro lado, el apéndice B explica en detalle el procedimiento que se utiliza para calcular el baricentro de un conjunto de histogramas basado en las distancias de Mallows y de Wasserstein.

Capítulo 2

Estado del Arte del Análisis de Datos Simbólicos

Lee, lee, lee. Lee de todo - basura, clásicos, bueno y malo, y observa cómo lo hicieron. Haz como el carpintero que trabaja como aprendiz y estudia a su maestro. ¡Lee! Lo absorverás. Después escribe. Si es bueno, te darás cuenta. Si no lo es, tíralo por la ventana.

William Faulkner

En este capítulo se realiza una revisión del análisis de datos simbólicos. Esta disciplina tiene como finalidad extraer información de datos representados por medio de intervalos, listas de valores, variables modales, histogramas, etc.

El capítulo revisará a fondo las técnicas propuestas para analizar datos descritos mediante variables de intervalo y de histograma. El capítulo también recogerá otros métodos no provenientes del análisis de datos simbólicos pero que también trabajan con intervalos e histogramas.

2.1. Introducción

Los datos simbólicos son un paradigma de representación de la información que surge a finales de los ochenta (Diday, 1987) bajo la premisa de que las variables clásicas, i.e., aquellas que a cada individuo le asignan un único valor, no son capaces de representar con fidelidad algunas situaciones. Por ejemplo, mediante una variable clásica se puede describir el peso o la altura de un caballo determinado, pero, ¿qué sucede si se quiere describir a la especie"caballo andaluz"? En ese caso, las variables clásicas no permiten recoger la riqueza de información que la realidad ofrece.

Entre las variables simbólicas se incluyen las listas de valores, los intervalos, las variable modales y las variables de histograma. Siguiendo con el

ejemplo anterior, para describir la capa de los caballos andaluces se puede emplear la lista {tordo, bayo} y para describir su peso en kilogramos el intervalo $[400, 450]$. Si se optase por emplear una variable modal podríamos describir con mayor precisión la capa de los caballos andaluces dando un peso asociado a los valores 'tordo'y 'bayo'que representase la frecuencia con que se dan estas capas en dichos caballos. De la misma forma, para la magnitud peso, en lugar de un intervalo con el rango se puede utilizar un histograma que represente la distribución de frecuencias del peso en la especie de los caballos andaluces. Resulta obvio que las variables clásicas no son capaces de presentar tanto nivel de detalle de forma sintetizada.

En el ejemplo, la forma de construir la descripción simbólica del concepto "caballo andaluz" es mediante el resumen de una serie de observaciones individuales de dicho concepto. Esta forma de construir los conceptos tiene un nexo muy claro con las ideas aristotélicas sobre la percepción de la realidad. Aristóteles afirmaba que, mediante los sentidos, el hombre sólo puede captar lo individual, es decir, las formas sensibles de las cosas concretas. Pero que, gracias al entendimiento, el hombre es capaz de captar la esencia de esas cosas concretas. De esta forma, el hombre concibe el concepto universal "caballo" a partir de la percepción de una serie de caballos concretos. Según Aristóteles, posteriormente el hombre aplica los conceptos universales que se ha formado a cada uno de los individuos concretos de los cuales puede tener conocimiento por medio de sus sentidos.

Los datos simbólicos, a diferencia de los clásicos, permiten representar conceptos de una manera sintética y descriptiva. Sin embargo, el término *concepto* hay que entenderlo en un sentido amplio, ya que no sólo hace referencia a conceptos genéricos como "caballo" u "hombre", sino a conceptos más concretos como "mujer de entre 20 y 30 años" o "comprador de un determinado modelo de coche", que son más interesantes desde la perspectiva de las aplicaciones prácticas.

La característica fundamental de los datos simbólicos es que permiten la descripción de elementos o fenómenos donde exista una variabilidad interna. Los conceptos implican variabilidad ya que las distintas realizaciones de ese concepto pueden ser algo diferentes entre sí. La variabilidad surge de manera natural al agregar observaciones. Por agregación se entiende la recopilación de observaciones que satisfacen un requisito que les permite ser agrupadas. La agregación puede ser:

- Contemporánea, si se recopilan observaciones recogidas en un mismo instante temporal o cuando el instante temporal no es relevante.
- Temporal, si el criterio de agregación es el tiempo y se recopilan observaciones ocurridas a lo largo de una unidad de tiempo, e.g. un día.

Los ejemplos dentro del área de la estadística o de la minería de datos son incontables, e.g. las características resumidas de todas las transacciones ban-

carias o de todas las compras realizadas por un individuo en un establecimiento, un resumen de los datos monitorizados de un paciente a lo largo de un periodo de tiempo, la caracterización de distintos segmentos de clientes o de consumidores, las características de una cartera de valores, etc. En todos estos ejemplos de agregación, la representación simbólica de la información permiten obtener descripciones más fieles. Además, si el número de elementos a considerar es enorme o el propósito es el de estudiar la magnitud de forma agregada, puede ser la única alternativa razonable.

Hoy en día, con la popularización de la informática y de las bases de datos, el tamaño de los conjuntos de datos ha aumentado notablemente y la agregación se presenta como una herramienta necesaria para poder ordenar y extraer el conocimiento de dichos conjuntos.

La agregación permite distinguir entre observaciones de primer nivel y de segundo nivel. Las de primer nivel representan a individuos y las de segundo nivel representan a colectivos. También es posible que haya observaciones de tercer nivel o incluso de niveles superiores. Este puede ser el caso de los datos recogidos por un instituto estadístico que puede agregar los datos individuales por ciudades y a continuación por regiones, y analizar posteriormente los datos entre las distintas regiones. Los datos simbólicos permiten representar observaciones de segundo nivel y de niveles superiores.

Los datos simbólicos también permiten representar dependencias lógicas, taxonómicas o jerárquicas. Una dependencia lógica puede ser, por ejemplo, una regla que se añada para mantener la integridad al agregar los datos, por ejemplo, si la edad de un mujer es menor de catorce años, no puede tener hijos, esto evitaría que al representar el concepto "mujer" se pudiese interpretar que una mujer menor de 14 años puede tener hijos. Las taxonomías permiten representar variables en forma de árbol invertido donde cada nivel representa un nivel de generalidad: las hojas representan el menor nivel y la raíz representa el mayor. Por su parte, las jerarquías de variables permiten establecer relaciones madre-hija entre variables, de forma que una variable hija sólo está operativa dependiendo del resultado de la variable madre en el nivel superior. Por ejemplo, la variable hija 'edad del conyuge' sólo tiene sentido si la variable madre 'casado' es cierta.

2.1.1. Diferenciación entre dato simbólico, número borroso y número con incertidumbre asociada

El paradigma simbólico de representación de la información hace frente a una importante limitación que sufre el paradigma clásico: la incapacidad de representar variabilidad asociada a una observación. Sin embargo, no es el único paradigma que trasciende a la forma clásica de representar las observaciones. El paradigma borroso y el paradigma de representación de números con incertidumbre también se encuentran dentro de esa categoría. Por ello, es conveniente establecer la diferencia entre estos tres paradigmas.

El paradigma simbólico, en el que se sitúa esta tesis, permite representar observaciones que conllevan una variabilidad interna. De esta manera, si en un individuo se observa como valor de una determinada magnitud un intervalo simbólico, $[a, b]$ con $a \leq b$, la interpretación correcta consiste en considerar que el individuo toma o puede tomar varios valores dentro de ese intervalo. Por tanto, el intervalo acota los valores que el individuo puede tomar. Dentro del paradigma simbólico, se asume normalmente que la distribución dentro del intervalo es uniforme.

El paradigma borroso de representación de la información que tiene su origen en la década de los 60 (Zadeh, 1965) no pretende representar variabilidad, sino la imprecisión, como contraposición a la exactitud de los valores en el paradigma clásico. La imprecisión se representa mediante una función de pertenencia cuyo dominio es el rango de la variable y que toma valores entre 0 y 1. En dicha función el 0 representa el mínimo grado posible (el individuo no "pertenece" a ese valor o etiqueta lingüística) y 1 el máximo (el individuo "pertenece" a ese valor o etiqueta lingüística con el máximo nivel de presunción). Si el número borroso es expresado como un intervalo, sobre él no se impone ninguna función de distribución ya que los números borrosos no tienen un nexo con los conceptos de frecuencia o de probabilidad, sino con el concepto de posibilidad. Existen números borrosos más sofisticados que los intervalos como los triangulares, los trapezoidales, los campaniformes o, incluso, números cuya función de pertenencia no es convexa.

El análisis de intervalos (Moore, 1966) es otro paradigma de representación de la información cuyo objetivo es representar la incertidumbre asociada a una observación. En este área, si el valor de una determinada magnitud se expresa como un intervalo, lo que se quiere representar es que el valor real (y único) de la magnitud se encuentra acotado por el intervalo dado. En otras palabras, dicho intervalo es una estimación pesimista del rango en el que se encuentra el valor real. Una aplicación típica de éste paradigma es el manejo y análisis de datos cuyos valores contienen imprecisiones debidas a errores de medida. Si la información de la incertidumbre es más detallada, puede representarse mediante una función de distribución que podría estar en forma de histograma. Este enfoque se abordará con mayor profundidad en la sección 2.7.

La relación entre el intervalo (del análisis de intervalos) y un número borroso es bastante estrecha. El primero representa incertidumbre proveniente del proceso de medida, mientras que el segundo normalmente representa incertidumbre que proviene de estimaciones de los expertos. El intervalo puede verse como un caso particular de número borroso y un número borroso puede descomponerse en una familia de intervalos anidada (llamados α-cortes). Las técnicas del análisis de intervalos no fueron diseñadas para trabajar con los datos borrosos, sin embargo, debido a la descomposición de un número borroso en α-cortes, se usan a menudo para tratar α-corte a α-corte los da-

tos borrosos. La relación entre ambas áreas es clara y cada vez más fluida, como demuestra que se hayan dedicado o se vayan a dedicar a su estudio un número especial de la revista *Fuzzy Sets and Systems* (Lodwick y Jamison, 2003), un futuro número especial de *Reliable Computing*, revista de referencia en el análisis de intervalos, o sesiones dentro de conferencias tales como la *International Conference on Fuzzy Systems FUZZ-IEEE'2008*.

2.1.2. El Análisis de Datos Simbólicos

Tal y como se ha mencionado, el paradigma de los datos simbólicos es una extensión del paradigma de datos clásicos que tiene como objetivo plasmar situaciones reales que, por su complejidad y riqueza de información intrínseca, no pueden ser reflejadas adecuadamente mediante el paradigma clásico. Su principal característica es que pueden describir la variabilidad inherente a cada observación. Esto es posible porque su estructura es más compleja que la de los datos clásicos (e.g. los intervalos se caracterizan mediante dos valores que representan los extremos del intervalo). Al tener una estructura distinta que la de los datos clásicos, las técnicas de análisis del paradigma clásico no son válidas para analizar los datos simbólicos. Por ello, es necesario desarrollar un nuevo catálogo de métodos que sean capaces de extraer el conocimiento de este nuevo tipo de datos. Éste es el propósito del análisis de datos simbólicos.

Una de las aplicaciones más claras del análisis de datos simbólicos es la extracción de conocimiento de las grandes bases de datos. En los últimos tiempos, con el advenimiento de la era informática, la capacidad de almacenamiento de datos se ha disparado y las bases de datos pueden ser prácticamente tan grandes como se desee. En consecuencia, los conjuntos de datos que se pueden extraer de ellas pueden tener fácilmente cientos de variables y miles de registros. Afortunadamente, paralelamente a la capacidad de almacenamiento, también ha ido aumentando la capacidad de procesamiento de información y los ordenadores son capaces de trabajar con conjuntos de datos bastante grandes. Sin embargo, aunque en muchos casos sea computacionalmente posible abordar el análisis de los datos almacenados con las técnicas tradicionales, Billard y Diday (2003) se preguntan si es realmente apropiado. Siguiendo con el ejemplo de las tarjetas de créditos, cabe preguntarse si las transacciones que se añaden un tiempo más tarde, ¿pueden o deben considerarse como pertenecientes a la misma población? ¿qué pasa si los patrones de comportamiento han cambiado?

Más allá de estas preguntas, resulta incuestionable que las nuevas bases de datos presentan una oportunidad para el desarrollo de nuevas metodologías que permitan extraer el conocimiento subyacente. La aproximación que plantea el análisis de datos simbólicos consiste en resumir dichas bases de datos mediante un procedimiento de agregación y analizar el resultado de la agregación. Más concretamente, en el análisis de datos simbólicos los in-

dividuos se agregan en grupos que obedecen a un determinado criterio (e.g. varones de una determinada franja de edad, habitantes de una determinada región, transacciones realizadas en día festivo, etc.) y que reciben el nombre de objetos simbólicos. Los objetos simbólicos son representados mediante variables simbólicas que son capaces de recoger la variabilidad inherente a un grupo de individuos y son analizados con técnicas especialmente preparadas para trabajar con ellos. En Bock y Diday (2000) se pueden encontrar más detalles sobre la metodología que propone el análisis de datos simbólicos.

En algunos casos, el resultado de técnicas de análisis (simbólicas o no) puede ser expresado mediante datos simbólicos y ser el punto de partida de subsiguientes análisis. Por ejemplo, consideremos el caso de los datos del censo de un determinado país en el que se pretende formar clusters de gente desempleada y caracterizar dichos clusters mediante objetos simbólicos. A partir de las características de los clusters formados (i.e. de los valores simbólicos que describen los clusters), se pueden identificar los individuos que forman parte de dichos clusters en otros países y comparar ambas poblaciones mediante otras técnicas no simbólicas. Los trabajos de Laaksonen (2008) y de Mas y Olaeta (2008) contienen ejemplos de aplicaciones de análisis de datos simbólicos sobre datos reales.

El catálogo de métodos simbólicos desarrollados hasta el momento es escaso, aunque ha aumentado mucho en los últimos años. Por ello Billard y Diday (2003) exhortan a la comunidad científica para que desarrolle más métodos rigurosos desde el punto de vista matemático. Posteriormente, Billard y Diday (2006b) afirman que los métodos simbólicos desarrollados hasta el momento son intuitivamente correctos ya que funcionan y subsumen a los métodos clásicos, es decir, que, si se analizan datos clásicos con dichos métodos simbólicos, los resultados que se obtienen son los mismos que los que se conseguirían empleando los métodos clásicos.

Palumbo y Irpino (2005) consideran que muchas de las técnicas de análisis de datos simbólicos desarrolladas hasta el momento, son adaptaciones de técnicas ya existentes para datos clásicos. Sin embargo, afirman que dada su particular naturaleza, los datos simbólicos necesitan también de técnicas de análisis y de estadísticos que no tienen por qué tener un equivalente en el análisis de datos clásico. Estos autores reflexionan sobre la evolución del análisis de datos simbólicos y afirman que, en un primer momento, algunas de las técnicas propuestas recurren a una codificación de los datos simbólicos que les permitan ser tratados por los métodos tradicionales. Sin embargo, los autores consideran que el reto es desarrollar métodos numéricos y estadísticos que estén diseñados específicamente para trabajar con datos simbólicos.

2.1.3. Hitos en la historia del análisis de datos simbólicos

El padre del análisis de datos simbólicos es Edwin Diday. Antes del nacimiento de la disciplina, Diday había trabajado extensamente en el área

de los métodos de análisis *cluster*. Una de sus preocupaciones era la caracterización de los clusters que se obtenían en los análisis (Diday, 1976) y el llamado *clustering* conceptual (Michalski, Diday y Step, 1982). Estas ideas fueron seguramente el germen que dio lugar al análisis de datos simbólicos (Diday, 1987).

La disciplina recibe un notable impulso con el proyecto SODAS (*Symbolic Official Data Analysis System*), proyecto Esprit europeo llevado a cabo entre 1996 y 1999. La finalidad de este proyecto fue extender los métodos de análisis de datos clásicos a los datos simbólicos, ya que estos últimos son más completos que los clásicos a la hora de describir conceptos y grupos de individuos. Para ello, desarrollaron una metodología que abarcaba desde la extracción de datos simbólicos de grandes bases de datos, pasando por la visualización de datos simbólicos hasta su análisis mediante distintas técnicas. Dicha metodología quedó recogida en el software SODAS 1.0 [1].

En el proyecto SODAS participaron quince instituciones de nueve países europeos. El campo de aplicación principal del proyecto eran las estadísticas oficiales y por ello el proyecto contó con la colaboración de tres institutos estadísticos: EUSTAT del País Vasco, INE de Portugal y ONS de Inglaterra. Entre los frutos más significativos del proyecto SODAS cabe destacar la primera obra en recoger el estado del arte del análisis de datos simbólicos (Bock y Diday, 2000).

El proyecto ASSO (*Analysis System of Symbolic Official data*), desarrollado entre enero de 2001 y diciembre de 2003, supone la continuación del proyecto SODAS. ASSO formó parte del quinto Programa Marco de I+D de la Unión Europea. El objetivo de ASSO fue desarrollar métodos, metodologías y herramientas software para el análisis multidimensional de datos complejos (numéricos y no numéricos) procedentes de las bases de datos de los institutos y oficinas estadísticas y de la administración pública.

Entre los frutos del proyecto ASSO cabe destacar el nacimiento de la revista *Electronic Journal of Symbolic Data Analysis* (EJSDA) y la segunda versión del software de análisis de datos simbólicos SODAS [2]. En 2004, bajo el auspicio de la International Federation of Classification Societies (IFCS), se fundó un grupo transversal dedicado al análisis de datos simbólicos, cuyo objetivo es promover la materia mediante la creación de escuelas de verano, la difusión de material didáctico, el mantenimiento del EJSDA y la organización de talleres y sesiones especiales en conferencias promovidas por la IFCS u otras sociedades de clasificación, como se hizo en la conferencia del IFCS 2006 celebrada en Ljubljana.

Se han publicado artículos introductorios a la materia en revistas como *Journal of The American Statistical Association* (Billard y Diday, 2003) e *Intelligent Data Analysis* (Diday y Esposito, 2003). Además, desde el naci-

[1] Disponible en `http://www.ceremade.dauphine.fr/%7Etouati/sodas-pagegarde.htm`

[2] Disponible en `http://www.info.fundp.ac.be/asso/sodaslink.htm`

miento de la disciplina han ido apareciendo nuevos métodos de análisis de datos simbólicos en numerosos congresos internacionales y en revistas científicas entre las que cabe destacar *Pattern Recognition Letters* (Irpino, 2006; de Carvalho, Souza, Chavent y Lechevallier, 2006c), *Computational Statistics and Data Analysis* (Groenen, Winsberg, Rodriguez y Diday, 2006; Lima Neto y de Carvalho, 2008), *Advances in Data Analysis and Classification* (González-Rodríguez, Blanco, Corral y Colubi, 2007), *Discrete Applied Mathematics* (Diday y Vrac, 2005), *Intelligent Data Analysis* (Appice, d'Amato, Esposito y Malerba, 2006), *IEEE Transactions on Systems, Man and Cybernetics* (Ichino y Yaguchi, 1994) o *Pattern Recognition* (Gowda y Ravi, 1995), entre otras.

En consecuencia, el catálogo de métodos que trabajan con datos simbólicos esta creciendo en los últimos tiempos. Un hecho significativo en este aspecto es el número especial de la revista *Computational Statistics* dedicado al análisis de datos de intervalo (Palumbo, 2006) o el número especial de *Intelligent Data Analysis* dedicado a los datos simbólicos y a los datos espaciales (Brito y Noirhomme-Fraiture, 2006). Respecto a los métodos desarrollados, hay que decir que gran parte de ellos trabajan exclusivamente sobre variables de intervalos, mientras que los métodos dedicados a otras variables simbólicas, como por ejemplo los histogramas, son menores en número. La principal razón puede ser que el intervalo es el dato simbólico cuantitativo más sencillo.

En los últimos tiempos, han aparecido dos nuevos libros que revisan la disciplina (Billard y Diday, 2006b; Diday y Noirhomme, 2008). Estos dos libros son complementarios y suponen una actualización del primer volumen que trató a fondo esta materia (Bock y Diday, 2000). El primero realiza una revisión de la definición de los datos simbólicos y de las principales técnicas de análisis de los mismos (estadísticos descriptivos, análisis de *clusters*, de componentes principales y regresión lineal). El segundo aborda la relación entre los datos simbólicos y las bases de datos, explica el uso del software SODAS desarrollado para el análisis de datos simbólicos y amplía el catálogo de técnicas adaptadas para tratar con datos simbólicos entre las que se incluyen los perceptrones multicapa, los mapas de Kohonen, el análisis canónico y los árboles de decisión Bayesianos.

2.1.4. Situación de la tesis dentro del análisis de datos simbólicos

Aunque el análisis de datos simbólicos es una disciplina muy reciente, el conjunto de métodos simbólicos está aumentando notablemente en los últimos tiempos. Sin embargo, como es natural existen algunas áreas dentro de la minería de datos y del aprendizaje estadístico donde aún no hay desarrollos. Al comenzar esta tesis, una de estas áreas vírgenes era la de la predicción de series temporales valoradas mediante variables simbólicas.

Esta tesis fue planteada con el objetivo de comenzar a realizar desarrollos en dicho campo. Más concretamente, el objetivo es desarrollar métodos de predicción para series temporales valoradas mediante variables de intervalo y de histograma. Tal y como se comentará en el próximo capítulo, recientemente han aparecido algunas propuestas de métodos para la predicción de series temporales de intervalos. Esto es, sin duda, una buena señal que indica que la disciplina prosigue su avance y que los investigadores trabajan por ampliar sus límites.

En este capítulo, que revisa los logros del análisis de datos simbólicos, se prestará especial atención a los desarrollos para variables simbólicas de intervalos e histogramas, que, por otro lado, son las que han centrado la investigación en el área hasta el momento. Además, también se mostrarán otros conceptos no provenientes del análisis de datos simbólicos que también trabajan sobre intervalos o histogramas y que, o bien ofrecen una perspectiva diferente de los mismos, o bien son necesarios para los desarrollos que se explicarán más adelante.

2.2. Las variables simbólicas

Como ya se ha dicho, las variables clásicas asignan a cada individuo un único valor (ya sea categórico o cuantitativo). Sin embargo, las variables simbólicas pueden contener variación interna y pueden estar estructuradas. Es precisamente la presencia de variación interna la que hace que estas variables resulten adecuadas para representar conceptos, clases, o grupos de individuos. Como contraprestación, las variables simbólicas son más complejas que las clásicas, por lo que los métodos de análisis clásicos no pueden ser aplicados y se requieren métodos nuevos para tratar con este tipo de datos.

Variable aleatoria clásica. Un valor clásico o realización de una variable aleatoria Y en el individuo $w_u, u = 1, ..., m$, será denotado mediante x_i que representa a un único valor del dominio $\mathcal{Y}$. Si la variable es cuantitativa el dominio son los números reales, i.e. $Y(w_u) = x_i \in \Re$. Si la variable es cualitativa el dominio es un conjunto de categorías finito.

Como ejemplos, se puede considerar la variable $Y_1(w_u)$ que representa el peso del individuo w_u en kg, de forma que $Y_1(w_u) = 75$, y la variable Y_2 que representa el color del pelo de w_u, $Y_2(w_u) =$ moreno.

Variable aleatoria simbólica multivalorada. Una variable aleatoria multivalorada Y es aquella que para cada individuo $w_u, u = 1, ..., m$, toma como posibles valores uno o más valores de su dominio $\mathcal{Y}$, i.e. $Y(w_u) = \xi_u = \{a_{l_u}\}$, donde $a_{l_u} \in \mathcal{Y}$ con $l_u = 1, ..., k_u$. La lista completa de posibles valores $\mathcal{Y}$ es finita y debe estar definida, ya sean valores cuantitativos o cualitativos.

Un ejemplo podría ser la variable $Y(w_u)$ que representa las ciudades en las que ha residido el individuo w_u a lo largo de su vida, e.g. $Y(w_u) = \{\text{Madrid, Burgos, Santander}\}$.

Variable aleatoria simbólica de intervalo. Una variable aleatoria de intervalo es aquella que para cada individuo $w_u, u = 1, ..., m$, toma valores en forma de intervalo, i.e. $Y(w_u) = \xi_u = [a_u, b_u] \subset \Re^1$, donde $a_u \leq b_u$, y $a_u, b_u \in \Re^1$.

En principio, el intervalo puede ser cerrado o abierto en cualquiera de sus límites, i.e. $(a_u, b_u), [a_u, b_u], [a_u, b_u], \text{o}(a_u, b_u]$. El concepto de intervalo abierto o cerrado está muy ligado con el concepto de bola abierta o cerrada que se usa en topología. En esta disciplina, una bola abierta (resp. cerrada) es un conjunto de puntos que distan de otro punto menos (resp. no más) que un cierto radio. Un intervalo, que queda definido por sus extremos, puede, alternativamente, definirse por el centro (o punto medio del intervalo) y el radio del intervalo. Siguiendo la notación que se suele utilizar en topología para representar las bolas, el intervalo se reescribiría como

$$Y(w_u) = [a_u, b_u] = \bar{B}(c_u, r_u), \tag{2.1}$$

$$Y(w_u) = (a_u, b_u) = B(c_u, r_u), \tag{2.2}$$

donde $c_u = (b_u + a_u)/2$ es el centro del intervalo y $r_u = (b_u - a_u)/2$ su radio. De forma general, no se hará distinción sobre si los extremos del intervalo son abiertos o cerrados y se representará a los intervalos de la siguiente forma

$$Y(w_u) = \langle c_u, r_u \rangle. \tag{2.3}$$

Un ejemplo de variable de intervalo sería la variable $Y(w_u)$ que representase el rango del gasto medio mensual en euros en ocio del colectivo w_u. Si w_1 representa, pongamos por caso, al grupo de varones entre 12 y 16 años de Madrid un posible valor sería $Y(w_1) = [50, 120] = \langle 85, 35 \rangle$.

Variable aleatoria simbólica modal. Una variable aleatoria modal es aquella que para cada individuo $w_u, u = 1, ..., m$, toma valores de la siguiente forma $Y(w_u) = \xi_u = \{\eta_u, \pi_u\}$ donde η_u es un valor que pertenece al dominio $\mathcal{Y}$ de valores cuantitativos o cualitativos y cuya cardinalidad puede ser finita o infinita; π_u es una medida no-negativa asociada al valor η_u y que típicamente representa su peso, probabilidad, frecuencia relativa, aunque también podría representar su grado de necesidad o de posibilidad.

Por ejemplo, la variable $Y(w_u)$ en un estudio de mercado puede representar el producto elegido por el individuo w_u y su grado de satisfacción, representado mediante un entero entre 1 y 5, $Y(w_u) = \{\text{Coche modelo X}, 4\}$.

Variable aleatoria simbólica modal multivalorada. Sea $\mathcal{Y}$ un dominio de valores cuantitativos o cualitativos y cuya cardinalidad puede ser finita o infinita, una variable aleatoria modal multivalorada, Y, es aquella que toma como valor un subconjunto de $\mathcal{Y}$ donde cada uno de los valores de ese subconjunto tiene una medida asociada. Esto es, el valor de la variable para el individuo $w_u, u = 1, ..., m$, es $Y(w_u) = \xi_u = \{\eta_{uk}, \pi_{uk}\}$ con $k = 1, ..., s_u$. De esta manera, $\{\eta_1, ..., \eta_{s_u}\} \subseteq \mathcal{Y}$.

Por ejemplo, la variable $Y(w_u)$ puede representar el grado de concordancia con la idea w_u de los entrevistados en una encuesta de opinión. De esta forma, un posible valor sería $Y(w_u) = \{$Muy de acuerdo, 0.3; De acuerdo, 0.3; Neutral, 0.2; En contra, 0.1; Muy en contra, $0.1\}$. Como se puede ver la variable modal multivalorada es, en este caso, un distribución de frecuencias de una variable cualitativa.

Variable aleatoria simbólica de histograma. Una variable aleatoria de histograma es aquella variable cuantitativa en la que un individuo $w_u, u = 1, ..., m$, toma valores en $\Re$ de la siguiente forma $Y(w_u) = \xi_u = \{[a_{uk}, b_{uk}), \pi_{uk}\}$ con $k = 1, ..., s_u$, donde s_u es un número finito de intervalos que forman el soporte del histograma para la observación w_u; y donde π_{uk} es el peso en forma de probabilidad o frecuencia asignado al intervalo $[a_{uk}, b_{uk})$ con $\sum_{k=1}^{s_u} \pi_{uk} = 1$.

Un ejemplo de variable simbólica de histograma puede ser el valor $Y(w_u)$, donde w_u representa una acción bursátil en un determinado periodo de tiempo, y $Y(w_u)$ es una variable de histograma que representase la distribución de la cotización de w_u en euros en el periodo considerado. Un posible valor sería $Y(w_u) = \{[22, 23), .16; [23, 24), .34; [24, 25), .3; [25, 26), .2\}$.

Desde esta perspectiva se puede apreciar como las variables simbólicas subsumen a las variables clásicas. En realidad, un valor cuantitativo clásico puede considerarse como un caso particular de una variable de intervalo (i.e. un intervalo en el que los extremos coinciden en el mismo punto), o de una variable de histograma (i.e. un histograma donde el único intervalo tiene probabilidad 1, y sus extremos coinciden en el mismo punto). A su vez, el intervalo puede considerarse como un caso particular de una variable de histograma, es decir, un histograma con un único intervalo con probabilidad 1. De manera análoga, un valor cualitativo clásico puede considerarse como un caso particular de una variable simbólica modal simple o multivalorada.

Los datos simbólicos, además de permitir representar la variabilidad interna de las observaciones, también permiten representar información estructurada. Más concretamente, permite taxonomías, e.g. el color es "claro" si es "amarillo", "blanco" o "rosa"; dependencias jerárquicas, e.g. la variable "marca del coche" sólo debe aparecer si la variable "¿tiene coche?" es verdadera,

por tanto, ambas variables están relacionadas jerárquicamente; dependencias lógicas, e.g. si la edad de un individuo es menor de dos meses, la altura de dicho individuos es forzosamente menor de 75cm.

En el paradigma clásico, en la tabla de datos a cada individuo (fila de la tabla), le corresponde un único valor de cada variable (columna). En el paradigma simbólico, la tabla de datos es similar, pero las celdas contienen datos simbólicos. Cada fila contendrá la descripción simbólica de un individuo y dichos individuos pueden ser entes colectivos o conceptos.

2.2.1. Descripción virtual de una observación simbólica

Para realizar análisis estadístico de individuos descritos con variables simbólicas es necesario tener en cuenta que las variables simbólicas acarrean de forma implícita una variabilidad. Por ejemplo, una observación simbólica en forma de intervalo quiere decir que todos los posibles valores de dicha variable están contenidos dentro del intervalo. Puede darse el caso de que al considerar un individuo descrito por dos o más variables simbólicas no todas las posibles combinaciones de tuplas de valores sean válidas, ya que pueden ser incongruentes o no coherentes con lo que pretenden representar. Esta idea se expresa más formalmente a continuación mediante el concepto de descripción virtual.

Descripción simbólica. La descripción simbólica de una observación $w_u \in E$, donde $u = 1, ..., m$ y E es un conjunto de individuos, es dada por el vector descripción $\boldsymbol{d}_u = (\xi_{u1}, ..., \xi_{up})$, o, de forma más general por $\boldsymbol{d} \in (D_1, ..., D_p)$ en el espacio $\mathcal{D} = \times_{j=1}^{p} D_j$, donde la realización de la variable Y_j puede ser un valor clásico, x_j, o un valor simbólico, ξ_j.

Descripción individual. Son aquellos vectores descripción, denotados por $\boldsymbol{x}$, para los cuales cada D_j es un conjunto de un único valor, i.e., $\boldsymbol{x} = (x_1, ..., x_p) \equiv \boldsymbol{d} = (\{x_1\}, ..., \{x_p\})$, con $\boldsymbol{x} \in \mathcal{X} = \times_{j=1}^{p} \mathcal{Y}_j$.

Por ejemplo, consideremos el caso de dos variables (Y_1, Y_2) donde Y_1 representa la marca de coche de un individuo con valores $\mathcal{Y}_1 = \{\text{Seat, Citroen...}\}$, y Y_2 representa si el individuo tiene garaje con valores $\mathcal{Y}_2 = \{\text{No} = 0, \text{Sí} = 1\}$. El valor simbólico $(Y_{u1}, Y_{u2}) = \xi_u = (\{\text{Opel, Audi}\}, \{1\})$ tiene asociadas dos descripciones individuales $\boldsymbol{x} = (\text{Opel}, 1)$ y $\boldsymbol{x} = (\text{Audi}, 1)$.

Dependencia lógica. Una dependencia lógica puede expresarse como una regla v,

$$v : [x \in A] \Rightarrow [x \in B] \tag{2.4}$$

para $A \subseteq D$, $B \subseteq D$, y $x \in \mathcal{X}$ y donde v es una aplicación de $\mathcal{X}$ en $\{0, 1\}$ con $v(x) = 1$ (0) si la regla (no) es satisfecha por x. De ahí se deduce que

un vector descripción individual $\boldsymbol{x}$ satisface la regla v si y sólo si $x \in A \cap B$ o $x \notin A$. Las reglas normalmente afectan a más de una variable.

Descripción virtual. La descripción virtual $vir(\boldsymbol{d})$ del vector descripción $\boldsymbol{d}$ es el conjunto de todos los vectores descripción individuales $\boldsymbol{x}$ que satisfacen todas las reglas de dependencia lógica v definidas sobre $\mathcal{X}$. Al conjunto de reglas lógicas se le denota mediante $V_{\mathcal{X}}$. El concepto de descripción virtual se formaliza de la siguiente manera:

$$vir(\boldsymbol{d}) = \{x \in D; v(x) = 1, \forall v, v \in V_{\mathcal{X}}\}. \tag{2.5}$$

Esto quiere decir que a la hora de realizar un análisis estadístico sobre datos simbólicos sólo deben tenerse en cuenta aquellas descripciones individuales que cumplen las reglas de dependencia lógica definidas sobre las variables consideradas. Es decir, sólo deben tenerse en cuenta aquellas descripciones individuales que pertenecen a la descripción virtual del vector descripción simbólico.

2.3. Estadísticos descriptivos univariantes

El análisis descriptivo básico de una variable aleatoria suele incluir la representación del histograma de frecuencias y el cálculo de estadísticos como la media muestral y la varianza. El contexto en el que se van a definir estos estadísticos es el siguiente.

Cada individuo $w_u \in E$ con $u = 1, ..., m$ tomará como valor $\boldsymbol{\xi}_u = (\xi_{u1}, ..., \xi_{up})$ en la variable aleatoria $\boldsymbol{Y} = (Y_1, ..., Y_p)$. Para agilizar la notación, en lo subsiguiente, se identificará $w_u \in E$ como $u \in E$.

El cálculo de los estadísticos en las variables simbólicas se ve afectado por la presencia de reglas de dependencia. Se pueden consultar los detalles en el capítulo 3 de Billard y Diday (2006b) y en Billard y Diday (2006a), para las variables de intervalo. Para el propósito de esta tesis basta decir que si tenemos p variables simbólicas medidas sobre el individuo u y un conjunto de reglas $v = (v_1, v_2, ...)$ para calcular los estadísticos, sólo se deben tener en cuenta aquellos vectores de descripción individuales $x \in vir(d_u)$ que satisfacen todo el conjunto de reglas v.

2.3.1. Variables de intervalo

Los estadísticos descriptivos univariantes para variables de intervalo fueron propuestos por Bertrand y Goupil (2000). Dada una variable aleatoria de intervalo $Y_j \equiv Z$, la realización de Z en el individuo $u \in E$ es el valor

de intervalo $Z(u) = [a_u, b_u]$. Se asume que los vectores de descripción individuales $x \in vir(d_u)$ se distribuyen de manera uniforme sobre $Z(u)$. Por lo tanto, para cada ξ,

$$P\{x \leq \xi | x \in vir(d_u)\} = \begin{cases} 0, & \xi < a_u, \\ \frac{\xi - a_u}{b_u - a_u,} & a_u \xi < b_u, \\ 1, & \xi \geq b_u. \end{cases} \tag{2.6}$$

Los vectores de descripción individuales x toman valores en $\bigcup_{u \in E} vir(d_u)$. Además, hay que considerar que se asume que cada objeto es igualmente probable con probabilidad $1/m$.

Función de distribución empírica. Dada la fórmula (2.6), la función de distribución empírica, $F_Z(\xi)$, es la función de distribución de una mixtura de m distribuciones uniformes $\{Z(u), u = 1, ..., m\}$, tal que

$$\begin{aligned} F_Z(\xi) &= \frac{1}{m} \sum_{u \in E} P\{x \leq \xi | x \in vir(d_u)\} \\ &= \frac{1}{m} \left(g_Z(\xi) + \sum_{\xi \in Z(u)} \left(\frac{\xi - a_u}{b_u - a_u} \right) \right), \end{aligned} \tag{2.7}$$

donde $g_Z(\xi)$ es el número de intervalos de la variable Z para los cuales ξ es mayor o igual que su extremo superior.

Función de densidad empírica. La función de densidad empírica de Z es

$$f(\xi) = \frac{1}{m} \sum_{u : \xi \in Z(u)} \left(\frac{1}{b_u - a_u} \right); \tag{2.8}$$

o, equivalentemente,

$$f(\xi) = \frac{1}{m} \sum_{u \in E} \frac{I_u(\xi)}{\|Z(u)\|}, \text{ con } \xi \in \Re, \tag{2.9}$$

donde $I_u(\cdot)$ es la función que indica si ξ está o no en el intervalo $Z(u)$ y donde $\|Z(u)\|$ es la longitud del intervalo. Nótese que el sumatorio se realiza sólo sobre aquellos individuos u para los que $\xi \in Z(u)$.

Histograma. Para construir el histograma de Z es necesario definir el intervalo $I = [min_{u \in a}, max_{u \in b} b_u]$ que abarca todos los valores observados de Z en el dominio $\mathcal{X}$, y dividir I en r intervalos $I_g = [\zeta_{g-1}, \zeta_g)$, con $g =$

$1, ..., r-1$, e $I_r = [\zeta_{r-1}, \zeta_r]$. Entonces la frecuencia observada para el intervalo es I_g es

$$\frac{f_g = \sum_{u \in E} \|Z(u) \cap I_g\|}{\|Z(u)\|,} \tag{2.10}$$

y la frecuencia relativa es

$$p_g = f_g/m. \tag{2.11}$$

En este contexto p_g se interpreta como la probabilidad o la frecuencia relativa de que un vector descripción individual arbitrario esté contenido en el intervalo I_g. El histograma de Z es la representación gráfica de

$$\{(I_g, f_g), g = 1, ..., r\}, \text{donde el área de } I_g \text{ es } p_g = (\xi_g - \xi_{g-1}) f_g. \tag{2.12}$$

Bertrand y Goupil (2000) indican que usando la ley de los grandes números, la distribución límite verdadera de Z como $m \to \infty$ es sólo aproximada por la distribución exacta $f(\xi)$ de la fórmula (2.9) ya que ésta depende de la veracidad del hecho de que realmente el interior de cada intervalo se distribuya según una distribución uniforme.

Media muestral simbólica. La media muestral simbólica de una variable aleatoria de intervalo Z se calcula como sigue a partir de la fórmula de la función de densidad empírica (2.9)

$$\begin{aligned}
\bar{Z} &= \int_{-\infty}^{\infty} \xi f(\xi) d\xi, \\
&= \frac{1}{m} \sum_{u \in E} \int_{-\infty}^{\infty} \frac{I_u(\xi)}{\|Z(u)\|} \xi d\xi \\
&= \frac{1}{m} \sum_{u \in E} \frac{1}{b_u - a_u} \int_{\xi \in Z(u)} \xi d\xi \\
&= \frac{1}{2m} \sum_{u \in E} \frac{b_u^2 - a_u^2}{b_u - a_u} \\
&= \frac{1}{2m} \sum_{u \in E} (b_u + a_u).
\end{aligned} \tag{2.13}$$

Esta fórmula es consistente con la afirmación de que dentro de cada intervalo los valores se distribuyen según una uniforme, es decir, $Z(u) \sim U(a_u, b_u)$. Por tanto, $E(Z(u)) = \frac{a_u + b_u}{2}$. Puede verse claramente como la media de una variable de intervalo es la media de los centros, $c_u = \frac{a_u + b_u}{2}$, de cada una de sus realizaciones.

Varianza muestral simbólica. La varianza muestral simbólica de una variable aleatoria de intervalo Z se calcula de la siguiente manera

$$\begin{aligned} S^2 &= \int_{-\infty}^{\infty} (\xi - \bar{Z})^2 f(\xi) d\xi, \\ &= \left(\int_{-\infty}^{\infty} \xi^2 f(\xi) d\xi \right) - \bar{Z}^2. \end{aligned} \tag{2.14}$$

A continuación, utilizamos la fórmula de la función de densidad empírica (2.9) para calcular lo siguiente

$$\begin{aligned} \int_{-\infty}^{\infty} \xi^2 f(\xi) d\xi &= \frac{1}{m} \sum_{u \in E} \int_{-\infty}^{\infty} \xi^2 \frac{I_u(\xi)}{\|Z(u)\|} d\xi \\ &= \frac{1}{m} \sum_{u \in E} \int_{a_u}^{b_u} \frac{\xi^2}{b_u - a_u} d\xi \\ &= \frac{1}{3m} \sum_{u \in E} \frac{b_u^3 + a_u^3}{b_u + a_u}. \end{aligned} \tag{2.15}$$

A partir de las fórmulas (2.13), (2.14) y (2.15) obtenemos

$$S^2 = \frac{1}{3m} \sum_{u \in E} (b_u^2 + b_u a_u + a_u^2) - \frac{1}{4m^2} \left[\sum_{u \in E} (b_u + a_u) \right]^2. \tag{2.16}$$

La varianza se comporta de acuerdo con la afirmación de que dentro de cada intervalo los valores se distribuyen según una uniforme, es decir, $Z(u) \sim U(a_u, b_u)$, por lo que se cumple que $Var(Z(u)) = \frac{(b_u - a_u)^2}{12}$.

En Billard y Diday (2000) la definición de varianza que ofrecen los autores es la siguiente

$$S^2 = \frac{1}{4m} \sum_{u \in E} (b_u + a_u)^2 - \frac{1}{4m^2} \left[\sum_{u \in E} (b_u + a_u) \right]^2. \tag{2.17}$$

Esta definición puede verse como la varianza de los centros de los intervalos expresada como la diferencia entre el momento de orden 2 y el cuadrado del valor esperado (i.e. de la media). Al manejar sólo los centros esta definición no tiene en cuenta la longitud de los intervalos.

Tanto la definición de varianza de (2.16), como la de (2.17) subsumen a la definición de varianza para datos clásicos. Esto quiere decir que si utilizamos datos clásicos en ambas definiciones obtenemos el mismo resultado que se obtiene con la varianza clásica. Según la definición (2.17), el valor de la varianza de un conjunto de intervalos iguales es cero, i.e. sea la variable Z tal que $\forall u \in E$, $Z(u) = [a, b]$, entonces $S_Z^2 = 0$. Este resultado es coherente con

el hecho de que los intervalos son todos iguales, sin embargo, esta definición no sirve describir la variabilidad interna presente en cada intervalo; es decir, la variabilidad debida a la longitud del intervalo. Dicha variabilidad si queda recogida mediante la definición (2.16) que recoge la varianza de cada intervalo modelándola como la varianza de una distribución uniforme.

2.3.1.1. Estadísticos basados en distancias

Chavent y Saracco (2008) proponen medidas de tendencia central y de dispersión para intervalos definidas por medio de distancias. Para ello, muestran la relación que existe entre los estadísticos de tendencia central para números reales y unas distancias determinadas. Dicha relación dice que el estadístico de tencencia central $\hat{c}$ pueden obtenerse resolviendo la siguiente minimización

$$\hat{c} = \arg\min_{c\in\Re} S_p(c) = \arg\min_{c\in\Re} \|\boldsymbol{z} - \boldsymbol{c}\|_p, \tag{2.18}$$

donde $\boldsymbol{z} \in \Re^m$ es un vector de m observaciones, $\|\cdot\|_p$ es la norma L_p en $\Re^m$ y $\boldsymbol{c} \in \Re^m$ es un vector con todos los elementos idénticos, i.e. $\boldsymbol{c} = c\boldsymbol{I}_m$, donde $\boldsymbol{I}_m$ es el vector unidad de dimensión m y $c \in \Re$. De forma que

- Si $p = 1$, entonces $\hat{c}$ es la mediana muestral del vector de observaciones $\boldsymbol{z}$ y $\frac{S_1(\hat{c})}{m}$ es la desviación absoluta media respecto a la mediana.
- Si $p = 2$, entonces $\hat{c}$ es la media muestral del vector de observaciones $\boldsymbol{z}$ y $\frac{S_2(\hat{c})}{\sqrt{m-1}}$ es la desviación típica muestral.
- Si $p = \infty$, entonces $\hat{c}$ es el rango medio del vector de observaciones $\boldsymbol{z}$ y $2S_\infty(\hat{c})$ es el rango muestral.

Chavent y Saracco (2008) extienden esta idea para desarrollar estadísticos de tendencia central para intervalos, tal y como se muestra a continuación.

El intervalo $\hat{C} = [\hat{\alpha}, \hat{\beta}]$, que denota una medida de tendencia central para la variable aleatoria de intervalo Z definida para todos los m individuos de una muestra, se obtiene como resultado de la minimización de la medida de dispersión $S_p(C)$

$$\hat{C} = \arg\min_C S_p(C) = \arg\min_C \left(\sum_{u=1}^{m} (d(C, Z(u)))^p\right)^{\frac{1}{p}}, \tag{2.19}$$

donde $C = [\alpha, \beta]$ es un intervalo, $d(C, Z(u))$ es una distancia para intervalos que hace la función que hacía el valor absoluto de la diferencia en el caso de los números reales, y p es el orden de la norma. Si consideramos que los intervalos $Z(u)$ y C, pueden reescribirse en notación centro-radio como $Z(u) = \langle c_u, r_u\rangle$ y $C = \langle\gamma, \rho\rangle$, y consideramos distintas distancias y órdenes p obtenemos los siguiente:

- Si $p = 1$ y la distancia que se utiliza es la distancia de Hausdorff, $d(C, Z(u)) = |\gamma - c_u| + |\rho - r_u|$, entonces $\hat{C} = \langle \hat{\gamma}, \hat{\rho} \rangle$ es un intervalo mediano cuyo centro y radio se obtienen como

$$\hat{\gamma} = \text{mediana}(c_u), \text{ con } u = 1, ..., m \tag{2.20}$$

$$\hat{\rho} = \text{mediana}(r_u), \text{ con } u = 1, ..., m. \tag{2.21}$$

- Si $p = 2$ y la distancia que se utiliza es la distancia de Hausdorff, $d(C, Z(u)) = |\gamma - c_u| + |\rho - r_u|$, entonces $\hat{C} = [\hat{\alpha}, \hat{\beta}]$ es un intervalo de tendencia central cuyo mínimo y máximo se obtienen resolviendo un problema de minimización de una función cuadrática. Pueden consultarse más detalles en Chavent y Saracco (2008).

- Si $p = \infty$ y la distancia que se utiliza es la distancia de Hausdorff, $d(C, Z(u)) = |\gamma - c_u| + |\rho - r_u|$, entonces $\hat{C} = [\hat{\alpha}, \hat{\beta}]$ es un intervalo de tendencia central cuyo mínimo y máximo se obtienen como

$$\hat{\alpha} = \frac{\text{máx}_u(a_u) - \text{mín}_u(a_u)}{2} \tag{2.22}$$

$$\hat{\beta} = \frac{\text{máx}_u(b_u) - \text{mín}_u(b_u)}{2}. \tag{2.23}$$

- Si $p = 2$ y la distancia que se utiliza es una distancia L_2 del tipo, $d(C, Z(u)) = \sqrt{(\gamma - c_u)^2 + (\rho - r_u)^2}$, entonces $\hat{C} = \langle \hat{\gamma}, \hat{\rho} \rangle$ es el intervalo medio cuyo centro y radio se obtienen como

$$\hat{\gamma} = \frac{1}{m} \sum_{u=1}^{m} c_u \tag{2.24}$$

$$\hat{\rho} = \frac{1}{m} \sum_{u=1}^{m} r_u. \tag{2.25}$$

El intervalo $\hat{C} = [\hat{\alpha}, \hat{\beta}]$ se puede hallar de forma equivalente como

$$\hat{\alpha} = \frac{1}{m} \sum_{u=1}^{m} a_u \tag{2.26}$$

$$\hat{\beta} = \frac{1}{m} \sum_{u=1}^{m} b_u. \tag{2.27}$$

2.3.2. Variables de histograma

Se considera una variable aleatoria de histograma $Y \equiv Z$ y el individuo u con $u \in E = [1, ..., m]$, el valor $Z(u)$ de la variable para el individuo u es un conjunto de intervalos $\xi_{uk} = [a_{uk}, b_{uk})$ con $k = 1, ..., s_u$, donde cada intervalo tiene una probabilidad asociada p_{uk}, tal que $\sum_{k=1}^{s_u} p_{uk} = 1$.

Por analogía al caso de las variables de intervalo, se asume que dentro de cada intervalo ξ_{uk} cada vector de descripción individual $x \in vir(d_u)$ está distribuido de manera uniforme dentro de dicho intervalo. Los estadísticos descriptivos univariantes para variables de histograma fueron propuestos por Billard y Diday (2003).

Función de distribución empírica. La función de distribución empírica, $F_Z(\xi)$, es la función de distribución de una mixtura de m histogramas $w_u, u = 1, ..., m$, donde cada uno de los histogramas se descompone en s_u distribuciones uniformes $I_{uk} \sim U(a_{uk}, b_{uk})$ con $k = 1, ..., s_u$, y donde cada distribución uniforme tiene una probabilidad asociada p_{uk}. La función de distribución empírica se representa de la siguiente manera

$$\begin{aligned} F_Z(\xi) &= \frac{1}{m} \sum_{u \in E} P\{x \leq \xi | x \in vir(d_u)\} \\ &= \frac{1}{m} \sum_{u \in E} \left(\sum_{\xi \geq b_{uk}} p_{uk} + \sum_{\xi \in I_{uk}} p_{uk} \left(\frac{\xi - a_{uk}}{b_{uk} - a_{uk}} \right) \right). \end{aligned} \quad (2.28)$$

Función de densidad empírica. Según Billard y Diday (2003), la función de densidad empírica de Z es

$$f(\xi) = \frac{1}{m} \sum_{u \in E} \sum_{k=1}^{s_u} \frac{I_{uk}(\xi)}{\|Z(u;k)\|} p_{uk}, \text{ con } \xi \in \Re, \quad (2.29)$$

donde $I_u(\cdot)$ es la función binaria que vale 1 si ξ está en el intervalo $Z(u)$ y 0 en caso contrario, y donde $\|Z(u;k)\|$ representa la longitud del intervalo $Z(u;k)$.

Histograma. El histograma que se obtiene a partir de un conjunto de histogramas observados se construye de la siguiente manera. Sea $I = [min_{k,u \in E} a_{ku}, max_{k,u \in E} b_{ku}]$ el intervalo que abarca todos los valores observados de Z en el dominio $\mathcal{X}$, y sea una partición de I en r intervalos, $I_g = [\zeta_{g-1}, \zeta_g)$, con $g = 1, ..., r-1$, e $I_r = [\zeta_{r-1}, \zeta_r]$. Entonces la frecuencia observada para el intervalo es I_g es

$$O_Z(g) = \sum_{u \in E} \pi_Z(g;u), \quad (2.30)$$

donde

$$\pi_Z(g;u) = \sum_{k \in Z(g)} \frac{\|Z(k;u) \cap I_g\|}{\|Z(k;u)\|} p_{uk}, \quad (2.31)$$

donde $Z(k;u)$ es el intervalo $[a_{uk}, b_{uk})$, y donde el conjunto $Z(g)$ representa todos los intervalos $Z(k;u) \equiv [a_{uk}, b_{uk})$ que se solapan con I_g, para un individuo u dado. Por tanto, el cociente del sumatorio representa la fracción del intervalo $Z(k;u)$ que se solapa con I_g.

De lo anterior se deduce que $\sum_{g=1}^{r} O_Z(g) = m$. De ahí que la frecuencia relativa para el intervalo I_g sea $p_g = O_Z(g)/m$. El conjunto de valores $\{(p_g, Ig), g = 1, ..., r\}$ representa el histograma de frecuencias relativas construido para la variable de histograma Z.

Media muestral simbólica Dada la fórmula de la función de densidad empírica (2.9), la media muestral simbólica para histogramas se define como

$$\bar{Z} = \frac{1}{m} \sum_{u \in E} \left[\sum_{k=1}^{s_u} (b_{uk} + a_{uk}) p_{uk} \right]. \tag{2.32}$$

Al igual que en el caso de las variables de intervalo, esta fórmula es consistente con la afirmación de que dentro de cada intervalo los valores se distribuyen según una uniforme.

Varianza muestral simbólica. La varianza muestral simbólica de una variable aleatoria de histograma Z se calcula de la siguiente manera

$$S^2 = \frac{1}{3m} \sum_{u \in E} \left[\sum_{k=1}^{s_u} (b_{uk}^2 + b_{uk} a_{uk} + a_{uk}^2) p_{uk} \right] - \frac{1}{4m^2} \left[\sum_{u \in E} \sum_{k=1}^{s_u} (b_{uk} + a_{uk}) p_{uk} \right]^2. \tag{2.33}$$

La varianza se comporta de acuerdo con la afirmación de que dentro de cada intervalo los valores se distribuyen según una uniforme.

En Billard y Diday (2002) la definición de varianza que ofrecen los autores es la siguiente

$$S^2 = \frac{1}{4m} \sum_{u \in E} \left[\sum_{k=1}^{s_u} (b_{uk} + a_{uk}) p_{uk} \right]^2 - \frac{1}{4m^2} \left[\sum_{u \in E} \sum_{k=1}^{s_u} (b_{uk} + a_{uk}) p_{uk} \right]^2. \tag{2.34}$$

Como se puede ver, los estadísticos descriptivos para histogramas subsumen a los estadísticos descriptivos para intervalos. Esto es correcto ya que un intervalo no es más que un caso particular de un histograma. Para el caso de la varianza, la varianza para histogramas de (2.33) subsume a la definición de varianza para intervalos de (2.16) y a la definición de varianza para datos clásicos. De igual manera, la definición de (2.34) , subsume a la de intervalos de (2.17), que a su vez, subsume a la definición de varianza para intervalos datos clásicos. Al igual que como se ha explicado para el caso de los intervalos, la diferencia entre ambas definiciones radica en que la definición (2.33) sí

recoge la variabilidad interna de cada observación (i.e. de cada histograma), mientras que la definición de (2.34) no lo hace.

2.4. Estadísticos descriptivos bivariantes

Cuando se estudian dos variables de manera simultánea, interesa conocer cómo es su distribución conjunta, esto se puede hacer con ayuda del histograma conjunto de ambas variables. Por su parte, los estadísticos descriptivos bivariantes que caracterizan cómo es la relación entre las dos variables interés son las medidas de dependencia.

Los estadísticos descriptivos bivariantes para variables simbólicas de intervalo y de histograma fueron propuestos por Billard y Diday (2003). Al igual que en el caso univariante, el cálculo de estos estadísticos se ve afectado por la presencia de reglas de dependencia. Si tenemos p variables y un conjunto de reglas $v = (v_1, v_2, ...)$, y $R(u)$ representa el rectángulo p-dimensional de la observación u, para calcular los estadísticos sólo debe tenerse en cuenta la región $V(u) \subseteq R(u)$ que satisface todo el conjunto de reglas. Para obtener más detalles se puede consultar el capítulo 4 de Billard y Diday (2006b) y el artículo Billard y Diday (2006a) que desarrolla esta cuestión para variables de intervalo.

2.4.1. Variables de intervalo

Consideremos dos variables aleatorias simbólicas de intervalo $\boldsymbol{Y} = \{Y_{j_1}, Y_{j_2}\}$ valoradas para cada individuo $w_u \in E$ con $u = 1, ..., m$. Cada una de estas variables Y_j tomará valores en un subconjunto $\mathcal{Y}_j$ de la recta real $\Re$, $\mathcal{Y}_j \subseteq \Re$, con $\mathcal{Y}_j = \{[\alpha, \beta], -\infty < \alpha, \beta < \infty\}$. Por coherencia con la notación anterior, se renombrará a las variables de la siguiente manera $Y_{j_1} \equiv Z_1$ y $Y_{j_2} \equiv Z_2$. Dichas variables toman como valor para cada individuo $w_u \in E$ un rectángulo $\boldsymbol{Z}(u) = Z_1(u) \times Z_2(u) = ([a_{u1}, b_{u1}], [a_{u2}, b_{u2}])$. Se asume que los vectores descripción individuales $x \in vir(d_u)$ se distribuyen uniformemente sobre los intervalos $Z_1(u)$ y $Z_2(u)$.

Función de distribución conjunta empírica. La función de distribución conjunta empírica, $F_Z(\xi_1, \xi_2)$, se expresa de la siguiente forma

$$\begin{aligned} F_Z(\xi_1, \xi_2) &= \frac{1}{m} \sum_{u \in E} P\{x_1 \leq \xi_1, x_2 \leq \xi_2 | (x_1, x_2) \in vir(d_u)\} \\ &= \frac{1}{m} \left(g_Z(\xi_1, \xi_2) \right. \\ &+ \left. \sum_{(\xi_1 < b_{u1} \text{ o } \xi_2 < b_{u2})} \left(\frac{\xi_1 - a_{u1}}{b_{u1} - a_{u1}} \right) \left(\frac{\xi_2 - a_{u2}}{b_{u2} - a_{u2}} \right) \right), \end{aligned} \quad (2.35)$$

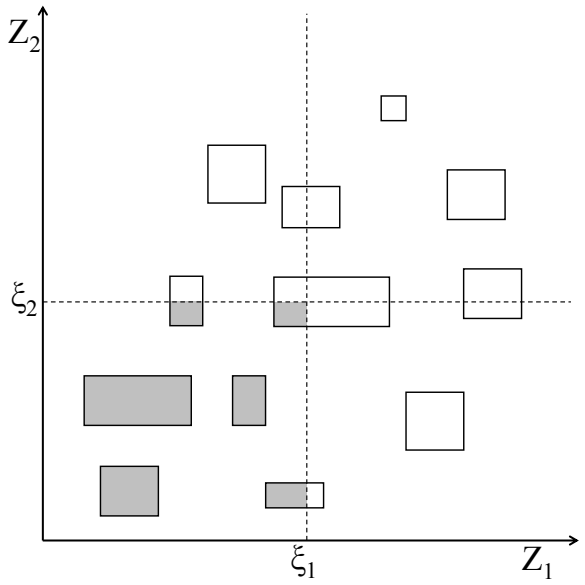


Figura 2.1: Función de distribución conjunta empírica de dos variables de intervalo Y_1 y Y_2 para el valor (ξ_1, ξ_2)

donde $g_Z(\xi_1, \xi_2)$ es el número de rectángulos $\boldsymbol{Z}(u)$ para los cuales $\xi_1 \geq b_{u1}$ y $\xi_2 \geq b_{u2}$. La figura 2.1 ilustra el concepto de función de distribución conjunta empírica.

Función de densidad conjunta empírica. Se define como

$$f(\xi_1, \xi_2) = \frac{1}{m} \sum_{u \in E} \frac{I_u(\xi_1, \xi_2)}{\|Z(u)\|}, \tag{2.36}$$

donde $I_u(\xi_1, \xi_2)$ es la función booleana que indica si el valor (ξ_1, ξ_2) se encuentra o no dentro del rectángulo $Z(u)$ y donde $\|Z(u)\|$ es el área de dicho rectángulo.

Histograma conjunto. De forma análoga al caso univariante (ver sección 2.3.1), se deben dividir los intervalos I_1 e I_2 que abarcan todos los valores observados en Z_1 y Z_2 en r_1 y r_2 intervalos,$I_{1,g_1} = [\xi_{1,g_1-1}, \xi_{1,g_1})$ e $I_{2,g_2} = [\xi_{2,g_2-1}, \xi_{2,g_2})$, respectivamente con $g_1 = 1, ..., r_1$ y $g_2 = 1, .., r_2$. Esto da lugar a una partición del espacio $\Re^2$ en una rejilla donde cada celda es un rectángulo $R_{g_1 g_2} = [\xi_{1,g_1-1}, \xi_{1,g_1}) \times [\xi_{2,g_2-1}, \xi_{2,g_2})$. La frecuencia absoluta asociada al rectángulo $R_{g_1 g_2}$ se calcula de la siguiente manera

$$f_{g_1 g_2} = \sum_{u \in E} \frac{\boldsymbol{Z}(u) \cap R_{g_1 g_2}}{\|\boldsymbol{Z}(u)\|}. \tag{2.37}$$

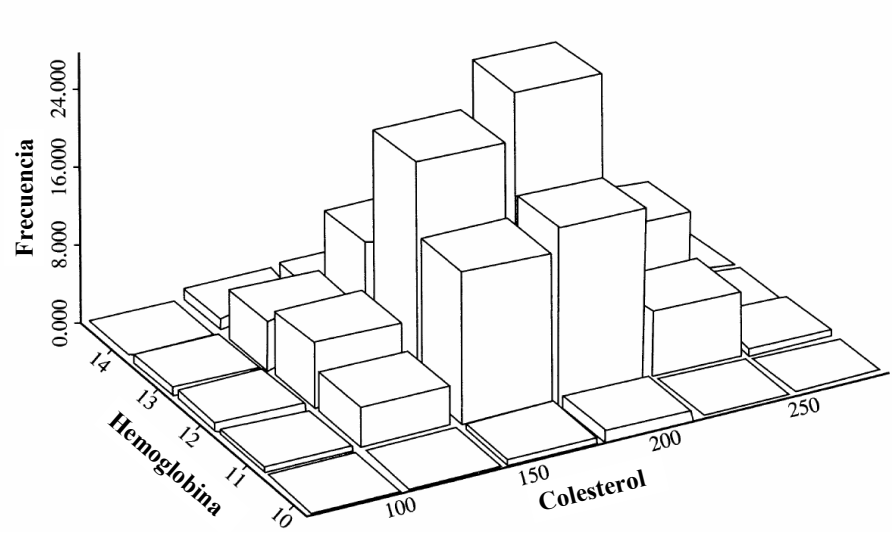


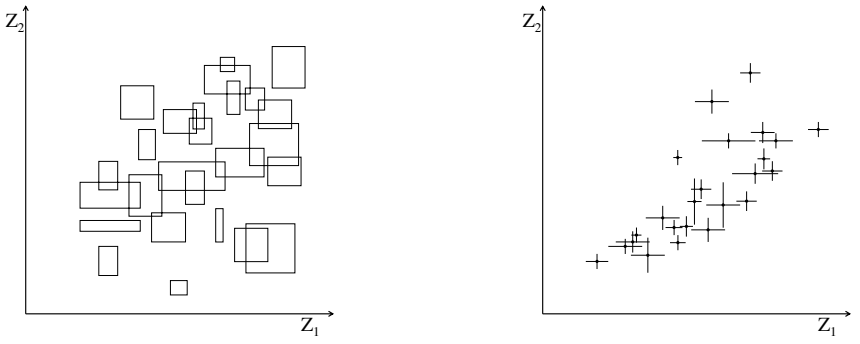
Figura 2.2: Histograma conjunto de dos variables de intervalo, hemoglobina y colesterol, extraído de Billard y Diday (2006b).

Cada término del sumatorio mide la fracción del área del rectángulo observado $\boldsymbol{Z}(u)$ que se solapa con la celda $R_{g_1 g_2}$. Conviene tener en cuenta que $f_{g_1 g_2}$ es el número de observaciones dentro de $R_{g_1 g_2}$, pero que no tiene por qué ser necesariamente un entero (salvo en el caso de los datos clásicos). La frecuencia relativa es simplemente

$$p_{g_1 g_2} = \frac{f_{g_1 g_2}}{m}. \tag{2.38}$$

El histograma conjunto consiste en representar gráficamente en tres dimensiones el rectángulo $R_{g_1 g_2}$ con una altura proporcional a $p_{g_1 g_2}$. Dicha representación no difiere de la que se haría en el caso de los datos clásicos. Un ejemplo de histograma conjunto construido a partir de dos variables de intervalo puede verse en la figura 2.2.

Diagrama de dispersión. La representación gráfica de dos variables de intervalos puede hacerse mediante un diagrama de dispersión en dos dimensiones. Los individuos se representan mediante rectángulos cuyos lados son paralelos a los ejes. Otra forma de representar diagramas de dispersión de intervalos, consiste en representar cada individuos mediante una cruz cuya intersección es el centro del rectángulo. Esta representación enfatiza que un intervalo se puede expresar como centro y rango, ya que el centro de la cruz corresponde con el centro de los dos intervalos de los individuos y los brazos



(a) Rectángulos (b) Cruces

Figura 2.3: Gráficos de dispersión bivariantes para variables de intervalo mediante rectángulos y mediante cruces

de la cruz corresponden con el radio de los intervalos en cada una de las dos dimensiones. La figura 2.3 muestra un ejemplo de ambas representaciones. En ella se aprecia que, pese a que son más complejas que la nube de puntos de los diagramas de dispersión clásicos, ambas muestran la información con claridad.

La extensión de los diagramas de dispersión a las tres dimensiones se puede realizar de manera sencilla. En dicho diagrama, los individuos deben ser representados como hiperrectángulos con los lados paralelos a los ejes. La figura 2.4 muestra un ejemplo de dicho diagrama. El diagrama de dispersión de cruces también puede extenderse a las tres dimensiones sin gran dificultad.

2.4.1.1. Las medidas de dependencia

En la estadística clásica la medida más popular de dependencia entre dos variables aleatorias cuantitativas es la covarianza. La covarianza mide cuánto varían conjuntamente las dos variables. Si las dos variables tienden a variar juntas (i.e. cuando una se sitúa por encima de su valor esperado, la otra también), la covarianza es positiva; en caso contrario, la covarianza es negativa. Otras medidas de dependencia como la función de correlación producto-momento o el coeficiente de regresión emplean la covarianza para obtener sus valores. Existen otras medidas de dependencia como la rho de Spearman que no utilizan la covarianza para medir la dependencia. En esta sección se revisarán las medidas de dependencia propuestas para un par de variables de intervalo $Y_{j_1} \equiv Z_1$ y $Y_{j_2} \equiv Z_2$.

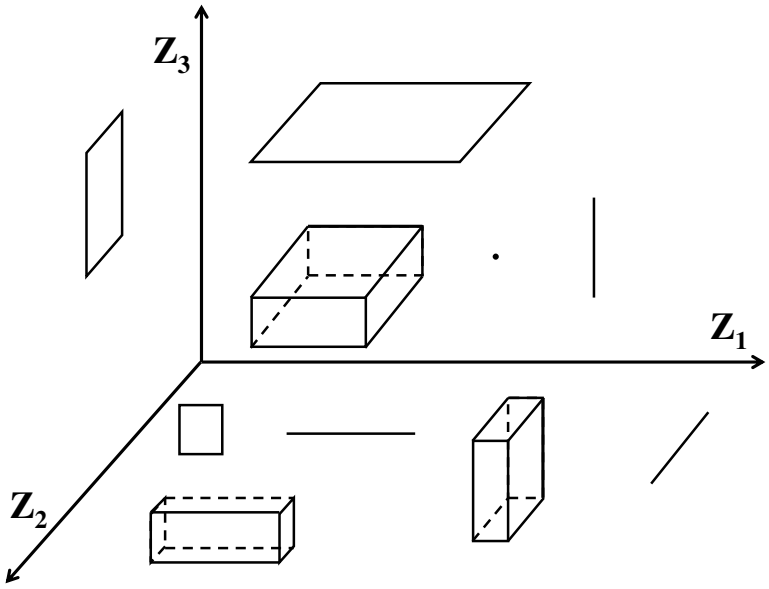


Figura 2.4: Gráfico de dispersión de tres variables de intervalo Z_1, Z_2 y Z_3 con los distintos tipos de hiperrectángulos que se pueden presentar.

Función de covarianza empírica. Según Billard y Diday (2006b), viene dada por la siguiente expresión:

$$Cov(Z_1, Z_2) = \frac{1}{3m} \sum_{u \in E} G_1 G_2 (Q_1 Q_2)^{1/2}, \tag{2.39}$$

donde, para $j = 1, 2$,

$$Q_j = (a_{uj} - \bar{Z}_j)^2 + (a_{uj} - \bar{Z}_j)(b_{uj} - \bar{Z}_j) + (b_{uj} - \bar{Z}_j)^2, \tag{2.40}$$

$$G_j = \begin{cases} -1, & \text{si } \bar{Z}_{uj} \leq \bar{Z}_j, \\ 1, & \text{si } \bar{Z}_{uj} > \bar{Z}_j, \end{cases} \tag{2.41}$$

y donde $\bar{Z}_j$ es la media definida en la ecuación (2.13) y donde $\bar{Z}_{uj}$ es el centro (y la media, asumiendo que un intervalo se comporta como una distribución uniforme) de la observación para el individuo u de la variable Z_j, i.e. $\bar{Z}_{uj} = (a_{uj} + b_{uj})/2$.

Cuando $Z_1 = Z_2$ esta medida pasa a ser la varianza de Z_1 presentada en la ecuación (2.16) . Además, si en esta medida se utilizan valores clásicos, i.e. intervalos tal que $a_{uj} = b_{uj}, \forall u, j$, el valor de la covarianza resultante es el que se obtiene con la fórmula para datos clásicos.

Otra definición de covariaza. En Billard y Diday (2000) se plantea una definición de covarianza distinta a la mostrada en (2.39)

$$Cov(Z_1, Z_2) = \frac{1}{4m} \sum_{u \in E} (b_{1u} + a_{1u})(b_{2u} + a_{2u}) - \bar{Z}_1 \bar{Z}_2, \tag{2.42}$$

donde $\bar{Z}_j$ es la media de Z_j que se calcula mediante la ecuación (2.13). Esta definición de covarianza está en función del momento de orden 2 de los centros de los intervalos y no tiene en cuenta la longitud de los mismos. Esta definición se corresponde con la definición de varianza mostrada en la ecuación (2.17). Estas definiciones no tienen en cuenta la variabilidad interna existente en los intervalos considerados (i.e. la longitud de los mismos).

Coeficiente de correlación. Se define de igual manera que para los datos clásicos.

$$r(Z_1, Z_2) = \frac{Cov(Z_1, Z_2)}{S_{Z_1} S_{Z_2}}, \tag{2.43}$$

donde $Cov(Z_1, Z_2)$ es la covarianza de la ecuación (2.39) y S_{Z_j} con $j = 1, 2$ es la desviación típica obtenida mediante la ecuación (2.16).

Coeficiente rho de Spearman. Billard (2004) extiende el coeficiente rho de Spearman para modelar dependencias entre variables de intervalo. Para ello, emplea el concepto de cópulas, que fue introducido por Sklar (1959). Sea $H(y_1, y_2)$ una función de distribución conjunta (Y_1, Y_2) y sea $F_j(y_j)$ la función de distribución marginal de Y_j con $j = 1, 2$. Entonces, existe una función $C(y_1, y_2)$ llamada cópula que satisface

$$H(y_1, y_2) = C(F_1(y_1), F_2(y_2)). \tag{2.44}$$

Es importante darse cuenta de que las funciones $H(\cdot)$ y $F(\cdot)$ son funciones de distribución acumuladas en lugar de funciones de densidad. Las cópulas permiten incorporar la dependencia en la relación existente entre la función de distribución conjunta y la función de distribución marginal, ya que dos distribuciones marginales dadas pueden provenir de infinitas distribuciones conjuntas. Deheuvels (1979) desarrolla un método para calcular la función cópula empírica que es adaptado por Billard (2004) para datos de tipo intervalo.

Sean dos variables de intervalo Z_j con $j = 1, 2$ y sea $X_{uj} = (a_{uj}, b_{uj})/2$ el punto medio (o centro) del intervalo $[a_{uj}, b_{uj}]$ para la variable Z_j y para la observación $u = 1, ..., m$. Sea W_{jk} el valor (X_{uj}) que ocupa la posición k_j-ésima al considerar los valores de X_{uj} ordenados de menor a mayor, $k_j = 1, ..., m$ y $j = 1, 2$. Dados estos elementos, se definen los rectángulos

$$R(w_{1k_1} w_{2k_2}) = (q_1, w_{1k_1}) \times (q_2, w_{2k_2}), \tag{2.45}$$

para $k_j = 1, ..., m$, donde

$$q_1 \leq min_u\{a_{u1}\}, q_2 \leq min_u\{a_{u2}\}, \tag{2.46}$$

y para $k_j = m + 1$, el rectángulo es

$$R(w_{1,m+1}, w_{2,m+1}) = (q_1, t_1) \times (q_2, t_2) \tag{2.47}$$

donde

$$t_1 \geq max_u\{b_{u1}\}, t_2 \geq max_u\{b_{u2}\}). \tag{2.48}$$

Dados estos conceptos la cópula empírica para datos de intervalo viene dada por

$$C'(w_{1k_1}, w_{2k_2}) = \frac{1}{m} \sum_{u \in E} \frac{\|Z(u) \cap R(w_{1k_1}, w_{2k_2})\|}{\|Z(u)\|}, \tag{2.49}$$

para $k_1, k_2 = 1, ..., m + 1$, donde $Z(u)$ es el rectángulo $(a_{u1}, b_{u1}) \times (a_{u2}, b_{u2})$ asociado con la observación u y donde $\|A\|$ es el área del rectángulo A. Una vez determinada la cópula, se puede estimar la rho de Spearman, la cual se define como una función de la diferencia entre la probabilidad de concordancia y la probabilidad de discordancia. Nelsen (1999) muestra que la relación entre la rho de Spearman y las cópulas para variables clásicas es la siguiente

$$\rho = 12 \int \int C(y_1, y_2) dy_1 dy_2. \tag{2.50}$$

Si a esta fórmula se le aplica el concepto de cópula empírica para datos de intervalo definida en la ecuación (2.50), se obtiene que la rho de Spearman empírica para dos variables de intervalo Z_1 y Z_2 viene dada por

$$r = \frac{12}{m^2 - 1} \left\{ \left[\sum_{k_1=1}^{m+1} \sum_{k_2=1}^{m+1} C'(w_{1k_1}, w_{2k_2}) \right] \right. \tag{2.51}$$

$$\left. - \left[\sum_{k_1=1}^{m+1} F_1(w_{1k_1}) \right] \left[\sum_{k_2=1}^{m+1} F_2(w_{2k_2}) \right] \right\}, \tag{2.52}$$

donde $\{F_j(w_{jk_j}), k_j = 1, ..., m + 1\}$ es la función de distribución marginal de Y_j con $j = 1, 2$.

2.4.2. Variables de histograma

Para el caso de las variables de histograma se procede de forma análoga a como se procedió con las variables de intervalo. Se consideran dos variables aleatorias simbólicas de histograma $\boldsymbol{Y} = \{Y_{j_1}, Y_{j_2}\}$ valoradas para cada individuo $w_u \in E$ con $u = 1, ..., m$. Cada una de estas variables Y_j tomará

valores en $\mathcal{Y}_j \subseteq \Re$. Por coherencia con la notación univariante, se renombrarán las variables de la siguiente manera $Y_{j_1} \equiv Z_1$ y $Y_{j_2} \equiv Z_2$. El valor de estas variables para cada individuo $w_u \in E$ es un histograma

$$Z_j(w_u) = \{[a_{ujk}, b_{ujk}), p_{ujk}, k = 1, ..., s_{uj}\}, \tag{2.53}$$

donde los intervalos $[a_{ujk}, b_{ujk})$ tienen una frecuencia relativa asociada p_{ujk} tal que $\sum_{k=1}^{s_{uj}} p_{ujk} = 1$, con $u = 1, ..., m$, $j = 1, 2$ y $k = 1, ..., s_{uj}$. Puede verse que la variable de intervalo es un caso particular de la variable de histograma donde $s_{uj} = 1$ y $p_{ujk} = 1$, $\forall u, j, k$. Por tanto, los estadísticos de histograma que se van a mostrar subsumen a los estadísticos de intervalo ya mostrados.

Función de distribución conjunta empírica. La función de distribución conjunta empírica, $F_Z(\xi_1, \xi_2)$, se representa de la siguiente forma

$$\begin{aligned} F_Z(\xi_1, \xi_2) &= \frac{1}{m} \sum_{u \in E} P\{x_1 \leq \xi_1, x_2 \leq \xi_2 | (x_1, x_2) \in vir(d_u)\} \\ &= \frac{1}{m} \left(\sum_{\xi_1 \geq b_{u1k}, \xi_2 \geq b_{u2k}} p_{u1k} p_{u2k} \right. \\ &+ \left. \sum_{(\xi_1, \xi_2) \in \boldsymbol{Z}(u)} p_{u1k} p_{u2k} \left(\frac{\xi_1 - a_{u1}}{b_{u1} - a_{u1}} \right) \left(\frac{\xi_2 - a_{u2}}{b_{u2} - a_{u2}} \right) \right) \end{aligned} \tag{2.54}$$

Función de densidad conjunta empírica. La función de densidad conjunta empírica para las variables (Z_1, Z_2) en el valor (ξ_1, ξ_2):

$$f(\xi_1, \xi_2) = \frac{1}{m} \sum_{u \in E} \left(\sum_{k_1=1}^{s_{u1}} \sum_{k_2=1}^{s_{u2}} \frac{I_{k_1 k_2}(\xi_1, \xi_2)}{\| Z_{k_1 k_2}(u) \|} p_{u1k_1} p_{u2k_2} \right), \tag{2.55}$$

donde $I_{k_1 k_2}(\xi_1, \xi_2)$ es la función booleana que indica si el valor (ξ_1, ξ_2) se encuentra o no dentro del rectángulo $Z_{k_1 k_2}(u)$ y donde $\| Z_{k_1 k_2}(u) \|$ es el área de dicho rectángulo.

Histograma conjunto. De forma análoga al caso univariante (ver sección 2.3.2), se toman los intervalos I_1 e I_2 que abarcan todos los valores observados en Z_1 y Z_2, y se dividen en r_1 y r_2 intervalos, $I_{1,g_1} = [\xi_{1g_1-1}, \xi_{1g_1})$ e $I_{2,g_2} = [\xi_{2g_2-1}, \xi_{2g_2})$, respectivamente con $g_1 = 1, ..., r_1$ y $g_2 = 1, .., r_2$. Con ello se divide el espacio $\Re^2$ en una rejilla donde cada celda es un rectángulo $R_{g_1 g_2} = [\xi_{1g_1-1}, \xi_{1g_1}) \times [\xi_{2g_2-1}, \xi_{2g_2})$. La frecuencia absoluta asociada al rectángulo $R_{g_1 g_2}$ se calcula de la siguiente manera

$$f_{g_1 g_2} = \sum_{u \in E} \sum_{k_1 \in Z(g_1)} \sum_{k_2 \in Z(g_2)} \frac{\| \boldsymbol{Z}(k_1, k_2; u) \cap R_{g_1 g_2} \|}{\| \boldsymbol{Z}(k_1, k_2; u) \|} p_{u1k_1} p_{u2k_2}, \tag{2.56}$$

donde $Z(g_j)$ con $j = 1, 2$, representa el conjunto de todos los intervalos $Z(k_j; u) \equiv [a_{ujk_j}, b_{ujk_j})$ que se solapan con la celda $R_{g_1g_2}$ para cada individuo u. Cada término del sumatorio mide la fracción del área del subrectángulo observado $\boldsymbol{Z}(k_1, k_2; u)$ que se solapa con la celda $R_{g_1g_2}$. La frecuencia relativa es sencillamente

$$p_{g_1g_2} = \frac{f_{g_1g_2}}{m}. \tag{2.57}$$

El histograma conjunto consiste en representar gráficamente en tres dimensiones el paralelepípedo que tiene como base el rectángulo $R_{g_1g_2}$ y una altura proporcional a su frecuencia $p_{g_1g_2}$.

Diagrama de dispersión. La representación gráfica de individuos descritos mediante dos variables de histograma da lugar a representaciones prácticamente ininteligibles. Uno de los problemas es que se requieren tres dimensiones para hacerlos. Además, existe el problema del posible solapamiento entre histogramas. Todo ello hace muy complicado extraer información de estos diagramas mediante la mera inspección visual.

2.4.2.1. Las medidas de dependencia

Sean dos variables de histograma $Y_{j_1} \equiv Z_1$ y $Y_{j_2} \equiv Z_2$, las medidas de dependencia propuestas consisten en la función de covarianza y en el coeficiente de correlación.

Función de covarianza empírica. Billard y Diday (2006b) proponen la siguiente función de covarianza para variables de histograma

$$cov(Z_1, Z_2) = \frac{1}{3m} \sum_{u \in E} \sum_{k_1=1}^{s_{u1}} \sum_{k_2=1}^{s_{u2}} p_{1uk_1} p_{2uk_2} G_1 G_2 [Q_1 Q_2]^{1/2}, \tag{2.58}$$

donde para $j = 1, 2$,

$$Q_j = (a_{ujk_j} - \bar{Z}_j)^2 + (a_{ujk_j} - \bar{Z}_j)(b_{ujk_j} - \bar{Z}_j) + (b_{ujk_j} - \bar{Z}_j)^2, \tag{2.59}$$

$$G_j = \begin{cases} -1, & \text{si } \bar{Z}_{uj} \leq \bar{Z}_j, \\ 1, & \text{si } \bar{Z}_{uj} > \bar{Z}_j, \end{cases} \tag{2.60}$$

y donde $\bar{Z}_j$ es la media de una variable de histograma Z_j tal y como se define en la ecuación (2.32) y donde $\bar{Z}_{uj}$ es la media del valor de la variable de histograma Z_j observado para el individuo u asumiendo que dentro de cada intervalo los puntos se distribuyen según una uniforme, es decir,

$$\bar{Z}_{uj} = \frac{1}{2} \sum_{k_j=1}^{s_{uj}} p_{ujk_j} (a_{ujk_j} + b_{ujk_j}). \tag{2.61}$$

Otra definición de covarianza. La definición de covarianza mostrada en la fórmula (2.58) es la más reciente, pero anteriormente (Billard y Diday, 2002, 2003) los mismos autores habían propuesto otra definición de covarianza para las variables de histograma

$$\begin{aligned} cov(Z_1, Z_2) &= \frac{1}{4m} \sum_{u \in E} \left\{ \sum_{k_1=1}^{s_{u1}} \sum_{k_2=1}^{s_{u2}} p_{1uk_1} p_{2uk_2} (a_{1uk_1} + b_{1uk_1})(a_{2uk_2} + b_{2uk_2}) \right\} \\ &- \bar{Z}_1 \bar{Z}_2, \end{aligned} \tag{2.62}$$

donde $\bar{Z}_j$ es la media de Z_j que se calcula mediante la ecuación (2.32).

Como sucede en el caso de los intervalos, esta definición de covarianza no tiene en cuenta la variabilidad interna de cada histograma.

2.5. Modelos de regresión lineal

Los modelos de regresión lineal planteados para datos simbólicos se basan, todos ellos, en el modelo de datos clásico. Por ello, conviene hacer una breve introducción a la teoría básica del análisis de regresión lineal.

Dado un conjunto de variables independientes $X_1, ..., X_p$ y una variable dependiente Y, el modelo de regresión lineal múltiple clásico se define como

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \varepsilon \tag{2.63}$$

o, vectorialmente, como

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2.64}$$

donde el vector de observaciones es $\boldsymbol{Y} = (Y_1, ..., Y_n)'$, el vector de coeficientes es $\boldsymbol{\beta} = (\beta_0, ..., \beta_p)'$, el vector error es $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_n)'$ y la matriz de regresión $\boldsymbol{X}$ es una matriz de $n \times (p+1)$ tal que

$$\begin{pmatrix} 1 & X_{11} & \dots & X_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & X_{n1} & \dots & X_{np} \end{pmatrix}. \tag{2.65}$$

Además, los términos de error ε_i con $i = 1, ..., n$ satisfacen $E(\varepsilon_i) = 0$, $Var(\varepsilon_i) = 0$ y $Cov(\varepsilon_i, \varepsilon_{i'}) = cte$ para todo $i \neq i'$. El estimador de mínimos cuadrados de los parámetros $\boldsymbol{\beta}$ viene dado, si $\boldsymbol{X}$ no es singular, por

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}. \tag{2.66}$$

Cuando $p = 1$, la ecuación (2.66) se simplifica a

$$\hat{\beta}_1 = \frac{Cov(X, Y)}{Var(X)} \text{ y } \beta_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \tag{2.67}$$

donde $\bar{Y}$ y $\bar{X}$ son la media de las variables Y y X, respectivamente; $Var(X)$ es la varianza de la variable X y $Cov(X, Y)$ es la covarianza entre las variables X e Y.

2.5.1. Variables de intervalo

Hasta el momento se han planteado numerosos enfoques para abordar la regresión lineal de datos de intervalo. En esta sección se resumirán todos ellos.

El enfoque simbólico o de los centros. Billard y Diday (2000) plantean el primer modelo de regresión para variables simbólicas de intervalo. Este modelo se basa en el modelo de regresión para variables clásicas. Sea una variable dependiente $Y(u)$ y una variable dependiente $X(u)$ que toma valores para cada individuo $w_u \in E$ con $u = 1, ..., m$, el análisis de regresión para variables de intervalo se desarrolla tal y como se ha mostrado en las ecuaciones (2.63-2.66) para la metodología clásica, pero considerando que las variables consideradas son de intervalo. Los parámetros de dicho modelo se obtienen de la siguiente forma:

$$\hat{\beta}_1 = \frac{Cov(X,Y)}{Var(X)} \text{ y } \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}, \tag{2.68}$$

donde $Cov(X,Y)$ y $Var(X)$ vienen dados por las fórmulas (2.42) y (2.17), respecivamente, y $\bar{X}$ es la media que viene dada por la fórmula (2.13). La ecuación de regresión resultante trata con valores clásicos es decir, la entrada y la salida, son valores clásicos. Para generar intervalos, es necesario *alimentar* el modelo con los valores mínimo y máximo de la variable dependiente.

Las definiciones de covarianza y varianza sobre las que se sustenta este modelo sólo tienen en cuenta los centros de los intervalos y no la longitud de los mismos. Por tanto, este modelo es totalmente equivalente a realizar una regresión lineal entre los centros de los intervalos de las variables dependiente e independiente.

Billard y Diday (2006b) plantean una ligera variante de este modelo donde simplemente cambian las fórmulas de $Cov(X,Y)$ y de $Var(X)$ para recoger la variabilidad interna de los intervalos, es decir, utilizan las fórmulas (2.39) y (2.16). Esta aproximación no sería equivalente a realizar una regresión de los centros. Sin embargo, la esencia del modelo no cambia.

Una alternativa similar es propuesta por Brito (2007) que propone un modelo de regresión basado en un índice de dispersión y un índice de co-dispersión que desempeñarían el mismo papel que el que juegan la varianza y a la covarianza en el modelo de regresión de Billard y Diday. El índice de dispersión para una variable de intervalo $X(u) = [a_u, b_u]$ que toma valores en $u = 1, ..., m$ individuos es definido como

$$\tilde{S}_X^2 = \frac{1}{m} \sum_{u \in E} \frac{(a_u - \bar{X})^2 + (b_u - \bar{X})^2}{m}, \tag{2.69}$$

donde $\bar{X}$ es la media de la variable X tal y como se define en la fórmula (2.13). De forma análoga, el índice de co-dispersión de dos variables de intervalo

$Y(u) = X_1(u) = [a_{u1}, b_{u1}]$ y $X(u) = X_2(u) = [a_{u2}, b_{u2}]$ que toman valores en $u = 1, ..., m$ individuos se define como

$$\tilde{S}_{X_1X_2} = \frac{1}{m} \sum_{u \in E} \frac{(a_{u1} - \bar{X}_1)(a_{u2} - \bar{X}_2) + (b_{u1} - \bar{X}_1)(b_{u2} - \bar{X}_2)}{2}. \quad (2.70)$$

Dados estos índices, los coeficientes del modelo de regresión se obtienen de forma análoga a la mostrada en la fórmula (2.68). Además, dichos coeficientes minimizan la siguiente expresión $\sum_{u \in E}[(a_{u2} - \hat{\beta}_0 - \hat{\beta}_1)^2 + (b_{u2} - \hat{\beta}_0 - \hat{\beta}_1)^2]$ siendo X_1 la variable independiente y X_2 la variable dependiente del modelo de regresión. De acuerdo con Brito (2007), los resultados obtenidos son similares a los obtenidos con el enfoque de Billard y Diday (2003).

El enfoque mínimo-máximo. Billard y Diday (2000) y Billard y Diday (2002) también esbozan una manera obvia de realizar la regresión lineal entre datos de intervalo. Ésta consiste en ajustar un modelo de regresión clásico para los mínimos y otro para los máximos, teniendo en cuenta que el intervalo resultante no será válido cuando el mínimo sea mayor que el máximo. Si renombramos la variable dependiente e independiente de la siguiente manera, $Y(u) = X_1(u) = [a_{u1}, b_{u1}]$ y $X_2(u) = [a_{u2}, b_{u2}]$, dicho enfoque, al que llamaremos mínimo-máximo, podría formalizarse mediante las siguientes ecuaciones

$$a_{u2} = \beta_0^a + \beta_1^a a_{u1} + \varepsilon_u^a, \quad (2.71)$$

$$b_{u2} = \beta_0^b + \beta_1^b b_{u1} + \varepsilon_u^b. \quad (2.72)$$

El criterio a minimizar en este caso sería el siguiente

$$\sum_{u=1}^{m} ((\varepsilon_u^a)^2 + (\varepsilon_u^b)^2), \quad (2.73)$$

donde $\varepsilon_u^a = a_{u2} - \hat{a}_{u2}$ y $\varepsilon_u^b = b_{u2} - \hat{b}_{u2}$, siendo $\hat{a}_{u2}$ y $\hat{b}_{u2}$ los valores estimados por el modelo para a_{u2} y b_{u2}, respectivamente.

El enfoque min-max asume independencia entre las observaciones de los extremos inferior y superior de los intervalos, lo cual normalmente no será cierto, ya que $a_{uj} \geq b_{uj}$ por definición. Este enfoque es conceptualmente más pobre que el enfoque simbólico presentado en Billard y Diday (2000) porque simplemente supone una *recodificación* del intervalo para hacer que el modelo de regresión clásico sea capaz de tratarlo. Sin embargo, intuitivamente es de esperar que esta alternativa produzca modelos mejor ajustados.

El enfoque centro-rango. Otra forma de modelar la relación lineal entre variables de intervalo, consiste en trabajar de manera independiente los centros y los rangos, siendo el rango del intervalo su longitud. Este enfoque ha

sido desarrollado progresivamente en los trabajos de Lima Neto, de Carvalho y Tenorio (2004), de Carvalho, Lima Neto y Tenorio (2004), Lima Neto, de Carvalho y Freire (2005) y finalmente en en el artículo de Lima Neto y de Carvalho (2008). Las ideas básicas del enfoque serán mostradas a continuación.

Sea una variable dependiente de intervalo $Y(u) = X_1(u) = [a_{u1}, b_{u1}]$, y una variable dependiente $X_2(u) = [a_{u2}, b_{u2}]$ que toman valores para cada individuo $w_u \in E$ con $u = 1, ..., m$, el centro y el rango de los intervalos del individuo w_u se define como

$$c_{uj} = (b_{uj} + a_{uj})/2 \text{ y } l_{uj} = b_{uj} - a_{uj}, \tag{2.74}$$

respectamente, con $j = 1, 2$. Dicho esto, el enfoque centro-rango de regresión de intervalos consiste en las siguientes dos ecuaciones de regresión clásica

$$c_{u2} = \beta_0^c + \beta_1^c c_{u1} + \varepsilon_u^c, \tag{2.75}$$

$$l_{u2} = \beta_0^l + \beta_1^l l_{u1} + \varepsilon_u^l. \tag{2.76}$$

En este enfoque, los coeficientes del modelo se estiman mediante el método de los mínimos cuadrados. El criterio a minimizar es el siguiente

$$\sum_{u=1}^{m}((\varepsilon_u^c)^2 + (\varepsilon_u^l)^2), \tag{2.77}$$

donde $\varepsilon_u^c = c_{u2} - \hat{c}_{u2}$ y $\varepsilon_u^l = l_{u2} - \hat{l}_{u2}$, siendo $\hat{c}_{u2}$ y $\hat{l}_{u2}$ los valores estimados por el modelo para c_{u2} y l_{u2}, respectivamente.

En este modelo no se considera ninguna relación de dependencia entre los centros y los rangos de los intervalos. Los autores plantean el enfoque exclusivamente como un problema de optimización y sin considerar los supuestos probabilísticos que implica el modelo de regresión clásico para el caso de los datos de intervalo.

En Lima Neto y de Carvalho (2008) se realiza una comparación entre el enfoque centro-rango (2.75) planteado en dicho artículo, el enfoque mínimo-máximo (2.71) y el enfoque de los centros (2.68). La comparación la realizan por medio de una simulación de Monte Carlo sobre conjuntos de datos sintéticos que son generados de acuerdo a distintas configuraciones. En dichas configuraciones, los centros de las variables dependiente e independiente están linealmente relacionados mediante una relación lineal que varía según la configuración y que se ve afectada por un término de error cuya variabilidad será alta o baja, dependiendo también de la configuración. En unas configuraciones, los rangos son independientes y se generan mediante distribuciones uniformes cuya variabilidad depende de la configuración. Mientas que en otras, el rango de la variable dependiente depende linealmente del centro de la variable independiente, siendo afectada esta relación por otro término de

error cuya variabilidad puede ser alta o baja, según la configuración. Además, existen configuraciones con una y con tres variables independientes.

Como medidas de calidad del ajuste de la regresión utilizan la raíz cuadrada del error cuadrático medio obtenido (RECM) de los mínimos y los máximos del intervalo, es decir

$$RECM_{min} = \sqrt{\frac{\sum_{u=1}^{m}(a_{u2} - \hat{a}_{u2})^2}{m}} \text{ y } RECM_{max} = \sqrt{\frac{\sum_{u=1}^{m}(b_{u2} - \hat{b}_{u2})^2}{m}} \tag{2.78}$$

y el coeficiente de correlación de los mínimos y de los máximos

$$r^2_{min} = \left(\frac{Cov(\boldsymbol{a}_2, \hat{\boldsymbol{a}}_2)}{S(\boldsymbol{a}_2)S(\hat{\boldsymbol{a}}_2)}\right)^2 \text{ y } r^2_{max} = \left(\frac{Cov(\boldsymbol{b}_2, \hat{\boldsymbol{b}}_2)}{S(\boldsymbol{b}_2)S(\hat{\boldsymbol{b}}_2)}\right)^2, \tag{2.79}$$

donde $\boldsymbol{a}_2 = (a_{12}, ..., a_{m2})'$, $\hat{\boldsymbol{a}}_2 = (\hat{a}_{12}, ..., \hat{a}_{m2})'$, $\boldsymbol{b}_2 = (b_{12}, ..., b_{m2})'$, $\hat{\boldsymbol{b}}_2 = (\hat{b}_{12}, ..., \hat{b}_{m2})'$ y $S(X)$ es la desviación estándar de la variable X y $Cov(X, Y)$ es la covarianza entre las variables X e Y. El resultado de estas medidas en las diferentes simulaciones realizadas para cada configuración es promediado. Para comparar los distintos enfoques dos a dos utilizan un test estadístico t para muestras pareadas a un nivel de significación del 1 % sobre cada una de las medidas consideradas.

Las conclusiones que se obtienen mediante la simulación del método de Monte Carlo son que, de acuerdo con las medidas de calidad establecidas, el enfoque centro-rango supera al enfoque simbólico y al enfoque mínimo-maximo. Para el caso en el que los centros y los rangos son independientes, la mejora se hace más patente cuando hay tres variables dependientes y cuando la relación lineal es más clara (i.e. la variabilidad del error es menor). Cuando existe relación lineal entre centro y rango, el enfoque centro-rango supera claramente al enfoque simbólico, pero el rendimiento en comparación con el del enfoque mínimo-máximo está más igualado.

Cuando se compara el enfoque mínimo-máximo con el enfoque simbólico, tanto para los casos en los que existe relación lineal entre el centro y el rango, como en los casos en los que no, los resultados no son tan concluyentes ya que en términos de $RECM_{min}$ y $RECM_{max}$ el enfoque min-max es claramente mejor que el simbólico, mientras que en términos de r^2_{min} y r^2_{max} los resultados no son concluyentes.

En Lima Neto et al. (2005) se compara el enfoque simbólico (o de los centros) con el enfoque centro-rango, pero añaden a los modelos de regresión restricciones para que el modelo garantice que el valor pronosticado para el extremo inferior de un intervalo no sea mayor que el pronosticado para el extremo superior, i.e. para que $\hat{a}_2 \leq \hat{b}_2$. El experimento para determinar el rendimiento de los enfoques es similar al explicado anteriormente. Las conclusiones que se extraen del mismo son que el uso de restricciones en el

enfoque centro-rango no empeora el rendimiento si lo comparamos con el enfoque centro-rango sin restricciones. Sin embargo, el añadir restricciones al enfoque simbólico, perjudica el rendimiento del enfoque.

2.5.2. Variables de histograma

El modelo de regresión para variables de histograma es propuesto en Billard y Diday (2002) y sigue un planteamiento muy similar al mostrado en Billard y Diday (2000) para variables de intervalo. Es decir, sigue la metodología clásica mostrada en las ecuaciones (2.63-2.66), pero considerando que las variables consideradas son de histograma. Los parámetros de la recta de regresión se obtienen según:

$$\hat{\beta}_1 = \frac{Cov(X,Y)}{Var(X)} \text{ y } \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}, \tag{2.80}$$

donde $Cov(X,Y)$ y $Var(X)$ son la covarianza (2.62) y la varianza (2.34) para variables de histograma, y $\bar{X}$ es la media que viene dada por la fórmula (2.32). La ecuación de regresión resultante trata con valores clásicos es decir, la entrada y la salida, son valores clásicos. Los autores *alimentan* el modelo con intervalos, es decir, con valores mínimo y máximo en la variable dependiente, con lo que obtienen un intervalo en la variable de salida. Sin embargo, no abordan el problema de utilizar como entrada un histograma, es decir, un conjunto de intervalos cada uno de ellos con un peso asociado.

Billard y Diday (2006b) desarrollan una ligera variación del modelo donde las fórmulas de la varianza y de la covarianza son (2.33) y (2.58), respectivamente. Sin embargo, tampoco ofrece soluciones sobre cómo emplear el modelo utilizando histogramas. Sin duda, la naturaleza del dato de histograma hace que sea notablemente más complejo plantear un modelo de regresión para ellos que para intervalos.

2.6. Otras técnicas de análisis de datos simbólicos.

Además de los modelos de regresión, se han desarrollado otras técnicas para el análisis de datos simbólicos de intervalo y de histograma. Tal y como se ha comentado, hasta el momento el desarrollo de técnicas para datos de intervalo está más desarrollada que las de histogramas. Esto resulta natural debido a la mayor complejidad que implica el histograma.

Este apartado no pretende ser exhaustivo citando todos los métodos propuestos, sino ofrecer una panorámica general de las áreas donde más se ha investigado. Bock y Diday (2000) y Billard y Diday (2006b) recogen algunos de las métodos desarrolladas pero existen muchos más diseminados en revistas y congresos.

2.6.1. Variables de intervalo

El área en la que más aportaciones sobre datos simbólicos se han realizado es el análisis cluster (o análisis de conglomerados). El análisis cluster tiene como objetivo la formación de grupos, de forma que los elementos de un grupo se parezcan entre sí y que no se parezcan a los elementos de otros grupos de acuerdo con un criterio de disimilaridad. En Bock y Diday (2000) se muestran técnicas de clustering divisivo y piramidal. Sin embargo, la metodología sobre la que existen más publicaciones es el clustering dinámico para datos de intervalo. Entre los trabajos más recientes se encuentran el de de Carvalho, de Souza, Chavent y Lechevallier (2006a) o el de de Carvalho, Brito y Bock (2006b) donde se abordan distintas formas de estandarizar las variables de intervalo. Otras técnicas de clustering adaptadas a los datos de intervalo incluyen el clustering borroso de k-medias (D'Urso y Giordani, 2006) o un enfoque basado en *rough sets* (Asharaf, Narasimha Murty y Shevade, 2006).

Otra técnica sobre la que se ha trabajado abundantemente es el análisis de componentes principales (ACP). El ACP es una técnica de análisis multivariante que se emplea para reducir la dimensionalidad (i.e. el número de variables) de un conjunto de datos, pero manteniendo la mayor parte de la estructura de varianza-covarianza de las variables originales. Las componentes principales se obtienen mediante una combinación lineal de las variables originales. En Chouakria, Cazes y Diday (2000) se muestran dos métodos para realizar el ACP en conjuntos de datos descritos por variables de intervalo: el método de los vértices y el método de los centros. El método de los vértices representa los intervalos mediante sus extremos superiores e inferiores y supone una carga computacional excesiva si el número de variables consideradas es muy grande. En ese caso, los autores recomiendan emplear el método de los centros con el que se obtienen resultados similares. Otros enfoques interesantes son los que se centran en la representación centro-radio del intervalo como los propuestos por Palumbo y Lauro (2003) y D'Urso y Giordani (2004), o el que emplea la aritmética de intervalos (ver Secc. 2.7.1.1) propuesto por Gioia y Lauro (2006). Giordani y Kiers (2004) desarrollan una extensión del análisis de componentes principales para datos de intervalos en tres vías (*three-way data*), es decir, una serie de variables medidas para un conjunto de individuos, en unas determinadas ocasiones que pueden caracterizarse por el instante o por las condiciones de medidas.

Otra técnica que ha sido adaptada para manejar datos de intervalos ha sido el escalamiento multidimensional. Denoeux y Masson (2000) y Groenen et al. (2006) han propuesto métodos para ello. El escalamiento multidimensional requiere una matriz donde se encuentren cuantificadas las diferencias entre pares de objetos. Los métodos propuestos permiten que dichas diferencias se representen mediante intervalos. Esto puede suceder en el caso de que los jueces, en lugar de dar un valor preciso para cuantificar la diferencia

entre los pares, prefieran dar un intervalo de valores. También puede darse el caso de que se disponga de un grupo de jueces y que se quiera resumir las diferencias dadas para cada par de objetos mediante un intervalo (el rango o el rango intercuartílico). También puede suceder que los datos originales se encuentren en forma de intervalo y que, por tanto, la diferencia entre los objetos pueda representarse mediante un intervalo.

Otro gran grupo de técnicas dentro del análisis multivariante y de la minería de datos es el de las técnicas de clasificación. Entre los métodos propuestos para datos de intervalo caben destacar distintas familias. Por un lado los árboles de decisión, como los basados en el criterio de Gini (Perinel, 1999) o los basados en el criterio de Kolmogorov-Smirnov (Mballo y Diday, 2006) o el propuesto por Limam, Diday y Winsberg (2003) que combina un criterio de homogeneidad y de discriminación para describir las clases de particiones hechas a priori. Por otro lado, también merece la pena destacar las adaptaciones del análisis discriminante para intervalos tanto en su variante lineal (Ishibuchi, Tanaka y Fukuoka, 1990; Nivlet, Fournier y Royer, 2001; Duarte Silva y Brito, 2006) como en su variante no lineal mediante maquinas de soporte vectorial (Angulo, Anguita, González-Abril y Ortega, 2008).

En el campo de las redes neuronales artificiales, Muñoz San Roque et al. (2007) proponen una adaptación del perceptrón multicapa que maneja datos en forma de intervalo tanto en las entradas como en la salida, pero donde los pesos son números escalares y no intervalos. Simonoff (1996) advierte que, si se consideran pesos con forma de intervalo, la amplitud del intervalo final aumenta considerablemente. Patiño-Escarcina, Callejas Bedregal y Lyra (2004) plantean un perceptrón de una capa cuyas entradas son intervalos y cuya salida es binaria. La aplicación de dicho modelo, como es natural, es clasificación binaria. Un modelo similar a éste es propuesto en Ishibuchi y Tanaka (1991). Puede encontrarse más información sobre perceptrones que manejan datos de intervalo en Muñoz San Roque et al. (2007). Más recientemente, Rossi y Conan-Guez (2008) han adaptado los perceptrones multicapa para trabajar con datos simbólicos. Lo que estos autores proponen es, en esencia, una recodificación de los datos simbólicos para que puedan ser tratados por la arquitectura clásica del perceptrón multicapa. Para el caso concreto de los datos de intervalo proponen emplear el extremo inferior y superior de los intervalos o el centro y la longitud de los mismos. En el ejemplo que desarrollan, prueban el enfoque simbólico utilizando datos de intervalo codificados mediante sus extremos superiores e inferiores da buenos los resultados, sin embargo la codificación mediante la media y la desviación típica de los datos originales obtiene mejor rendimiento en todos los casos trabajados. Sorprende que en dicho trabajo no se aborde el problema con la codificación mediante el centro y la longitud, la cual es bien conocida en el contexto simbólico.

Otro tipo de red neuronal que ha sido adaptado para trabajar con datos

de intervalo son los mapas autoorganizados (o mapas de Kohonen). El Golli, Conan-Guez y Rossi (2004) afirman que los mapas autoorganizados están basados en el concepto de centro de gravedad y que dicho concepto no es aplicable a muchos tipos de datos complejos, por ello proponen adaptarlos mediante el concepto de distancia. Con ello, sólo es necesario definir una distancia entre el tipo de datos que se estén manejando, que en dicho artículo son intervalos, para poder obtener mapas autoorganizados. Más recientemente, Bock (2008) propone un enfoque distinto para realizar mapas autoorganizados de intervalos en el que se realiza una adaptación del método más ortodoxa.

Chuang (2008) plantea un modelo de redes de regresión para intervalos basadas en vectores soporte. En dicho modelo tanto las entradas como las salidas son intervalos. Un modelo similar es desarrollado por Zhao, Liu y He (2006). Por su parte, Zhao, He y Chen (2005) proponen el uso de de una maquinas de vectores soporte entrenada mediante datos clásicos para clasificar datos de intervalo, para ello emplean aritmética de intervalos (ver Secc. 2.7.1.1). Angulo et al. (2008) van un paso más allá y proponen una adaptación de la máquina de soporte vectorial para trabajar con datos de intervalo, incluyendo la fase de entrenamiento; y utilizando para ello la aritmética de intervalos. Do y Poulet (2003) proponen la adaptación de la función kernel de base radial para tratar intervalos y proponen una adaptación de las máquinas de vectores soporte cuya única diferencia con el método clásico es el uso de dicha función kernel.

2.6.2. Variables de histograma

Tal y como se ha mencionado, el desarrollo de técnicas de análisis para datos de histograma es notablemente menor. Una de las posibles razones es que la complejidad del histograma respecto al intervalo es notablemente mayor. De hecho, algunas de las técnicas propuestas para datos de histograma requieren una acotación de la definición de histograma dada (ver Secc. 2.2) de forma que el conjunto de intervalos en que se divide el soporte de la variable de histograma, sea el mismo para todos los individuos y que, de esta forma, se pueda prescindir de la información asociada a los intervalos y manejar sólo la información de los pesos.

Entre los métodos de clustering para datos de histograma, cabe destacar el trabajo de Irpino y Verde (2006b) y de Irpino y Verde (2006a). El primero de ellos plantea una metodología para realizar clustering jerárquico, y el segundo para realizar clustering dinámico. Ambos toman como base la distancia de Mallows (Mallows, 1972) para funciones de densidad. Meneses y Rodríguez-Rojas (2006) adaptan el algoritmo de k-medias para trabajar con datos de histograma y de Souza, de Carvalho y Pizzato (2006) proponen un algoritmo de clustering dinámico para distintos tipos de variables entre las que se encuentran los histogramas. Sin embargo, ambos trabajos consi-

deran que en los histogramas los pesos se distribuyen sobre un conjunto de categorías.

Rodríguez et al. (2000) propone un método para realizar análisis de componentes principales para datos de histograma, que también es aplicable a datos de intervalo y a variables reales clásicas. De nuevo, en este método es necesario trabajar con histogramas cuya partición del soporte de la variable sea fija para todos los individuos. Esto es así, porque el método posteriormente ignora la partición y simplemente trabaja con la función de densidad acumulada del histograma.

Groenen y Winsberg (2006) definen un método de escalamiento multidimensional donde las diferencias entre los elementos considerados son expresadas mediante histogramas. Para que tal cosa suceda, puede darse el caso de que se tengan varios jueces y que se quieran agregar sus valoraciones por medio de un histograma, o que las diferencias entre los elementos requieran de cierto grado de borrosidad o de imprecisión que pueda expresarse mediante un histograma. La aproximación de este trabajo es una extensión del escalamiento multidimensional para intervalos propuesto por Groenen et al. (2006) para el caso en que las diferencias entre objetos se representan mediante histogramas.

Entre las técnicas de clasificación, Mballo y Diday (2004) extienden el criterio de Kolmogorov-Smirnov para inducir árboles de decisión en individuos descritos mediante datos de histograma y de intervalo. En esta técnica, los histogramas son una distribución de pesos asignada a un conjunto de categorías.

En el ámbito de las redes neuronales, la adaptación de los perceptrones multicapa propuesta por Rossi y Conan-Guez (2008) consiste en, como se ha dicho, recodificar los datos simbólicos para que puedan ser tratados por la arquitectura clásica del perceptrón multicapa. Los autores explican cómo recodificar variables modales y dicha recodificación puede utilizarse para datos de histogramas si en todos ellos la partición del rango de la variable es la misma.

El Golli et al. (2004) extienden los mapas autoorganizados para datos simbólicos mediante el uso de distancias, sorteando de esta forma la definición del concepto de centro de gravedad para datos simbólicos que sería necesario para realizar una adaptación más directa. Aunque en dicho artículo sólo se utiliza para intervalos, no parece difícil su extensión al terreno de los histogramas utilizando para ello alguna distancia adecuada.

En el área de tratamiento digital de imágenes es habitual el uso de histogramas. En esos casos, el histograma sirve para presentar de forma resumida los valores que toma en la imagen analizada una determinada característica, por ejemplo, el brillo. El rango de valores enteros es de 0 a 255, y el histograma representa el número de pixels de la imagen que toman cada uno de los valores del rango. Los histogramas en estos casos se emplean para tareas de

segmentación y de clasificación. Existe mucha literatura al respecto, como, por ejemplo, los trabajos de Puzicha, Hofmann y Buhmann (1999), Rubner, Tomasi y Guibas (2000) y Arifin y Asano (2006).

Otra área donde es habitual el manejo de histogramas es en el procesado de señales acústicas y musicales. Los histogramas permiten resumir características de la música tales como el tono o como el ritmo ignorando el orden temporal en el que se han producido y concentrándose en la frecuencia de aparición de la característica que se esté midiendo. Los datos en forma de histograma se utilizan también para la agrupación o clasificación de música o de cualquier otro tipo de audio. Algunos trabajos que emplean histogramas son Tzanetakis y Cook (2002) o Tzanetakis, Ermolinskyi y Cook (2003).

2.7. El cálculo con intervalos y con distribuciones de probabilidad

Típicamente se considera que existen dos tipos de incertidumbre:

- Incertidumbre estocástica (también conocida como irreducible o de primer orden): intrínseca al fenómeno que se está observando y que refleja la aleatoriedad del mismo.
- Incertidumbre epistémica (también conocida como reducible o de segundo orden): motivada por la incompletitud o la inexactitud de la información.

El límite entre ambas es, en algunos casos, borroso. Esto es debido a que un mayor conocimiento del fenómeno puede reducir lo que en un principio fue considerado como incertidumbre estocástica. Disquisiciones filosóficas aparte, los métodos estadísticos y de análisis de datos asumen tradicionalmente que los valores que manejan son precisos, i.e. ignoran la incertidumbre epistémica. En algunos casos, prescindir de esta incertidumbre puede no acarrear consecuencias. Sin embargo, si se requiere una gran exactitud en los cálculos que se están realizando, es absolutamente imprescindible tener en cuenta la incertidumbre de los datos recogidos de cara a operar con ellos y a tomar decisiones basadas en los resultados obtenidos. Los principales campos donde estos aspectos resultan cruciales son fiabilidad, análisis de riesgos y evaluación de la seguridad de los sistemas.

En los últimos tiempos, existe un interés creciente por manejar la incertidumbre epistémica. Existen distintos enfoques para reflejar dicha incertidumbre. Uno de ellos son los datos borrosos o difusos. Otro enfoque, que guarda una mayor relación con la tesis, es el del cálculo con intervalos y con distribuciones de probabilidad. Al igual que en el análisis de datos simbólicos, en estas disciplinas los datos se representan en forma de intervalos o de distribuciones. Sin embargo, en ellas, el intervalo o la distribución de

probabilidad es una descripción de la incertidumbre asociada al valor real. Esto lo diferencia de los datos simbólicos, donde el intervalo o el histograma pretenden reflejar la variabilidad observada del fenómeno.

A continuación, se revisarán algunos conceptos básicos sobre el cálculo con intervalos y con distribuciones de probabilidad. También se repasarán aquellas contribuciones que tienen más conexión con el contenido de esta tesis y con el análisis de datos simbólicos.

2.7.1. El cálculo con intervalos

El nacimiento de esta disciplina se debe a Ramon E. Moore (Moore, 1966) y a su estudio del truncamiento en los errores de redondeo en los cálculos realizados por los ordenadores. El objetivo que Moore perseguía se puede explicar mediante el siguiente ejemplo. Consideremos un ordenador que sólo puede almacenar números con dos dígitos decimales y sin dígitos enteros, e.g. $0.88, 0.23, etc.$ Supongamos que el valor a almacenar es $x = \frac{1}{3}$, dependiendo del tipo de redondeo que haga la máquina el valor almacenado será $x_M = 0.33$ o $x_M = 0.34$, donde x_M denota la representación que la máquina hace de x. Moore afirmaba que $x = \frac{1}{3}$ no podía ser almacenado de forma exacta usando una representación de coma flotante, sin embargo, el valor exacto podía ser acotado mediante un intervalo lo más pequeño que la representación de la máquina permita, en el ejemplo sería $x \in [0.33, 0.34]$. De esta forma, si queremos obtener el valor $f(x) = x^2$, Moore afirmaba que la aritmética de coma flotante producirá un valor u otro dependiendo del tipo de redondeo que se aplique. Sin embargo, sí es posible dar como solución un intervalo donde el valor real esté contenido con absoluta seguridad; en el ejemplo sería $x^2 \in [0.33^2, 0.34^2]$. En la década de los 60, este problema se veía agravado por el hecho de que cada procesador usaba un número de bytes para almacenar los números y sus propias reglas de redondeo. Moore desarrolló la aritmética de intervalos para que, con independencia de las características del procesador, los ordenadores pudiesen realizar operaciones asegurando que el resultado de dichas operaciones se encontrase con absoluta certeza dentro de un intervalo lo más acotado posible.

Con el estándar del IEEE para aritmética en coma flotante (IEEE 754) y el espectacular aumento de la capacidad de proceso de los ordenadores, la aritmética de intervalos dejó de emplearse para su propósito original y comenzó a ser utilizada por científicos que empleaban datos provenientes de sensores cuyas medidas no estaban exentas de cierta imprecisión. En la década de los 90, la disciplina resurgió, impulsada desde EEUU, como una herramienta que permitía resolver una gran cantidad de problemas en los que el denominador común era la imprecisión asociada con los valores que se manejan. Una de las principales áreas de aplicación son los sistemas electrónicos, sin embargo, tal y como muestra Kearfott y Kreinovich (1996), hay muchas más en campos como el control de calidad, la economía, la bioin-

formática, la geofísica, la medicina, la mecánica cuántica y la inteligencia artificial.

Antes de describir algunos de los conceptos sobre esta disciplina, es necesario hacer una consideración sobre la notación.

Consideración sobre la notación. En el análisis de intervalos no existe una notación estándar y, en consecuencia, cohabitan diferentes propuestas de notación, muchas de las cuales son incompatibles entre sí. Kearfott, Nakao, Neumaier, Rump, Shary y van Hentenryck (2005) realizan una propuesta de estandarización. Dicha propuesta tiene el fin de encajar perfectamente con la notación matemática tradicional y en especial con el análisis numérico y la optimización, de forma que que las fórmulas resultantes sean sencillas y fáciles de leer incluso para los legos en la materia. A continuación, se muestra un resumen de la propuesta

x	representa a un número real
$\boldsymbol{x} = [\underline{x}, \overline{x}]$	representa a un intervalo donde $\underline{x} \leq \overline{x}$
mid $\boldsymbol{x} = \frac{\underline{x}+\overline{x}}{2}$	es el punto medio del intervalo $\boldsymbol{x}$
rad $\boldsymbol{x} = \frac{\overline{x}-\underline{x}}{2}$	es el radio del intervalo $\boldsymbol{x}$
X	representa a una matriz o vector de números reales
$\boldsymbol{X}$	representa a una matriz o vector de intervalos

2.7.1.1. La aritmética de intervalos

La aritmética de intervalos propuesta por Moore (1966) permite realizar operaciones aritméticas con los intervalos. La premisa fundamental sobre la que se sustenta es la siguiente: dados dos intervalos $\boldsymbol{a}$ y $\boldsymbol{b}$, y un operador aritmético $\square$, entonces $\boldsymbol{a}\square\boldsymbol{b}$ es el intervalo más pequeño posible que contiene $a\square b$, $\forall a \in \boldsymbol{a}$ y $\forall b \in \boldsymbol{b}$. De acuerdo con esta premisa, la suma, resta, multiplicación y cociente de intervalos se definen de la siguiente manera

$$\boldsymbol{a} + \boldsymbol{b} = [\underline{a} + \underline{b}, \overline{a} + \overline{b}]$$

$$\boldsymbol{a} - \boldsymbol{b} = [\underline{a} - \overline{b}, \overline{a} - \underline{b}]$$

$$\begin{aligned} \boldsymbol{a} \cdot \boldsymbol{b} \quad = \quad & [\text{mín}\{\underline{a} \cdot \underline{b}, \underline{a} \cdot \overline{b}, \overline{a} \cdot \underline{b}, \overline{a} \cdot \overline{b}\}, \\ & \text{máx}\{\underline{a} \cdot \underline{b}, \underline{a} \cdot \overline{b}, \overline{a} \cdot \underline{b}, \overline{a} \cdot \overline{b}\}] \end{aligned}$$

$$\boldsymbol{a}/\boldsymbol{b} = \boldsymbol{a} \cdot (1/\boldsymbol{b}), \text{ con } 1/\boldsymbol{b} = [1/\overline{b}, 1/\underline{b}]$$

Es importante tener en cuenta que la aritmética de intervalos subsume a la aritmética clásica, es decir, si en una operación de la aritmética de intervalos tomamos como operandos números reales considerándolos como intervalos de longitud cero, y operamos con ellos, obtenemos el mismo resultado que si

hubiésemos hecho dicha operación utilizando la aritmética clásica con dichos números reales.

En la aritmética de intervalos, la suma y la multiplicación cumplen siempre la propiedad asociativa y conmutativa, pero no así la propiedad distributiva. En su lugar, la propiedad subdistribuiva rige en la aritmética de intervalos. Dicha propiedad se describe de la siguiente manera

$$\boldsymbol{a}(\boldsymbol{b}+\boldsymbol{c}) \subseteq \boldsymbol{ab}+\boldsymbol{ac}$$

Esta propiedad quiere decir que el intervalo resultante de la parte izquierda de la expresión será más conciso o igual de conciso que el intervalo resultante de la parte derecha, es decir, contendrá menor incertidumbre.

Las características de la aritmética de intervalos son las siguientes:

- Corrección: El intervalo resultante de evaluar una expresión aritmética es un intervalo que contiene todos los posibles resultados de las evaluaciones de dicha expresión utilizando la aritmética clásica y los valores puntuales contenidos dentro de los intervalos que son argumentos de la expresión.

- Totalidad: Cualquier conjunto de valores puntuales, pertenecientes cada uno de ellos a sus respectivos intervalos, da lugar a una solución válida.

- Optimalidad: El intervalo solución obtenido no es más ancho de lo necesario.

La aritmética de intervalos no asume ningún tipo de dependencia entre los operandos. Sin embargo, cabe preguntarse qué sucede si existe algún tipo de dependencia entre los operandos que intervienen en una operación aritmética de intervalos. Por dependencia se entiende el hecho de que exista una restricción sobre los posibles pares de valores de los intervalos. Por ejemplo, que valores altos de un intervalo operando sólo tengan sentido con valores altos de otro intervalo operando. Lo que sucede en esos casos es que el intervalo resultante puede ser menos amplio que en el caso que no asume dependencia. Por lo que el resultado es más preciso. Ferson y Kreinovich (2006) muestran diversos modelos de dependencia entre los intervalos, a través de una definición de correlación que relaciona los valores de un intervalo con los de otro. El problema que existe a la hora de aplicar estos modelos en un problema real es conocer el modelo de correlación y el valor de la misma. Los autores sugieren el uso de un intervalo de valores para el coeficiente de regresión y aconsejan asumir que no hay dependencia en los casos en los que no se esté seguro del modelo a elegir, ya que esa opción es la más conservadora.

2.7.1.2. Estadísticos para intervalos de incertidumbre

Consideremos unos intervalos (de forma más general, podrían considerarse unos conjuntos aleatorios) $\boldsymbol{x}_1, \boldsymbol{x}_2..., \boldsymbol{x}_n$, de forma que los valores reales $x_1, x_2..., x_n$ cumplen que $x_i \in \boldsymbol{x}_i$, $\forall i \in \{1, n\}$. El objetivo es conocer el valor de los estadísticos que caracterizan la distribución de la variable considerada. Sin embargo, al no estar disponibles los datos reales, sino intervalos que contienen los valores reales, los estadísticos que se pueden obtener también sufren esa imprecisión y, por tanto, sólo pueden obtenerse intervalos que contengan el valor real del estadístico.

Formalmente, dicho estadístico se representa como $\boldsymbol{y} = S_n(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$. El resultado $\boldsymbol{y}$ es un intervalo, si la función S_n que caracteriza el estadístico es continua, como en el caso de la varianza y de la media. Cuando $n \to \infty$, el intervalo $\boldsymbol{y}$ tiende a un intervalo límite L que contiene el valor del estadístico en la distribución que ha generado los valores reales. Pueden encontrarse más detalles sobre estadísticos para conjuntos aleatorios y sus propiedades asintóticas en Li, Ogura y Kreinovich (2002).

El problema que surge al calcular estadísticos para intervalos es que la complejidad computacional aumenta drásticamente y que el tiempo de proceso requerido es considerable o, en algunos casos, no asumible. Una excepción a este fenómeno, es la media

$$\bar{\boldsymbol{x}} = \frac{\boldsymbol{x}_1 + \boldsymbol{x}_2 + \ldots + \boldsymbol{x}_n}{n}, \tag{2.81}$$

donde las operaciones empleadas vienen descritas según la aritmética de intervalos (ver Sección 2.7.1.1). Gioia y Lauro (2005) enumeran las propiedades que cumple este estadístico, como, por ejemplo, el criterio de internalidad según el cual la media se encuentra contenida entre el mínimo y el máximo de los valores empleados para calcularla o la propiedad según la cual si se transforman linealmente un conjunto de intervalos y se calcula su media, el valor que se obtiene es el mismo que hallando en primer lugar la media de los intervalos y después aplicando la transformación lineal.

El cálculo de la media utilizando aritmética de intervalos da lugar al mínimo intervalo que contiene todas las medias. Sin embargo, el cálculo de la varianza con aritmética de intervalos y siguiendo la fórmula clásica,

$$S^2 = \frac{(\boldsymbol{x}_1 - \bar{\boldsymbol{x}})^2 + (\boldsymbol{x}_2 - \bar{\boldsymbol{x}})^2 + \ldots + (\boldsymbol{x}_n - \bar{\boldsymbol{x}})^2}{n}, \tag{2.82}$$

da lugar a un intervalo con un ancho excesivo. Esto es debido a que en dicha fórmula cada variable $\boldsymbol{x}_i$ aparece dos veces (una de forma directa y otra en la fórmula de $\bar{\boldsymbol{x}}$. Otra forma de hallar el valor de la varianza es mediante la resolución de un problema de optimización con restricciones. Sin embargo, el resultado de esos métodos es exponencial en n.

Ferson, Ginzburg, Kreinovich, Longpré y Aviles (2005) proponen una serie de algoritmos para calcular el limite inferior del intervalo varianza en

tiempo cuadrático, i.e. su complejidad es $O(n^2)$. Sin embargo, tal y como demuestran Ferson, Ginzburg, Kreinovich, Longpré y Aviles (2002), el cálculo del límite superior de la varianza es NP-difícil, es decir, no existe un algoritmo que garantice calcular dicho valor en todas las situaciones posibles en un tiempo razonable. Ferson et al. (2005) proponen un algoritmo para calcular su valor en tiempo cuadrático siempre y cuando los intervalos reducidos no intersecten entre sí. Los intervalos reducidos se obtienen transformando los intervalos originales de la siguiente forma

$$[\text{mid } \boldsymbol{x}_i - \text{rad } \boldsymbol{x}_i/n, \text{mid } \boldsymbol{x}_i - \text{rad } \boldsymbol{x}_i/n]. \tag{2.83}$$

Xiang (2006) propone un intervalo para el caso en que, para un valor de k fijo, no más de k intervalos reducidos intersecten entre sí que tiene una complejidad de $O(n \cdot \log(n))$.

Ferson et al. (2005) demuestran que el cálculo de los límites superiores e inferiores de los intervalos covarianza y correlación es un problema NP-difícil. Beck, Kreinovich y Wu (2004) proponen un algoritmo para el cálculo de la covarianza con una complejidad de $O(n^3)$. Por su parte, Ferson et al. (2005) muestran que el cálculo de la mediana para intervalos tienen una complejidad $O(n \cdot \log(n))$, siempre que no más de k intervalos intersecten entre sí. Una revisión más detallada sobre los algoritmos desarrollados para el cálculo de estadísticos con incertidumbre en forma de intervalo puede encontrarse en Kreinovich et al. (2006).

Paralelamente a estos desarrollos, Gioia y Lauro (2005) también proponen un método para el cálculo de la varianza, de la covarianza y del coeficiente de correlación para intervalos. Su propuesta es similar a la realizada en los los trabajos antes mencionados, sin embargo, no ofrece detalles sobre los algoritmos de optimización utilizados para hallar los límites inferior y superior de los estadísticos, ni de la complejidad de estos algoritmos, ni de su comportamiento ante casos desfavorables como cuando los intervalos considerados intersectan mucho entre sí. Los ejemplos que aparecen en dicho trabajo ilustran el funcionamiento de los estadísticos propuestos ante distintas situaciones, pero corresponden a situaciones relativamente favorables donde el número de intervalos es relativamente bajo y éstos no intersectan apenas por lo que los tiempos de cálculo requerido no han debido ser altos.

En Canal y Marques Pereira (1998) se propone otra forma de calcular la varianza para intervalos. Este enfoque está basado en el cálculo de dos índices escalares asociados al intervalo: el rango y el módulo. Según estos autores, el rango es una medida escalar entre 0 y 1 que determina lo separado que se puede considerar el intervalo del valor cero; mientras que el módulo es una medida escalar que estima la diferencia entre el intervalo considerado y el intervalo $[0, 0]$, con lo que se puede decir que viene a ser similar al valor absoluto en números reales. La definición de varianza propuesta está basada en el módulo, el cual se aplica como valor absoluto y cuya definición

requiere del cálculo del rango que a su vez depende de un parámetro que fija el analista. En este trabajo no queda suficientemente claro qué pretende reflejar esta definición de varianza, ni la utilidad práctica de la misma. Esta debe ser la razón de que no se hayan encontrado trabajos donde se aplique o se continúe esta línea de investigación.

2.7.1.3. Modelos de regresión lineal

A continuación se revisan los principales aproximaciones al problema de la regresión lineal desde el área del análisis de intervalos.

El enfoque basado en el análisis de intervalos. Marino y Palumbo (2002) proponen un método para realizar regresión lineal para datos de intervalos que se encuadra dentro del análisis de intervalos. Por tanto, en él los intervalos representan una acotación del valor real de la variable, el cual no puede ser determinado con exactitud debido a, por ejemplo, la imprecisión proveniente de las herramientas de medida o de otras fuentes.

Empleando la notación del análisis de intervalos descrita anteriormente, el modelo de regresión propuesto por Marino y Palumbo (2002) se denota como

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.84}$$

donde $\boldsymbol{Y}$ y $\boldsymbol{\varepsilon}$ son vectores de $n \times 1$ intervalos, $\boldsymbol{X}$ es una matriz de intervalos de $n \times p+1$ y $\boldsymbol{\beta}$ es el vector de coeficientes de $p+1 \times 1$ intervalos. Es decir, todos los términos del modelo son intervalos. Al aplicar mínimos cuadrados sobre dicho modelo se obtiene

$$Q = min_{\beta \in \Re} \left(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\right)^2 \equiv min(\boldsymbol{\varepsilon})^2. \tag{2.85}$$

La solución a esta ecuación requiere

$$\frac{dQ}{d\beta} = -2\left(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\right)\boldsymbol{X}. \tag{2.86}$$

Mediante la generalización de las reglas de la diferenciación a la aritmética de intervalos, se obtiene que el vector de coeficientes en forma de intervalo $\boldsymbol{\beta}$ se determina encontrando la solución del siguiente sistema lineal

$$\begin{aligned} (\boldsymbol{X}'\boldsymbol{X})\boldsymbol{\beta} &= \boldsymbol{X}'\boldsymbol{Y} \\ \boldsymbol{\beta} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{Y}) \end{aligned} \tag{2.87}$$

La estimación de $\boldsymbol{\beta}$ requiere calcular $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ lo cual es posible si se da la condición de regularidad fuerte.

Definición 1 *Una matriz de intervalos $\boldsymbol{A}$ es **regular** si toda matriz clásica $A \in \boldsymbol{A}$ es no-singular i.e. tiene inversa.*

Definición 2 *Una matriz de intervalos* $\boldsymbol{A}$ *es* ***fuertemente regular*** *si cumple*

$$\rho\left(|mid(\boldsymbol{A}^{-1})|rad(\boldsymbol{A})\right) < 1 \tag{2.88}$$

donde $\rho(\cdot)$ *es el radio espectral, que para una matriz cuadrada de datos clásicos* M *se define como*

$$\rho(M) = max\{|\lambda|\} \text{ donde } \lambda \text{ es un autovalor de } M. \tag{2.89}$$

La regularidad fuerte implica que en un espacio p-dimensional, para cualquier dirección genérica, la variabilidad de los centros de los intervalos prevalece sobre la variabilidad de los rangos de los mismos. Sin embargo, en algunos casos se puede obtener la solución de un sistema lineal, aunque no se satisfaga la condición de la regularidad fuerte.

Marino y Palumbo (2002) afirman que en los problemas estadísticos la condición de regularidad fuerte no es habitual y plantean un algoritmo de programación lineal para hallar la solución de la ecuación (2.87) para todos los casos (incluyendo los casos en los que se dé la regularidad fuerte) y dando un $\boldsymbol{\beta}$ cuyos intervalos sean lo más reducidos posibles.

Marino y Palumbo (2002) muestran un ejemplo del ámbito químico en el que tienen dos variables clásicas relacionadas linealmente sobre las que añaden cierta imprecisión generada de forma artificial para crear una versión del conjunto de datos en forma de intervalo y mostrar la utilidad de su método.

El fin de dicho método es modelar la variabilidad debida a una relación lineal existente entre dos variables a pesar de que la incertidumbre intrínseca de los datos obliga a expresarlos en forma de intervalo. Este enfoque difiere del enfoque simbólico, en el cual la variabilidad de los datos (i.e. el rango de los intervalos) no significa imprecisión en las medidas, sino variabilidad; y dicha variabilidad debe ser recogida por el modelo puesto que es posible que ella condicione la relación lineal. Marino y Palumbo (2003) trabaja con el mismo método sobre un ejemplo relacionado con la permeabilidad del suelo.

Corsaro y Marino (2006) realiza un revisión sobre el estado del arte de los sistemas lineales de intervalos y muestra tres métodos para resolverlos: el método de Krawczyk, el método de Intervalos Gauss-Seidel y el método de Eliminación Gauss de Intervalo. Siendo los dos primeros iterativos y el tercer un método directo. El artículo realiza una serie de simulaciones para determinar cuál de los métodos funciona mejor. El problema que resuelven las simulaciones es el del ajuste de un modelo de regresión lineal de intervalos. En dicho ejemplo, los autores primero consideran datos clásicos, luego introducen perturbaciones en la variable dependiente y, por último, en la independiente. Con ello pretenden también estudiar cómo se propaga la incertidumbre de los datos en la solución. Las conclusiones del estudio son que el método de intervalos Gauss-Seidel es el que obtiene un mejor compromiso entre precisión y eficiencia.

El enfoque basado en el conjunto de rectas de regresión. Otro modelo de regresión lineal dentro del análisis de intervalos es planteado por Gioia y Lauro (2005). En él, como es natural en el análisis de intervalos, se considera que el intervalo es una representación que contiene el valor real. El modelo de regresión que plantean estos autores pretende acotar todas las posibles rectas de regresión que se pueden dar entre los infinitos puntos contenidos dentro de los pares de intervalos considerados.

De manera más formal, sea $\boldsymbol{X}$ la variable de intervalo independiente e $\boldsymbol{Y}$ la variable de intervalo dependiente, sean $\boldsymbol{x}_i = [\underline{x}_i, \overline{x}_i]$ e $\boldsymbol{y}_i = [\underline{y}_i, \overline{y}_i]$ los valores observados para el individuo i con $i = 1, ..., n$, y sean $x_i \in \boldsymbol{x}_i$ e $y_i \in \boldsymbol{y}_i$ los infinitos valores contenidos dentro de los intervalos $\boldsymbol{x}_i$ e $\boldsymbol{y}_i$, respectivamente, el modelo propuesto por Gioia y Lauro (2005) viene determinado por dos parámetros en forma de intervalo $\boldsymbol{a}$ y $\boldsymbol{b}$ tal que

$$\boldsymbol{y}_i = \boldsymbol{a} + \boldsymbol{b}\boldsymbol{x}_i, \tag{2.90}$$

donde $\boldsymbol{a}$ y $\boldsymbol{b}$ son dos intervalos que acotan el conjunto de constantes y de pendientes de las rectas de regresión que se dan entre los infinitos posibles pares de puntos (x_i, y_i) contenidos dentro de los intervalos considerados.

Los autores determinan el valor de los parámetros resolviendo un problema de optimización con el objetivo de buscar los límites inferiores y superiores de los intervalos $\boldsymbol{a}$ y $\boldsymbol{b}$. Para probar la validez del enfoque lo prueban en el conjunto de datos empleado en Marino y Palumbo (2003) y obtienen unos parámetros en forma de intervalo más estrechos, i.e. con mayor precisión.

El enfoque basado en la teoría de conjuntos aleatorios. Otro enfoque que aborda la regresión para intervalos se ha planteado desde la teoría de conjuntos aleatorios. Según dicha teoría, los intervalos son un caso particular de los conjuntos aleatorios compactos convexos. Otros conjuntos aleatorios convexos más generales que los intervalos son los hipercubos y las esferas. La teoría de conjuntos aleatorios es, además, coherente con la aritmética de intervalos. Desde este área, Gil, Lubiano, Montenegro y López-García (2002) proponen emular el modelo de regresión lineal para intervalos mediante una transformación afín entre la variable de intervalo independiente y la dependiente. Este trabajo, basado en el modelo desarrollado por Diamond (1990), permite encontrar soluciones óptimas para la transformación afín mediante un procedimiento de optimización mínimo cuadrática que minimiza los errores cometidos. Dichos errores se estiman como la distancia entre los intervalos observados y estimados medida mediante una métrica generalizada en el espacio de intervalos compactos no vacíos. La transformación afín con la que emulan la regresión es la siguiente

$$\boldsymbol{y} = a\boldsymbol{x} + \boldsymbol{b}, \tag{2.91}$$

es decir, el intervalo $\boldsymbol{x}$ es multiplicado por el coeficiente escalar a y sumado, mediante la suma de Minkowski, con el intervalo $\boldsymbol{b}$ para obtener el intervalo

$\boldsymbol{y}$. Las soluciones que se obtienen con dicho enfoque son únicas o, en algunos casos, dobles, i.e. hay dos posibles soluciones. Una diferencia importante con el enfoque de Marino y Palumbo (2003) es que en dicho modelo el coeficiente era un valor de intervalo y no un escalar lo que afecta a la interpretación de dicho modelo. Conviene reseñar que Gil et al. (2002) emplea una métrica para cuantificar la exactitud del modelo y define un coeficiente de determinación para estudiar la relación entre los conjuntos aleatorios considerados.

González-Rodríguez, Colubi, Coppi y Giordani (2006) demuestran mediante una serie de ejemplos y de simulaciones que el modelo óptimo formulado según (2.91) y estimado mediante mínimos cuadrados no produce valores de $\boldsymbol{y}$ adecuados. Esto es debido a que pueden darse casos en los que $\boldsymbol{y} -_H a^*\boldsymbol{x}$ no exista, siendo a^* el valor óptimo estimado para el coeficiente a y siendo $\boldsymbol{a} -_H \boldsymbol{b} = [\underline{a} - \underline{b}, \overline{a} - \overline{b}]$ una operación de aritmética de intervalos conocida como la diferencia de Hukuhara que sólo está definida si $\underline{a} - \underline{b} \leq \overline{a} - \overline{b}$. Este problema quiere decir que los valores obtenidos por el modelo estimado no son válidos.

Para superar esos problemas González-Rodríguez et al. (2006) proponen el siguiente modelo

$$\boldsymbol{y}|\boldsymbol{x} = a\boldsymbol{x} + \boldsymbol{\epsilon}_x, \tag{2.92}$$

donde $\boldsymbol{\epsilon}_x$ representaría un intervalo residual tal que, para todos los individuos i del conjunto considerado, cumple que $\boldsymbol{y}_i -_H \hat{a}\boldsymbol{x}_i = \boldsymbol{\epsilon}_{x,i}$, siendo $\hat{a}$ el valor estimado para el coeficiente a y siendo $E(\boldsymbol{\epsilon}_x,) = \boldsymbol{b}$, de forma que la función de regresión de la población vendría dada por $E[\boldsymbol{y}|\boldsymbol{x}] = ax + \boldsymbol{b}$ para todo $x \in \bigcup_i \boldsymbol{x}_i$.

Los autores desarrollan una variante del método de mínimos cuadrados para estimar el modelo con estas restricciones y, según demuestran empíricamente, dicho modelo mejora los resultados del modelo sin restricciones en términos del error cuadrático medio. El modelo obtenido por este método asegura que $\boldsymbol{y} -_H \hat{a}\boldsymbol{x}$ se cumple al menos para todos los intervalos utilizados para estimar el modelo. González-Rodríguez et al. (2007) extienden el modelo de González-Rodríguez et al. (2006) para proponer un modelo de regresión múltiple para intervalos.

Gil, González-Rodríguez, Colubi y Montenegro (2007) parten del modelo de regresión propuesto por González-Rodríguez et al. (2006) y proponen un test para comprobar si las variables de intervalo de $\boldsymbol{x}$ y de $\boldsymbol{y}$ son independientes, es decir, si $a = 0$ en la ecuación (2.92). En dicho trabajo se demuestra que $a = 0$ si y sólo si se da simultáneamente que los centros de $\boldsymbol{x}$ y de $\boldsymbol{y}$ y los radios (o rangos) de $\boldsymbol{x}$ y de $\boldsymbol{y}$ son independientes. Este resultado confirma la idea intuitiva de que es más correcto trabajar con los centros y los radios de los intervalos, en lugar de con sus extremos inferiores y superiores.

2.7.2. El cálculo con distribuciones de probabilidad

El cálculo con distribuciones de probabilidad es en realidad una sofisticación con respecto al análisis y al cálculo de intervalos. El objetivo en ambos campos es similar: realizar cálculos con valores cuya magnitud no conocemos con absoluta certeza o precisión. La diferencia es que en el análisis de intervalos sólo se conocen los límites entre los que se encuentran cada uno de los valores estudiados, mientras que al emplear distribuciones de probabilidad se tiene información más detallada sobre esta incertidumbre.

Si las distribuciones de probabilidad de entrada están definidas completamente y se conocen perfectamente las relaciones de dependencia entre ellas, se pueden emplear métodos basados en simulaciones de Monte Carlo. Sin embargo, tal y como indica Ferson (1996), estos métodos no son válidos si no se conoce la relación de dependencia entre las distribuciones o si éstas no están completamente especificadas.

Sin embargo, existen otros enfoques que permiten operar aritméticamente con las distribuciones de manera directa y evitando las tediosas simulaciones. Una primera aproximación a la materia es planteada por Colombo y Jaarsma (1980) que proponen un método para operar con variables aleatorias representadas mediante histogramas. Dicho método supone una extensión de la aritmética de intervalos de Moore (1966) para realizar operaciones con histogramas. Esta idea se basa en que los histogramas son, al fin y al cabo, un conjunto de intervalos con una probabilidad (o frecuencia) asociada. Posteriormente, Moore (1984) propone un método muy similar al método propuesto por Colombo y Jaarsma (1980) para operar con la unión de intervalos disjuntos pero contiguos y con una probabilidad asociada. Otro trabajo relacionado con estos enfoques es el de Kaplan (1981) que aproxima las funciones de densidad continuas mediante una secuencia de funciones delta.

El trabajo doctoral de Robert C. Williamson sobre aritmética probabilística, publicado en el International Journal of Approximate Reasoning (Williamson y Downs, 1990), constituye otro hito notable dentro de la disciplina. En él se propone caracterizar el resultado de dichas operaciones no mediante un histograma sino mediante dos recubrimientos (*envelopes*). Los recubrimientos contendrían la función de probabilidad solución teniendo en cuenta los errores de representación, es decir, los errores debidos a la aproximación de las funciones mediante conjuntos finitos de puntos o de coeficientes. Esta representación mediante recubrimientos encaja con el concepto de límites de dependencia que son la aproximación inferior y superior de la función resultado cuando no se tiene información sobre la dependencia entre los operandos.

Berleant y Zhang (2004) proponen un método similar que permite calcular las funciones resultado de operar con variables aleatorias que pueden: ser independientes, tener una relación de dependencia caracterizada de forma completa o parcial, o que no se puede caracterizar por ser desconocida.

Otros métodos similares son propuestos por Li y Mac Hyman (2004) y por Lodwick y Jamison (2003).

Por el momento, no hay investigaciones sobre el cálculo de estadísticos con este tipo de datos. Sin embargo, como se puede anticipar, el cálculo de esos estadísticos o, mejor dicho, de la distribución de probabilidad de los mismos no es trivial y ha de suponer una carga computacional considerable.

Capítulo 3

Las Series Temporales Simbólicas

Todas las teorías son legítimas
y ninguna tiene importancia.
Lo que importa es lo que se hace con ellas.

Jorge Luis Borges

En este capítulo se abordarán las series temporales simbólicas, es decir, las series temporales donde la variable observada a lo largo del tiempo es una variable simbólica. Este nuevo tipo de serie temporal permite representar la variabilidad que se dé en cada instante. Dicha variabilidad puede representarse, por ejemplo, mediante un histograma o un intervalo. En el capítulo se definirán estas series temporales, se explicarán cómo se pueden obtener y los métodos que se han propuesto en la literatura para predecirlas. Antes, el capítulo realiza un repaso de otras aproximaciones que están relacionadas de alguna manera con las series temporales simbólicas. Estas aproximaciones incluyen las series temporales multivariantes, las predicciones de intervalo y de densidad, las series temporales valoradas mediante alfabetos de símbolos y las series temporales de gráficos de velas o *candlesticks*.

3.1. Introducción

Tal y como se ha indicado en el capítulo 2, el análisis de datos simbólicos es un área novedosa dentro de la estadística y de la minería de datos. Su objetivo es el de extraer información sobre datos expresados mediante variables simbólicas, como por ejemplo variables de intervalo y variables de histograma. Estas variables son más complejas que las variables clásicas que describen los elementos mediante un único valor, pero describen de manera más fiel la realidad al poder representar variabilidad.

Al ser un área que tiene menos de 20 años y con un objetivo tan ambicioso, existen muchas posibilidades de desarrollo dentro de ella. En un principio las contribuciones en el área se centraron en el análisis *cluster* y en los datos de intervalo, ya que éstos entrañan una menor complejidad. Con el paso de los años el catálogo de métodos se ha ido ampliando y es más habitual encontrar métodos propuestos para otros datos simbólicos, aunque las técnicas desarrolladas para datos de histogramas siguen siendo más escasas. Pese a esta evolución, tal y como afirman Billard y Diday (2003), es necesario y deseable que el repertorio de métodos que tratan este tipo de datos continúe ampliándose.

El objetivo de esta tesis es el de hacer confluir el área de predicción de series temporales y el del análisis de datos simbólicos. En la fecha en la que se empezó a trabajar en la tesis, el año 2004, no existía en la literatura especializada ninguna propuesta que abordase la predicción o el análisis de series temporales desde la perspectiva del análisis de datos simbólicos. Sin embargo, durante estos años han surgido algunas propuestas para predecir series temporales de intervalos y, más recientemente, para predecir series temporales de histogramas.

En este capítulo se definirá el concepto de serie temporal simbólica el cual engloba a las series temporales de intervalo y de histograma. También se explica cómo se pueden obtener este tipo de series y se estudia su conexión con otras aproximaciones del análisis de series temporales en las que la información que se maneja va más allá del valor puntual. Por último, se detallarán los métodos propuestos para la predicción de series temporales simbólicas al margen de la tesis.

3.2. El concepto de serie temporal simbólica

Antes de entrar a definir el concepto de serie temporal simbólica es preciso definir el concepto de serie temporal clásica. Una serie temporal es el resultado de observar los valores de una variable a lo largo del tiempo, normalmente, con una frecuencia uniforme. La figura 3.1 muestra un ejemplo de una serie temporal. Para formular una definición formal de serie temporal se requiere del concepto de proceso estocástico.

Un proceso estocástico es un conjunto de variables aleatorias X_t donde el índice t toma valores en un conjunto T de instantes temporales para los que el proceso está definido. Formalmente se denota como

$$\{X_t\} \text{ tal que } t \in T \text{ con } T \subseteq \mathbb{R}. \tag{3.1}$$

Cada una de las variables aleatorias del proceso se rige según su propia función de distribución de probabilidad, pudiendo dichas funciones de distribución estar correlacionadas.

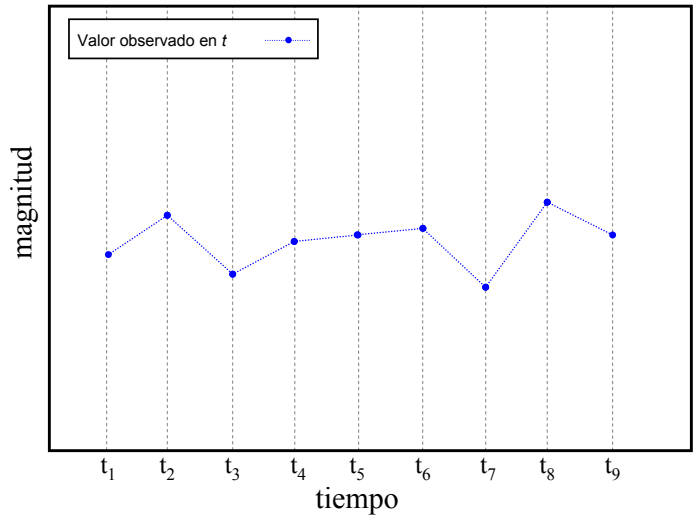


Figura 3.1: Serie temporal clásica: valores observados x_t

En los procesos estocásticos que consideraremos, T es discreto, los valores $t \in T$ están espaciados de forma uniforme a lo largo del tiempo y el valor observado de la variable aleatoria X_t en t, que se denota como x_t, no depende en ningún caso de los valores futuros. Una serie temporal $\{x_t\}$ es la realización finita de un proceso estocástico $\{X_t\}$ de estas características. En otras palabras, es una muestra de tamaño uno del conjunto de variables aleatorias $\{X_t\}$ que definen el proceso estocástico. La figura 3.2 ilustra este concepto.

En las series temporales normalmente sólo se dispone de una realización del proceso estocástico a estudiar, es decir, la serie temporal $\{x_t\}$ suele ser la única observación disponible de $\{X_t\}$. Este hecho complica su análisis y lo diferencia de otras áreas de la estadística donde se cuenta con un conjunto de realizaciones de las variables aleatorias de interés. Al disponer de una única realización del proceso estocástico subyacente resulta más complicado determinar las propiedades de dicho proceso, que es precisamente el objetivo del análisis de series temporales. Para poder estimar dichas propiedades, es necesario que el proceso sea estacionario o, dicho informalmente, que el proceso subyacente no cambie a lo largo del tiempo. Más detalles sobre este punto pueden encontrarse en textos básicos sobre la materia como Chatfield (2001b) y Peña (2005).

Según el concepto clásico de serie temporal, las observaciones de una serie son valores puntuales, es decir, cada instante temporal t son descritas mediante un único valor x_t de la variable X_t. Dichos valores puntuales no son capaces de representar la variabilidad de la observación en el instante t. Esto

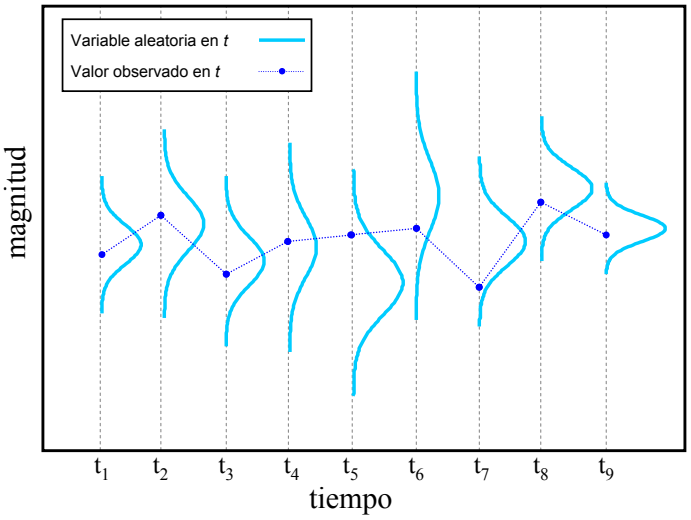


Figura 3.2: Proceso estocástico clásico: variables aleatorias X_t y valores observados x_t

no es necesario en muchos casos, pero existen otros, como cuando se estudia la evolución en el tiempo de una variable medida en conjunto de individuos, donde sí es aconsejable poder plasmar la variabilidad de la observación en t de alguna forma.

Para los casos en que la variabilidad sea importante, esta tesis propone emplear series temporales donde cada observación sea descrita mediante un rango de valores observados (es decir, un intervalo que contiene todos los valores observados) o una distribución de frecuencias de los valores observados recogida en forma de histograma; es decir, donde la variable observada sea una variable simbólica de intervalo o de histograma.

3.3. Definición de serie temporal simbólica

Una serie temporal simbólica es una serie temporal donde las observaciones son representadas mediante un dato simbólico. Tal y como se mencionó en el capítulo 2, existen varios tipos de datos simbólicos. Sin embargo, aunque esta tesis se circunscribe al ámbito de las series temporales valoradas mediante intervalos y mediante histogramas, también podrían considerarse, por ejemplo, series temporales valoradas mediante listas de valores.

Para proponer una definición más formal de serie temporal simbólica es preciso definir previamente el concepto de variable simbólica aleatoria y de proceso estocástico simbólico.

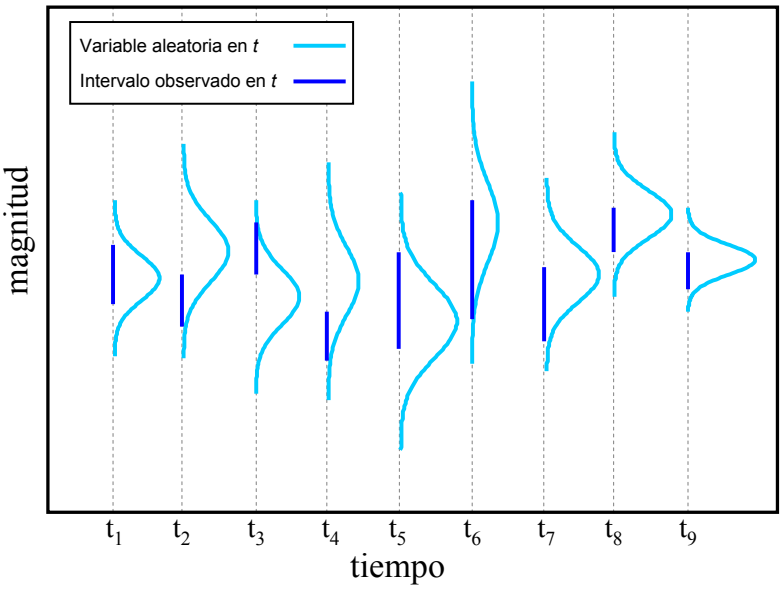


Figura 3.3: Proceso estocástico clásico: representación de las variables aleatorias X_t y de los intervalos observados x_t

Variable aleatoria simbólica. Una variable aleatoria simbólica es una función que asigna un valor simbólico a cada elemento del espacio muestral. En el caso de las variables simbólicas de intervalo, a cada elemento del espacio muestral se le asociará un intervalo, y en el caso de las variables simbólicas de histograma un histograma. Pueden verse las definiciones concretas de cada una de estas variables aleatorias simbólicas en el apartado 2.2.

Cada variable aleatoria simbólica se rige según su propia distribución. El cálculo de las funciones de densidad y de distribución empíricas para las variables aleatorias simbólicas de intervalo y de histograma fue definido en los apartados 2.3.1 y 2.3.2, respectivamente.

Proceso estocástico simbólico. Un proceso estocástico simbólico es definido por un conjunto de variables aleatorias simbólicas $\{X_t\}$, donde cada variable X_t está indexada por un índice t, tal que $t \in T$, con $T \subseteq \mathbb{R}$, i.e. T denota los instantes para los que el proceso está definido.

Serie temporal simbólica. Una serie temporal simbólica es la realización de un proceso estocástico simbólico. Es decir, es una muestra de tamaño uno del vector de variables aleatorias simbólicas que caracteriza el proceso estocástico simbólico.

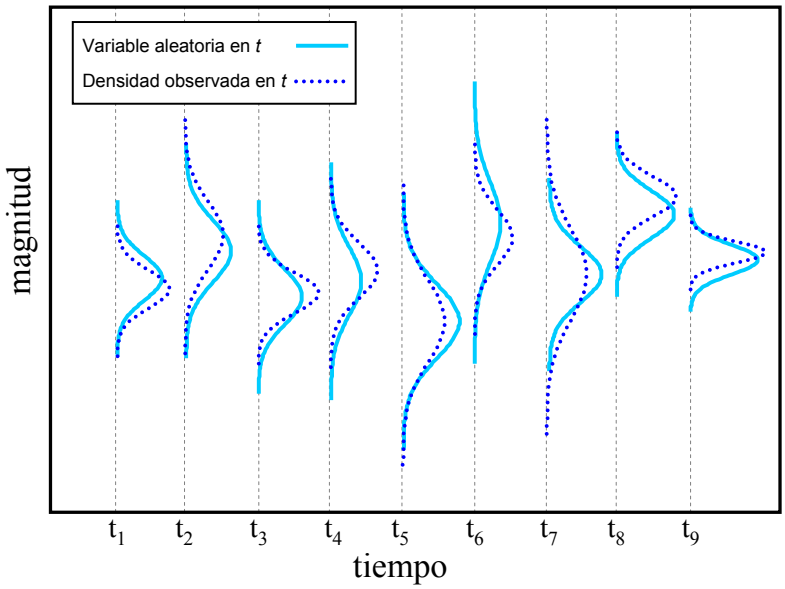


Figura 3.4: Proceso estocástico clásico: representación de las variables aleatorias X_t y de las densidades observadas x_t

La figura 3.3 ilustra la idea de que una serie temporal simbólica de intervalos es una realización de un proceso estocástico simbólico caracterizado en cada instante por una variable aleatoria y que tiene como valor observado un intervalo. La figura 3.4 hace lo propio para el caso en el que las observaciones son funciones de densidad.

Por el momento, dentro del análisis de datos simbólicos no se ha desarrollado una teoría que estudie los procesos estocásticos de naturaleza simbólica. Sin ese paso previo, no puede definirse el concepto de estacionariedad para este tipo de procesos, ni están claras las implicaciones de dicha propiedad de cara a la elaboración de predicciones.

Las primeras aportaciones que han aparecido a lo largo de estos años dedicadas a las series temporales simbólicas no abordan el tema desde la perspectiva de los procesos estocásticos simbólicos, sino que toman una perspectiva más pragmática centrándose únicamente en la predicción de las series. En esta tesis se adoptará también dicha perspectiva.

3.4. Origen de las series temporales simbólicas

Pese a que hasta ahora no haya existido una literatura que se dedicase al análisis de las series temporales simbólicas, éstas son más frecuentes en

la vida real de lo que se podría pensar en un primer momento. Las series temporales valoradas mediante variables simbólicas pueden surgir en alguno de los siguientes supuestos:

1. Cuando se mide el valor de una variable en una población de individuos a lo largo del tiempo, pero el interés no recae sobre los valores individuales, sino sobre el comportamiento de la población como conjunto. En este caso, las variables simbólicas resultan una herramienta óptima para agregar la información individual concentrándose, por ejemplo, en el rango de los valores observados en cada instante, o en un boxplot o un histograma que resuma la distribución de valores observados.

2. Cuando la variable se observa con una determinada frecuencia (por ejemplo, cada minuto), pero se quiere analizar su comportamiento en una frecuencia menor (por ejemplo, cada día) pero recogiendo información sobre la variabilidad de los valores observados entre cada instante de la frecuencia de interés (en este caso, entre cada día). Las variables simbólicas sirven, en este caso, para resumir las observaciones de una forma que recoja más información que una serie temporal clásica.

Las situaciones aquí descritas corresponden con casos de agregación: el punto 1 con agregación temporal y el punto 2 con agregación contemporánea. La agregación es, en muchos casos, una estrategia necesaria a la hora de abordar el análisis de datos en la vida real. Hoy en día, debido a la gran proliferación de sistemas informáticos y de las bases de datos es habitual encontrarse con una gran cantidad de datos almacenados. En algunos de los casos, la frecuencia con la que se registran dichos datos o la gran cantidad de individuos para los que se almacena información hace inabordable su análisis sin una estrategia adecuada. La agregación es una alternativa a considerar para analizar dicha información.

Si a la hora de agregar la información, en lugar de usar series temporales simbólicas, se emplean series temporales clásicas, la información que se recoge es menor. Si dicha simplificación es excesiva, las series temporales simbólicas ofrecen una alternativa interesante para reflejar más fielmente la realidad al recoger la variabilidad o la incertidumbre del fenómeno considerado. En los casos en los que usar una serie temporal clásica sea suficiente, las series temporales simbólicas pueden ofrecer una visión complementaria.

A continuación, se muestran una serie de ejemplos de cada tipo de agregación con el fin de ilustrar situaciones donde las series temporales simbólicas resultan útiles.

3.4.1. Ejemplos de agregación contemporánea

Un caso muy interesante a tener en cuenta en el que la agregación contemporánea resulta útil es el caso de un instituto nacional de estadística que

registra una serie de variables para todos o para una muestra de los habitantes de un país a lo largo del tiempo. Para cada variable, los datos de los habitantes opueden ser resumidos mediante su media o su mediana o mediante cualquier otro estadístico. Un buen ejemplo de este caso lo constituye la renta per capita ya que es interesante el estudio de su evolución a lo largo del tiempo es interesante, pero lo sería aún más conocer la evolución de la distribución de los ingresos de los habitantes del país y no solo de su media.

En realidad, el caso prototípico de los datos que maneja un instituto de estadística, es extensible a las empresas demoscópicas que realizan sondeos de opinión o paneles de consumidores (Baltagi, 2004). Si éstos se repiten a lo largo del tiempo y se quiere estudiar su evolución temporal, las series temporales simbólicas pueden resultar una herramienta muy útil. De hecho, en este tipo de datos no se puede trabajar con los datos desagregados, es decir, con series temporales clásicas, ya que los individuos para los cuales se recoge la información no tienen por qué ser los mismos en los diferentes instantes en los que se recoge la información. Por tanto, la agregación es necesaria y el uso de intervalos, gráficos de caja o histogramas, puede complementar muy bien al uso de otros estadísticos como la media.

Otro caso donde la agregación contemporánea resulta útil es el de las redes de estaciones meteorológicas o medioambientales. En estos casos, normalmente se dispone de una serie de estaciones situadas en puntos estratégicos formando una malla que, o bien se distribuye uniformemente en el espacio, o bien que cubre todos los puntos considerados relevantes. Si el objetivo no está en estudiar el comportamiento individual de cada una de las estaciones, sino el comportamiento global, la agregación puede resultar muy útil y el uso de las series temporales simbólicas es una alternativa a tener en cuenta.

En el ámbito financiero, las series temporales simbólicas pueden utilizarse para agregar información de, por ejemplo, los rendimientos de un conjunto de acciones a lo largo del tiempo. Es interesante destacar que este enfoque ha sido planteado en el trabajo de González-Rivera et al. (2008) donde se muestra una serie temporal de histogramas que representa la distribución de los rendimientos de las acciones que conforman el índice bursátil S&P500.

3.4.2. Ejemplos de agregación temporal

En agregación temporal, uno de los ámbitos que mejor ilustra el uso de series temporales simbólicas es el de la climatología. En climatología, variables como la temperatura en una determinada localidad, se miden de manera continua o casi continua. Sin embargo, el interés en las mismas reside en estudiar su comportamiento con una frecuencia fija (normalmente, diaria), utilizándose únicamente sus valores mínimo y máximo. En estos casos, las series temporales de intervalos surgen de forma natural al considerar los intervalos a lo largo del tiempo.

Otra área donde se puede aplicar este tipo de agregación es la hidrología, donde variables como el caudal de los ríos suelen ser registradas mediante sus valores mínimo y máximo en un periodo determinado.

En las finanzas también es habitual agregar temporalmente la información. En este ámbito, los valores de las cotizaciones de las acciones, de las divisas y de los índices se actualizan con una frecuencia alta, pero son típicamente resumidos mediante cuatro valores: apertura, cierre, mínimo y máximo. En estos cuatro valores subyacen dos intervalos: el del mínimo y el del máximo, y el de la apertura y el cierre (que es un intervalo que tiene una dirección, pues indica si el valor ha aumentado o ha decrecido en el periodo considerado).

En general, cualquier serie de datos que sea registrada de forma continua es susceptible de ser agregada temporalmente. Ejemplos de estos tipos de datos son el tráfico de las redes de ordenadores, las búsquedas de información en la web y los datos provenientes de sensores. Del análisis de este tipo de datos se encarga el área del *data stream mining* (Gama y Gaber, 2007), sin embargo, también son susceptibles de ser representados mediante series temporales simbólicas. De hecho, en muchos casos, estos datos sólo pueden ser leídos una única vez o una cantidad limitada de veces debido a restricciones de almacenamiento, por ello el hecho de usar datos simbólicos para representarlos puede ofrecer soluciones interesantes. Hébrail y Lechevallier (2007) muestran una primera aproximación simbólica a este tema.

3.5. Aproximaciones que van más allá de las series temporales clásicas

Tal y como se indica en el apartado anterior, las series temporales clásicas presentan limitaciones a la hora de representar algunas situaciones que surgen en la vida real. Pese a estas limitaciones, hasta el momento no se ha planteado el uso de las series temporales simbólicas. Sin embargo, si existen una serie de propuestas que de una u otra manera van más allá del concepto de serie temporal clásica y que encierran algún tipo de relación con las series temporales simbólicas. En este apartado se repasan algunas propuestas existentes en la literatura de las series temporales donde se intenta superar las limitaciones que impone la serie temporal clásica.

En primer lugar se analizan las aproximaciones que se encuentran dentro de lo que se podría llamar el análisis clásico de series temporales, es decir, aquellas aproximaciones que se basan en el concepto de serie temporal como realización de un proceso estocástico. Entre ellas se encuentran:

- Las series temporales multivariantes, que proporcionan la posibilidad de incluir otras variables para explicar y predecir el comportamiento de la variable que se está analizando.

- Las predicciones de intervalo y de densidad, que reflejan la necesidad de obtener predicciones que ofrezcan más información que la mera estimación puntual, informando sobre la incertidumbre que rodea a una predicción.
- La agregación de series temporales.

Ninguna de estas aproximaciones trasciende el concepto de serie temporal clásico. Sin embargo, existen otras aproximaciones que sí suponen una mayor ruptura con dicho concepto al plantear una representación de la información temporal que trasciende la secuencia de valores puntuales, son las siguientes:

- Las series temporales basadas en alfabetos de símbolos.
- Las series temporales de gráficos de velas.
- Las series temporales valoradas mediante variables simbólicas, como las variables de histograma o de intervalo.

Todas ellas serán tratadas en este apartado. Sin embargo, antes de continuar es necesario resolver la ambigüedad que surge con el término simbólico. Tal y como se dijo en la sección 1.3, existen dos áreas, el análisis simbólico y el análisis de datos simbólicos, que hacen uso del término *simbólico* de una forma ligeramente distinta. En la primera, el término viene dado porque las series temporales clásicas son discretizadas por medio de un alfabeto de símbolos. En la segunda, el término simbólico se utiliza para referirse a datos que se expresan mediante intervalos, variables modales, listas de valores categóricos o cuantitativos, histogramas, etc. Se podría decir que el análisis de datos simbólicos subsume al análisis simbólico, pero realmente las dos áreas son distintas.

3.5.1. Aproximaciones desde el ámbito de las series temporales clásicas

3.5.1.1. Los intervalos y las densidades de predicción

En un primer momento, las predicciones de las series temporales clásicas eran dadas como un único valor, sin mayor información sobre la precisión de dicho valor. En algunos casos, además, dichas predicciones eran representadas con muchos dígitos ofreciendo una precisión que era, en el fondo, espuria. Resulta obvio que la predicción puntual es muy útil y que en algunos casos puede ser la única información que el usuario final emplee en la toma de decisiones, pero toda predicción conlleva asociada una incertidumbre que no conviene pasar por alto. Para reflejar dicha incertidumbre se utilizan los intervalos y las densidades de predicción.

La incertidumbre en las predicciones proviene del desconocimiento de cuatro factores:

- Las innovaciones futuras.
- La distribución que siguen las innovaciones.
- Los verdaderos valores de los parámetros del modelo considerado.
- El modelo que ha generado los datos.

El primer factor es inevitable ya que los valores futuros dependen de innovaciones futuras que no son conocidas. Como consecuencia natural de este factor, la incertidumbre crece con el horizonte de predicción. Los otros tres factores pueden decrecer al aumentar el tamaño muestral.

Las buenas prácticas de predicción recomiendan representar verazmente la incertidumbre mediante un intervalo o una densidad de predicción. En líneas generales, estas representaciones sirven para:

- Informar sobre la incertidumbre que rodea a una predicción.
- Permitir la elaboración de distintas estrategias de acuerdo con el rango de posibles valores que se estima tendrá la variable observada en el futuro.
- Comparar predicciones de distintos métodos con mayor detalle.

Intervalo de predicción. Los intervalos de predicción indican la probabilidad de que el valor pronosticado se encuentre entre dos límites determinados, e.g. el intervalo de predicción del 95 %. El intervalo de predicción depende muy estrechamente del tamaño de los errores de predicción. A continuación se muestra la fórmula que se emplea habitualmente para hallarlos.

Sea $\hat{x}_N(h)$ el valor pronosticado para el instante $N+h$ por un modelo que emplea datos hasta el instante N y sea $e_N(h)$ el error de esa predicción, el intervalo de predicción con probabilidad asociada de $100(1-\alpha)\,\%$ viene dado por

$$\hat{x}_N(h) \pm z_{\alpha/2}\sqrt{Var[e_N(h)]}, \tag{3.2}$$

donde $z_{\alpha/2}$ denota el punto porcentual de la distribución normal estándar con una proporción $\alpha/2$ por encima de él, y donde la fórmula para hallar $Var[e_N(h)]$ viene dada por el modelo.

En realidad, dependiendo del método de predicción, de la estrategia y del contexto del problema, la forma de calcular los intervalos de predicción puede variar bastante. Chatfield (2001a) presenta una muy buena revisión sobre el tema en la que aborda cómo calcular los intervalos dependiendo de la situación. A continuación, se muestra de forma resumida un esquema con las distintas alternativas

- Si se conoce el modelo probabilístico que ha generado la serie temporal y se obtienen predicciones óptimas que minimizan el Error Cuadrático Medio, entonces es posible hallar $Var[e_N(h)]$ y emplear la fórmula (3.2). Este es el caso de los modelos ARIMA para los cuales hay fórmulas desarrolladas para estimar los intervalos de predicción.

- Si se asume, aunque no se comprueba, que el método empleado es el correcto. El lector interesado puede encontrar una discusión sobre el tema, especialmente para el caso de los alisados exponenciales, en el trabajo de Chatfield (2001a).

- Si no se conoce el modelo o no hay fórmulas teóricas disponibles

  - Se pueden usar fórmulas *aproximadas* y genéricas para distintos modelos. Tienen la ventaja de que son sencillas de utilizar pero el inconveniente de que son bastante imprecisas, por lo que no deben ser una opción a considerar.
  - Los intervalos de predicción se pueden calcular mediante enfoques empíricos que suponen una gran carga computacional y que se basan en las propiedades de las distribuciones observadas de los errores, siguiendo enfoques como el de Williams y Goodman (1971).
  - Los intervalos de predicción se pueden hallar mediante métodos de remuestreo o simulación, especialmente para series no demasiado largas, ya que son métodos que requieren mucha carga computacional. La simulación requiere considerar que el modelo considerado es el correcto y, a partir de dicho modelo, generar datos pasados y futuros para estudiar las propiedades de los intervalos de predicción sobre dichos datos. El remuestreo, en lugar de muestrear las innovaciones de una distribución paramétrica que se asume como cierta, muestrea de la distribución empírica de los errores pasados ya ajustados, y aproxima la distribución teórica de las innovaciones mediante la distribución empírica de los residuos observados.
  - Dado un modelo adecuado, el enfoque Bayesiano permite calcular la distribución completa de un valor futuro, a partir de la cual se pueden obtener los intervalos de predicción.
  - También se puede estimar el intervalo de predicción mediante juicios. Según Webby y O'Connor (1996), los intervalos de predicción hallados con este método suelen ser a menudo demasiado estrechos, lo que denota un exceso de confianza en el método utilizado. Según estos autores, hasta el momento, los resultados en este área no son demasiado esperanzadores.

Densidad de predicción. Los intervalos de predicción simplemente acotan la predicción, pero no informan sobre en qué parte del intervalo es más probable que se encuentre la predicción. Para eso se emplean las densidades de predicción. La densidad de una predicción proporciona una descripción completa de la distribución de probabilidad de los posibles valores de la predicción. En otras palabras, es una estimación de la variable aleatoria que dará lugar a la predicción. La predicción de la densidad subsume al intervalo de predicción, ya que una vez que se conoce la densidad de una predicción, el proporcionar intervalos de predicción con distinta probabilidad asociada resulta inmediato.

Tay y Wallis (2000) revisan las aplicaciones de las densidades de predicción en los campos de la macroeconomía y las finanzas. En macroeconomía, las densidades de las predicciones se emplean para dar estimaciones más completas de magnitudes cuya predicción es vital, como la inflación; y en finanzas para caracterizar la incertidumbre asociada a predicciones del rendimiento de una acción o de una cartera de acciones, donde la densidad es habitualmente *no normal* debido a la falta de simetría o al exceso de curtosis.

En los modelos lineales con innovaciones distribuidas según una normal, se suele considerar que la densidad de la predicción es una normal que tiene como media la estimación puntual y como varianza la varianza del error de predicción. Sin embargo, eso obedece a una situación que en los contextos financieros no suele darse. En esos casos donde la distribución de los errores de predicción no sigue una normal, la estimación de la densidad de predicción es más complicada. En la literatura sobre el tema hay una gran cantidad de métodos propuestos para estimar densidades que no tienen necesariamente por qué ser normales. Una alternativa es, por ejemplo, utilizar algún enfoque que realice una mixturas de Gaussianas, como los descritos en Weigend y Shi (2000), o mediante métodos empíricos que siguen un enfoque no paramétrico, técnicas de simulación, aproximaciones Bayesianas, etc. En el volumen 19 del Journal of Forecasting, el número 4 está dedicado a la predicción de densidades en economía y finanzas (Timmermann, 2000). Dentro de ese número pueden verse muy distintos modelos y un artículo de revisión sobre el tema a nivel global (Tay y Wallis, 2000).

Un método no paramétrico de predicción de densidades. Una manera obvia de obtener una distribución futura de la predicción, es estimar la distribución de todos los valores históricos de la serie (mediante un histograma o mediante un estimador más sofisticado) y asumir que dicha distribución es constante, y que es útil para caracterizar el futuro. Sin embargo, es una forma muy tosca de estimar la distribución, ya que la densidad pronosticada apenas cambia con el tiempo o lo hace muy lentamente.

Pasley y Austin (2004) proponen un método más sofisticado que, como se verá más adelante, se ha considerado interesante para la tesis. Se trata

de un método no paramétrico para predecir la densidad de un valor futuro, que utiliza la idea de usar valores históricos para estimar la densidad futura, pero que la refina ya que genera la densidad de la predicción utilizando sólo un subconjunto de los valores históricos que el método considera adecuados.

En líneas generales, este método busca en el pasado de la serie fragmentos o secuencias de la propia serie (i.e. observaciones consecutivas) que se parezcan a la secuencia actual. Una vez que identifica las secuencias similares a la actual utiliza el siguiente valor de cada una de ellas para construir un histograma que presenta como predicción de la densidad del futuro valor de la secuencia actual. Para agilizar el proceso, el método consta de una fase de aprendizaje en la que agrupa en clusters secuencias de la serie similares entre sí, y una fase de predicción propiamente dicha en la que se buscan los clusters más similares a la secuencia actual. Para pasar de la predicción en forma de distribución a la predicción puntual, aconsejan tomar la media o la moda de la predicción en forma de distribución.

Como se puede ver el método que proponen es una variante del algoritmo de Farmer-Sidorowich (Farmer y Sidorowich, 1987) o de la aplicación del algoritmo de k-NN en series temporales (Yakowitz, 1987). El método resulta adecuado para contextos como el financiero donde se dispone de series temporales de alta frecuencia que, además, son el resultado de procesos caóticos o no-lineales.

3.5.1.2. Las series temporales multivariantes

Las series temporales multivariantes tienen como objetivo estudiar la evolución de un conjunto de dos o más series temporales y sus interrelaciones a lo largo del tiempo. Como en el caso univariante, existen dos caminos para predecir de series temporales multivariantes: utilizar un método *ad hoc* que no necesite la formulación explícita de un modelo, o proponer un modelo que recoja las interacciones entre el conjunto de series considerado y utilizar dicho modelo para generar predicciones. Entre las aproximaciones *ad hoc* que permiten predecir un conjunto de series temporales sin asumir un modelo, puede citarse la técnica de los k-vecinos simultáneos utilizada en Fernández-Rodríguez, Sosvilla-Rivero y Andrada-Félix (1999). Sin embargo, la mayoría de las aproximaciones existentes para predecir series temporales multivariantes provienen del terreno de la econometría y, en consecuencia, consideran de forma explícita un modelo.

El proceso de modelado multivariante se complica notablemente ya que debe recoger la dependencia serial de las series estudiadas y las posibles interdependencias que puedan existir entre ellas. En economía, donde se suelen aplicar estos modelos, suelen existir teorías que relacionan un conjunto de variables, por lo que lo natural es considerarlas todas en el modelo. Por ejemplo, el consumo doméstico está íntimamente ligado con variables como los ingresos, los tipos de interés y el gasto en productos de inversión.

Los modelos multivariantes tienen dos objetivos principales:

1. Obtener predicciones más precisas que las que se obtienen mediante técnicas univariantes.
2. Conocer en profundidad la estructura subyacente del fenómeno que se quiere estudiar.

Sin embargo, el primer objetivo no se cumple siempre. Los modelos multivariantes suelen obtener mejores ajustes que los modelos univariantes, pero son varias las razones por las que las predicciones fuera de los datos empleados en la estimación empeoran a menudo:

- Al haber más parámetros a estimar, hay más posibilidades de que la variación muestral incremente la incertidumbre en los parámetros y que ello afecte a las predicciones.
- Al considerarse más variables, existen más posibilidades de que haya errores de medida o valores atípicos.
- Los datos multivariantes observados pueden no ser adecuados para ajustar un modelo multivariante.
- El cálculo de predicciones de la variable dependiente a menudo requiere los valores futuros de las variables explicativas (los cuales no están disponibles en el momento de hacer la predicción). Si para este propósito se emplean valores de las variables explicativas se empeora inevitablemente la calidad de la predicción.
- Los modelos multivariantes deben estar bien especificados y son más vulnerables a las malas especificaciones y a los cambios estructurales. A esto hay que añadir que el modelado multivariante es notablemente más complejo que el univariante. A menudo existe el riesgo de incluir variables que dan un mejor ajuste (pero que empeoran la predicción) y de omitir variables realmente relevantes.

El objetivo debe ser siempre determinar un modelo parsimonioso que considere todas las variables importantes. Sin embargo, no es un objetivo trivial. Los modelos que trabajan con series temporales multivariantes pueden dividirse en los siguientes grupos:

- Los modelos de una sola ecuación.
    - Modelos de regresión múltiple: muy empleados en la práctica porque permiten trabajar con datos que no son necesariamente series temporales. Sin embargo, se desaconsejan para modelar series temporales (Chatfield, 2001c). Las razón principal es que un modelo de regresión no tiene en cuenta que las observaciones de las

variables están ordenadas en el tiempo, por lo que hay que hacer una serie de modificaciones sobre él para que pueda trabajar correctamente con series temporales.

- Modelos de función de transferencia o modelos de regresión dinámica: es una clase de modelos de regresión más general y más adecuada para predecir una serie temporal. La serie de interés se ve afectada por los valores de otra serie explicativa pero no al revés (existe causalidad, pero en un sólo sentido, sin retroalimentación).

- Modelos vectoriales autorregresivos: Son modelos en los que se considera un conjunto de variables que están interrelacionadas entre sí. Puede darse el caso de que haya variables que se retroalimenten unas a otras (a las que se les llama endógenas) y otras que afecten a las variables endógenas pero que no se vean afectadas por ellas (estas variables reciben el nombre de exógenas). Lütkepohl (2005) trata en profundidad este tipo de modelos y distingue los siguientes tipos.

  - Modelos VAR: que sólo tienen componente autorregresiva.
  - Modelos VARMA: con componente autorregresiva y de media móvil.
  - Modelos VECM: que se usan cuando las series consideradas están cointegradas, es decir, no son estacionarias pero una combinación lineal entre ellas sí da lugar a una serie estacionaria. Los modelos de corrección del error (ECM) y su extensión vectorial (VECM) permiten modelar relaciones de cointegración entre variables.
  - Modelos ARCH y GARCH multivariantes: Constituyen una extensión de los modelos heterocedásticos condicionales autorregresivos propuestos por Engle (1982) para el caso univariante.

- Otros modelos: Chatfield (2001c) cita además otras familias de modelos que tratan series temporales multivariantes entre las que se incluyen: los modelos econométricos (similares a los VAR, pero que pretenden reflejar una teoría econométrica y que contienen típicamente más ecuaciones), los modelos espacio-estado multivariantes (extensión de los modelos espacio-estado univariantes que añade variables explicativas en la parte derecha de la ecuación), el análisis de intervención (para recoger en los modelos la acción de eventos aislados que van a ocurrir en un determinado instante temporal) y los modelos multivariantes no lineales (como por ejemplo, generalizaciones de los modelos de umbral para el caso multivariante).

3.5.1.3. Agregación de series temporales

En predicción, y más especialmente en el área de economía, es habitual encontrar variables que son observadas a lo largo del tiempo en una serie de

individuos o regiones y donde interesa conocer el comportamiento agregado y no el individual. En este caso se habla de agregación contemporánea. También existe la agregación temporal que se da cuando el interés reside no en saber el valor de una magnitud en un determinado instante, sino su valor agregado en el tiempo, e.g., el total de ventas anuales, en lugar de las ventas mensuales.

Si el número de valores a agregar es muy alto, la agregación es la forma razonable de abordar el problema. El valor agregado que se utiliza para resumir la información puede ser una media, un total, un valor muestreado de entre los valores observados, etc. Evidentemente, en el proceso de agregación se pierde información, pero si se opta por agregar, se supone que es porque la pérdida de información no es relevante o, al menos, es asumible.

A la hora de obtener una predicción de la magnitud agregada, puede que exista la posibilidad de, o bien modelar los datos de forma desagregada, hacer las predicciones y luego agregar, o bien agregar primero, y, a continuación, predecir y modelar la serie agregada. Ante estas dos posibilidades surge la duda de cuál de los dos enfoques resulta más adecuado. La resolución de estas cuestiones depende de si la agregación es temporal o contemporánea.

Agregación temporal. Dada una serie temporal $\{y_t\}$, con $t = 1, ..., n$, la definición general de una serie temporal agregada temporalmente cada k instantes, $\{y^*_{t'}\}$, con $t' = 0, k, 2k, 3k..., n/k$, es la siguiente :

$$y^*_{t'} = \sum_{j=0}^{k-1} \omega_j y_{t-j}. \tag{3.3}$$

calculándose el valor $y^*_{t'}$ en los instantes de $t = 1, ..., n$ que son múltiplos de k. Esta fórmula es simplemente una combinación lineal que recoge las formas clásicas de agregación temporal como los totales, las medias aritméticas o las medias ponderadas, pero también otras tales como el muestreo de la serie cada k periodos (aunque el muestreo no suela considerarse como agregación temporal).

Las características de los procesos generadores de los datos agregados y de los datos desagregados no tienen por qué ser las mismas, como resulta obvio. Brannas y Ohlsson (1999) muestran un ejemplo basado en series temporales de desempleo donde las no-linealidades que aparecen en la serie de frecuencia mensual desaparecen al agregar la serie utilizando la media para construir series trimestrales y anuales. Rossana y Seater (1995) apunta que en las series económicas agregadas anualmente se pierde la variación relativa al ciclo de negocio que sí aparece en series desagregadas de frecuencia mensual, por lo que, según los autores, no son útiles para estudiar los ciclos económicos subyacentes. Una parte de la literatura dedicada a la agregación ha tratado de dilucidar la relación existente entre el proceso generador de la serie agregada

y el de la desagregada. Silvestrini y Veredas (2005) presentan un estudio a fondo del comportamiento de la agregación temporal para procesos generados a partir de modelos ARIMA-GARCH y muestran cómo derivar el modelo con los datos agregados a partir del modelo desagregado.

Agregación contemporánea. Una de las cuestiones centrales dentro de este área trata de dilucidar si es mejor predecir las series temporales desagregadas y agregar la predicción, o agregar las series temporales desagregadas y a continuación predecir. Granger (1990) presenta una revisión sobre el tema. La literatura teórica muestra que, si el proceso generador de datos es conocido, la predicción de los datos desagregados supera a la predicción que trabaja directamente sobre los datos agregados. Si el proceso generador no es conocido y el modelo debe ser estimado, los resultados dependen del propio proceso, aunque parece más razonable modelar directamente los datos agregados. En la práctica el proceso generador de datos es (normalmente) desconocido por lo que la cuestión permanece abierta.

Todo hace pensar que la respuesta depende del grado de interdependencia existente entre las variables consideradas. Una parte de la literatura trata el caso en que dicha interdependencia sea debida a un factor común. Las conclusiones en ese caso apuntan a que lo mejor es trabajar con las series desagregadas pero teniendo en cuenta el factor común (Granger, 1987; Zellner y Tobias, 2000). Otra parte de la literatura (Lütkepohl, 1987) estudia la interdependencia desde el punto de los modelos VARMA, lo que permite el tratamiento de la dependencia temporal y de sección cruzada. Sin embargo, a medida que el número de series a considerar aumenta, los modelos se hacen más difíciles de estimar (la llamada "*maldición de la dimensionalidad*") y esto afecta inexorablemente a la precisión de las predicciones. Para evitar este problema en datos geográficos, Giacomini y Granger (2004) proponen el concepto de autocorrelación espacial, que surge cuando las observaciones en una región están correlacionadas con las regiones vecinas pero no con las regiones más lejanas. En este trabajo, se propone un modelo autorregresivo para los datos desagregados al cual imponen una matriz de autocorrelación espacial. Los resultados muestran que este modelo mejora los resultados obtenidos agregando o ignorando la autocorrelación espacial.

Por otro lado, el trabajo de Hendry y Hubrich (2006) estudia si se mejoran las predicciones de la serie agregada al incluir datos desagregados en el modelo construido para predecir la serie agregada, respecto a modelos sin esos datos o a modelos desagregados. Las conclusiones que obtienen indican que sí existe una mejora en la precisión de las predicciones, aunque no en todos los casos.

3.5.2. Aproximaciones al margen de las series temporales clásicas

3.5.2.1. Las series temporales valoradas mediante alfabetos de símbolos

El objetivo de este área es analizar los patrones que pueden aparecer en series temporales que son muy complejas o de un tamaño abrumador. Para ello, lo que hace es traducir estas series a una secuencia de símbolos y analizar posteriormente dicha secuencia. La secuencia de símbolos representa de una manera sencilla e ilustrativa la serie original, es más sencillas de tratar a nivel computacional que la propia serie original, y es poco sensible a los errores de medida en la serie original. Daw et al. (2003) ofrecen una revisión muy interesante del estado del arte de la materia.

El proceso de simbolización o de definición de los símbolos tiene sus raíces en la teoría de la información y puede realizarse de varias maneras. Una forma simple consiste en dividir el rango de las observaciones en un número finito de regiones y asociar a cada región un símbolo concreto. También se puede trabajar con el rango de alguna transformada de los datos originales, como la de los valores diferenciados. Esto es útil en casos en los que la serie no es estacionaria o es más interesante trabajar con la serie de los cambios en el tiempo que con los valores de la propia serie. El alfabeto de símbolos puede tener sólo dos elementos. A medida que el tamaño aumenta, las probabilidades de que el ruido afecte al proceso de simbolización también crecen. Este proceso requiere también la definición de la localización de las particiones. Al tratar datos experimentales, esto supone una complejidad añadida ya que tanto el proceso generador de la serie, como el nivel de ruido son desconocidos. Entre las opciones *ad hoc* se encuentra coger el punto central de la partición, la media o la mediana de los datos; también se pueden generar particiones de igual tamaño o de igual probabilidad, etc. En otros casos, las particiones vienen fijadas por el propio problema, como en algunos sistemas químicos donde hay umbrales predefinidos. En cualquier caso, es recomendable analizar la sensibilidad de los resultados ante distintas particiones.

Otro enfoque diferente para el proceso de simbolización consiste en particionar el espacio de fases. En este caso los símbolos representan distintas regiones o un subconjunto de dicho espacio. Las secuencias observadas de símbolos son segmentos de trayectorias que enlazan regiones separadas del espacio de fases.

Hébrail y Hugueney (2001) plantean otra manera de realizar el proceso de simbolización. En ella, los símbolos representan distintos episodios de la serie temporal, siendo un episodio una secuencia de valores consecutivos de dicha serie. Para hallar el alfabeto de símbolos que representará la serie, realizan un *clustering* de los episodios de igual longitud de toda la serie mediante

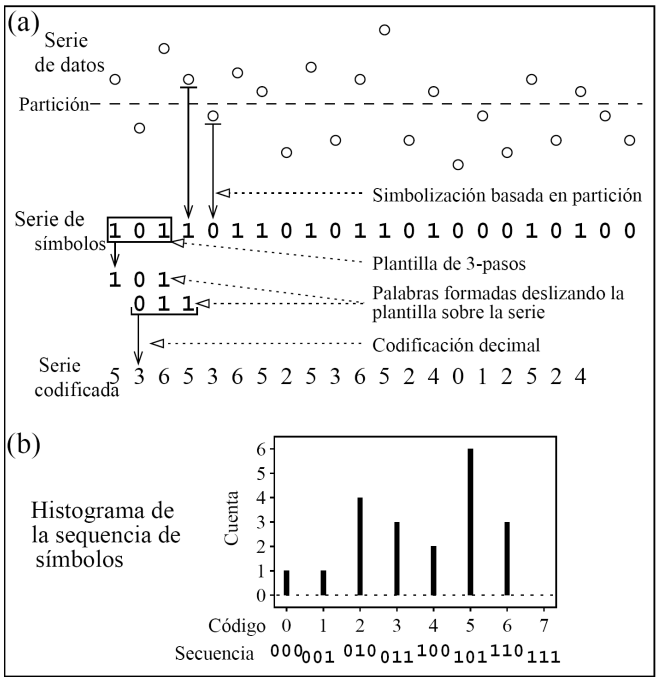


Figura 3.5: Arriba: Proceso de simbolización (a). Abajo: Histograma de la secuencia simbólica (b). (Daw et al., 2003)

un mapa auto-organizado (Kohonen, 1995). En este caso, la construcción del alfabeto depende en gran medida de la serie que se esté analizando con lo que se pretende captar mejor su estructura subyacente. Con esta técnica reducen una serie de consumo eléctrico de cerca de nueve mil datos a una serie de veintiseis símbolos con un alfabeto de cardinalidad tres.

Tras el proceso de simbolización, se ha de definir la secuencia de símbolos, agrupando símbolos consecutivos en orden temporal. Esto se hace mediante una ventana deslizante, tal y como muestra la figura 3.5. A cada secuencia posible se le representa mediante un identificador.

Una vez obtenida la secuencia de símbolos, ésta debe ser analizada. Dicho análisis entronca con el campo de la dinámica simbólica y con los modelos de Markov. Puede encontrarse más información sobre estos temas en el libro de Kitchens (1998). Además del siempre necesario análisis visual de la serie de símbolos, se pueden utilizar herramientas como el histograma (ver figura 3.5.b o aplicar técnicas de análisis basadas, o bien en la estadística (e.g. la norma euclídea y el estadístico de chi-cuadrado), o bien en la teoría de la información (e.g. la medida de entropía de Shannon). Con las primeras se pueden observar las diferencias entre dos secuencias de histogramas y con las segundas estudiar la complejidad de las mismas.

La secuencia de símbolos también puede ser transformada en una má-

quina de estados finita o en una cadena de Markov finita y ser analizada. Por ejemplo, Keogh, Lonardi y Chiu (2002) proponen un algoritmo que tiene como objetivo descubrir patrones anómalos o *sorprendentes* en una base de datos de series temporales. Para ello, emplean series temporales simbólicas y modelos de Markov.

En general, las series temporales simbólicas se emplean cuando la estructura dinámica de la serie original no puede ser captada satisfactoriamente por funciones senoidales o mediante herramientas como las transformadas de Fourier. Sus aplicaciones pueden encontrarse en campos aparentemente tan dispares como la astrofísica, la geofísica, biología, finanzas, medicina, dinámica de fluidos, química, sistemas mecánicos, sistemas de control y monitorización en tiempo real, etc. Por ejemplo, Brida y Punzo (2003) muestran una aplicación de las series temporales simbólicas para analizar la evolución del crecimiento económico en las cuatro macro-regiones italianas.

3.5.2.2. La predicción basada en gráficos de velas

Tal y como se indican Engle y Russell (2009), las series temporales financieras poseen unas características muy particulares que hacen que sean difíciles de tratar: ingente cantidad de datos, la frecuencia con la que se producen los valores no es constante a lo largo de las sesiones, existen patrones intra-diarios, etc. Por ello, para facilitar su visualización y su comprensión muchas veces se utilizan simplemente las series de los cierres diarios, semanales o mensuales, con la consiguiente pérdida de información de valores intra-diarios, intra-semanales o intra-mensuales, respectivamente.

Los gráficos de velas o *candlesticks* mitigan esta pérdida y ofrecen más información que el valor de cierre del bien que se esté analizando. Más concretamente, proporcionan los valores de apertura y de cierre y mínimo y máximo del periodo considerado (día, semana o mes). Su representación gráfica ofrece esta información de una manera muy clara, tal y como muestra la ilustración 3.6. La representación más habitual del *candlestick* es la 3.6.d en la que el intervalo entre los valores de apertura y de cierre es una caja blanca (o verde) si $cierre > apertura$, y negra (o roja) si $apertura > cierre$.

Los *candlesticks* son una herramienta básica a la hora de realizar un análisis técnico sobre la evolución de un bien financiero. Son gráficos de fácil lectura que reflejan bien la psicología del mercado y la variabilidad del bien financiero considerado. Existe toda una teoría sobre los *candlesticks* que se dedica a identificar las figuras más típicas y a ayudar a la toma de decisiones de compra-venta basándose en el estudio de la secuencia de *candlesticks* (Morris, 2006). Sin embargo, los trabajos que se dedican al análisis de series de *candlesticks* usando técnicas estadísticas o de *soft-computing* son, por el momento, escasos.

Lee y Jo (1999) desarrolla un sistema que recoge mediante un conjunto de reglas el conocimiento experto sobre los *candlesticks*. Los datos reales son

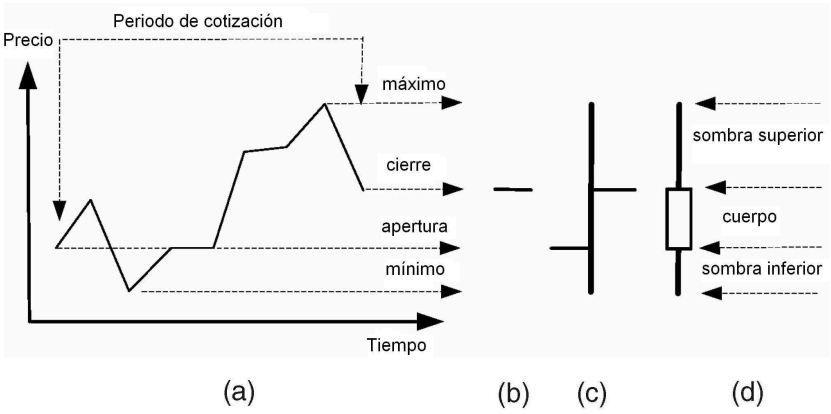


Figura 3.6: a) Fluctuación durante el periodo de cotización. b) Valor de cierre. c) *Candlestick* esquemático. d) *Candlestick* con caja. (Lee et al., 2006)

analizados a la luz de la base de conocimiento (del conjunto de reglas) con el fin de extraer órdenes concretas de compra-venta. Este es el primer trabajo que conocido que analiza a los *candlesticks* de forma rigurosa, sin embargo, no lo hace desde la perspectiva de las series temporales.

Fiess y MacDonald (2002) analizan si el análisis técnico basado en los valores del *candlestick*, es decir, en los valores de apertura, cierre, máximo y mínimo, tiene alguna fundamentación desde el punto de vista econométrico. En su artículo, estos autores analizan las propiedades de dichas series temporales para el caso del cambio de divisas y concluyen que estas series ofrecen información relevante a la hora de predecir los valores de cierre. En el cambio de divisas, el cierre de una sesión coincide con la apertura de la siguiente porque las divisas cotizan en distintos lugares alrededor del globo durante las 24 horas del día. Por esta razón, los autores descartan analizar la serie de los valores de apertura y de cierre y emplean solo los de cierre. Para modelar la dinámica y las interrelaciones entre las tres series restantes aplican una serie de técnicas multivariante con el fin de obtener información clave sobre la volatilidad y el nivel de las series consideradas. Según el artículo, existe una relación estructural estable entre las tres series y hay evidencia de causalidad de Granger, por lo que al considerarlas es posible obtener mejores predicciones de volatilidad y de los niveles de la serie, al menos en el caso analizado.

Lee et al. (2006) representan las series de *candlesticks* como series temporales borrosas. Para ello caracterizan la longitud de las *sombras superior e inferior* y del *cuerpo* del *candlestick* (ver ilustración 3.6) mediante un número borroso. También caracterizan con números borrosos la posición de los valores de apertura y cierre del *candlestick* actual en relación con el *candlestick* anterior (es decir, si los valores actuales están en la parte de la sombra inferior o del cuerpo o de la sombra superior del *candlestick* anterior) y la

tendencia que llevaba la serie. La variable a predecir no es el *candlestick* de la próxima sesión, sino la variación en % entre el precio de cierre de la sesión actual y el de n sesiones futuras. Para realizar la predicción de la variación en % borrosifican también la variable de salida. El algoritmo de predicción, en primer lugar, refina el conjunto de variables de entrada para quedarse sólo con las más relevantes a la hora de determinar la variación futura. Esto lo hace con ayuda de un árbol de decisión ID3. La predicción se realiza con ayuda de la regla de Bayes para determinar, dado el patron actual, cuál es su variación futura más probable de acuerdo con los valores históricos. En el artículo también se realiza una comparativa utilizando un conjunto de datos de referencia entre el método propuesto y otros métodos de predicción de series temporales borrosas de la literatura. Según los resultados que se obtienen, el método propuesto queda en muy buen lugar.

Otro trabajo de predicción basada en *candlesticks* es el de Izumi, Yamaguchi, Mabu, Hirasawa y Hu (2006). En este trabajo, los autores utilizan los *candlesticks* para representar la evolución del comportamiento diario de un bien financiero a lo largo del tiempo y emplean programación de redes genéticas para modelar estrategias de compra-venta de dicho bien. Las redes que utilizan son grafos dirigidos donde existen dos tipos de nodos: de juicio y de procesamiento. Los nodos de juicio representan seis posibles tipos de estado según la posición del *candlestick* actual respecto del inmediatamente anterior, prestando atención a dos factores: el color del cuerpo del *candlestick* y si existe un salto o ventana entre ellos (es decir, si no existe solapamiento entre ellos). Los momentos en los que se produce un salto son los momentos en los que, según este enfoque, se deben dar las órdenes de compra-venta. Los nodos de procesamiento modelan las decisiones de compra y de venta. La transición entre nodos se realiza de un día a otro. Una sucesión de nodos de juicio finalizados por un nodo de procesamiento representan una estrategia de compra-venta. La configuración final de la red deberá determinar la dinámica que han seguido las acciones. Durante el periodo de entrenamiento, la configuración de la red se optimiza para obtener el máximo de beneficios. El método propuesto es comparado con otros métodos utilizados para predecir series temporales financieras como son los perceptrones multicapa y las redes neuronales de ramas múltiples (*multi-branch*). La comparación consiste en medir los beneficios obtenidos por cada uno de los modelos. Los resultados indican que el método propuesto supera habitualmente al perceptrón multicapa y que se comporta de forma similar a la red neuronal de múltiples ramas.

En el trabajo de Torkamani, Asgari y Lucas (2006) se estudian las dinámicas de las series temporales de *candlesticks* desde la perspectiva de los sistemas caóticos. Según la teoría del caos, una serie aparentemente aleatoria puede estar gobernada por un sistema caótico no lineal relativamente simple y, por tanto, determinista. Torkamani et al. (2006) adaptan el método

de Grassberger-Procaccia (Grassberger y Procaccia, 1983) para estimar la dimensión caótica de una serie a las series de *candlesticks*. Dicho trabajo muestra que, en las tasas de cambio de divisas, la series de *candlesticks* tienen una dimensión caótica menor que las series de los valores de cierre. Esto quiere decir que el sistema dinámico que rige las series de *candlesticks* es menos complejo que el que está tras las series de los valores de cierre.

Como se puede ver, el interés por el análisis estadístico y la aplicación de técnicas de *soft-computing* sobre los gráficos *candlesticks* está despertando en los últimos años. Los resultados son prometedores y muestran un ejemplo práctico de series temporales que van más allá de la valor puntual, donde la información adicional puede conducir a obtener predicciones que aporten más información y que permitan tomar mejores decisiones.

3.6. Relación de las series temporales simbólicas con las otras aproximaciones que van más allá del valor puntual

Tal y como ha mostrado el apartado anterior, existen otras aproximaciones que, de una manera o de otra, van más allá de lasrepresentaciones que ofrecen las series temporales clásicas. Estas aproximaciones son, en general, notablemente diferentes a las series temporales valoradas mediante variables simbólicas. Sin embargo, existen algunas interrelaciones entre ambos enfoques que son dignos de mención.

Las predicciones de intervalo y de densidad vienen a reflejar la incertidumbre que existe en torno la predicción dada en forma de valor puntual. Pero, ¿y si existe imprecisión en los propios valores de la serie temporal? Esto puede suceder en casos en los que, por ejemplo, el instrumento de medida sea impreciso y que dicha imprecisión afecte a cada valor recogido, o si para cada valor de la serie se toman varias medidas que no coinciden. En estos casos, las series temporales valoradas mediante variables simbólicas permitirían representar esa imprecisión o incertidumbre.

Dentro, de los métodos para predecir la densidad futura, el trabajo de Pasley y Austin (2004) presenta ciertas analogías con el enfoque de esta tesis. En primer lugar, considera los histogramas como herramienta para representar las distribuciones de predicción y los histogramas son un tipo de variable simbólica. En segundo lugar, su técnica (similar al k-NN) trocea la serie en segmentos no tienen por qué corresponder con unidades de tiempo. Sin embargo, en esta tesis los segmentos corresponden a unidades de tiempo (ver el item 2 de la lista de la sección 3.4). La principal diferencia radica en que el objetivo del método de Pasley y Austin es caracterizar la incertidumbre de la predicción mediante una densidad y no predecir la densidad real de la magnitud considerada en el instante siguiente.

Por su parte, las series temporales multivariantes permiten reflejar la evolución de un conjunto de variables de interés. Puede darse el caso de que esas variables representen una misma magnitud, pero medida sobre distintos sujetos y que el interés residiese en predecir el comportamiento del conjunto o de un agregado. Tal y como muestran Giacomini y Granger (2004), este problema tienen dos respuestas dentro del área de las series temporales clásicas: modelar las series individuales y luego agregar o hacerlo a la inversa. El resultado de la agregación puede ser un total, una media o una mediana, típicamente. Estos valores pueden ofrecer información muy valiosa, pero, como es bien conocido, también pueden enmascarar características importantes del conjunto de datos considerado. En este sentido, si la agregación se realiza utilizando variables simbólicas, como el intervalo del rango o del recorrido intercuartílico, o como el histograma o el gráfico de cajas de los datos, la información agregada que se obtiene puede ser notablemente más completa. Esta agregación fue comentada en el punto 1 de la lista de la sección 3.4.

Otro punto de conexión entre las series temporales valoradas mediante variables simbólicas y las series multivariantes es que algunos métodos que se utilizan para predecir las series multivarantes sirven para predecir a las series temporales simbólicas, como las series temporales de intervalos y las de gráficos de cajas. Estas series pueden ser descompuestas en un conjunto de series multivariantes: las series temporales de intervalos se pueden descomponer en las series de los mínimos y de los máximos, y las series temporales de gráficos de cajas se pueden descomponer en cinco series temporales clásicas, una por cada cuartila considerada. En estos dos supuestos, la descomposición da lugar a series clásicas que presentan una clara interrelación (p.ej. la serie de los mínimos siempre es menor que la serie de los máximos) por lo que parece lógico considerar su predicción mediante modelos multivariantes como los modelos VAR.

Las series temporales basadas en alfabetos de símbolos comparten con las series temporales basadas en variables simbólicas un afán por simplificar la serie temporal original y representarla de una manera más tratable que conserve la información relevante. Sin embargo, la forma de construir las series en ambos casos es notablemente diferente. Una excepción a este hecho la constituye el proceso de simbolización de Hébrail y Hugueney (2001) que divide la serie original en segmentos y representa dichos segmentos mediante símbolos. Este procedimiento es muy similar a la agregación temporal que se realiza para construir series temporales basadas en variables simbólicas (ver el punto 1 de la lista de la sección 3.4).

El uso de gráficos de vela en economía supone un proceso de simbolización en el que se pretende simplificar el contenido intradiario (o intra-'periodo de interés') y sustituirlo por un símbolo. En este caso el símbolo no es arbitrario, sino que tiene información cuantitativa (los valores de apertura, de cierre, mínimo y máximo) que está expresada en la misma escala que los

datos originales. Los valores numéricos determinan la forma del símbolo. La conexión con las series temporales valoradas mediante variables simbólicas es evidente, ya que éstas también permiten resumir contenido intradiario (o intraperiodo) y hacerlo, por ejemplo, mediante un intervalo o mediante un histograma. Realmente el gráfico de velas está compuesto por dos intervalos: el que forman el mínimo y el máximo, y el que delimitan la apertura y el cierre (éste último tiene un sentido asociado según la apertura sea mayor que el cierre o viceversa). Este hecho hace que los gráficos de velas puedan considerarse como un tipo de dato simbólico que sirve para resumir información financiera.

3.7. Métodos para el análisis temporal de datos simbólicos

Hasta el momento, si descontamos los métodos que se van a proponer en esta tesis, han sido pocos los métodos desarrollados para predecir series temporales valoradas mediante una variable simbólica. Las primeras propuestas que se han realizado se centran en las series temporales de intervalo, lo cual resulta natural ya que los intervalos son notablemente más sencillos que los histogramas.

3.7.1. Métodos para las series temporales de intervalos

En los últimos años, han surgido algunas propuestas que adaptan métodos clásicos a la predicción de series temporales de intervalo. Dichas propuestas, que serán descritas en este apartado, son un modelo ARMA y un modelo híbrido que combina el modelo ARMA con una red neuronal artificial (RNA).

Por otro lado, en finanzas los intervalos de los valores mínimo y máximo de las cotizaciones son registrados periódicamente y son publicados en la prensa y en los sitios web especializados. Pese a este hecho, no abundan los trabajos académicos que investigan estos intervalos. En este apartado se reseñarán también los trabajos más significativos en el área.

3.7.1.1. Una extensión de los modelos ARMA a las series temporales de intervalos

Teles y Brito (2005) proponen una extensión de los modelos ARMA para trabajar con series temporales de intervalos. Ellos asumen que los procesos que generan los mínimos y los máximos de los intervalos tienen los mismos parámetros pero diferentes medias, siendo la media de la ecuación de los máximos mayor que la media de la ecuación de los mínimos. En su modelo también se asume que las series de los residuos de los dos procesos gene-

radores (el de los mínimos y el de los máximos) son ruido blanco y son independientes y con distinta varianza. Los autores consideran además dos enfoques: en el primero de ellos utilizan todas las observaciones de los mínimos y de los máximos del intervalo doblando el tamaño muestral usado para la estimación; en el segundo muestran que la serie temporal de la diferencia entre el mínimo y el máximo (i.e. la serie temporal de los rangos de los intervalos) sigue otro proceso ARMA de igual orden y parámetros.

Según los autores, los resultados obtenidos indican que los procedimientos propuestos funcionan correctamente tanto en términos de estimación del modelo, como en términos de precisión en la predicción. En cuanto a la estimación, el método que maneja los mínimos y los máximos supera ampliamente al otro enfoque. Sin embargo, en cuanto a precisión en la predicción, ambos métodos obtienen resultados muy similares.

3.7.1.2. Una modelo híbrido ARMA+RNA para la predicción de series temporales de intervalos

En Maia et al. (2006a) y Maia, de Carvalho y Ludermir (2006b), los autores proponen una extensión de los modelos ARMA y de los modelos híbridos para predecir series temporales de intervalo.

La extensión de los modelos ARMA es distinta a la propuesta por Teles y Brito (2005) ya que modelan las series temporales de los centros y de los radios de los intervalos, en lugar de los mínimos y los máximos. El modelo híbrido que proponen se basa en el modelo desarrollado por Zhang (2003) para series temporales clásicas. Dicho modelo combina un modelo ARMA y un perceptrón multicapa para obtener las predicciones. Con ellos, el modelo híbrido pretende recoger los comportamientos lineal y no lineal del proceso generador de la serie. El modelo híbrido considera que la serie temporal se descompone en una componente lineal L_t y en otra no lineal N_t de forma que $\{X_t\} = \{L_t + N_t\}$.

Tanto para el modelo ARMA, como para el modelo híbrido, los autores trabajan con las series de los centros $\{X_{t,C}\}$ y de los radios $\{X_{t,R}\}$ de los intervalos. Además, las modelan de manera independiente sin imponer ninguna restricción adicional sobre el modelo.

El modelo ARMA que proponen consiste en

$$X_{t,C} = \phi_{0,C}+\phi_{1,C}X_{t-1,C}+...+\phi_{p,C}X_{t-p,C}+Z_{t,C}-\theta_{1,C}Z_{t-1,C}-...-\theta_{q,C}Z_{t-q,C} \tag{3.4}$$

$$X_{t,R} = \phi_{0,R}+\phi_{1,R}X_{t-1,R}+...+\phi_{p,R}X_{t-p,R}+Z_{t,R}-\theta_{1,R}Z_{t-1,R}-...-\theta_{q,R}Z_{t-q,R} \tag{3.5}$$

donde $\{Z_{t,C}\}$ y $\{Z_{t,R}\}$ son procesos aleatorios con media cero y varianza constante. A partir de los valores del centro y del radio pronosticados por el modelo, resulta trivial obtener las predicciones del mínimo y el máximo del

intervalo

$$\hat{X}_{t,L} = \hat{X}_{t,C} - \hat{X}_{t,R} \text{ y } \hat{X}_{t,U} = \hat{X}_{t,C} + \hat{X}_{t,R}. \tag{3.6}$$

Por su parte, el modelo híbrido que los autores proponen considera que tanto la serie de los centros, como la de los radios se pueden descomponer en una componente lineal y otra no lineal de la siguiente manera

$$\{X_{t,C}\} = \{L_{t,C} + N_{t,C}\} \text{ y } \{X_{t,R}\} = \{L_{t,R} + N_{t,R}\} \tag{3.7}$$

Las componentes lineales de ambas series serían estimadas por los modelos ARMA de los descritos anteriormente. Los residuos de las series de los centros y de los radios a partir de los modelos ARIMA recogen el comportamiento no lineal de ambas series, es decir, $Z_{t,C} = N_{t,C}$ y $Z_{t,R} = N_{t,R}$. Las series de residuos son modeladas cada una de ellas por un perceptrón multicapa. El perceptrón multicapa de una capa oculta que utilizan para modelar cada una de las series es el que se propone habitualmente

$$X_t = \alpha_0 + \sum_{j=1}^{q} \alpha_j \cdot g(\beta_{0j} + \sum_{i=1}^{p} \beta_{ij} X_{t-i}) + \varepsilon_t, \tag{3.8}$$

donde α_j y β_{ij} son parámetros del modelo, p es el número de nodos de entrada y q es el número de nodos de la capa oculta. El algoritmo de entrenamiento utilizado es el de *backpropagation* y la función de activación que emplean es la función logística (o sigmoidea). Una vez que se han entrenado adecuadamente los dos perceptrones, el de los residuos de los centros y el de los residuos de los radios, proporcionarán las predicciones de las componentes no lineales de las series de los centros y de los radios.

Los autores prueban estos modelos tanto con datos sintéticos como con datos reales. En los datos sintéticos, consideran cuatro series temporales: dos donde los centros son generados mediante procesos AR(1) y dos donde son generados mediante procesos no lineales de diferente complejidad. En todos los casos los radios son generados mediante distribuciones uniformes. Los datos reales que utilizan pertenecen a una serie de temperaturas mínimas y máximas mensuales de una estación meteorológica de China y realizan predicciones para seis y doce periodos con ambos modelos. En todas las series que los autores manejan, el modelo híbrido obtiene mejores predicciones que el modelo ARMA.

3.7.1.3. El uso de los valores mínimos y máximos en datos financieros

En el ámbito de las finanzas, los valores mínimos y máximos de las cotizaciones han gozado siempre de gran popularidad. La razón es que estos valores permiten hacerse una idea de la volatilidad de la cotización y, por ello, son empleados por los analistas técnicos y por el inversor de a pie para

guiarse en la toma de decisiones. De hecho, estos valores se emplean para el cálculo de indicadores técnicos como el ADX (*Average Directional movement indeX*) o el Oscilador Estocástico y para el estudio de las tendencias y de las líneas de soporte y resistencia de las cotizaciones. Puede consultarse más información sobre el tema en algún manual de análisis técnico como, por ejemplo, el de Bernstein (1995).

Pese a esta popularidad entre los analistas técnicos, las investigaciones académicas se centran habitualmente en el tratamiento de la serie de los valores de cierre de sesión. Dicho hecho resulta sorprendente porque es obvio que tanto el mínimo como el máximo ofrecen una información que no está recogida en el valor de cierre. Con todo, los trabajos académicos que consideran los valores mínimo y máximo en finanzas constituyen un pequeño margen. A continuación, se reseñarán los más significativos.

El modelado de las series temporales de los mínimos y los máximos en finanzas. El trabajo de Fiess y MacDonald (2002), que ya ha sido citado en el apartado 3.5.2.2, estudia la relación entre las series de los valores mínimos, máximos y de cierre en series temporales de divisas. Según dicho estudio, hay evidencias de cointegración entre las tres variables para las series de cambio de divisas entre el Dólar-Yen, Libra-Dólar y Marco-Dólar. Además, en el contexto de un modelo vectorial de corrección del error hay evidencias de que los valores del mínimo, del máximo y de cierre son significativos a la hora de predecir los valores de cierre. Para determinarlo utilizan el test de causalidad de Granger. Es importante destacar que estos resultados apuntan a que el análisis técnico y basado en gráficos de velas en realidad intenta sacar partido de forma intuitiva a dicha relación de causalidad.

Cheung (2007) considera, con buen criterio, que en finanzas las series temporales de los mínimos y los máximos tienden a no distanciarse mucho a lo largo del tiempo y que, por tanto, tiene sentido considerar que están cointegradas. El trabajo de Cheung se centra en el modelado de las series de los índices DJI (*Dow Jones Industrial*), S&P500 y NASDAQ, y demuestra que en dichos casos la relación de cointegración es cierta. En base a dicha relación de cointegración, el autor plantea un modelo VECM para los mínimos y los máximos y lo extiende para añadir otras variables como los valores de apertura y cierre y el volumen de negociación. A la luz de este trabajo, al incorporar dichas variables la capacidad explicativa del modelo aumenta. La variable menos relevante resulta ser el volumen de negociación.

Cheung (2007) también muestra que el modelado de la serie temporal de los rangos (siendo el rango la diferencia entre el máximo y el mínimo) utilizando únicamente los valores pasados de dicha serie no es la mejor opción para caracterizar el intervalo de variabilidad diaria y que se obtiene un modelo mejor al incorporar información de los mínimos y de los máximos. Desde el punto de vista de esta tesis, este resultado es lógico porque el rango

del intervalo no tiene información sobre la posición del mismo, para ello es necesario considerar el centro y el rango del intervalo de forma conjunta. Sin embargo, Cheung (2007) no llega a considerar el intervalo como tal, ni estudia la capacidad predictiva de los modelos considerados. Estos aspectos sí serán estudiados en el desarrollo de esta tesis.

Determinación del instante en el que se producen los valores mínimo máximo en las series temporales financieras. Otro trabajo interesante dentro del ámbito financiero es el llevado a cabo por Mok, Lam y Li (2000). En dicho trabajo, el interés no está en determinar la magnitud de los valores mínimo y máximo de una serie financiera, sino en determinar cuándo es más probable que sucedan estos valores.

En las series temporales financieras los valores del mínimo y el máximo diarios son puntos críticos dentro de la serie temporal de precios intradiarios. Dichos valores representan los dos grandes puntos de inflexión en la serie de precios diarios y vienen determinados por las fuerzas de la oferta y la demanda. Sin embargo, en el momento en que estos valores se producen, es imposible saber a ciencia cierta si dichos valores serán finalmente el mínimo o el máximo diario.

Mok et al. (2000) estudian este fenómeno en los mercados de futuros donde, según afirman, el mínimo y el máximo se alcanzan habitualmente al principio o al final de la sesión. En dicho trabajo, los autores demuestran que el hecho de que los mínimos y los máximos se den al principio o al final de la sesión, concuerda con el hecho de que el movimiento de la serie intradiaria se rija según un camino aleatorio.

El trabajo de Mok et al. (2000) no tiene como objetivo predecir la magnitud del mínimo y del máximo. La conclusión que obtienen sirve para corroborar de forma estadísticamente rigurosa un fenómeno observado en los mercados de futuros, que los mínimos y los máximos se obtienen típicamente al principio o al final de la sesión, pero no es de gran utilidad a la hora de fijar una estrategia inversora. En realidad, determinar el momento exacto en el que se producirán los valores mínimos y máximos para una serie intradiaria de cotizaciones parece una labor muy complicada. Sin embargo, dicha información puede obtenerse de forma indirecta a través de la predicción de los valores del mínimo y del máximo. Además, el contar con predicciones más o menos fiables de dichos valores resulta más útil a la hora de la toma de decisiones, porque también informa de la magnitud que va a alcanzar la cotización.

El uso de los valores mínimo y máximo de las series temporales financieras para la elaboración de estimadores de volatilidad. La volatilidad es un concepto fundamental en finanzas. En un principio, se asumía que la volatilidad era constante. Sin embargo, dicha consideración era

excesivamente simplista y alejada de la realidad. Hoy en día se asume que la volatilidad varía a lo largo del tiempo y que es hasta cierto punto predecible. Por ello, existen modelos para recoger la volatilidad estocástica. Sin embargo, pese a la gran cantidad de modelos disponibles, la estimación de la volatilidad continúa siendo un asunto complicado. El lector interesado puede encontrar una introducción a los modelos de volatilidad estocástica en el artículo de Alizadeh, Brandt y Diebold (2002).

Para el propósito de esta tesis, es interesante mencionar una familia de modelos de volatilidad basados en el rango. El rango en este caso es la diferencia entre valores del máximo y del mínimo del periodo considerado y actúa como representante de la volatilidad. Parkinson (1980) propone el primer estimador basado en el rango. Garman y Klass (1980) y Beckers (1983) proponen otros que combinan la información del rango con la de cierre. Yang y Zhang (2000) proponen un estimador donde también tienen en cuenta el valor de apertura.

Alizadeh et al. (2002) demuestran teórica y empíricamente que los estimadores basados en el rango son estimadores muy eficientes en comparación con otros estimadores tradicionales y que son robustos ante el ruido que introduce la microestructura del mercado. El hecho de que sean estimadores eficientes resulta perfectamente lógico ya que una serie de cierres no contiene información de la variabilidad intradiaria, y si dos cierres consecutivos son prácticamente idénticos los estimadores que se basan únicamente en esa información obtendrán como resultado una volatilidad baja, cuando la realidad ha podido ser bien distinta debida a las fluctuaciones intradiarias. Esta idea es corroborada por Fiess y MacDonald (2002) que afirma que los estimadores de volatilidad basados en el rango permiten recoger características de la microestructura del mercado de las series de divisas tales como los puntos en los que el mercado se da la vuelta. Estos autores afirman que estos valores no sólo pueden generar mejores predicciones de volatilidad, sino también mejores predicciones del nivel de la serie.

Corrado y Truong (2007) emplean un estimador de la volatilidad basado en el rango para construir predicciones de la volatilidad condicional de series de índices bursátiles. Para ello, incorporan las estimaciones de la volatilidad como variables explicativas a un modelo GARCH. El resultado del estudio indica que las estimaciones obtenidas con el estimador basado en el rango mejoran la precisión de las predicciones.

3.7.2. Métodos para la predicción de series temporales de histogramas

Entre los métodos de predicción para series temporales de histograma planteados al margen de esta tesis sólo se encuentra el trabajo de Maté y González-Rivera (2007). Estos autores plantean un método para predecir series temporales de histograma basado en el análisis de componentes princi-

pales para datos de histograma desarrollado en Rodríguez et al. (2000). Dicho método de predicción usa como variables de entrada un conjunto de retardos de la serie temporal de histogramas sobre las cuales realiza un análisis de componentes principales y usa la transformada de la primera componente principal como predicción para el día siguiente. Este método es aplicado para predecir series temporales de histogramas donde cada histograma representa la distribución de los rendimientos geométricos diarios de una acción que se han dado cada mes entre noviembre de 2004 y abril de 2007. Aplican el método sobre las 35 acciones del IBEX y sobre el propio índice y obtienen unos resultados según los cuales su método funciona mejor que el método ingenuo en 20 de las 36 acciones.

Capítulo 4

Predicción de Series Temporales de Intervalos

Es fácil predecir un futuro
e imposible predecir el futuro.

David Donnelly

Las series temporales de intervalos son un tipo de series temporales simbólicas donde cada observación es descrita mediante un intervalo o rango de valores. Como cada observación está representada mediante un dato simbólico, i.e. un intervalo, lo natural es abordarlas desde la perspectiva de los datos simbólicos. Sin embargo, para representar un intervalo sólo hacen falta dos números reales (sus extremos inferior y superior, o, alternativamente, su centro y su radio), por lo que las series temporales de intervalos también pueden ser abordadas desde el ámbito de las series temporales clásicas utilizando métodos univariantes o multivariantes.

En este capítulo se presentarán las series temporales de intervalos y se estudiarán distintos aspectos relativos a ellas entre los que se encuentran: su definición, cómo se obtienen, distintas aproximaciones a la medición del error en este tipo de serie temporal y un conjunto de métodos y de estrategias para predecirlas. La capacidad de predicción de los métodos presentados en el capítulo será probados sobre un conjunto de series temporales provenientes de distintos ámbitos.

4.1. Introducción

En una serie temporal clásica, cada periodo está representado como un único valor. Este enfoque es útil para representar una gran multitud de situaciones que se dan en la práctica. Sin embargo, como se ha mencionado en el capítulo anterior, existen fenómenos que no pueden ser representados adecuadamente por este tipo de series temporales.

Las series temporales de intervalos (STI) permiten representar situaciones en las que las observaciones se ven afectadas de variabilidad, lo que hace que sea más adecuado representarlas por medio de un intervalo. Los intervalos de la STI también podrían representar imprecisión o incertidumbre.

En este capítulo se abordarán la predicción de STI y se sentarán las bases para poder trabajar y predecir series temporales de histograma. En primer lugar, se definirá qué son las STI.

4.2. Definición de Serie Temporal de Intervalos

Definición. Una serie temporal de intervalos, $\{[X]_t\}$, puede definirse como una secuencia de rangos de valores que son observados en instantes sucesivos en el tiempo denotados por $t = 1, ..., n$, y donde cada rango se representa mediante un intervalo de la forma

$$[X]_t = [X_{t,L}, X_{t,U}], \tag{4.1}$$

con $-\infty < X_{t,L} \leq X_{t,U} < \infty$, siendo $X_{t,L}$ el extremo inferior del intervalo, i.e., el mínimo valor observado, y $X_{t,U}$ el extremo superior del intervalo, i.e., el máximo valor observado. Alternativamente, el intervalo $[X]_t$ puede representarse como

$$[X]_t = \langle X_{t,C}, X_{t,R} \rangle, \tag{4.2}$$

donde $X_{t,C} = (X_{t,L} + X_{t,U})/2$ es el punto medio del intervalo, i.e., el centro, y $X_{t,R} = (X_{t,U} - X_{t,L})/2$ es la mitad de la longitud del intervalo, i.e., el radio.

Como puede observarse, la notación que se emplea en este capítulo para representar los intervalos no coincide con la presentada en el capítulo 2, pero es la que se ha estimado como más adecuada y menos confusa para trabajar con intervalos en el contexto de las series temporales.

Desde el punto de vista del análisis de datos simbólicos, una STI puede definirse como la realización de un proceso estocástico donde cada instante temporal está representado por una variable aleatoria de intervalo (ver la definición de variable aleatoria de intervalo en el punto 2.2).

Otra forma alternativa de concebir una STI es como la realización de un proceso estocástico bivariante donde las dos variables pueden ser el mínimo y el máximo, o el centro y el radio. Cada una de estas variables daría lugar a su propia serie temporal.

Un intervalo queda perfectamente definido mediante dos valores: sus extremos superior e inferior o, de forma equivalente, su centro y su radio. En realidad, cualquier pareja tomada de entre estos cuatro posibles valores sirve para definir al intervalo ya que permiten obtener los otro dos valores, e.g. a partir del centro y del extremo inferior podemos obtener fácilmente el radio

y el extremo superior. Sin embargo, conceptualmente resulta más correcto tomar o bien los extremos, porque denotan los límites del intervalo, o bien el centro y el radio, porque hacen énfasis en dos características fundamentales en la estadística como son la tendencia central y porque además dichos valores están relacionados con el concepto de bola que se emplea en topología, como ya se vio en el apartado 2.2.

4.3. El interés de las series temporales de intervalos

Los intervalos permiten representar los siguientes casos.

- Cuando no se dispone de información precisa y la observación debe recoger dicha imprecisión en forma de intervalo. En estos casos, el intervalo recoge el rango de valores en el que se encuentra el valor real. Esto sucede cuando el instrumento de medida no es fiable.
- Cuando se resume mediante un intervalo el valor de una variable observada en un conjunto de individuos en un determinado instante.
- Cuando se resume mediante un intervalo una secuencia de valores observada en un individuo en un periodo de tiempo.

Respecto al primer caso, es complicado pensar una situación en la que sea necesario predecir una STI donde los intervalos representen la imprecisión de un instrumento de medida. Sin embargo, en el caso de que se presente una serie de esta naturaleza y se quiera predecir prescindiendo del intervalo las conclusiones que se pueden obtener pueden ser erróneas, especialmente si la imprecisión que rodea a las observaciones no es pequeña.

El segundo caso representa una situación en la que se realiza agregación contemporánea. En estos casos, lo habitual es tomar como resumen de las observaciones una medida de tendencia central, como sería la media o la mediana. La información que aporta el intervalo es complementaria a la que aportan las medidas de tendencia central, ya que describe la variabilidad de la muestra observada. Además, según el objetivo del análisis, puede plantearse un intervalo donde, para evitar la influencia de los valores extremos, los extremos sean los percentiles del 10 y del 90, o donde el intervalo represente el recorrido intercuartílico de las observaciones.

El tercer caso, representa una situación de agregación temporal. En estos casos, se puede proceder de varias formas:

- Se puede tomar el último valor observado de la secuencia o el primero o uno muestreado al azar. El problema que tiene esta aproximación es que ignora completamente la variabilidad del fenómeno. Si no existe variabilidad o ésta no es relevante, puede ser un enfoque apropiado,

pero si hay variabilidad, parece más adecuado recogerla mediante un intervalo.

- Si se trata de una variable en la que tenga sentido estudiar el total acumulado, como es el caso de una variable que mida el consumo o la producción de un bien, pueden calcularse los totales en cada periodo. En estos casos, los intervalos ofrecen una representación alternativa para informar de la variabilidad durante cada periodo.

- Existen otros casos en los que la representación natural de este tipo de series son los intervalos que recogen los valores mínimos y máximos observados en cada periodo. Los ejemplos más claros de este tipo de series se dan en las finanzas para reflejar el rango de precios de una acción o de un índice durante una sesión y en meteorología para reflejar las temperaturas mínimas y máximas.

Como ya se vio en el apartado 3.7.1.3, los intervalos de los valores mínimo y máximo juegan un papel muy relevante en finanzas porque son los puntos de inflexión en los precios de los bienes. Estos puntos vienen determinados por las fuerzas de la oferta y la demanda y pueden convertirse, mientras que no sean superados, en los valores de soporte y resistencia para las sesiones futuras. El conocimiento anticipado de dichos valores es, por tanto, de vital interés para los inversores, ya que marcan sobre los puntos de referencia para entrar o salir del mercado. De hecho, los inversores a menudo lanzan órdenes *stop-loss*[1] cerca de los valores mínimos y máximos recientes. Si se tiene una predicción fiable de los valores mínimos y máximos de la sesión, a lo largo de la propia sesión se pueden identificar los instantes en los que se alcanzan dichos valores y establecer una estrategia de compra o de venta con respecto a ellos. Además, son unos buenos indicadores de la volatilidad de un bien financiero. Pese a su relevancia y tal y como mostró el apartado el apartado 3.7.1.3, se ha trabajado poco en su predicción de los valores mínimo y máximo. Desde la perspectiva de esta tesis, la secuencia de valores mínimo y máximo puede ser considerada como una STI. Por tanto, los métodos de que van a ser propuestos en este capítulo pueden ser empleados para predecir estos valores y, de hecho, así se hará en el apartado 4.10.

4.4. Medidas de Error para series temporales de intervalos

En las series temporales clásicas, el error en un determinado instante se mide como la diferencia entre el valor pronosticado y el valor real. Para poder agregar el error cometido a lo largo del tiempo se toma el error elevado al

[1] Las órdenes *stop-loss* son órdenes de venta que obligan a vender (comprar, resp.) si las cotizaciones bajan (suben, resp.) o si pierden (alcanzan, resp.) un nivel prefijado

cuadrado o en valor absoluto para evitar el efecto de compensación que se da entre los errores negativos y positivos.

La adaptación al contexto de las STI del concepto de error como diferencia entre el valor observado y el pronosticado no es posible utilizando la aritmética de intervalos (ver más información sobre ella en la Secc. 2.7.1.1). Esto es debido a que el operador aritmético de diferencia entre dos intervalos no sirve para cuantificar la desviación que existe entre ellos (Palumbo y Lauro, 2003). Para explicarlo mejor emplearemos un ejemplo.

Consideremos el intervalo observado $[Y]_t = [A]$ y un intervalo que pronostica perfectamente el valor observado $[\hat{Y}]_t = [A]$, si empleamos la diferencia de intervalos tal y como se define en el apartado 2.7.1.1, entonces se cumple

$$[Y]_t - [\hat{Y}]_t = [0,0] \Leftrightarrow [A] = [a,a] \text{ con } a \in \Re. \tag{4.3}$$

Si el intervalo $[A]$ no es degenerado (i.e. no es un intervalo de longitud cero), lo que se obtiene es un intervalo centrado en el cero y de longitud igual al doble de la longitud de $[A]$; e.g. si $[A] = [1,2]$, el resultado es $[-1,1]$. Esto sucede porque la finalidad de la aritmética de intervalos es ofrecer un intervalo que acote todos los posibles resultados que se dan al considerar cada uno de los valores posibles en cada uno de los operandos.

Esto demuestra que la aritmética de intervalos no es útil para definir el concepto de error en una STI, por lo que son necesarias otras aproximaciones. A continuación, se propondrán dos de ellas, una basada en el uso de distancias para intervalos y otra que consiste en considerar el error cometido en las series temporales clásicas de los mínimos, los máximos, los centros y los radios.

4.4.1. Medidas de error basadas en el concepto de distancias

Una posible alternativa para medir el error en STI consiste en representar el concepto de error como la distancia entre el intervalo observado y el intervalo pronosticado.

Las distancias permiten cuantificar de forma objetiva la discrepancia entre dos intervalos. Además, por definición una función distancia es definida positiva lo cual facilita la agregación de errores sin necesidad de tomar sus valores absolutos o de elevarlos al cuadrado.

4.4.1.1. Distancias para datos de intervalos

A continuación, se muestran una serie de distancias para datos de intervalo que han sido planteadas en la literatura.

Distancia de Hausdorff. Hausdorff definió una métrica para conjuntos compactos. Un intervalo es un conjunto compacto definido por un par de valores ordenados, i.e. el extremo inferior y el extremo superior. Dados dos

intervalos $[A] = [A_L, A_U] = \langle A_C, A_R \rangle$ y $[B] = [B_L, B_U] = \langle B_C, B_R \rangle$, la métrica de Hausdorff para intervalos es

$$\begin{aligned} d_H([A],[B]) &= \text{máx}(|A_L - B_L|, |A_U - B_U|) \\ &= |A_C - B_C| + |A_R - B_R|. \end{aligned} \tag{4.4}$$

Esta medida cumple las propiedades para ser considerada una distancia y se emplea habitualmente en el análisis de intervalos (Moore, 1979). Es interesante resaltar que, si se consideran dos intervalos degenerados, i.e., $[A] = [x, x]$ y $[B] = [y, y]$, entonces obtenemos $d_H([A],[B]) = |x - y|$, que es la topología habitual que se emplea en la recta real.

La distancia de Hausdorff puede interpretarse como la distancia de Manhattan entre dos intervalos, donde cada intervalo es un elemento definido por su centro y por su radio. De forma equivalente, también puede interpretarse como el máximo de las distancias entre los extremos inferiores o entre los extremos superiores de los intervalos.

Distancia de Ichino-Yaguchi. Ichino y Yaguchi (1994) proponen una métrica de Minkowski generalizada para un espacio multidimensional donde las variables pueden ser de distintos tipos (entre ellos, se encuentran los intervalos). Dicha métrica se basa en el uso de los operadores cartesianos de unión e intersección, los cuales se definen de la siguiente forma para dos intervalos $[A]$ y $[B]$

$$[A] \oplus [B] = [A] \cup [B] = [min(A_L, B_L), max(A_U, B_U)] \text{ y} \tag{4.5}$$

$$[A] \otimes [B] = [A] \cap [B]. \tag{4.6}$$

Dados estos operadores, la distancia de Ichino-Yaguchi entre $[A]$ y $[B]$ se define como

$$\begin{aligned} d_{IY}([A],[B]) &= w([A] \cup [B]) - w([A] \cap [B]) \\ &\quad + \gamma\big(2w([A] \cap [B]) - w([A]) - w([B])\big), \end{aligned} \tag{4.7}$$

donde $w([X])$ es la longitud del intervalo [X], i.e. $w([X]) = X_U - X_L$, y donde $\gamma \in [0, 0.5]$ controla el efecto de la cercanía de los extremos interiores y de los extremos exteriores de los intervalos $[A]$ y $[B]$. Esta medida cumple las propiedades para ser considerada una distancia. Los autores sugieren el valor $\gamma = 0.5$ como apropiado, lo que da como resultado

$$d_{IY}^{\gamma=0.5}([A],[B]) = w([A] \cup [B]) - \frac{1}{2}\big(w([A]) + w([B])\big), \tag{4.8}$$

Al tomar el valor de $\gamma = 0.5$ la distancia se vuelve más fácilmente interpretable ya que se convierte en

$$d_{IY}^{\gamma=0.5}([A],[B]) = \frac{1}{2}\big(|B_L - A_L| + |B_U - A_U|\big). \tag{4.9}$$

Esta distancia se puede interpretar como un medio de la distancia de Manhattan entre dos intervalos, donde cada intervalo es un elemento definido por su límite inferior y por su límite superior.

Distancia de De Carvalho. De Carvalho (1994) propone la siguiente normalización de la distancia de Ichino-Yaguchi

$$d_{DC}([A],[B]) = \frac{d_{IY}^{\gamma}([A],[B])}{w([A] \cup [B])}, \tag{4.10}$$

donde $d_{IY}^{\gamma}([A],[B])$ viene dado en (4.7). Esta medida es una distancia y toma valores en el rango $[0,1]$.

Si $\gamma = 0$, la medida toma su valor máximo cuando la intersección entre los intervalos es nula, sin tener en cuenta lo distantes que se encuentren éstos entre sí. Éste no es un comportamiento adecuado para medir errores en STI, por lo que se descarta $\gamma = 0$. Sin embargo, para $\gamma = 0.5$, la distancia presenta unas propiedades más interesantes; son las siguientes:

- $d_{DC}([A],[B]) = 1$ si y sólo si los intervalos son degenerados y diferentes entre sí, e.g $[A] = [3,3]$ y $[B] = [4,4]$;
- $d_{DC}([A],[B]) = 0.5$ si los intervalos considerados son adyacentes, e.g. $[A] = [1,3]$ y $[B] = [3,7]$, o si uno de los intervalos es degenerado y está contenido dentro del otro, e.g. $[A] = [1,4]$ and $[B] = [2,2]$;
- $d_{DC}([A],[B]) < 0.5$ si y sólo si $w([A] \cap [B]) > 0$;
- $d_{DC}([A],[B]) \leq 0.25$ si y sólo si $\frac{w([A]\cap[B])}{w([A]\cup[B])} \geq 0.5$.

Resulta evidente que en el contexto de la predicción de STI una distancia entre la predicción y la observación de $d_{DC} \geq 0.5$ denotaría una predicción muy pobre.

Distancia intervalar definida a partir de un núcleo. González, Velasco, Angulo, Ortega y Ruiz (2004) realizan un análisis de una serie de distancias para intervalos. En dicho análisis descartan la distancia de Hausdorff (4.4) por no tener en cuenta la distancia entre los extremos cercanos de los intervalos. Para mejorar esta distancia recurren a la definición de una función núcleo (o *kernel*). La función núcleo permite establecer similitudes entre los elementos originales a partir de sus transformados, lo que conlleva, en ocasiones, la posibilidad de definir una distancia entre los puntos origen. Para ello, el kernel emplea una aplicación definida desde el conjunto de trabajo (el cual puede no estar provisto de ninguna estructura matemática *a priori*) a un espacio vectorial dotado de un producto escalar, e.g. $\Re^n$. Puede consultarse el artículo de González et al. (2004) para obtener más detalles sobre el *kernel* y sobre la función de aplicación.

La distancia que se define como conclusión de. estudio es la siguiente

$$d_k([A],[B]) = \frac{1}{\sqrt{2}}\sqrt{(B_L - A_L)^2 + (B_U - A_U)^2} \quad (4.11)$$

$$= \sqrt{(B_C - A_C)^2 + (B_R - A_R)^2}. \quad (4.12)$$

Esta distancia puede interpretarse como la distancia euclídea entre dos intervalos, donde cada intervalo es un elemento definido por su centro y por su radio. En otras palabras, es una métrica de Minkowski de orden dos, mientras que la métrica de Hausdorff (4.4), lo era de orden uno.

La distancia definida a partir de un núcleo también guarda una relación similar con la distancia de Ichino-Yaguchi (4.9). Ambas distancias pueden considerarse como una métrica de Minkowski del siguiente tipo

$$d_{M_{LU}}([A],[B]) = \left[\frac{1}{2}(B_L - A_L)^q + \frac{1}{2}(B_U - A_U)^q\right]^{1/q}, \quad (4.13)$$

siendo $q = 1$ para la distancia de Ichino-Yaguchi y $q = 2$ para la distancia definida a partir de un núcleo.

González et al. (2004) plantean también una extensión de su medida (4.12) para el caso en el que los intervalos se encuentren dentro de un soporte compacto, i.e. de un intervalo cerrado.

4.4.1.2. Definición del Error Medio basado en una Distancia

Sea $\{[Y]_t\}$ la STI observada y $\{[\hat{Y}]_t\}$ la STI pronosticada, con $t = 1, ..., n$, el Error Medio basado en una Distancia se define como

$$EMD_X^q = \left(\frac{\sum_{t=1}^{n}(d_X([Y]_t, [\hat{Y}]_t))^q}{n}\right)^{\frac{1}{q}}, \quad (4.14)$$

donde d_X es una de las distancias que han sido consideradas como adecuadas en el apartado anterior, a saber: Hausdorff (4.4), Ichino-Yaguchi (4.9), De Carvalho (4.10) o la distancia definida a partir de un kernel (4.12). El parámetro q es el orden de la medida de error, de forma que si $q = 1$, el error en t será agregado como en el error absoluto medio, mientras que si $q = 2$, la medida resultante será similar a la raíz cuadrada del error cuadrático medio.

Cualquiera de estas distancias puede resultar adecuada, todo depende del propósito del análisis de si, por ejemplo, se considera adecuado usar una medida de error cuadrática o no, o de si se necesita una medida fácil de interpretar, etc.

4.4.2. Medidas de error estimadas sobre las series temporales clásicas

Las medidas de error para STI basadas en distancias miden las diferencias entre dos componentes de los intervalos, que son el centro y el radio, o bien el

mínimo y el máximo. Los dos pares de componentes están interrelacionados, pero no es posible a partir del error medido sobre uno de ellos hacerse una idea del error que se ha cometido en el otro par. Además, estas medidas agregan el error de los dos componentes considerados, de forma que tampoco permiten discernir en la predicción de cuál de ellos se está errando más.

Por tanto, si lo que se busca es un conocimiento más preciso del error que se comete en cada uno de los componentes, lo más razonable es calcular el error cometido en cada uno de ellos por separado. Para ello, basta con considerar las cuatro series temporales clásicas $\{X_{t,L}\}$, $\{X_{t,U}\}$, $\{X_{t,C}\}$ y $\{X_{t,R}\}$ y estimar el error que se ha cometido en cada una de ellas según sus predicciones, $\{\hat{X}_{t,L}\}$, $\{\hat{X}_{t,U}\}$, $\{\hat{X}_{t,C}\}$ y $\{\hat{X}_{t,R}\}$.

Puede pensarse que, como las cuatro series temporales están expresadas en la misma unidad, basta con emplear una medida de error dependiente de la escala de medida, como el Error Cuadrático Medio o el Error Absoluto Medio. Sin embargo, la magnitud de los errores cometidos en las cuatro series temporales puede ser bastante diferente, lo cual es habitual especialmente en el caso del error cometido en la serie de los radios. Esto puede dar lugar a interpretaciones erróneas como, por ejemplo, pensar que el radio se está prediciendo mejor que el resto de los componentes sólo porque la magnitud de su error es menor que la del error de los otros componentes.

Puede pensarse también que, para evitar el efecto de la escala de medida, puede resultar adecuado emplear el Error Porcentual Absoluto Medio. Sin embargo, esta medida toma valores inusualmente grandes si los valores de la serie son cercanos a cero y toma el valor infinito si hay algún instante en el que la serie vale cero. En algunas STI, los valores de los radios (o de alguna otra componente) pueden ser cercanos a cero o cero, por lo que esta medida no resultaría adecuada en esos casos. Además, la medida también plantea problemas de simetría ya que penaliza más los errores positivos que los errores negativos y no es adecuada si la serie que se está analizando no está expresada sobre una escala donde el cero significativo, como sucede con las temperaturas expresadas en ^oC.

Por ello, es necesaria otra medida de error independiente de la escala de medida. Hyndman y Koehler (2006) realizan un análisis en profundidad de este tipo de medidas, entre ellas, las medidas basadas en el error porcentual como el Error Porcentual Absoluto Medio y sus variantes simétricas, también las medidas basadas en errores relativos como el Error Relativo Absoluto Medio, y las medidas relativas como la U de Theil (Theil, 1966). Según Hyndman y Koehler todas las medidas de error existentes en la literatura presentan algún inconveniente, por ello estos autores optan por presentar una nueva medida de error basada en el concepto de error escalado.

Esta medida consiste en dividir el error cometido en el instante t entre el Error Absoluto Medio muestral cometido por un método de referencia, como puede ser, por ejemplo, el método ingenuo. Sea la serie temporal $\{y_t\}$ con

$t = 1, ..., m, m+1, ..., n$, donde m representa la longitud del periodo muestral y n es la longitud de la serie temporal, el error absoluto escalado en t es

$$q_t = \frac{|y_t - \hat{y}_t|}{\frac{1}{m-1}\sum_{i=2}^{m}|y_i - y_{i-1}|}. \tag{4.15}$$

La interpretación de q_t es sencilla. Si $q_t < 1$, entonces el método que se está analizando es más eficaz en la predicción que el método ingenuo en media durante el periodo muestral. Inversamente, si $q_t > 1$ el método analizado obtiene una predicción peor que las que ofrece en media el método ingenuo durante el periodo muestral.

Una vez definido el error escalado, Hyndman y Koehler (2006) proponen el Error Absoluto Escalado Medio ($EAEM$) como

$$EAEM = media(q_t). \tag{4.16}$$

Según Hyndman y Koehler (2006), los errores escalados deben convertirse en un enfoque estándar a la hora de comparar las predicciones obtenidas sobre series que se expresan en una escala diferente o con distintas unidades de medida. Las razones que alegan son que estos errores son fácilmente interpretables y a que son fiables en cualquier circunstancia, ya que no toman valores infinitos o valores sesgados.

4.4.2.1. Una medida de error escalada y cuadrática

Hyndman y Koehler (2006) afirman que se pueden desarrollar medidas similares al $EAEM$, como, por ejemplo, la raíz cuadrada del error cuadrático escalado medio ($RECEM$).

Resulta más natural usar este error que el $EAEM$ cuando los parámetros del modelo se han estimado mediante mínimos cuadrados como puede hacerse con los modelos ARIMA, los modelos VAR o el perceptrón multicapa, ya que el $RECEM$ es un error cuadrático y no absoluto el $EAEM$. A continuación se definirá el $RECEM$.

En primer lugar, el Error Cuadrático Escalado en t se define como

$$s_t = \frac{(y_t - \hat{y}_t)^2}{\frac{1}{m-1}\sum_{i=2}^{m}(y_i - y_{i-1})^2}. \tag{4.17}$$

El factor por el que se escala el error cuadrático en t es el Error Cuadrático Medio cometido en el periodo muestral. Dada la definición del Error Cuadrático Escalado en t, el $RECEM$ se define como

$$RECEM = \sqrt{media(s_t)}. \tag{4.18}$$

Es fácil comprobar que la ecuación (4.18) es equivalente a

$$RECEM = \frac{RECM}{RECM_m^*}, \tag{4.19}$$

donde $RECM_m^*$ es la raíz cuadrada del error cuadrático medio cometido por el método ingenuo en el periodo muestral y donde m es la longitud del periodo muestral. Por su parte, $RECM$ es el error cuadrático medio cometido por el método considerado en el periodo de interés.

Tanto el $RECEM$ (4.18) como el $RECEM$ (4.16) son consideradas como medidas adecuadas para estimar el error cometido en cada uno de los cuatro componentes de la STI, ya que son independientes de la escala de medida y tienen una interpretación clara.

4.5. Predicción mediante los modelos univariantes y multivariantes clásicos

La sencillez de los datos de intervalo permite que éstos sean representados únicamente mediante dos valores de entre estos cuatro: mínimo, máximo, centro y radio. Esto permite analizar los conjuntos de datos de intervalos aplicando técnicas multivariantes considerando que el intervalo es un individuo definido mediante dos componentes.

Esta idea puede extenderse al contexto de las STI, donde una STI de $\{[X]_t\}$ con $[X_t] = [X_{t,L}, X_{t,U}] = \langle X_{t,C}, X_{t,R} \rangle$ puede desglosarse en cuatro series temporales clásicas: la de los extremos inferiores $\{X_{t,L}\}$, la de los extremos superiores $\{X_{t,U}\}$, la de los centros $\{X_{t,C}\}$ y la de los radios $\{X_{t,R}\}$.

Por ello, una alternativa muy sencilla para predecir las STI consiste en predecir dos de las series de sus componentes con los métodos que se estimen necesarios. En un principio, resulta más natural analizar, o bien las series de los extremos superiores e inferiores, o bien las del centro y el radio. Sin embargo, es posible analizar cualquier posible pareja tomada de entre las cuatro series.

4.5.1. La aproximación univariante

En primer lugar, se debe determinar si se va a optar por la aproximación de los extremos del intervalo o por la aproximación del centro y el radio. Normalmente, las series de los extremos tendrán un comportamiento muy similar, mientras que la del centro y la del radio tendrán un comportamiento muy distinto y cada una requerirá un modelo de predicción diferente.

Una vez determinadas las dos series que se van a predecir, cada una de ellas debe ser estudiada de forma independiente y debe determinarse qué modelo resulta mejor para predecirla. Entre el catálogo de métodos a elegir, se puede optar por cualquiera de los métodos de predicción de series temporales univariantes habituales, como los alisados exponenciales (Gardner, 2006), los modelos ARIMA (Box y Jenkins, 1970), el método k-NN (Yakowitz, 1987), las redes neuronales (Zhang, Patuwo y Hu, 1998), etc.

A partir de las predicciones obtenidas para cada una de las dos series de los componentes se puede componer la STI pronosticada. Sin embargo, antes hay que comprobar que todas las predicciones obtenidas son en realidad intervalos. Para ello, en el caso de haber optado por pronosticar las series de los extremos, hay que comprobar si todos los intervalos cumplen la condición de que su extremo inferior es menor o igual que su extremo superior; en el caso de haber pronosticado las series de los centros y de los radios, hay que comprobar que todos los radios pronosticados son mayores que cero. Si existe alguna condición donde esto no se cumple se debe cuestionar la validez del modelo empleado.

La aproximación univariante es la más sencillo, pero también es la más pobre desde el punto de vista conceptual ya que no trabaja con los intervalos como tal, sino que los descompone en dos series sin estudiar, si quiera, la posible relación entre ambas series.

4.5.2. La aproximación multivariante

Por otro lado, además de considerar un enfoque univariante, también puede ser conveniente modelar las series mediante un modelo multivariante que no sólo considere a la propia serie, sino que incluya retardos de los otros componentes.

Si para modelar un componente de una STI se opta por utilizar modelos como, por ejemplo, los de función de transferencia, se pueden añadir como variables explicativas del modelos retardos de alguno de los otros componentes del intervalo; e.g. para modelar el radio se puede probar si se obtienen mejores predicciones al añadir retardos de la serie de los centros. Sin embargo, al incluir estas variables como explicativas se está asumiendo una relación de causalidad en un solo sentido, lo cual no está claro que suceda. Además, es posible que se estén añadiendo conviene tener en cuenta que en el ámbito de predicción suele ser mejor emplear modelos sencillos y parsimoniosos en lugar de usar modelos mejor especificados, ya que estos últimos no tienen por qué obtener mejores resultados de predicción (Allen y Fildes, 2001).

En la mayoría de las STI, las series temporales del mínimo, del máximo y del centro mostrarán una gran correlación. Estas series también pueden estar correlacionadas con la serie del radio, aunque normalmente lo estarán en mucha menor medida. En cualquier caso, es normal pensar que la pareja de series considerada se encuentre interrelacionada. Por ello, una forma de predecir las STI es hacer uso de modelos multivariantes que tengan en cuenta las interdependencias entre las dos series consideradas.

Una forma habitual de abordar la predicción de series temporales interdependientes que no requiere especificar la relación existente entre las variables que entran en juego consiste en el uso de modelos de vectores autorregresivos (modelos VAR). Estos modelos resultan muy útiles porque no requieren que

se establezcan relaciones de causalidad entre las variables que intervienen en ellos. En un modelo VAR, cada serie temporal es recogida mediante una ecuación en la que se incluyen sus propios retardos y los retardos del resto de variables consideradas. Puede verse una introducción de los mismos en el punto A.1 del apéndice A.

Es interesante saber que las predicciones que se obtienen estimando mediante mínimos cuadrados un modelo VAR de orden p para las series temporales de los mínimos y de los máximos son equivalentes a las que se obtienen estimando un modelo VAR del mismo orden para las series del centro y del radio. En realidad, dado uno de los modelos, se pueden hallar fácilmente las ecuaciones. Por tanto, basta con plantear uno de los dos modelos porque el otro será equivalente. Esto se explica con mayor detalle en el siguiente apartado.

4.5.2.1. Relación entre el modelo VAR del mínimo y del máximo y el modelo VAR del centro y del radio

Una STI $\{[X]_t\}$ donde $[X_t] = [X_{t,L}, X_{t,U}] = \langle X_{t,C}, X_{t,R} \rangle$ puede descomponerse en cuatro series temporales clásicas: la de los extremos inferiores $\{X_{t,L}\}$, la de los extremos superiores $\{X_{t,U}\}$, la de los centros $\{X_{t,C}\}$ y la de los radios $\{X_{t,R}\}$.

Dado $\{[X]_t\}$, el modelo VAR estimado mediante mínimos cuadrados ordinarios sobre la serie de los extremos inferiores y de los extremos superiores de $\{[X]_t\}$ será el siguiente

$$\boldsymbol{Y}_t = \boldsymbol{\Theta} + \boldsymbol{\Phi}_1 \boldsymbol{Y}_{t-1} + \ldots + \boldsymbol{\Phi}_p \boldsymbol{Y}_{t-p} + \boldsymbol{\epsilon}_t, \tag{4.20}$$

donde $\boldsymbol{Y}_{t-i}$ es el vector de los valores de las series retardados i periodos, con $i = 0, ..., p$, que viene dado por

$$\boldsymbol{Y}_{t-i} = \begin{pmatrix} X_{t-i,L} \\ X_{t-i,U} \end{pmatrix}, \tag{4.21}$$

$\boldsymbol{\Theta}$ es el vector de las constantes del modelo que viene dado por

$$\boldsymbol{\Theta} = \begin{pmatrix} \theta_L \\ \theta_U \end{pmatrix}, \tag{4.22}$$

$\boldsymbol{\Phi}_i$ es la matriz de coeficientes del modelo asociada al vector $\boldsymbol{Y}_{t-i}$ que viene dada por

$$\boldsymbol{\Phi}_i = \begin{pmatrix} \phi_{11}^{i)} \phi_{12}^{i)} \\ \phi_{21}^{i)} \phi_{22}^{i)} \end{pmatrix}, \tag{4.23}$$

y donde ϵ_t es el vector de residuos del modelo que viene dado por

$$\boldsymbol{\epsilon}_t = \begin{pmatrix} \epsilon_{t,L} \\ \epsilon_{t,U} \end{pmatrix}. \tag{4.24}$$

La ecuación del modelo VAR mostrado en (4.20) para la serie de los extremos inferiores es

$$X_{t,L} = \theta_L + \mathbf{\Phi}_{1,L}\boldsymbol{Y}_{t-1} + \ldots + \mathbf{\Phi}_{p,L}\boldsymbol{Y}_{t-p} + \epsilon_{t,L}, \tag{4.25}$$

donde $\mathbf{\Phi}_{i,L} = \left(\phi_{11}^{i)}\phi_{12}^{i)}\right)$; y para la serie de los extremos superiores es

$$X_{t,U} = \theta_U + \mathbf{\Phi}_{1,U}\boldsymbol{Y}_{t-1} + \ldots + \mathbf{\Phi}_{p,U}\boldsymbol{Y}_{t-p} + \epsilon_{t,U}, \tag{4.26}$$

donde $\mathbf{\Phi}_{i,U} = \left(\phi_{21}^{i)}\phi_{22}^{i)}\right)$.

De forma análoga, el modelo VAR estimado mediante mínimos cuadrados ordinarios sobre la serie de los centros y de los radios de $\{[X]_t\}$ es el siguiente

$$\boldsymbol{Z}_t = \mathbf{\Omega} + \mathbf{\Psi}_1\boldsymbol{Z}_{t-1} + \ldots + \mathbf{\Psi}_p\boldsymbol{Z}_{t-p} + \boldsymbol{\varepsilon}_t, \tag{4.27}$$

donde $\boldsymbol{Z}_{t-i}$ es el vector de los valores de las series retardados i periodos, con $i = 1, ..., p$, que viene dado por

$$\boldsymbol{Z}_{t-i} = \begin{pmatrix} X_{t-i,C} \\ X_{t-i,R} \end{pmatrix}, \tag{4.28}$$

$\mathbf{\Omega}$ es el vector de las constantes del modelo que viene dado por

$$\mathbf{\Omega} = \begin{pmatrix} \omega_C \\ \omega_R \end{pmatrix}, \tag{4.29}$$

$\mathbf{\Psi}_i$ es la matriz de coeficientes del modelo asociada al vector $\boldsymbol{Z}_{t-i}$ que viene dada por

$$\mathbf{\Psi}_i = \begin{pmatrix} \psi_{11}^{i)}\psi_{12}^{i)} \\ \psi_{21}^{i)}\psi_{22}^{i)} \end{pmatrix}, \tag{4.30}$$

y donde ε_t es el vector de residuos del modelo que viene dado por

$$\boldsymbol{\varepsilon}_t = \begin{pmatrix} \varepsilon_{t,C} \\ \varepsilon_{t,R} \end{pmatrix}. \tag{4.31}$$

La ecuación del modelo VAR mostrado en (4.27) para la serie de los centros es

$$X_{t,C} = \omega_C + \mathbf{\Psi}_{1,C}\boldsymbol{Z}_{t-1} + \ldots + \mathbf{\Psi}_{p,C}\boldsymbol{Z}_{t-p} + \varepsilon_{t,C}, \tag{4.32}$$

donde $\mathbf{\Psi}_{i,C} = \left(\psi_{11}^{i)}\psi_{12}^{i)}\right)$. Mientras que para la serie de los radios es

$$X_{t,R} = \omega_R + \mathbf{\Psi}_{1,R}\boldsymbol{Z}_{t-1} + \ldots + \mathbf{\Psi}_{p,R}\boldsymbol{Z}_{t-p} + \varepsilon_{t,R}, \tag{4.33}$$

donde $\mathbf{\Psi}_{i,R} = \left(\psi_{21}^{i)}\psi_{22}^{i)}\right)$.

Los parámetros del modelo VAR para los extremos inferiores y superiores pueden hallarse a partir de los parámetros del modelo VAR para el centro (4.32) y para el radio (4.33), y viceversa. Como ambas transformaciones siguen la misma filosofía basta con realizar la conversión del modelo de los extremos inferiores y superiores al modelo del centro y el radio, ya que la otra se realiza de forma similar.

Para realizar dicha conversión nos valemos de las siguientes equivalencias ya conocidas

$$X_{t,L} = X_{t,C} - X_{t,R} \text{ y } X_{t,U} = X_{t,C} + X_{t,R}. \tag{4.34}$$

Dadas estas equivalencias, formularemos la predicción del mínimo de la siguiente forma

$$X_{t,L} = X_{t,C} - X_{t,R}, \tag{4.35}$$

si reemplazamos $X_{t,C}$ por la expresión en (4.32) y $X_{t,R}$ por la expresión en (4.33), obtenemos lo siguiente

$$X_{t,L} = \omega_C - \omega_R + (\boldsymbol{\Psi}_{1,C} - \boldsymbol{\Psi}_{1,R})\, \boldsymbol{Z}_{t-1} + \ldots + (\boldsymbol{\Psi}_{p,C} - \boldsymbol{\Psi}_{p,R})\, \boldsymbol{Z}_{t-p} + \varepsilon_{t,C} - \varepsilon_{t,R}. \tag{4.36}$$

Si en esta ecuación sustituimos $\boldsymbol{Z}_{t-i}$ por su equivalente

$$\boldsymbol{Z}_{t-i} = \begin{pmatrix} \frac{X_{t-i,L} + X_{t-i,U}}{2} \\ \frac{X_{t-i,U} - X_{t-i,L}}{2} \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{-1}{2} & \frac{1}{2} \end{pmatrix} \begin{pmatrix} X_{t-i,L} \\ X_{t-i,U} \end{pmatrix}, \tag{4.37}$$

obtenemos el mínimo en función del mínimo y del máximo

$$\begin{aligned} X_{t,L} = \omega_C - \omega_R + (\boldsymbol{\Psi}_{1,C} - \boldsymbol{\Psi}_{1,R}) \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{-1}{2} & \frac{1}{2} \end{pmatrix} \boldsymbol{Y}_{t-1} + \ldots \\ + (\boldsymbol{\Psi}_{p,C} - \boldsymbol{\Psi}_{p,R}) \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{-1}{2} & \frac{1}{2} \end{pmatrix} \boldsymbol{Y}_{t-p} + \varepsilon_{t,C} - \varepsilon_{t,R}. \end{aligned} \tag{4.38}$$

Si procedemos de forma análoga para la predicción del máximo, tomando la siguiente expresión

$$X_{t,U} = X_{t,C} + X_{t,R}, \tag{4.39}$$

y reemplanzando en ella $\hat{X}_{t,C}$ por la expresión en (4.32) y $\hat{X}_{t,R}$ por la expresión en (4.33), y operamos como hemos hecho para la ecuación del mínimo, obtenemos el máximo en función del mínimo y del máximo

$$\begin{aligned} X_{t,U} = \omega_C + \omega_R + (\boldsymbol{\Psi}_{1,C} + \boldsymbol{\Psi}_{1,R}) \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{-1}{2} & \frac{1}{2} \end{pmatrix} \boldsymbol{Y}_{t-1} + \ldots \\ + (\boldsymbol{\Psi}_{p,C} + \boldsymbol{\Psi}_{p,R}) \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{-1}{2} & \frac{1}{2} \end{pmatrix} \boldsymbol{Y}_{t-p} + \varepsilon_{t,C} + \varepsilon_{t,R}. \end{aligned} \tag{4.40}$$

Las ecuaciones del modelo VAR del extremo inferior y del extremo superior mostradas en (4.38) y (4.40), que han sido obtenidas a partir del modelo VAR del centro y radio estimado mediante mínimos cuadrados, son idénticas a las que obtenemos si estimamos directamente el modelo VAR sobre las series de los extremos inferiores y de los extremos superiores.

Se puede proceder de forma similar para obtener las ecuaciones del modelo VAR del centro y radio a partir de las ecuaciones del modelo VAR del extremo inferior y del extremo superior. Por razones de espacio, el procedimiento no será desarrollado aquí.

Explicación. Como es bien sabido, el predictor que minimiza el error cuadrático medio de predicción de un valor futuro $t+k$ se obtiene tomando su esperanza condicionada a los datos observados hasta el instante t, es decir

$$\hat{Y}_{t+k} = \mu_{t+k|t} = E[Y_{t+k}|\{Y_t\}], \tag{4.41}$$

donde $\{Y_t\}$ con $t = 1, ..., n$ es la serie temporal de longitud n para la que se quiere prever el valor futuro. Al estimar un modelos VAR mediante mínimos cuadrados ordinarios, obtenemos un predictor óptimo que representa la esperanza de las variables del modelo condicionada a la información disponible.

Para nuestros modelos VAR tenemos las siguientes esperanzas

$$\begin{array}{l} E[X_{t+k,L}|\{\boldsymbol{Y}_t\}],\ E[X_{t+k,U}|\{\boldsymbol{Y}_t\}], \\ E[X_{t+k,C}|\{\boldsymbol{Z}_t\}],\ E[X_{t+k,R}|\{\boldsymbol{Z}_t\}], \end{array} \tag{4.42}$$

donde $\{\boldsymbol{Y}_t\}$ y $\{\boldsymbol{Z}_t\}$ con $t = 1, ..., n$ son las series temporales de los vectores $\boldsymbol{Y}_t = (X_{t,L}\ X_{t,U})^t$ y $\boldsymbol{Z}_t = (X_{t,C}\ X_{t,R})^t$, respectivamente.

Por otro lado, una de las propiedades de la esperanza es la siguiente

$$E[A + B] = E[A] + E[B]. \tag{4.43}$$

Dada esta propiedad y las equivalencias mostradas en (4.34) obtenemos las siguientes equivalencias para las esperanzas condicionadas

$$E[X_{t+k,L}|\{\boldsymbol{Y}_t\}] = E[X_{t+k,C}|\{\boldsymbol{Z}_t\}] - E[X_{t+k,R}|\{\boldsymbol{Z}_t\}], \tag{4.44}$$

$$E[X_{t+k,U}|\{\boldsymbol{Y}_t\}] = E[X_{t+k,C}|\{\boldsymbol{Z}_t\}] + E[X_{t+k,R}|\{\boldsymbol{Z}_t\}], \tag{4.45}$$

$$E[X_{t+k,C}|\{\boldsymbol{Z}_t\}] = \frac{E[X_{t+k,L}|\{\boldsymbol{Y}_t\}] + E[X_{t+k,U}|\{\boldsymbol{Y}_t\}]}{2} \text{ y} \tag{4.46}$$

$$E[X_{t+k,R}|\{\boldsymbol{Z}_t\}] = \frac{E[X_{t+k,U}|\{\boldsymbol{Y}_t\}] - E[X_{t+k,L}|\{\boldsymbol{Y}_t\}]}{2}. \tag{4.47}$$

Estas equivalencias explican que los modelos que se obtienen estimando el VAR mediante mínimos cuadrados para los extremos inferiores y superiores y para el centro y del radio son equivalentes y que, por tanto, también lo son sus predicciones.

En realidad, debido a las propiedades de la esperanza, esto se cumple para cualquier conjunto de variables que sea combinación lineal de un conjunto de variables originales. Por ello, un modelo VAR (p) estimado para, por ejemplo, las series del centro y del mínimo, también obtendrá las mismas predicciones que el modelo VAR (p) para las series del mínimo y del máximo o para las series del centro y del radio.

4.5.2.2. La cointegración entre las series temporales del mínimo y del máximo

A la hora de plantear un modelo VAR es importante determinar si las variables que lo integran se encuentran cointegradas o no. Si es así, el modelo debería recoger la relación de cointegración y ser planteado como un modelo vectorial de error corrección (VECM). Puede verse una breve introducción al concepto de cointegración y a los VECM en el punto A.1.1 de los apéndices.

Tal y como muestra Cheung (2007) para el ámbito financiero, es razonable pensar que en una STI las series de los extremos inferiores $\{X_{t,L}\}$ y superiores $\{X_{t,U}\}$ pueden estar cointegradas. Intuitivamente, estas series responden al concepto de cointegración ya que se comportan de forma similar a lo largo del tiempo y, aunque su separación pueda aumentar en algunos momentos, es de esperar que lo hace para reducirse posteriormente. En este caso, se da también la peculiaridad de que por definición las series de los extremos nunca se cruzan, pero esta restricción no tiene por qué cumplirse en otras series donde haya cointegración.

Curiosamente, en su trabajo Cheung (2007) compara el modelado de las series de los extremos con el modelado de la serie temporal de los rangos (siendo el rango la diferencia entre los extremos máximo y el mínimo), pero no considera el centro del intervalo. Al modelar el rango sin tener en cuenta el centro no tiene la información del intervalo completa. Por ello, su conclusión es que el rango por si solo tiene menos poder explicativo que un modelo que considere los máximos y los mínimos. Esta conclusión resulta obvia desde la perspectiva de las STI, perspectiva que no toma Cheung (2007).

Recordamos que dos series temporales $\{X_t\}$ y $\{Y_t\}$ están cointegradas si ambas series son integradas de orden d, $I(d)$, y existe una combinación lineal entre ellas que es de un orden de integración menor $I(d_1)$ con $d_1 < d$. En otras palabras, dichas series están cointegradas si podemos construir una serie $\{M_t\}$ que sea $I(d_1)$ y que se obtenga como

$$M_t = \alpha_1 Y_t + \alpha_2 X_t. \tag{4.48}$$

Dada la relación de cointegración definida por (α_1, α_2), cualquier relación del tipo $(c\alpha_1, c\alpha_2)$ es también de cointegración para $c \neq 0$. En las STI, se ha comprobado empíricamente que la relación de cointegración suele ser

$$M_t = X_{t,U} - X_{t,L}, \tag{4.49}$$

donde M_t es estacionario. La parte derecha de la ecuación representa al rango de la serie, que suele ser estacionario. En realidad, si en lugar de considerar la relación de cointegración definida por (1,-1), multiplicamos dicha relación por $c = 0.5$ obtenemos

$$M'_t = 0.5X_{t,U} - 0.5X_{t,L}. \tag{4.50}$$

En esta relación se puede ver que $\{M'_t\} = \{X_{t,R}\}$. En los casos en los que rija dicha relación de cointegración, las series temporales de los radios y de los rangos serán estacionarias. Al existir cointegración, tiene sentido incorporar dicha relación al modelo VAR y plantearlo como un modelo vectorial de corrección del error.

La figura 4.1 muestra un ejemplo de una STI real donde se da esta relación. La imagen muestra las series de los extremos inferiores y superiores que no son estacionarias y cuyo comportamiento tiende a no distanciarse demasiado y la serie del radio que sí es estacionaria. Evidentemente, pueden darse casos en los que no se cumpla esta relación. Por ejemplo, puede suceder que el radio del intervalo presente una cierta tendencia o que no tenga variabilidad constante y que, por tanto, no sea estacionario. Sin embargo, tal y como se verá más adelante en el apartado 4.10, en la práctica es muy habitual que las series de los extremos inferiores y superiores tengan el mismo orden de integración y que la serie de los radios sea estacionaria. En estos casos, entre ellas rige una relación como la expresada en (4.49).

Otro aspecto reseñable que se suele presentar en la práctica es que la serie temporal de los centros suele presentar el mismo orden de integración que la de los extremos superiores e inferiores, lo cual no es de extrañar ya que cumple la relación $X_{t,U} \geq X_{t,C} \geq X_{t,L}$.

La cointegración entre las series temporales de los precios mínimo, máximo y de cierre en finanzas. Fiess y MacDonald (2002) analizan la cointegración entre este tipo de series para el caso del cambio de divisas. Estos autores detectan que existe una relación de cointegración equivalente a la mostrada en (4.49) y otra del tipo

$$N_t = (X_{t,S} - X_{t,L}) - \alpha(X_{t,U} - X_{t,L}), \tag{4.51}$$

donde $X_{t,S}$ es el valor de cierre en el periodo t. Tal y como apuntan estos autores, esta relación guarda cierta similitud con el oscilador estocástico, que es una herramienta que utilizan los analistas técnicos para conocer la posición de una cotización con respecto a su máximo y a su mínimo en un periodo determinado. La fórmula del oscilador estocástico es la siguiente

$$O_t = \frac{X_{t,S} - X'_L}{X'_U - X'_L}, \tag{4.52}$$

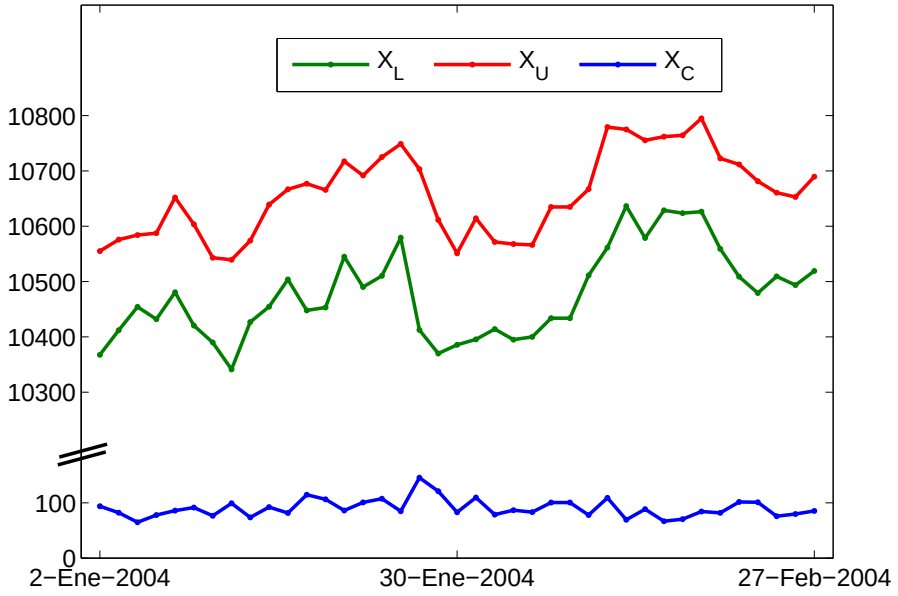


Figura 4.1: Series temporales de los extremos inferiores (X_L), superiores (X_U) y de los centros (X_C) del índice Dow Jones diario.

donde X'_U y X'_L son el valor máximo y mínimo obtenido durante el periodo de tiempo que se está analizando. La similitud con la ecuación (4.51) resulta evidente.

Según Fiess y MacDonald (2002), en los conjuntos de datos que analizan no son capaces de encontrar simultáneamente las relaciones de cointegración mostradas en (4.49) y en (4.51). Sin embargo, en un trabajo anterior (Fiess y MacDonald, 1999) sí que identificaban ambas relaciones entre las series de los valores de cierre diarios y las series de los máximos y los mínimos periódicos (pero no diarios). En esta tesis el propósito es la predicción de STI y por ello no se tendrá en cuenta la serie de los valores de cierre, pero resulta interesante conocer esta relación.

4.6. Predicción mediante alisados exponenciales

Los métodos de alisado permiten predecir una serie temporal como un promedio ponderado de n valores pasados de la propia serie. Tal y como se muestra en el punto A.2 de los apéndices, los alisados se subdividen en dos grupos: las medias móviles y los alisados exponenciales. Los primeros se emplean para eliminar las fluctuaciones espurias de la serie y los segundos como métodos de predicción. En este apartado, se abordarán los segundos, que son los que interesan de cara a cumplir el objetivo de la tesis.

Los alisados exponenciales, pese a estar basados en un concepto muy sencillo como es el del promedio ponderado, obtienen resultados más que aceptables en las competiciones de predicción y son el método a batir a la hora de generar predicciones automáticas (Gardner, 2006). Su sencillez conceptual y su notable capacidad predictiva les hacer ser unos candidatos óptimos para ser adaptados al contexto de las STI.

Para realizar la adaptación se empleará la aritmética de intervalos propuesta por Moore (1966), pueden encontrarse más detalles sobre la aritmética de intervalos en el punto 2.7.1.1. La mayoría de los métodos que se propondrán en este apartado han sido presentados en Arroyo, Muñoz San Roque, Maté y Sarabia (2007).

4.6.1. La adaptación del alisado mediante la aritmética de intervalos

A continuación, se va a mostrar cómo adaptar las ecuaciones de las medias móviles y del alisado exponencial en forma recursiva mostradas en el apartado A.2 del apéndice A. La adaptación sólo requiere reemplazar los valores reales por intervalos y la aritmética clásica por la aritmética de intervalos.

Adaptación de la media móvil. Dada un serie temporal de intervalos $\{[X]_t\}$, la predicción para el instante $t+1$ producida por una media móvil de orden q es una suma ponderada de los últimos q intervalos de la serie

$$[\hat{X}]_{t+1} = \omega_1[X]_t + \omega_2[X]_{t-1} + \ldots + \omega_q[X]_{t-(q-1)}, \tag{4.53}$$

de manera que $\sum_{i=1}^{q} \omega_i = 1$ y $\omega_i \geq 0, \forall i$.

En una media móvil simple de orden n se asigna el mismo peso a cada valor $\omega_i = \frac{1}{q}$. En la media móvil con pesos aritméticamente decrecientes, los pesos son $\omega_i = \frac{q-i+1}{\sum_{i=1}^{q} i}$ y en la media móvil con pesos exponencialmente decrecientes, los pesos son $\omega_i = \alpha(1-\alpha)^{i-1}$ donde $\alpha = \frac{2}{q+1}$.

Adaptación del alisado exponencial. Tal y como se muestra en el apéndice A.2, la ecuación del alisado exponencial simple para series temporales clásicas puede formularse de dos maneras equivalentes entre sí: la recursiva y en forma de corrección del error. Sin embargo, si adaptamos ambas expresiones empleando la aritmética de intervalos, el resultado no es equivalente.

La fórmula del alisado exponencial simple para STI en forma de corrección del error viene dada por

$$[\hat{X}]_{t+1} = [\hat{X}]_t + \alpha[E]_t, \tag{4.54}$$

donde $[E]_t$ representa el error cometido en el instante t, $[E]_t = [X]_t - [\hat{X}]_t$, y donde $\alpha \in [0,1]$. Mientras que su equivalente recursiva se expresa como

$$[\hat{X}]_{t+1} = \alpha[X]_t + (1-\alpha)[\hat{X}]_t. \tag{4.55}$$

Según la propiedad subdistributiva de la aritmética de intervalos (mostrada en el apartado 2.7.1.1), la relación entre ambas fórmulas es la siguiente

$$\alpha[X_t] + (1-\alpha)[\hat{X}]_t \subseteq [\hat{X}]_t + \alpha([X_t] - [\hat{X}_t]). \tag{4.56}$$

Recordamos que la propiedad subdistributiva implica que dados tres intervalos $[A]$, $[B]$ y $[C]$ se cumple que

$$[A]([B] + [C]) \subseteq [A][B] + [A][C], \tag{4.57}$$

y que la propiedad subdistributiva se torna en distributiva si $[A]$ es un valor real y no un intervalo. Teniendo en cuenta esta propiedad, es fácil demostrar que

$$\alpha[X]_t + (1-\alpha)[\hat{X}]_t \subseteq \alpha[X]_t - \alpha[\hat{X}]_t + [\hat{X}]_t = [\hat{X}]_t + \alpha([X]_t - [\hat{X}]_t). \tag{4.58}$$

Esto quiere decir que la expresión recursiva del alisado produce predicciones más concisas y que, por tanto, es más adecuada para adaptar el alisado exponencial para predecir STI. Por ello, para generar predicciones para una STI se aconseja emplear el alisado exponencial según la fórmula recurrente (4.55).

Además, el resultado que se obtiene prediciendo según la fórmula recurrente es el mismo que se obtiene mediante su equivalente media móvil de pesos exponencialmente decrecientes; al igual que sucede en los alisados exponenciales para series temporales clásicas. Sin embargo, esto no se cumple si se emplea la ecuación en forma de corrección del error. Más formalmente, si en la ecuación recursiva se sustituye sucesivamente el término $[\hat{X}]_t$ se obtiene la expresión

$$[\hat{X}]_{t+1} = \sum_{j=1}^{t} \alpha(1-\alpha)^{j-1}[X]_{t-(j-1)} \tag{4.59}$$

que puede ser reescrita como una media móvil de pesos exponencialmente decrecientes de orden q

$$[\hat{X}]_{t+1} = \frac{[X_t] + (1-\alpha)[X]_{t-1} + \ldots + (1-\alpha)^{q-1}[X]_{t-(q-1)}}{1 + (1-\alpha) + \ldots + (1-\alpha)^{q-1}}, \tag{4.60}$$

donde $\alpha = \frac{2}{q+1}$, ya que si q, es lo suficientemente grande, entonces $1 + (1-\alpha) + \ldots + (1-\alpha)^{q-1} \simeq \alpha^{-1}$.

Tal y como se indicó en el apartado 2.7.1.1, la aritmética de intervalos subsume a la aritmética clásica. Por tanto, los métodos de alisado basados en aritmética de intervalos generalizan los métodos de alisado sobre números reales para poder tratar con STI.

A continuación, se va a analizar el efecto del alisado sobre una STI.

4.6.2. Análisis del efecto del alisado basado en aritmética de intervalos

Cuando se realiza el alisado de una serie temporal clásica, el objetivo es eliminar las fluctuaciones de la misma y obtener una representación que tenga una desviación típica menor. El alisado de una STI empleando aritmética de intervalos cumple el mismo objetivo. A continuación, se analiza cómo se comporta el promedio de intervalos y la ecuación recursiva de alisado para poder entender el efecto que tendrán sobre una STI.

Promedio de intervalos. Dado un conjunto de intervalos $[X]_i$ con $i = 1, ..., q$, el intervalo resultado de promediar dicho conjunto se calcula como

$$[\bar{X}] = \frac{[X]_1 + ... + [X]_q}{q}, \tag{4.61}$$

donde las operaciones aritméticas requeridas son las descritas para la aritmética de intervalos en el apartado 2.7.1.1.

El promedio de un conjunto de intervalos presenta las siguientes características

$$\begin{aligned} \bar{X}_L = \frac{X_{L,1} + X_{L,2} + ... + X_{L,n}}{n}, \quad \bar{X}_U = \frac{X_{U,1}+X_{U,2}+...+X_{U,n}}{n} \\ \bar{X}_C = \frac{X_{C,1} + X_{C,2} + ... + X_{C,n}}{n}, \text{ y } \quad \bar{X}_R = \frac{X_{R,1}+X_{R,2}+...+X_{R,n}}{n}. \end{aligned} \tag{4.62}$$

Esto quiere decir que la media de un conjunto de intervalos es igual a la media de sus componentes. En el apartado 4.8.2.1, se mostrará la relación entre el intervalo promedio y el concepto de baricentro.

La ecuación recursiva de alisado para STI. Por simplificar, reescribiremos la ecuación recursiva del alisado exponencial mostrada en (4.55) como

$$[C] = \alpha[A] + (1 - \alpha)[B], \tag{4.63}$$

donde $[A] = [X]_t$, $[B] = [\hat{X}]_t$ y $[C] = [\hat{X}]_{t+1}$.

El intervalo que resulta al aplicar la ecuación recursiva del alisado exponencial presenta las siguientes características

$$\begin{aligned} C_L = \alpha A_L + (1 - \alpha)B_L, \quad C_U = \alpha A_U + (1 - \alpha)B_U, \\ C_C = \alpha A_C + (1 - \alpha)B_C, \quad \text{y } C_R = \alpha A_R + (1 - \alpha)B_R. \end{aligned} \tag{4.64}$$

Este resultado se ajusta perfectamente al comportamiento que cabría esperar del alisado exponencial ya que todos los componentes se ven afectados por el alisado de igual forma.

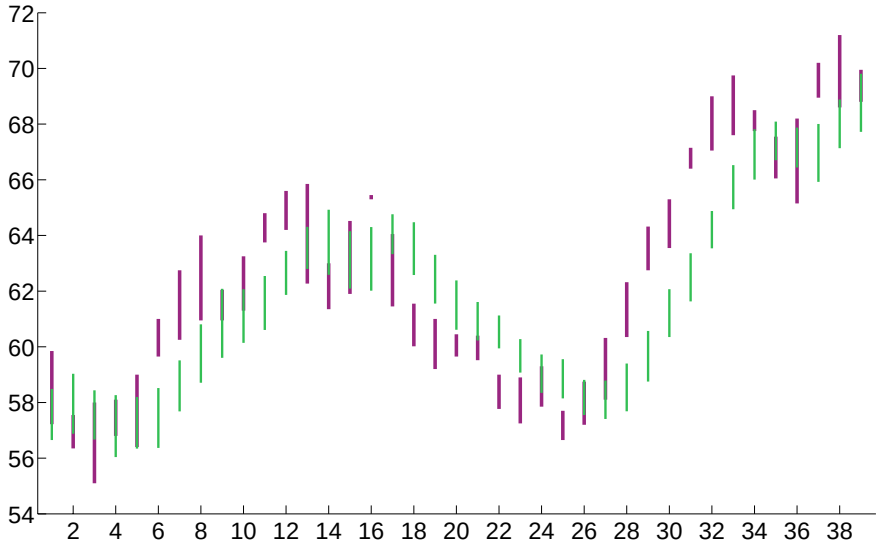


Figura 4.2: STI real (trazo morado) y suavizada (trazo verde) empleando el alisado exponencial basado en aritmética de intervalos con $\alpha = 0.4$.)

El efecto de alisado en una STI. Por último, se va a mostrar cuál es el resultado que se obtiene al aplicar la ecuación recursiva del alisado exponencial (4.55) sobre una STI. En la figura 4.2 se muestra la serie original y la serie alisada con $\alpha = 0.4$. Puede apreciarse que en la serie alisada las posiciones de los intervalos fluctúan menos que la serie original. También se puede ver que la serie alisada presenta una menor variabilidad en el ancho de los intervalos. Ambos efectos, es decir, el alisado en los centros y en los radios de los intervalos, se corresponden a lo que cabe esperar al realizar el alisado de la serie. Los valores extremos en ambas características de los intervalos se compensan con otros valores no tan extremos y, como resultado, se suavizan.

4.6.3. Métodos de alisado exponencial basados en la aritmética de intervalos

Dada la STI $\{[X]_t\}$ con $t = 1, ..., n$, se van a proponer un conjunto de métodos de alisado exponencial. Entre ellos, se encuentran: un alisado simple para las STI en las que no haya ni tendencia, ni estacionalidad; un alisado que permite recoger la tendencia de la serie y otro alisado que recoge la tendencia de la serie en forma atenuada; y dos alisados con estacionalidad, uno para el caso en el que la estacionalidad sólo afecte a la posición del intervalo y otro para el caso en el que también afecte al ancho del mismo. En todos los casos, la predicción se obtiene de forma aditiva, i.e., sumando cada una de los componentes de la serie.

Estos métodos son una adaptación de los alisados para series temporales clásicas, puede verse una descripción de éstos en el apéndice A.2. La notación de los métodos propuestos respeta la notación propuesta por Gardner (2006). La principal diferencia es que algunos de los términos serán del tipo intervalo y que la aritmética que rige en las ecuaciones es la aritmética de intervalos propuesta por Moore (1966).

Tal y como se ha dicho en el apartado 4.6.1, la aritmética de intervalos no permite adaptar de forma satisfactoria la ecuación del alisado exponencial en forma de corrección del error al contexto de las STI. Por ello, los métodos que se presentarán en este apartado están expresados en forma recursiva.

4.6.3.1. Alisado exponencial simple

El alisado exponencial simple (AES) para STI se define como

$$[\hat{X}]_{t+1} = \alpha[X_t] + (1-\alpha)[\hat{X}]_t, \tag{4.65}$$

donde $\alpha \in [0,1]$. La inicialización del método precisa del valor $[\hat{X}]_1$, que puede ser el primer valor observado, la media, empleando aritmética de intervalos, de los tres o cuatro primeros valores de la serie o puede ser estimada mediante *backcasting* (o proyección inversa). El *backcasting* consiste en invertir la serie, de forma que $\{[X_{t'}]\}$ con $t' = n, ..., 1$, y predecirla hasta obtener la predicción para $t' = 1$ que se empleará como valor inicial

4.6.3.2. Alisado exponencial con tendencia

El alisado exponencial simple no es capaz de predecir adecuadamente series donde la posición del intervalo siga una tendencia. Para ello, es necesario un método que sea capaz de alisar dicha tendencia e incorporarla a la predicción, de forma similar a como hace el alisado exponencial propuesto por Holt (1957) para las series temporales clásicas.

El método que se propondrá alisará tanto el nivel de la STI que se representará como un intervalo, como la tendencia de la serie que recogerá los cambios en la posición del intervalo mediante un número real. Para representar la posición del intervalo se ha utilizará el centro del intervalo. Podría considerarse también el extremo inferior o superior del mismo, pero éstos pueden tomar valores extremos que desvirtúen la posición del histograma.

El alisado exponencial con tendencia aditiva (AET) consta de una ecuación para suavizar el nivel de la STI, $[S]_t$, y otra suavizar la tendencia en t, T_t. La predicción se obtiene como la suma de la componente de la tendencia y la componente del nivel. Más concretamente, el AET se define como

$$\begin{aligned} [S]_t &= \alpha[X]_t + (1-\alpha)([S]_{t-1} + T_{t-1}), && (4.66)\\ T_t &= \gamma(S_{t,C} - S_{t-1,C}) + (1-\gamma)T_{t-1}, && (4.67)\\ [\hat{X}]_{t+m} &= [S]_t + mT_t, && (4.68) \end{aligned}$$

donde $\alpha, \gamma \in [0,1]$, $S_{t,C}$ es el centro del intervalo $[S]_t$, y m es un factor multiplicador que sirve para obtener la predicción futura para el periodo $t+m$. Pueden tomarse $T_1 = X_{2,C} - X_{1,C}$ y $[S]_1 = [X]_1$ como valores iniciales o pueden obtenerse mediante *backcasting*.

4.6.3.3. Alisado exponencial con tendencia atenuada

Según Gardner y McKenzie (1985), en las series temporales clásicas, el alisado exponencial de Holt tiende a sobreestimar el valor de la tendencia real. Por ello, estos autores proponen utilizar un parámetro adicional, ϕ, para atenuar la tendencia.

Tomando como referencia el AET definido en el apartado anterior, la adaptación de este nuevo modelo es directa. El alisado exponencial con tendencia atenuada aditiva (AETA) se define como

$$[S]_t = \alpha[X]_t + (1-\alpha)([S]_{t-1} + \phi T_{t-1}), \tag{4.69}$$

$$T_t = \gamma(S_{t,C} - S_{t-1,C}) + (1-\gamma)\phi T_{t-1}, \tag{4.70}$$

$$[\hat{X}]_{t+m} = [S]_t + \sum_{i=1}^{m} \phi^i T_t, \tag{4.71}$$

donde $\alpha, \gamma \in [0,1]$, $\phi \geq 0$, $S_{t,C}$ es el centro del intervalo $[S]_t$, y m indica el número de periodos futuros para los que se calculará la predicción. Si $\phi = 0$, el método es idéntico al AES mostrado en la ecuación (4.65). Si $\phi \in (0,1)$, la tendencia será atenuada en mayor medida cuanto más próximo sea ϕ a cero. Si $\phi = 1$ el método es el AET presentado en el apartado anterior. Por último, si $\phi > 1$, la predicción tiene tendencia exponencial.

4.6.3.4. Alisado exponencial con estacionalidad

En una STI pueden considerarse al menos dos tipos de estacionalidad. Una en la que la variación estacional sólo afecta a la localización del intervalo, y otra en la que la estacionalidad afecta a todo el intervalo. A continuación, se mostrarán dos ejemplos uno con cada uno de los dos tipos de estacionalidad.

En la STI mostrada en la figura 4.3, se representan las temperaturas mensuales mínima y máxima registradas en China entre enero de 1952 y diciembre de 1957. En la serie mostrada, la estacionalidad es muy patente y afecta a todo el intervalo, tanto a su localización, como a su ancho.

La figura 4.4 muestra la STI de los niveles mínimo y máximo del cauce del río islandés Jökulsá á Fjöllum entre enero de 1972 y diciembre de 1974. Si se observa la serie con atención se verá que el ancho de los intervalos, no siguen claramente un patrón estacional, ya que existe una gran diferencia entre el ancho de los intervalos de algunos meses, e.g. los meses de enero, julio y noviembre. Sin embargo, la posición del intervalo sí parece seguir más fielmente un patrón estacional.

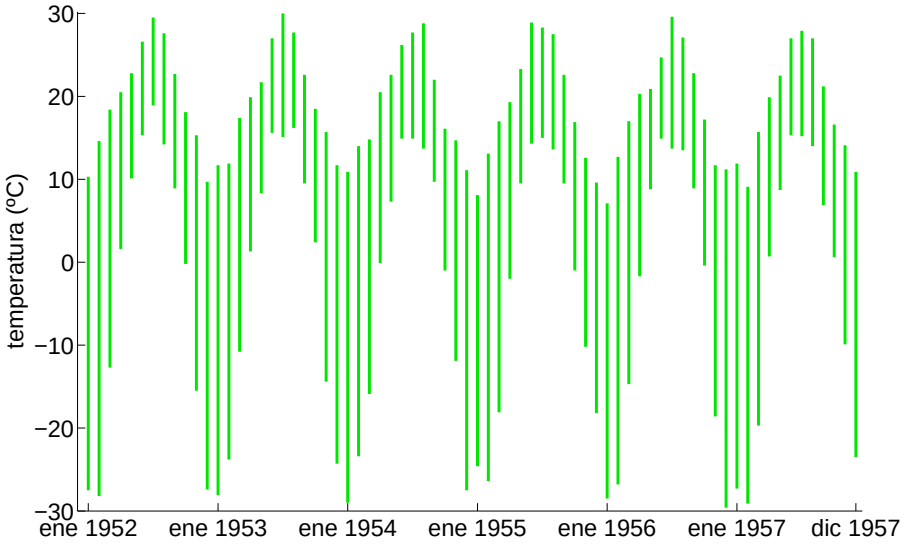


Figura 4.3: STI que representa las temperaturas mensuales mínima y máxima registradas en China.

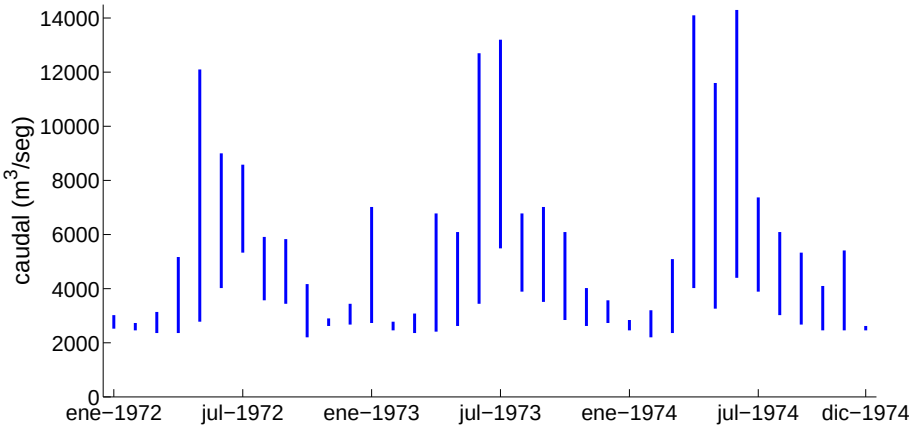


Figura 4.4: STI que representa el cauce mínimo y máximo mensual del río Jökulsá á Fjöllum.

Cada uno de los tipos de estacionalidad, da lugar a un alisado exponencial con estacionalidad diferente.

Estacionalidad que sólo afecta a la posición del intervalo. En esta aproximación, la componente estacional en t se representa como un número real, I_t, que representa los cambios en la localización del intervalo a lo largo del ciclo estacional. La localización del intervalo, al igual que cuando se modela la tendencia, está representada por el centro del intervalo. Por su parte el nivel de la serie en t, es un intervalo desestacionalizado que se representa como $[S]_t$. Las predicciones se obtienen como la suma del intervalo del nivel y de la componente estacional.

El alisado exponencial con estacionalidad clásica aditiva (AEEc) se representa como

$$\begin{aligned} [S_t] &= \alpha([X]_t - I_{t-p}) + (1-\alpha)[S]_{t-1}, && (4.72)\\ I_t &= \delta(X_{t,C} - S_{t,C}) + (1-\delta)I_{t-p}, && (4.73)\\ [\hat{X}]_{t+1} &= [S]_t + I_{t-p+1}, && (4.74) \end{aligned}$$

donde $\alpha, \delta \in [0,1]$ y p es la longitud del ciclo estacional. Para inicializar la serie son necesarios los p primeros periodos de la misma. La inicialización puede realizarse de distintas formas, una de ellas es mediante un *backcasting* cuyos valores iniciales sean a su vez

$$\begin{aligned} [S]_{n-(p-1)} &= \frac{[X]_n + [X]_{n-1} + ... + [X]_{n-(p-1)}}{p},\\ I_n &= X_{n,C} - S_{n-(p-1),C},\\ I_{n-1} &= X_{n-1,C} - S_{n-(p-1),C}, \dots\\ I_{n-(p-1)} &= X_{n-(p-1),C} - S_{n-(p-1),C}, \end{aligned} \qquad (4.75)$$

Estacionalidad que afecta a todo el intervalo. En el segundo enfoque, el nivel de la serie, S_t, se representa como un valor real y la estacionalidad, $[I]_t$, se representa como un intervalo.

El alisado exponencial con estacionalidad aditiva en forma de intervalo (AEEi) se define como

$$\begin{aligned} S_t &= \alpha(X_{c,t} - I_{c,t-p}) + (1-\alpha)S_{t-1}, && (4.76)\\ [I]_t &= \delta([X]_t - S_t) + (1-\delta)[I]_{t-p}, && (4.77)\\ [\hat{X}]_{t+1} &= S_t + [I]_{t-p+1} && (4.78) \end{aligned}$$

donde $\alpha, \delta \in [0,1]$, y p es la longitud del ciclo estacional. Al igual que en el método anterior, los p primeros valores son necesarios para inicializar el

método. Una posible forma de realizar la inicialización consiste en realizar *backcasting* utilizando como valores iniciales

$$\begin{aligned} S_{n-(p-1)} &= \frac{X_{n,C} + X_{n-1,C} + ... + X_{n-(p-1),C}}{p}, \\ [I]_n &= [X]_n - S_{n-(p-1)}, \\ [I]_{n-1} &= [X]_{n-1} - S_{n-(p-1)}, \dots \\ [I]_{n-(p-1)} &= [X]_{n-(p-1)} - S_{n-(p-1)}, \end{aligned} \tag{4.79}$$

4.7. Predicción mediante el perceptron multicapa para intervalos

Muñoz San Roque et al. (2007) proponen un perceptrón multicapa que permite tratar datos de intervalo tanto de entrada, como de salida. Dicho perceptrón recibe el nombre de iMLP, acrónimo de *interval Multi-Layer Perceptron*, y es una adaptación al contexto de los datos de intervalo del perceptrón multicapa (Bishop, 1995). La principal diferencia reside en que, como el iMLP trabaja con intervalos, las operaciones que se realizan en él se rigen según la aritmética de intervalos propuesta por Moore (1966). Pueden consultarse los principios básicos de dicha aritmética en el apartado 2.7.1.1 de esta tesis.

A continuación, se detalla la estructura y el proceso de aprendizaje en el iMLP y cómo se realizan predicciones con él.

4.7.1. Estructura del iMLP

La estructura de un iMLP con n entradas en forma de intervalo, una capa oculta con h neuronas y una salida en forma de intervalo ($m = 1$) se muestra en la figura 4.5. La generalización de esta estructura para permitir más capas ocultas o más neuronas de salida es directa.

Sean n intervalos de entrada $[X]_i = \langle X_{i,C}, X_{i,R} \rangle = [X_{i,C} - X_{i,R}, X_{i,C} + X_{i,R}]$ con $i = 1, .., n$, la salida de la neurona j-ésima de la capa oculta es

$$[S]_j = w_{j0} + \sum_{i=1}^{n} w_{ji}[X]_i = \left\langle w_{j0} + \sum_{i=1}^{n} w_{ji} X_{i,C}, \sum_{i=1}^{n} |w_{ji}| X_{i,R} \right\rangle \tag{4.80}$$

donde los pesos w_{ji} no son intervalos, sino números reales, y donde $j = 1, ..., h$.

La activación de la neurona j-ésima se obtiene transformando el intervalo $[S]_j$ con una función no lineal como la tangente hiperbólica, tanh. Como

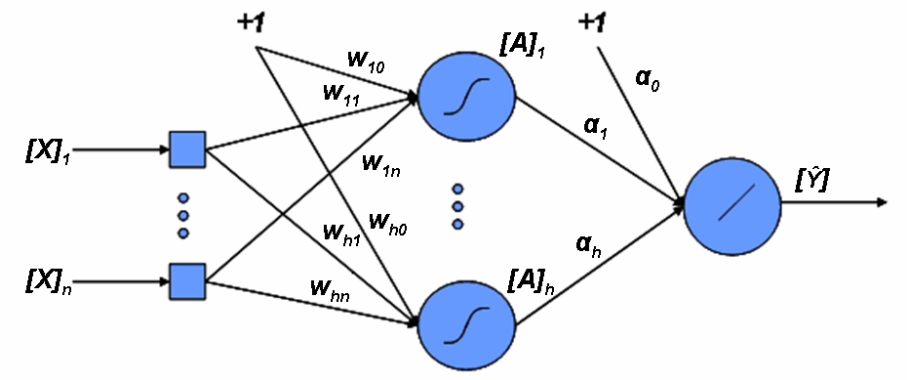


Figura 4.5: Estructura del iMLP con n entradas, una capa oculta de h neuronas y una neurona de salida

dicha función es monótona creciente, la salida de la neurona j-ésima de la capa oculta viene dada por

$$[A]_j = \tanh([S]_j) = [\tanh(S_{j,C} - S_{j,R}), \tanh(S_{j,C} + S_{j,R})] = \\ = \left\langle \frac{\tanh(S_{j,C}-S_{j,R})+\tanh(S_{j,C}+S_{j,R})}{2}, \frac{\tanh(S_{j,C}+S_{j,R})-\tanh(S_{j,C}-S_{j,R})}{2} \right\rangle .$$

La salida de la red, $[\hat{Y}]$, se obtiene realizando una combinación lineal de los valores devueltos por la capa oculta más la constante

$$[\hat{Y}] = \sum_{j=1}^{h} \alpha_j [A]_j + \alpha_0 = \left\langle \sum_{j=1}^{h} \alpha_j A_{j,C} + \alpha_0, \sum_{j=1}^{h} |\alpha_j| A_{j,R} \right\rangle , \tag{4.81}$$

donde los pesos α_j con $j = 0, ..., h$ son números reales.

4.7.2. Aprendizaje en el iMLP

El iMLP puede emplearse para aproximar funciones de intervalo. Para ello, sus pesos deben ser estimados mediante un procedimiento de aprendizaje supervisado que tenga como objetivo la minimización de una función de error de la forma

$$E = \frac{1}{n} \sum_{i=1}^{n} d([Y]_i, [\hat{Y}]_i) + \lambda \Phi(\hat{f}), \tag{4.82}$$

donde n el numero de elementos del conjunto de entrenamiento, $\lambda\Phi(\hat{f})$ es un término de regularización (Girosi, Jones y Poggio, 1995) y $d([Y]_i, [\hat{Y}]_i)$ es una medida de discrepancia entre el valor observado $[Y]_i$ y el valor pronosticado $[\hat{Y}]_i$. La medida de discrepancia aplicada en Muñoz San Roque et al. (2007) es la siguiente

$$d([Y]_i, [\hat{Y}]_i) = \beta(Y_{i,C} - \hat{Y}_{i,C})^2 + (1-\beta)(Y_{i,R} - \hat{Y}_{i,R})^2, \tag{4.83}$$

con $\beta \in [0,1]$. Si β toma valores más cercanos a cero, el aprendizaje se centra en estimar el centro del intervalo, y si toma valores más cercanos a uno, entonces se le asigna más peso al radio.

La minimización de la función de error se realiza mediante un método Quasi-Newton de baja memoria (Luenberg, 1984) donde los pesos de la capa oculta (w) y de la capa de salida (α) de la red se inicializan de forma aleatoria. El método requiere el cálculo del gradiente de la función de error con respecto a los pesos. Esto puede hacerse aplicando un procedimiento de retroprogación (*backpropagation*) similar al propuesto por Rumelhart, Hinton y Williams (1987) para el perceptrón multicapa clásico.

Las derivadas de la función de coste, omitiendo el término de regularización, con respecto a los pesos de la capa de salida son de la forma

$$\frac{\partial E}{\partial \alpha_j} = \frac{2}{n}\sum_{t=1}^{n}\left(\beta(\hat{Y}_{t,C} - Y_{t,C})\frac{\partial \hat{Y}_{t,C}}{\partial \alpha_j} + (1-\beta)(\hat{Y}_{t,R} - Y_{t,R})\frac{\partial \hat{Y}_{t,R}}{\partial \alpha_j}\right) \tag{4.84}$$

donde

$$\frac{\partial \hat{Y}_{t,C}}{\partial \alpha_j} = \begin{cases} 1, & \text{si } j = 0, \\ A_{jt,C}, & \text{si } j > 0. \end{cases} \tag{4.85}$$

y donde

$$\frac{\partial \hat{Y}_{t,R}}{\partial \alpha_j} = \begin{cases} 0, & \text{si } j = 0, \\ sgn(\alpha_j) A_{jt,R}, & \text{si } j > 0. \end{cases} \tag{4.86}$$

Por su parte, las derivadas de la función de coste con respecto a los pesos de la capa oculta vienen dadas por

$$\begin{aligned} \frac{\partial E}{\partial w_{ji}} &= \frac{2}{n}\sum_{t=1}^{n}\left(\beta(\hat{Y}_{t,C} - Y_{t,C})\frac{\partial \hat{Y}_{t,C}}{\partial w_{ji}} + \right. && (4.87) \\ &\quad \left. (1-\beta)(\hat{Y}_{t,R} - Y_{t,R})\frac{\partial \hat{Y}_{t,R}}{\partial w_{ji}}\right) && (4.88) \\ &= \frac{2}{n}\sum_{t=1}^{n}\left(\beta(\hat{Y}_{t,C} - Y_{t,C})\frac{\partial \hat{Y}_{t,C}}{\partial a_{jt,C}}\frac{\partial a_{jt,C}}{\partial w_{ji}} + \right. && (4.89) \\ &\quad \left. (1-\beta)(\hat{Y}_{t,R} - Y_{t,R})\frac{\partial \hat{Y}_{t,R}}{\partial a_{jt,R}}\frac{\partial a_{jt,R}}{\partial w_{ji}}\right), && (4.90) \end{aligned}$$

donde

$$\frac{\partial \hat{Y}_{t,C}}{\partial a_{jt,C}} = \alpha_j, \tag{4.91}$$

$$\begin{aligned} \frac{\partial a_{jt,C}}{\partial w_{ji}} &= \frac{1}{2}\tanh'(s_{jt,C} + s_{jt,R})(X_{it,C} + sgn(w_{ji})X_{it,R}) && (4.92)\\ &+\frac{1}{2}\tanh'(s_{jt,C} - s_{jt,R})(X_{it,C} - sgn(w_{ji})X_{it,R}), && (4.93) \end{aligned}$$

y donde

$$\frac{\partial \hat{Y}_{t,R}}{\partial a_{jt,R}} = |\alpha_j|, \tag{4.94}$$

$$\begin{aligned} \frac{\partial a_{jt,R}}{\partial w_{ji}} &= \frac{1}{2}\tanh'(s_{jt,C} + s_{jt,R})(X_{it,C} + sgn(w_{ji})X_{it,R}) && (4.95)\\ &-\frac{1}{2}\tanh'(s_{jt,C} - s_{jt,R})(X_{it,C} - sgn(w_{ji})X_{it,R}) && (4.96) \end{aligned}$$

4.7.3. El iMLP como método de predicción

El perceptrón multicapa clásico permite resolver problemas de aproximación funcional. En los problemas de aproximación funcional se cuenta con un conjunto de muestras con valores de entrada y de salida y el objetivo es determinar de la forma más precisa posible la función que rige la relación entre las entradas y las salidas. La predicción de series temporales puede enfocarse como un problema de aproximación funcional que puede resolverse mediante un perceptrón multicapa. Ver Zhang et al. (1998) para una revisión en profundidad del tema de la predicción con redes neuronales.

Al ser una extensión del perceptrón multicapa clásico, el iMLP puede emplearse para aproximar funciones donde las entradas y salidas sean intervalos y, por tanto, puede utilizarse para predecir STI. En un contexto univariante, dada la STI $\{[X_t]\}$ con $t = 1, \ldots, n$, la función a estimar por el iMLP es

$$[X]_{t+1} = f([X]_t, [X]_{t-1}, \ldots, [X]_{t-l}), \tag{4.97}$$

donde $[X]_t$ es el intervalo observado en el instante t y l es el número de retardos que se incluyen en el modelo.

Las entradas de la función pueden afinarse más ya que es posible que no sea necesario incluir todos los retardos entre t y $t - l$ en el modelo. Adicionalmente, es importante reseñar que se pueden incorporar al iMLP variables de entrada que no sean retardos de la serie $\{[X_t]\}$, sino valores retardados de otras series que pueden ser de intervalo o no.

A la hora de ajustar el modelo se recomienda, al igual que en el caso de los perceptrones multicapa clásicos, dividir la STI en dos partes: periodo de

entrenamiento y periodo de prueba. El periodo de entrenamiento se emplea para ajustar los pesos del iMLP partiendo de unos pesos iniciales aleatorios, el periodo de prueba se utiliza para medir el error cometido por el iMLP ajustado durante las diferentes épocas del entrenamiento. La configuración de pesos que se emplea en el iMLP será la que menor error de prueba haya obtenido a lo largo de las distintas épocas del entrenamiento.

4.8. Predicción mediante el método de los k vecinos más cercanos

Tal y como se muestra en el apéndice A.3, la idea en la que se basa el método de predicción de los k vecinos más próximos (o k-NN según la abreviatura inglesa de *k-Nearest Neighbours*) consta de dos pasos:

1. Búsqueda de las k secuencias más similares a la actual: para lo cual se suele emplear la distancia euclídea.

2. Obtención de la predicción: para lo cual se suele realizar el promedio de los valores siguientes de las k secuencias determinadas en el paso anterior.

El k-NN es un método clásico de aprendizaje estadístico, pero su versatilidad permite aplicarlo a la predicción de series temporales. La adaptación que se propondrá a continuación permite utilizar el k-NN no solo para predecir STI, sino para realizar reconocimiento de patrones sobre datos de intervalos. Sin embargo, la explicación se referirá al contexto de las STI que es el que concierne a esta tesis.

4.8.1. El método de k-NN para predecir STI

A continuación, se explicarán con mayor detalle los dos pasos en los que consiste la predicción con k-NN utilizando STI y en el punto 4.8.2 se considerarán algunas alternativas para realizar ambos procesos.

4.8.1.1. Determinación de los vecinos más próximos

Para determinar la semejanza entro dos secuencias de una STI se puede emplear alguna de las distancias para intervalos que fueron mencionadas en el apartado 4.4.1.1. En principio, parece más adecuado emplear la distancia definida a partir de un kernel mostrada en la ecuación (4.12) ya que dicha distancia es una distancia de tipo euclídeo, que es el tipo de distancia que se suele emplear en el k-NN clásico. Si consideramos dicha distancia, el proceso a realizar sería el siguiente.

La STI $\{[X]_t\}$ con $t = 1, ..., n$ se transforma en una serie de vectores de intervalo d-dimensionales de la siguiente forma

$$[X]_t^d = ([X]_t, [X]_{t-1}, ..., [X]_{t-d+1}), \tag{4.98}$$

con $t = d, ...n$. A continuación, se calcula la distancia entre el último vector de la serie $[X]_n^d$ y el resto de vectores $[X]_t^d$ con $t = d, ..., n-1$ de la siguiente manera

$$D_k^q([X]_n^d, [X]_t^d) = \left(\frac{\sum_{i=1}^{d} (D([X]_{n-i+1}, [X]_{t-i+1}))^q}{d} \right)^{\frac{1}{q}}, \tag{4.99}$$

donde $D([X]_{n-i+1}, [X]_{t-i+1})$ es una distancia que viene a representar la función que hace el valor absoluto de la diferencia de dos valores al manejar números reales, y el parámetro q indica el orden de la distancia. Como distancia se puede utilizar la distancia definida a partir de un kernel mostrada en la ecuación (4.12), ya que es similar a la distancia euclídea. En cuanto al orden q, si $q = 2$ la discordancia entre los intervalos se agrega al cuadrado, lo que daría lugar a una medida de discrepancia similar a la raíz cuadrada del error cuadrático medio, y si $q = 1$, la medida de discrepancia resultante sería de la forma del error absoluto medio.

Una vez que se han calculado las $n - d$ distancias, se determinan los k vectores más próximos al vector $[X]_n^d$. Dichos vectores se denotarán como $[X]_{T_p}^d$ con $p = 1, ..., k$.

4.8.1.2. Obtención de predicciones

En el k-NN que se emplea en las series temporales clásicas, que fue explicado en el apartado A.3 del apéndice, las predicciones se calculan como la media (ponderada o no) de los valores siguientes de cada una de las k secuencias vecinas.

De cara a adaptar el k-NN para predecir STI, se puede sustituir la media por el intervalo medio calculado utilizando aritmética de intervalos. El intervalo medio o promedio, ver ecuación (4.61), ya fue empleado como herramienta para realizar las medias móviles y el alisado de STI que desarrollados en el apartado 4.6.

En el k-NN para STI, la predicción $[\hat{X}]_{n+1}$ será el intervalo promedio que se obtiene al realizar la media ponderada de los intervalos siguientes de las k secuencias más similares a la actual que fueron determinadas en el paso anterior. Más formalmente, la predicción $[\hat{X}]_{n+1}$ se obtiene como

$$[\hat{X}]_{n+1} = \sum_{p=1}^{k} \omega_p \cdot [X]_{T_p+1}, \tag{4.100}$$

donde $[X]_{T_p+1}$ es el siguiente intervalo de la secuencia $[X]^d_{T_p}$, y ω_p es el peso asignado al vecino p tal que satisface $\omega_p \geq 0$ y $\sum_{p=1}^{k} \omega_p = 1$. Las operaciones se realizan conforme a la aritmética de intervalos mostrada en el apartado 2.7.1.1.

Respecto a las pesos, se puede optar por un esquema de ponderación que asigne a todos los vecinos el mismo peso, i. e. $\omega_p = 1/k \ \forall p$, de forma que en la ecuación (4.100) se calcule el intervalo medio. Alternativamente, también se puede optar por asignar al intervalo p un peso inversamente proporcional a la distancia entre la secuencia actual y la secuencia vecina p tal que

$$\omega_p = \frac{\psi_p}{\sum_{l=1}^{k} \psi_l}, \tag{4.101}$$

con $\psi_p = (D^q_k([X]^d_n, [X]^d_{T_p}) + \xi)^{-1}$ y $p = 1, ..., k$, y donde $D^q_k([X]^d_n, [X]^d_{T_p})$ viene dado por la ecuación (4.99). La constante $\xi = 10^{-8}$ impide que el peso tome el valor infinito, si la distancia entre las secuencias consideradas es cero.

4.8.2. Alternativas al método de k-NN

Como ya se ha mencionado, en el k-NN clásico se usa la distancia euclídea como criterio para determinar los vecinos más cercanos y el promedio para obtener la predicción. En este apartado se propondrán métodos de k-NN alternativos que utilizan otras distancias y otras maneras de hallar la predicción. Para ello, es necesario aclarar antes la relación entre el promedio, la distancia euclídea y el concepto de baricentro.

4.8.2.1. Relación entre el promedio, el baricentro y la distancia euclídea

En física, el baricentro es el centro de masas de un sistema de partículas. En otras palabras, es un punto específico en el que la masa del sistema se comporta como si estuviese concentrada en dicho punto. Analíticamente, las coordenadas del baricentro se obtienen como el promedio ponderado de las coordenadas de las partículas, donde el peso asignado a cada partícula es proporcional a su masa.

Para una definición más formal del baricentro, consideremos un sistema de k partículas en un espacio unidimensional (consideraremos una única dimensión por simplificar) donde cada partícula p_i con $i = 1, ..., k$ está definida por un valor x_i y tiene una masa asociada w_i. El baricentro de este sistema al que denotaremos como p_B se halla de la siguiente forma

$$p_B = x_B = \frac{\sum_{i=1}^{k} x_i \cdot w_i}{\sum_{i=1}^{k} w_i}. \tag{4.102}$$

Si la masa w_i es la misma para todas las partículas, las coordenadas del baricentro se obtienen como la media aritmética de las coordenadas de las partículas en cada una de las dimensiones consideradas.

El concepto de baricentro está relacionado con el concepto de centroide que se emplea en el algoritmo de *clustering* de k medias. En dicho algoritmo, el centroide de un *cluster* se obtiene como la media de los puntos que pertenecen a ese *cluster*, es decir, como la media aritmética de sus coordenadas.

De forma equivalente, el baricentro p_B también cumple que es el punto que minimiza

$$p_B = \min_{p_B} \left(\sum_{i=1}^{k} \omega_i D_{Euclid}(p_B, p_i)^2 \right)^{\frac{1}{2}} = \min_{x_B} \left(\sum_{i=1}^{k} \omega_i (x_B - x_i)^2 \right)^{\frac{1}{2}}, \quad (4.103)$$

donde $\omega_i = \frac{w_i}{\sum_{i=1}^{k} w_i}$ es el peso de la partícula p_i. Es sencillo demostrar que la solución a dicho problema de minimización coincide con la mostrada en la ecuación (4.102).

La relación entre la media y el resultado de la minimización de la distancia euclídea también es expuesta por Chavent y Saracco (2008) y ha sido mostrada en el apartado 2.3.1.1 de esta tesis.

A continuación, se hace una traslación de estos conceptos al contexto de los datos de intervalo.

4.8.2.2. Relación entre el promedio y el baricentro de un conjunto de intervalos

Consideremos un sistema de k elementos en un espacio unidimensional donde cada elemento p_i está definido por el intervalo $[X]_i$ y tiene una masa asociada w_i con $i = 1, ..., k$. El baricentro en forma de intervalo de este sistema al que denotaremos como p_B se halla de la siguiente forma

$$p_B = [X]_B = \frac{\sum_{i=1}^{k} [X]_i \cdot w_i}{\sum_{i=1}^{k} w_i}, \quad (4.104)$$

donde las operaciones se rigen de acuerdo a la aritmética de intervalos (ver apartado 2.7.1.1). El intervalo $[X]_B$ de dicho baricentro presenta las siguientes características:

$$X_{L,B} = \frac{\sum_{i=1}^{k} X_{L,i} \cdot w_i}{\sum_{i=1}^{k} w_i}, \quad X_{U,B} = \frac{\sum_{i=1}^{k} X_{U,i} \cdot w_i}{\sum_{i=1}^{k} w_i},$$
$$X_{C,B} = \frac{\sum_{i=1}^{k} X_{C,i} \cdot w_i}{\sum_{i=1}^{k} w_i}, \text{ y} \quad X_{R,B} = \frac{\sum_{i=1}^{k} X_{R,i} \cdot w_i}{\sum_{i=1}^{k} w_i}. \quad (4.105)$$

En este caso, el baricentro es equivalente al que se obtiene al resolver este

problema de minimización

$$\min_{[X]_B} \left(\sum_{i=1}^{k} \omega_i D_k([X]_B, [X]_i)^2 \right)^{\frac{1}{2}}, \tag{4.106}$$

donde $D_k([X]_B, [X]_i)$ es la distancia definida a partir de un núcleo mostrada en la ecuación (4.12). Si en lugar de utilizar esta distancia, se emplean otras distancias se obtienen baricentros en forma de intervalo que presentan características diferentes. De ello se ocupará el siguiente apartado.

4.8.2.3. Métodos de k-NN para STI basados en otras distancias

En el apartado 4.8.1 se ha adaptado el método de k-NN a la predicción de STI de la siguiente forma

1. Determinando los vecinos más próximos mediante la distancia definida a partir de un núcleo

2. Calculando las predicciones mediante un promedio basado en aritmética de intervalos o, de forma equivalente, como el intervalo baricentro que minimiza la suma del cuadrado de la distancia definida a partir de un núcleo entre sí mismo y el resto de partículas del sistema.

Si tanto en la fase de la búsqueda de los vecinos más próximos, como en la del cálculo de la predicción se emplea otra distancia, el k-NN resultante presenta distintas propiedades. Una alternativa es considerar la distancia de Hausdorff. A continuación, se va a mostrar el efecto de calcular la predicción como el baricentro que minimiza la distancia de Hausdorff (4.4) entre sí mismo y el resto de intervalos.

En ese caso, el baricentro de un sistema de elementos $p_i = [X]_i$ con $i = 1, ..., k$ se calcula de la siguiente forma

$$\min_{[X]_B} \frac{\sum_{i=1}^{k} w_i D_H([X]_B, [X]_i)}{\sum_{i=1}^{k} w_i}. \tag{4.107}$$

El intervalo $[X]_B = \langle X_{C,B}, X_{R,B} \rangle$ que soluciona dicha minimización se obtiene como

$$X_{C,B} = \text{mediana}(X_{C,i}), \text{ con } i = 1, ..., k \tag{4.108}$$

$$X_{R,B} = \text{mediana}(X_{R,i}), \text{ con } i = 1, ..., k. \tag{4.109}$$

Este baricentro que se obtiene es un baricentro mediano, mientras que el baricentro que se obtenía en la ecuación (4.106) era, como se vio, un baricentro medio. Si calculamos el baricentro considerando la distancia de Ichino-Yaguchi definida en la ecuación (4.9) el resultado que se obtiene un baricentro

mediano pero con distintas características

$$X_{L,B} = \text{mediana}(X_{L,i}),\ \text{con } i = 1, ..., k \tag{4.110}$$
$$X_{U,B} = \text{mediana}(X_{U,i}),\ \text{con } i = 1, ..., k. \tag{4.111}$$

A la hora de emplear el método de k-NN, el analista debe decidir qué distancia prefiere utilizar prestando especial atención al tipo de predicción que quiere obtener. Por ejemplo, si se quiere generar predicciones que no tengan en cuenta el comportamiento extremo se debe optar por usar la distancia de Hausdorff o de Ichino-Yaguchi, ya que éstas generan predicciones por medio de una mediana.

La idea aquí presentada entronca con la de las medidas de tendencia central basadas en distancias propuestas por Chavent y Saracco (2008) y mostradas en el apartado 2.3.1.1 de esta tesis.

4.9. Elaboración y predicción de una STI

En este apartado se describirán los pasos para obtener una STI y predecirla de forma adecuada. Dichos pasos son los siguientes.

1. Selección de los datos iniciales. Para poder construir una STI es necesario contar con datos temporales obtenidos bajo algunas de estas circunstancias:

1. Se dispone de un conjunto de valores de una variable en los individuos de un conjunto a lo largo del tiempo. Para cada instante temporal, se construirá un intervalo que resuma el rango de los datos en el conjunto de individuos.

2. Se dispone de una serie temporal continua o de una frecuencia mayor a la que interesa considerar. Cada instante de la frecuencia deseada, se representa mediante un intervalo acotado por los valores mínimo y máximo obtenidos entre dos instantes consecutivos de la frecuencia deseada.

La primera circunstancia describe un caso de agregación contemporánea, mientras que el segundo ilustra un caso de agregación temporal.

2. Construcción de la STI. En algunos casos, puede ser deseable no manejar la STI donde cada intervalo refleje el rango de valores observados en cada instante, e.g., puede ser deseable eliminar los valores extremos. Para ello, se puede emplear el intervalo que acota el 95 % central de los valores observados en cada instante, o, si el interés reside en la parte central del conjunto de datos, el rango intercuartílico.

3. Análisis de la STI. Una vez construida la STI, se deben analizar sus características. Para ello, conviene representar gráficamente la STI y las series de sus componentes: mínimo, máximo, centro y radio. Lo habitual es que el comportamiento de las series del mínimo, del máximo y del centro se comporten de forma muy similar, mientras que la serie de los radios tengan un comportamiento diferente. Este análisis permite determinar si existe tendencia, estacionalidad, patrones que se repitan en el tiempo, etc. También hay que comprobar si las series del mínimo y del máximo se encuentran cointegradas (ver apartado 4.5.2.2), puesto que en ese caso conviene predecirlas mediante un modelo VECM. En este punto debe considerarse si tiene sentido o no transformar la STI. El tema de la transformación de una STI será abordado en el apartado 4.9.1.

4. Predicción de la STI. Para ello se puede emplear cualquiera de los métodos que fueron considerados apropiados en el punto anterior. Se aconseja dividir la serie en tres periodos:

1. inicialización: constará de tantos periodos como requiera el método de predicción que se va a utilizar.

2. entrenamiento: se empleará para estimar el modelo de predicción o para ajustar los parámetros del método de predicción que se esté empleando. Los parámetros de los métodos de predicción proporcionados se pueden determinar mediante una búsqueda en rejilla en el espacio s-dimensional, donde s es el número de parámetros, que tenga como objetivo encontrar la combinación de parámetros que produzca el menor error en el conjunto de entrenamiento.

3. prueba: permite comprobar el rendimiento de los métodos estimados en el entrenamiento y determinar cuál es el que mejor predice la STI.

Ante el amplio abanico de opciones a la hora de predecir una STI, resulta aconsejable probar, al menos, tres aproximaciones para predecir la serie: una que maneje las series de los mínimos y de los máximos, otra que maneje las series de los centros y de los radios, y otra que maneje el intervalo como tal.

4.9.1. Transformaciones sobre las STI

En la predicción de series temporales clásicas es habitual transformar la serie temporal original para obtener una serie temporal estacionaria. Algunas de las transformaciones más habituales son la diferenciación y la transformada logarítmica.

La diferenciación permite eliminar la tendencia estocástica de una serie y consiste en convertir la serie original X_t en la serie de incrementos Z_t donde

$$Z_t = X_t - X_{t-1}, \ \forall t. \tag{4.112}$$

Por su parte, el objetivo de realizar una transformación logarítmica es el de homogeneizar la varianza a lo largo de la serie original X_t. Para ello, se utiliza la función logaritmo de la siguiente forma

$$Z_t = \log(X_t),\ \forall t. \tag{4.113}$$

A la hora de trabajar con STI, puede resultar necesario recurrir a las transformaciones para facilitar la predicción de la serie. Si se ha optado por predecir la STI a partir de las series de dos de sus componentes, se puede emplear cualquiera de las transformaciones que se aplican en las series temporales clásicas, como la diferenciación o la transformación logarítmica. Sin embargo, si se está prediciendo la STI considerando el intervalo como un todo, es necesario definir nuevas transformaciones para datos de intervalo.

Para realizar la diferenciación de dos intervalos no se puede utilizar el operador resta que proporciona la aritmética de intervalos, ya que, como se indicó en el apartado 4.4, dicho operador no refleja de forma adecuada la diferencia existente entre dos intervalos. Además, en caso de realizar la diferenciación como $[Z]_t = [X]_t - [X]_{t-1}$, la serie resultante no puede volver a ser transformada en la serie original, i.e. $[X]_t \neq [Z]_t + [X]_{t-1}$. Por ello, es necesario utilizar otra aproximación.

Para realizar la diferenciación en una serie temporal de intervalos se propone realizar la diferenciación únicamente sobre el centro de los intervalos. De forma que, dada una STI $[X]_t$, su transformada mediante diferenciación $[Z]_t$ se obtenga como

$$[Z]_t = \langle X_{t,C} - X_{t-1,C}, X_{t,R} \rangle \forall t. \tag{4.114}$$

Esta diferenciación permite eliminar la tendencia estocástica que afecta a la posición del intervalo. Este concepto de tendencia ya fue empleado al definir los alisados exponenciales con tendencia aditiva para STI (ver el apartado 4.6.3.2). La tendencia en la posición del intervalo, i.e., en su centro, es la más habitual en los contextos prácticos, aunque teóricamente es posible pensar en STI cuya amplitud, i.e., cuyo radio, tenga también una tendencia.

Otro tipo de transformación que puede ser aplicada sobre una STI es la transformada logarítmica. La función logaritmo es una función monótona creciente y, por tanto, su adaptación al contexto de los intervalos es directa. Dada la STI $[X]_t$, su transformada logarítmica se calcula como

$$[Z]_t = \log([X]_t) = [\log(X_{t,L}), \log(X_{t,U})],\ \forall t. \tag{4.115}$$

Al ser la función logarítmica una función monótona creciente, la transformada de todo valor $x \in [X]_t$ va a ser un valor contenido dentro del intervalo resultante, i.e., $z \in [Z]_t$.

4.10. Ejemplos ilustrativos de la predicción de STI

A continuación, las técnicas de predicción propuestas a lo largo del capítulo van a ser aplicadas sobre un conjunto de STI reales. Las STI elegidas provienen de dos ámbitos: las finanzas y la meteorología. La razón por la que se han escogido estos campos es porque, en ellos, las STI aparecen de forma natural y porque los datos históricos suelen ser de acceso público y gratuito.

En el caso de la meteorología, se va a trabajar con la temperatura. Esta variable se recoge habitualmente mediante los valores mínimo y máximo registrados a lo largo de un periodo de tiempo, normalmente, un día. De igual forma, el pronóstico de las temperaturas se suele ofrecer en forma de intervalo, por lo que la utilidad de la predicción en forma de intervalo es evidente.

En finanzas, también es frecuente que se utilicen los mínimos y los máximos, ya sean diarios o semanales, del precio de una acción, del valor de un índice o del cambio entre divisas. El conocimiento de la predicción del intervalo de valores mínimo y máximo para el periodo siguiente permite hacerse una idea de la volatilidad en dicho periodo y es de gran ayuda a la hora de fijar una estrategia inversora.

Además de las finanzas y la meteorología, existen otros muchos ámbitos donde aplicar las STI, entre ellos: la hidrología, donde se registran los cauces mínimos y máximos registrados en los ríos; el medioambiente, para registrar los niveles de los contaminantes; y, en general, cualquier ámbito donde se registren datos de forma continua mediante sensores y donde interese analizar los valores mínimos y máximos de dichos datos. También es posible obtener STI mediante la agregación contemporánea de conjuntos de valores medidos en un conjunto de individuos. Un ejemplo podía ser el rango de valoraciones de un producto o un servicio recogido en un grupo de clientes o el intervalo de los precios del m^2 de vivienda libre en una determinada región.

4.10.1. Descripción de la metodología seguida en la predicción de cada STI

4.10.1.1. Métodos de predicción empleados

Para predecir cada una de las STI se van a emplear todas las aproximaciones presentadas en este capítulo:

- Técnicas que trabajan con el intervalo como una entidad en sí misma
  - los alisados exponenciales
  - el k-NN
  - el iMLP

- Técnicas que trabajan con el intervalo a partir de las series de dos de sus componentes (mínimo y máximo o centro y radio)
  - métodos univariantes
    - alisados exponenciales
    - método de k-NN
    - modelo ARIMA
    - perceptrón multicapa (o MLP según el acrónimo inglés)
    - modelo híbrido ARIMA+MLP Zhang (2003)
  - métodos multivariantes
    - modelo VAR: como ya se mencionó en el apartado 4.5.2.1, el VAR de orden p planteado sobre las series de los centros y de los radios obtiene las mismas predicciones que el VAR de orden p planteado sobre las series de los mínimos y de los máximos. Por ello, no se hará diferencia entre ambas posibilidades y sólo se hará referencia al modelo utilizado como VAR de orden p.
    - modelo VECM: al trabajar con las series de los mínimos y de los máximos es posible que exista una relación de cointegración entre ambas series (ver apartado 4.5.2.2). Según establecieron Engle y Granger (1987), si eso sucede, el modelo multivariante que se plantee puede recoger dicha relación y, por tanto, puede plantearse un modelo vectorial de corrección del error (VECM).

Para estimar los parámetros de cada uno de los métodos se han utilizado los datos del conjunto de entrenamiento. La estrategia que se ha empleado en cada uno de los métodos es la siguiente:

- Para los modelos ARIMA, VAR y VECM se han utilizado tres criterios: su capacidad predictiva, la no existencia de trazas de información lineal en los residuos y los criterios de información de Akaike y de Schwartz.
- Para los perceptrones multicapa se han utilizado como criterios su capacidad predictiva y el número de parámetros (variables de entrada y unidades de la capa oculta), optando, en caso de similar capacidad predictiva, por el modelo más parsimonioso.
- En el caso de los alisados y del k-NN, los parámetros han sido determinados como la solución de una búsqueda en rejilla por el espacio de parámetros con el objetivo de minimizar el error cuadrático en el periodo de entrenamiento.

En caso de ser necesario, se ha empleado algún tipo de transformación para de esa manera obtener predicciones más precisas de la serie temporal. La

diferenciación ha sido indispensable para predecir de forma precisa las series del contexto financiero analizadas con los métodos de k-NN y del iMLP, tanto en su variante clásica, como en la de STI. Sin diferenciar los resultados que obtenían eran muy pobres.

4.10.1.2. Presentación de los resultados

A la hora de mostrar los resultados se ha pretendido no resultar prolijo y presentar únicamente aquella información verdaderamente relevante. Para cada STI analizada, los resultados se mostrarán en dos tablas cuyo contenido es descrito a continuación.

En la primera de ellas, se mostrarán los resultados obtenidos por los métodos de predicción univariantes al pronosticar cada una de las series temporales que componen una STI (i.e. la serie de los extremos inferiores, la de los extremos superiores, la de los centros y la de los radios). Esa tabla sirve para determinar cuál es el mejor método para predecir cada una de esas series. Las predicciones de los métodos elegidos sirven para componer la predicción de la STI a partir de, o bien las predicciones de los extremos inferior y superior, o bien las predicciones del centro y del radio. Para evitar un excesivo nivel de detalle, en esta tabla no se mostrará información sobre el tipo de modelo ARIMA ha utilizado para cada serie, los parámetros empleados para cada alisado exponencial, los retardos incluidos en el MLP, si el método requería diferenciar la serie o no, etc.

En la segunda tabla, se mostrarán los resultados obtenidos por las distintas aproximaciones que permiten pronosticar la STI completa. Dichas aproximaciones incluyen las que consideran al intervalo como una entidad en si misma, las aproximaciones multivariantes y las aproximaciones univariantes de los extremos inferior y superior y del centro y el radio, cuyas predicciones han sido obtenidas mediante los métodos elegidos en la primera tabla. En esta segunda tabla, la información que se muestra por columnas corresponde al error cometido por cada aproximación en cada uno de los cuatro componentes del intervalo (extremos inferior y superior, centro y radio). Esto permite saber en qué componente se está errando más y en cual menos.

Con el fin de reducir su tamaño, en ambas tablas sólo se mostrarán los errores cometidos por cada uno de los métodos considerados en el periodo de prueba y no en el periodo de entrenamiento.

Medida de error mostrada. La medida de error que aparecerá en las tablas será la raíz cuadrada del error cuadrático escalado medio o $RECEM$, ver la ecuación (4.18), cometido en cada una de las series temporales de los componentes del intervalo, .i.e., en las series de los centros, de los radios y de los extremos inferiores y superiores. Para escalar el error, se utilizará el error cuadrático cometido por el método ingenuo en el periodo de entrenamiento.

Pese a que las series temporales de los cuatro componentes del intervalo están expresados en la misma unidad, resulta adecuado utilizar una medida que escale el error y que lo muestre eliminando el efecto de la unidad de medida. Esto es debido a que la magnitud del error cometido en la serie temporal del radio puede ser muy diferente a la de las otras tres series, lo cual puede dar una impresión errónea sobre la precisión en dicha serie.

Se ha descartado mostrar el error medio basado en una distancia de intervalos (4.14), porque esta medida, presenta el error cometido en los componentes de forma agregada y no permite determinar qué componente se ha pronosticado mejor o peor con cada método. Sin embargo, el error medio basado en la distancia definida sobre un kernel ha sido utilizado para ajustar los parámetros de los métodos de predicción específicos de STI, porque estos métodos trabajan con el intervalo como un todo y requieren de una medida que informe con un solo valor del error que han cometido. La razón por la que se ha usado la distancia definida sobre un kernel (4.12) es porque es cuadrática y la medida de error que se va a usar para comparar los resultados de los métodos, el $RECEM$, también es cuadrática.

4.10.2. Predicción del rango de valores diario del índice Dow Jones

En este ejemplo se va a predecir la serie de los valores mínimos y máximos diarios registrados por el índice Dow Jones Promedio Industrial. Este índice agrega el comportamiento en la bolsa de las treinta mayores empresas negociadas en la bolsa de Estados Unidos. El periodo considerado abarca desde el 01 de enero de 2004 hasta el 30 de diciembre de 2005, dos años completos en los que hubo 504 días de negociación. Los primeros 377 periodos de la serie, i.e. hasta el 30 de junio de 2005, han sido usados como conjunto de entrenamiento, mientras que los 127 periodos restantes han sido empleados como conjunto de prueba.

En la figura 4.6 puede verse la STI completa. En ella, se aprecia que durante la mayor parte de 2004 el índice tiene una tendencia descendente que se manifiesta como una sucesión de bajadas y rebotes. A finales de 2004, la serie asciende de forma vertiginosa hasta los valores con los que inició el año. Finalmente, durante 2005 la serie oscila en torno a los 10600 puntos.

En la tabla 4.1 se muestran los errores cometidos en la predicción de las series temporales de los componentes por cada uno de los métodos de predicción considerados. Los resultados de aquellos métodos que han obtenido mejor resultado (sin incluir al método ingenuo) son resaltados en negrita. Se puede apreciar, como en la predicción de las series de los extremos inferior y superior, ha resultado muy complicado batir al método ingenuo. En la serie de los centros, todos los métodos, a excepción del k-NN, superan el rendimiento del ingenuo, aunque no por una gran diferencia. Sin embargo, en la serie de los radios la mejora que se obtiene al utilizar un método distinto al

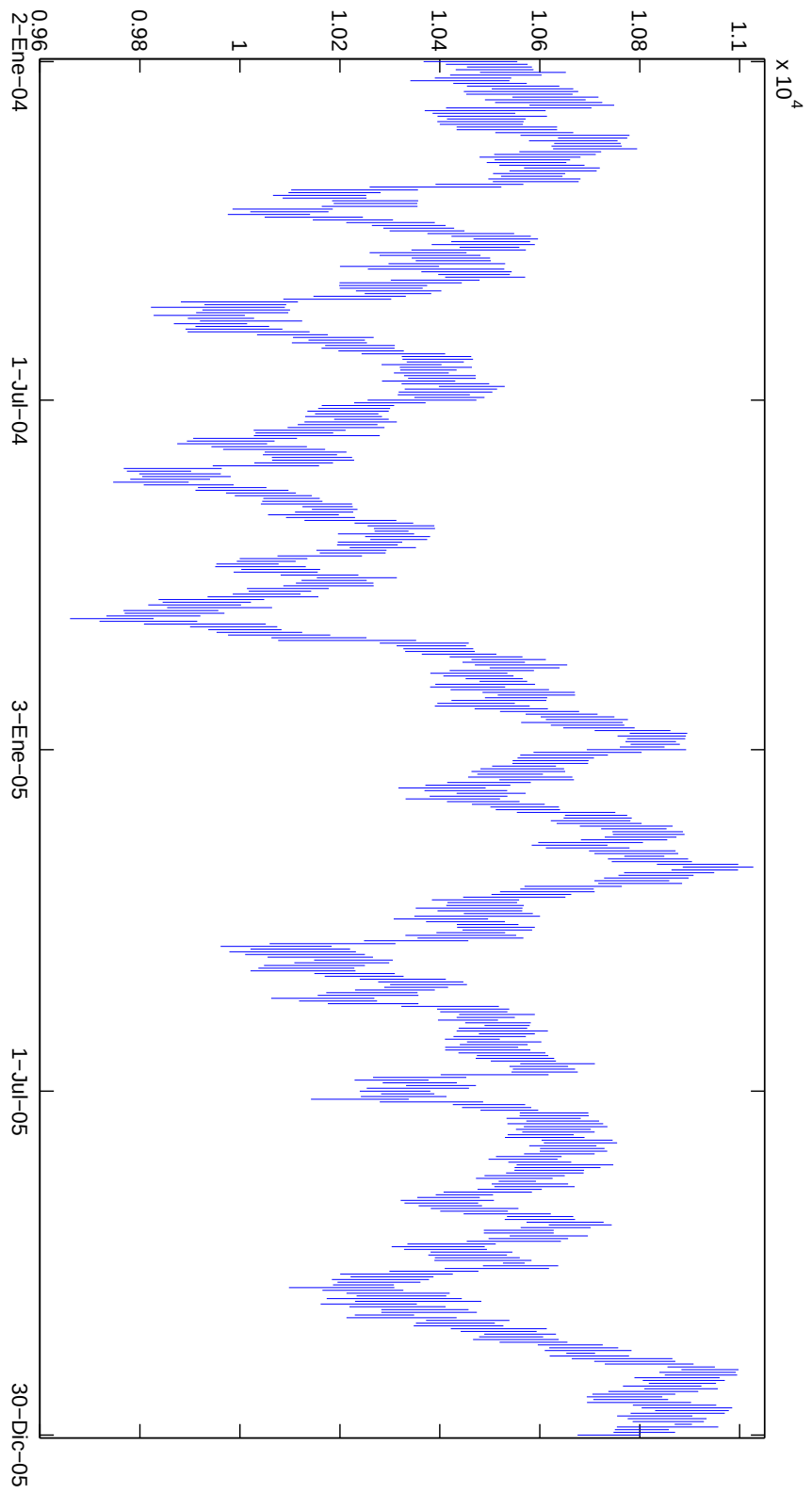


Figura 4.6: STI diaria del índice Dow Jones durante los años 2004 y 2005.

Tabla 4.1: *RECEM* obtenido por los métodos univariantes en el periodo de prueba de cada una de las series componentes de la STI del índice Dow Jones.

Modelos	ext.inf.	ext.sup.	centro	radio
Método ingenuo	0.9064	0.9426	0.8873	1.1469
alisados	0.9157	0.962	0.878	0.8659
k-NN	0.9703	0.9567	0.9241	**0.8612**
ARIMA	**0.9103**	0.9517	**0.8709**	0.8919
MLP	0.9349	**0.9382**	0.8851	0.8722
ARIMA+MLP	0.9155	0.9502	0.8753	0.8899

Tabla 4.2: *RECEM* obtenido por las distintas aproximaciones de predicción de STI en el periodo de prueba de las cuatro series de componentes de la STI del índice Dow Jones.

Modelo	ext.inf.	ext.sup.	centro	radio
Método ingenuo	0.9064	0.9426	0.8873	1.1469
VAR (2)	0.8659	0.8849	0.8770	0.8596
VECM (1) ext. inf-sup	0.8583	0.8911	0.8761	**0.8562**
AETA ($\alpha = .93$; $\gamma = 1$; $\phi = .35$)	0.8984	0.9280	0.8805	1.1120
iMLP (h=10; retardos=3)	0.8698	**0.8818**	0.8689	0.9197
k-NN (k=19; d=3)	0.8872	0.8955	0.8946	0.8657
Aproxim. univar. ext. inf-sup	0.9103	0.9382	**0.8641**	1.2663
Aproxim. univar. cen-rad	**0.8429**	0.8997	0.8709	0.8612

ingenuo es muy significativa. En ese caso, el k-NN es el método que mejor resultado obtiene.

En conclusión, dada esta tabla, para predecir según la aproximación de los extremos del intervalo se empleará un modelo ARIMA para el mínimo y un perceptrón multicapa (MLP) para el máximo. Mientras que en la aproximación centro-radio, se utilizarán las predicciones del modelo ARIMA para la serie del centro y las del k-NN para la del radio. A continuación, estas aproximaciones serán comparadas con los métodos que trabajan con las series de intervalos como un todo.

En la tabla 4.2 se muestran los resultados obtenidos por las distintas aproximaciones que permiten predecir la STI en su totalidad. En cada una de las columnas de esta tabla, se muestra el error cometido por dichos métodos en cada uno de los componentes de la serie. Comparando todas las aproximaciones, las que mejores resultados obtienen son la basada en los modelos univariantes centro-radio (centro mediante ARIMA y radio mediante MLP)

y los modelos multivariantes VAR y VECM. El iMLP obtiene también unos resultados destacables, pero su principal inconveniente es que no pronostica tan eficazmente el radio. Por su parte, el alisado y el k-NN para STI, aunque mejoran al método ingenuo, no son tan precisos como los métodos ya citados. La aproximación de los modelos univariantes aplicados sobre las series de los extremos obtiene peores resultados que el ingenuo en la predicción de los extremos inferiores y, especialmente, en la de los radios.

Para interpretar de forma adecuada las tablas, hay que tener en cuenta que el error que se muestra es el *RECEM*, que es el error cuadrático medio cometido por cada método en el conjunto de prueba dividido entre el error cometido por el **método ingenuo** en el conjunto de entrenamiento. Por ello, el hecho de que el *RECEM* del método ingenuo en la serie de los radios sea mayor que uno indica que en dicho periodo dicha serie se vuelve más impredecible para el método ingenuo. De hecho, es en dicha serie donde se ha obtenido un mayor margen de mejora con respecto al ingenuo, ya que los modelos multivariantes obtienen un *RECEM* casi tres décimas inferior al obtenido por el método ingenuo en dicho periodo. Sin embargo, en la serie de los centros, en el mejor de los casos, la mejora respecto al método ingenuo es sólo de dos centésimas.

4.10.3. Predicción del rango de valores diario del índice Standard & Poor's 500

El índice bursátil del Standard & Poor's 500 (S&P 500) incluye a las 500 empresas con mayor capitalización de EEUU y es considerado como un indicador fiable del estado de la economía estadounidense. El periodo analizado de la serie será el mismo que en el caso del Dow Jones, los años 2004 y 2005 al completo.

Tal y como se ve en la figura 4.7, durante 2004 el S&P 500 oscila en torno al nivel de los 1120 puntos, hasta que a finales de año asciende rápidamente hasta el nivel de los 1200 ptos. Durante la primera mitad de 2005, se mantiene en torno a dicho nivel, mientras que en la segunda mitad lo rebasa, entrando a final de año en otro periodo ascendente que lo consolida en los 1250 puntos. La serie tiene 504 periodos, de los cuales, los 377 primeros (que abarcan hasta el 30 de junio de 2005) se han utilizado para inicialización y entrenamiento y los restantes 127 (los últimos seis meses de 2005) para validar los métodos.

En la tabla 4.3 se muestra el error cometido por los métodos univariantes en el periodo de prueba en cada una de las series de los componentes. El método ingenuo se muestra difícil de batir en las series de los extremos inferior y superior. El perceptrón multicapa es el método que mejores resultados obtiene para estas series y para la de los radios, lo que parece indicar cierto comportamiento no lineal en dichas series. Por su parte, en el caso de la serie de los centros, el modelo ARIMA obtiene los mejores resultados.

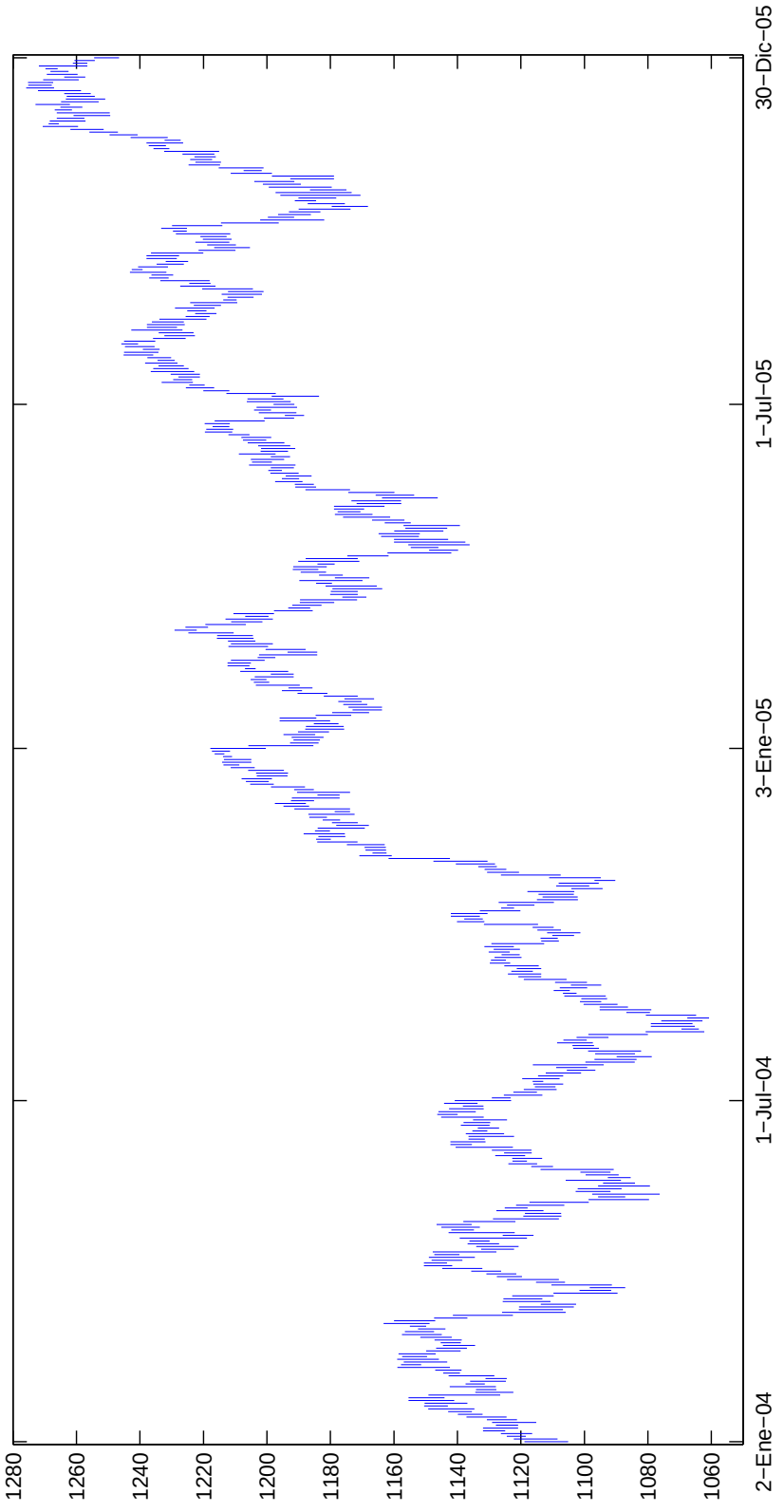


Figura 4.7: STI diaria del índice S&P 500 durante los años 2004 y 2005.

Tabla 4.3: *RECEM* obtenido por los métodos univariantes en el periodo de prueba de cada una de las series componentes de la STI del índice S&P 500.

Modelos	ext.inf.	ext.sup.	centro	radio
Método ingenuo	0.9674	0.9467	0.9673	0.9168
alisados	0.9533	0.9497	0.9378	0.7007
k-NN	0.9866	0.9518	0.9637	0.7309
ARIMA	0.9522	0.9467	**0.9248**	0.7050
MLP	**0.9477**	**0.9321**	0.9317	**0.6961**
ARIMA+MLP	0.9487	0.9416	0.9357	0.7106

Tabla 4.4: *RECEM* obtenido por las distintas aproximaciones de predicción de STI en el periodo de prueba de las cuatro series de componentes de la STI del índice S&P 500.

Modelo	ext.inf.	ext.sup.	centro	radio
Método ingenuo	0.9674	0.9467	0.9673	0.9168
VAR (2)	0.8684	0.9419	0.9550	0.6702
VECM (1) ext. inf-sup	0.8589	0.9131	0.9336	**0.6610**
AETA ($\alpha = .94$; $\gamma = 1$; $\phi = .4$)	0.9293	0.9402	0.9444	0.8925
iMLP (h=15; retardos=2)	0.8524	**0.8887**	**0.8951**	0.7626
k-NN (k=15; d=1)	**0.8231**	0.909	0.9034	0.6924
Aproxim. univar. ext. inf-sup	0.9477	0.9321	0.9141	1.0363
Aproxim. univar. cen-rad	0.8533	0.9153	0.9248	0.6961

Por tanto, en la aproximación de los extremos inferior y superior la STI se compondrá a partir de las predicciones de los perceptrones multicapa estimados para cada una de estas dos series. Mientras que para la aproximación centro-radio se utilizarán las predicciones del modelo ARIMA para los centros y del perceptrón multicapa para los radios. En la tabla 4.4, estos resultados son comparados con los obtenidos por otras aproximaciones que permiten predecir la STI. En dicha tabla se puede comprobar que el k-NN para STI, los modelos multivariantes VAR y VECM, y la aproximación centro-radio son los métodos que mejor se comportan. Todos ellos obtienen un rendimiento bastante similar y que mejora muy claramente al método ingenuo. El iMLP al igual que en el ejemplo del índice Dow Jones falla más en la predicción del radio. Tanto el k-NN, como el iMLP se han aplicado sobre la STI diferenciada. Por su parte, el alisado para STI también mejora el ingenuo, pero no tan ampliamente como estos otros métodos.

Al igual que en el caso del Dow Jones, la serie donde más ampliamente se ha mejorado al método ingenuo es la de los radios, donde el VECM mejora en más de veinticinco centésimas al método ingenuo. Sin embargo, la aproximación univariante que predice las series de los extremos inferior y superior obtiene peores predicciones que el método ingenuo en el radio de los intervalos. Además, resulta sorprendente que el error cometido por dicha aproximación en las predicciones de las series de los extremos, que son las series sobre las que dicha aproximación trabaja, sea mayor que el cometido por prácticamente el resto de métodos que utilizan otras estrategias para predecir la STI. En el ejemplo de la STI del Dow Jones, la aproximación univariante basada en las series de los extremos obtuvo resultados similares, lo que parece indicar que dicha aproximación no es adecuada para predecir estos índices bursátiles.

4.10.4. Predicción del rango de valores diario del cambio Euro-Dólar

En este caso se va a predecir la STI de los valores mínimo y máximo diarios del cambio de divisas Euro-Dólar. La serie abarca los años 2002 y 2003 y tiene 519 periodos que han sido divididos de la siguiente forma: el conjunto de entrenamiento lo forman las sesiones hasta el 30 de Junio de 2003 (388 periodos) y el conjunto de prueba lo forman las sesiones de los seis meses restantes de 2003 (131 periodos).

La figura 4.8 muestra gráficamente la STI. En ella, puede apreciarse que la cotización tiene una tendencia creciente durante los dos años. Dicha tendencia se inicia en febrero de 2002, se interrumpe entre junio y agosto de 2003 donde el valor de la cotización desciende, pero tras dicho periodo el crecimiento se retoma hasta final de año.

En la tabla 4.5 se muestran los resultados obtenidos por los métodos univariantes en cada una de las series de los componentes. Es interesante

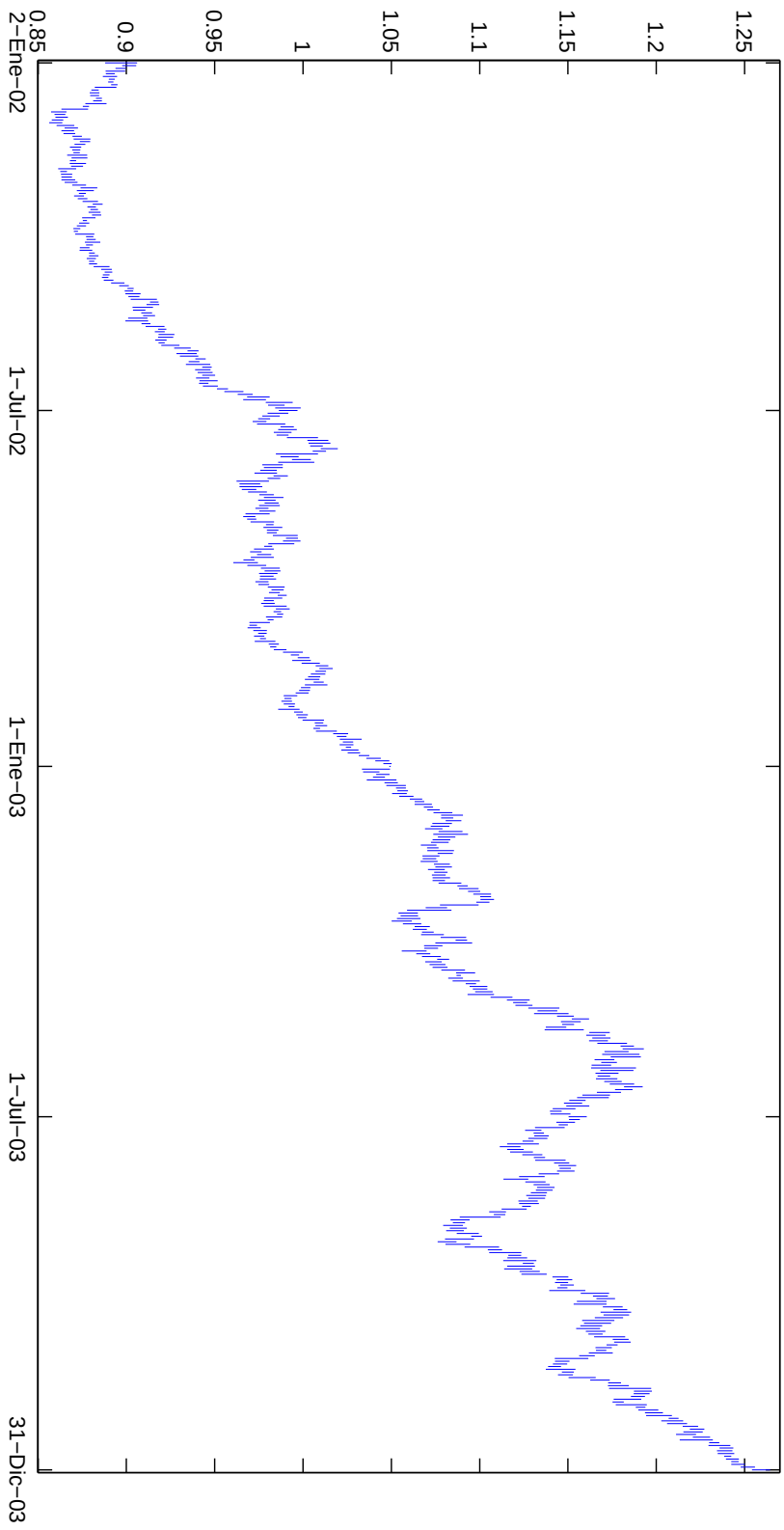


Figura 4.8: STI diaria del cambio de divisas € – $ durante los años 2002 y 2003.

Tabla 4.5: *RECEM* obtenido por los métodos univariantes en el periodo de prueba de cada una de las series componentes de la STI del cambio € – $.

Modelos	ext.inf.	ext.sup.	centro	radio
Método ingenuo	1.3343	1.1258	1.2258	1.2585
alisados	1.3252	1.1093	1.1486	0.8518
k-NN	1.3468	1.1147	1.1993	0.8739
ARIMA	**1.3159**	**1.1059**	**1.1424**	0.8463
MLP	1.3548	1.1157	1.1497	0.9318
ARIMA+MLP	1.3159	1.1119	1.1445	**0.8419**

Tabla 4.6: *RECEM* obtenido por las distintas aproximaciones de predicción de STI en el periodo de prueba de las cuatro series de componentes de la STI del cambio € – $.

Modelo	ext.inf.	ext.sup.	centro	radio
Método ingenuo	1.3343	1.1258	1.2258	1.2585
VAR (3)	1.1333	1.0773	1.1632	0.8645
VECM (1) ext. inf-sup	1.1410	1.0663	1.1582	0.8794
AETA ($\alpha = .86$; $\gamma = 1$; $\phi = .36$)	1.2668	1.0558	1.1648	1.1640
iMLP (h=15; retardos=2)	1.1848	1.0589	1.1551	0.9974
k-NN (k=14; d=1)	**1.1022**	**1.014**	**1.0832**	0.9636
Aproxim. univar. ext. inf-sup	1.3159	1.1059	1.1892	1.2996
Aproxim. univar. cen-rad	1.1123	1.0514	1.1394	**0.8419**

destacar que, tal y como muestra la tabla, en las series de los extremos, los métodos considerados obtienen una muy ligera mejora con respecto a los resultados del método ingenuo. La mejora es mayor en las series del centro y, especialmente, del radio.

La tabla 4.5 también muestra que el método ARIMA es el mejor para predecir las series temporales de los extremos inferiores y superiores y la de los centros, seguido de cerca por los métodos de alisado, por el modelo híbrido y por el MLP. Por su parte, el modelo híbrido es el mejor para los radios, donde la mejora que obtiene con respecto al método ingenuo es considerable. El modelo ARIMA se queda cerca, pero no llega a superar al ingenuo. Por tanto, para la aproximación que predice la STI a partir de las predicciones de los extremos se utilizarán las predicciones de los extremos inferiores y superiores obtenidas por los modelos ARIMA; mientras que para la aproximación centro-radio se utilizarán las predicciones de los centros del modelo ARIMA y las de los radios obtenidas por el modelo híbrido.

En la tabla 4.6 se muestra el resultado que obtienen las distintas aproximaciones de predicción de STI. En este caso, el mejor método es el k-NN

para STI. Dicho método trabaja con la STI diferenciada según explica el apartado 4.9.1, es decir, diferenciando únicamente la serie de los centros y no la de los radios que es donde, como se ve en la figura 4.8, se presenta la tendencia. Sin embargo, el k-NN para STI no obtiene los mejores resultados en la predicción del radio. La mejor aproximación para predecir dicha serie es la aproximación centro-radio donde la componente de los radios se pronostica con un modelo híbrido ARIMA+MLP. A nivel global, es decir, analizando los resultados en las cuatro componentes, la aproximación centro-radio y los modelos VAR y VECM obtienen también buenas predicciones. El iMLP sobre la STI diferenciada funciona ligeramente peor que las aproximaciones citadas y todavía algo peor que el iMLP funciona el alisado. De nuevo, la peor aproximación es la que predice las series de los extremos con métodos univariantes, que obtiene peores resultados que el método ingenuo en la serie de los radios.

Es interesante destacar que en todas las series de componentes, a excepción de la del radio, todos los métodos de predicción obtienen valores del *RECEM* mayores que uno en el periodo de prueba. Esto quiere decir que, en dicho periodo, estos métodos predicen en media peor de lo que lo hace en media también el método ingenuo durante el periodo de entrenamiento. Este resultado no es negativo, porque el *RECEM* del método ingenuo en las series de los componentes es también mayor que uno y mucho mayor que el obtenido por estos métodos. Esto indica que en el periodo de prueba la STI se vuelve más impredecible, pero, pese a ello, los métodos considerados obtienen mejores resultados que el método ingenuo. La componente donde mayor margen de mejora se obtiene es el radio.

4.10.5. Predicción del rango de valores diario del cambio Dólar-Yen

En este ejemplo se trabajará con la STI que resulta de considerar los valores mínimos y máximos diarios del cambio de divisas Dólar-Yen durante 2002 y 2003. La serie se ha dividido de forma que los primeros 389 periodos, que se corresponden con los dieciocho primeros meses, son empleados como conjunto de entrenamiento. Mientras que los 131 periodos restantes, que corresponden a los seis últimos meses de 2003, forman el conjunto de prueba.

La figura 4.9 muestra el comportamiento de la serie temporal. En ella puede verse como, durante dicho periodo, el Dólar se deprecia frente al Yen. Sin embargo, la depreciación se produce principalmente en dos momentos: en el segundo trimestre de 2002 y en el último trimestre de 2003. El resto del tiempo, la serie oscila en torno a una banda sin una tendencia clara.

En la tabla 4.7 se muestran los resultados obtenidos en el periodo de prueba por los métodos univariantes al predecir cada una de las series de los componentes. En este caso, el modelo ARIMA resulta ser la mejor alternativa para predecir todas las series de los componentes. En las series de los

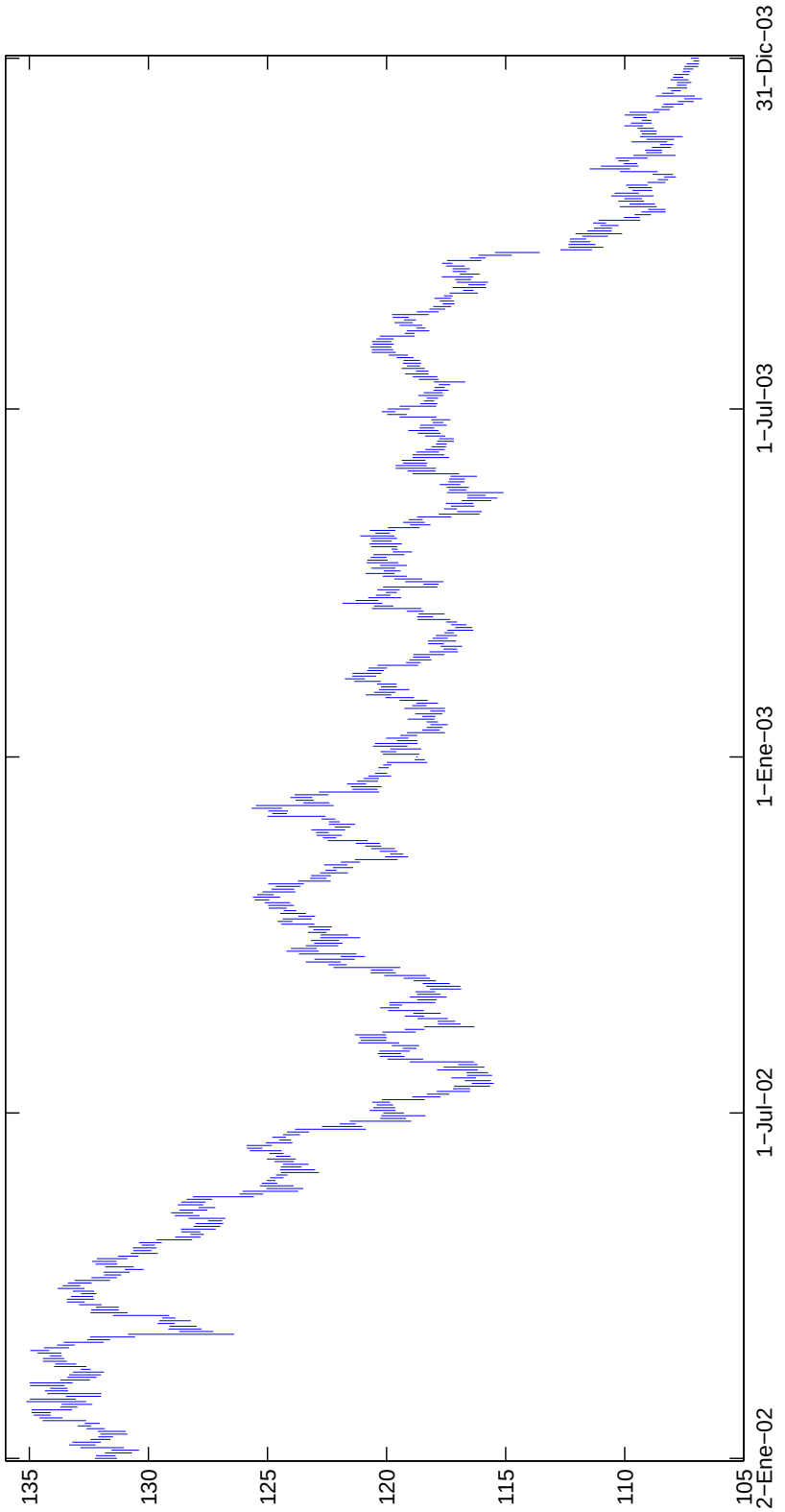


Figura 4.9: STI diaria del cambio de divisas \$ – ¥ durante los años 2002 y 2003.

Tabla 4.7: *RECEM* obtenido por los métodos univariantes en el periodo de prueba de cada una de las series componentes de la STI del cambio $\$ - ¥$.

Modelos	ext.inf.	ext.sup.	centro	radio
Método ingenuo	0.7433	0.8560	0.8214	0.7296
alisados	0.7424	0.8508	0.786	0.6308
k-NN	0.7433	0.8566	0.8217	0.6248
ARIMA	**0.7327**	**0.8445**	**0.7791**	**0.5633**
MLP	0.8366	0.8796	0.7904	0.5943
ARIMA+MLP	0.7432	0.8495	0.7835	0.5634

Tabla 4.8: *RECEM* obtenido por las distintas aproximaciones de predicción de STI en el periodo de prueba de las cuatro series de componentes de la STI del cambio $\$ - ¥$.

Modelo	ext.inf.	ext.sup.	centro	radio
Método ingenuo	0.7433	0.8560	0.8214	0.7296
VAR (3)	0.7219	0.7883	0.8057	0.5942
VECM (1) ext. inf-sup	0.7205	0.7562	0.7839	0.5979
AETA ($\alpha = .88$; $\gamma = 1$; $\phi = .42$)	0.7232	0.8147	0.7929	0.6930
iMLP (h=15; retardos=3)	**0.7028**	0.7642	**0.7683**	0.6259
k-NN (k=16; d=1)	0.7447	0.8021	0.8176	0.6357
Aproxim. univar. ext. inf-sup	0.7327	0.8445	0.7946	0.7632
Aproxim. univar. cen-rad	0.7103	**0.7437**	0.7791	**0.5633**

extremos, el método ingenuo es complicado de batir. Mientras que, al igual que sucede en los ejemplos anteriores, la serie donde se produce una mayor mejora es en la serie del radio. Tomando las series que se obtienen con los modelos ARIMA se componen las predicciones de las STI de la aproximación de los extremos inferior y superior, y de la aproximación centro y radio.

En la tabla 4.8 pueden verse los resultados obtenidos en la predicción de la STI con las diferentes aproximaciones. La aproximación centro-radio, donde ambas series se predicen con un modelo ARIMA, y el iMLP aplicado sobre la STI diferenciada son las que mejores predicciones generan. Los rendimientos de ambos métodos en todos los componentes es aproximadamente similar, excepto en los radios, donde el iMLP, aunque mejora ampliamente al ingenuo, obtienen unos resultados más discretos. Otro resultado interesante es que los modelos VAR y VECM, aunque mejora al método ingenuo, no consiguen mejores predicciones que las obtenidas con un modelo ARIMA para el centro y otro para el radio. El resultado es interesante porque los modelos VAR y VECM son modelos autorregresivos que recogen las interdependencias existentes entre las series consideradas, cosa que los ARIMA no

Tabla 4.9: *RECEM* obtenido por los métodos univariantes en el periodo de prueba de cada uno de las series de componentes de la STI de la temperatura de Pekín.

Modelos	ext.inf.	ext.sup.	centro	radio
Método ingenuo estacional	0.8570	0.9078	0.9442	0.6664
alisados	0.7907	0.6568	0.7346	0.5548
k-NN	0.6824	0.6572	0.7058	0.5649
ARIMA	**0.6309**	**0.6114**	**0.6541**	**0.5087**
MLP	0.7365	0.7250	0.8484	0.6488
ARIMA+MLP	0.6673	0.6189	0.6588	0.5088

hacen, pero pese a ello, obtienen peores resultados. Por su parte, los alisados y el k-NN para STI están un peldaño por debajo de las aproximaciones ya mencionadas, pero consiguen mejorar al método ingenuo. La aproximación que pronostica los extremos de las series mediante modelos univariantes es, de nuevo, la más floja, aunque sólo empeora al método ingenuo en la serie de los radios.

4.10.6. Predicción del rango de la temperaturas mensuales en Pekín

En este ejemplo se va a trabajar con una STI del ámbito meteorológico, donde cada intervalo representa un rango de temperaturas. El extremo inferior de los intervalos representa la media mensual de las temperaturas mínimas en Pekín, mientras que el extremo superior representa la media mensual de las temperaturas máximas en dicha ciudad. La serie abarca desde enero de 1952 hasta diciembre de 1988, ambos inclusive, por lo que consta de 456 periodos mensuales. Los primeros 324 periodos de la serie, i.e. los primeros 27 años, han sido usados como conjunto de entrenamiento, mientras que los 132 periodos restantes han sido empleados como conjunto de prueba. Los datos han sido obtenidos de la base de datos *Long-Term Instrumental Climatic Database of the People's Republic of China*[2].

Tal y como muestra la figura 4.10, la serie tiene una estacionalidad manifiesta que afecta a todo el intervalo y no solo a la posición del mismo. El patrón estacional es muy regular, tal y como puede esperarse de una serie temporal de temperaturas de frecuencia mensual. Esto implica que el método ingenuo con estacionalidad pronosticará mucho mejor que el método ingenuo normal. Por ello, se empleará dicho método como la referencia a batir.

En la tabla 4.9 se muestra el resultado obtenido por cada uno de los méto-

[2]Disponible en `http://dss.ucar.edu/datasets/ds578.5/data/` para usuarios registrados (el registro es gratuito)

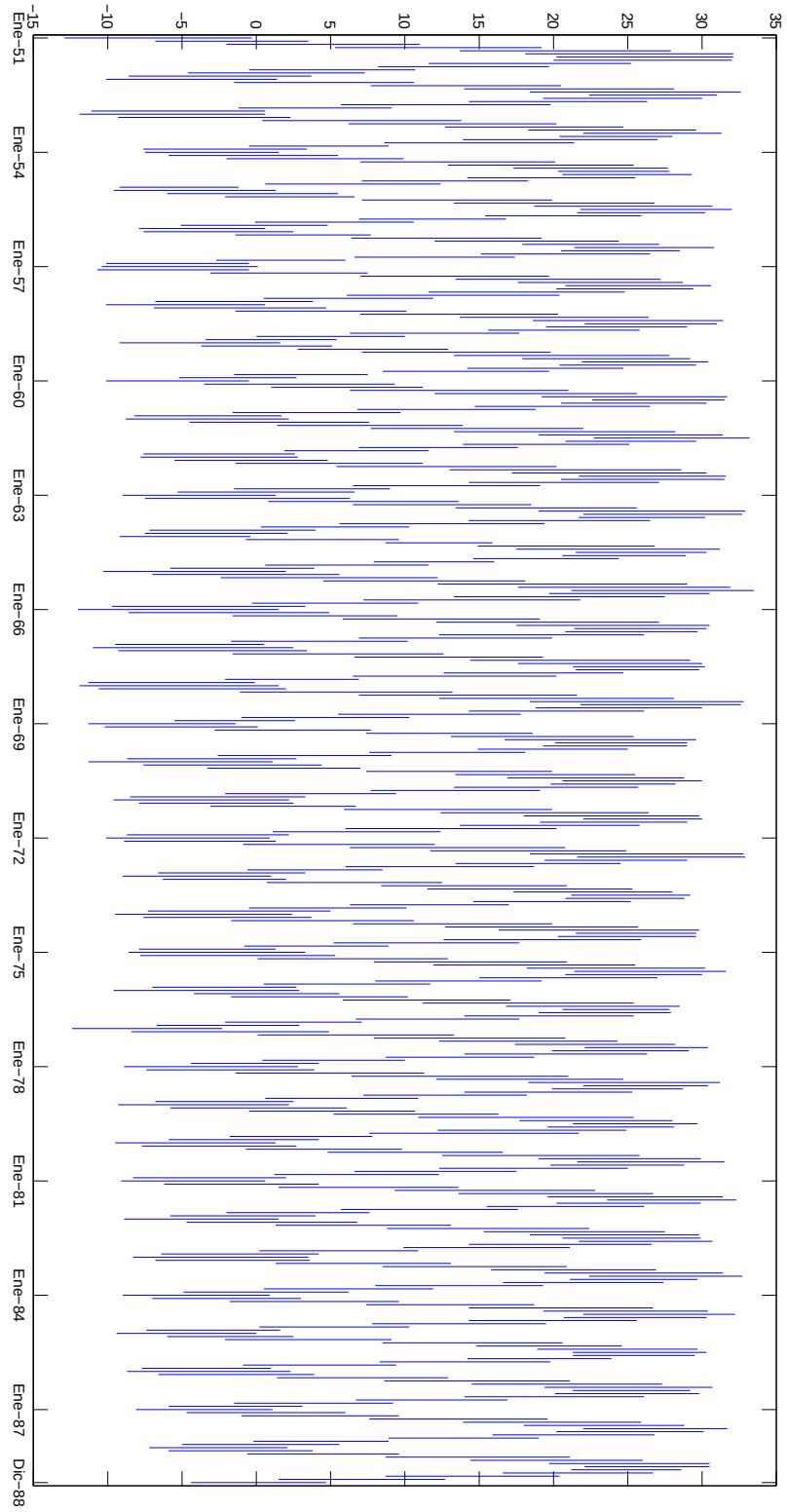


Figura 4.10: STI mensual de las temperaturas mínimas y máximas medias mensuales en Pekín entre 1952 y 1988.

Tabla 4.10: *RECEM* en el periodo de prueba en cada una de las cuatro series de componentes de la STI de la temperatura de Pekín.

Modelo	ext.inf.	ext.sup.	centro	radio
Método ingenuo estacional	0.8570	0.9078	0.9442	0.6664
VAR (12)	0.6824	0.6965	0.7218	0.5740
VECM (11) ext. inf-sup	0.6792	0.7065	0.7240	0.5921
AEEc ($\alpha = .07$; $\delta = .07$)	0.6886	0.6806	0.6637	0.7480
AEEi ($\alpha = .07$; $\delta = .07$)	0.6612	0.6248	0.6637	0.5431
iMLP (h=10; retardos=1, 11, 12)	0.7158	0.7226	0.7308	0.6819
k-NN pond. dist. (k=12; d=4)	0.7224	0.6412	0.7069	0.5372
Aproxim. univar. ext. inf-sup	**0.6309**	**0.6114**	**0.6509**	**0.4921**
Aproxim. univar. cen-rad	0.6391	0.6155	0.6541	0.5087

dos de predicción considerados al predecir cada una de las series temporales de los componentes. En este caso, al ajustar los métodos de predicción se ha tenido en cuenta la componente estacional de la serie. La tabla destaca en negrita aquellos métodos que mejor resultado han dado en cada serie. Tal y como puede verse, los modelos ARIMA con estacionalidad han sido los que mejores predicciones han obtenido en cada una de las cuatro series de los componentes. Sin embargo, el resto de métodos también funciona notablemente mejor que el ingenuo en todas las series. Esto quiere decir que, pese a la estabilidad del patrón estacional, el método ingenuo estacional ofrece una predicción muy susceptible de mejora.

A la vista de estos resultados, se han utilizado los modelos ARIMA para predecir los extremos, el centro y el radio de los intervalos, y para componer con ellos la predicción de la STI según la aproximación basada en los extremos y la basada en el centro y el radio. En la tabla 4.10 se muestran los resultados obtenidos por estas dos aproximaciones y por el resto de métodos que permiten predecir la STI en su totalidad.

La aproximación que pronostica de forma independiente las series de los extremos del intervalo es la que mejores resultados obtiene. Este resultado llama la atención ya que en las STI financieras trabajadas en los apartados anteriores esta aproximación era la que peores resultados obtenía en la mayoría de los casos. Por su parte, la aproximación que pronostica de forma independiente la serie de los centros y la de los radios también obtienen unos resultados notables en este ejemplo.

El alisado exponencial con estacionalidad de intervalo (AEEi) obtiene unos buenos resultados. Como era de esperar, este método funciona mejor que el alisado con estacionalidad clásica en la posición del intervalo (AEEc). Este resultado es lógico ya que tal y como muestra la figura 4.10 la estacio-

nalidad afecta al intervalo completo y no sólo a su posición. También resulta interesante comprobar como el AEEi obtiene mejores resultados en todos los componentes que los que se obtienen prediciendo las series temporales de los componentes por separado (fila "*alisados*" de la tabla 4.9). Este es un argumento a favor de los métodos de predicción que consideran al intervalo como una entidad en sí misma, en lugar de descomponerlo en componentes.

Por su parte, los modelos multivariantes VAR y VECM obtienen buenos resultados, pero se encuentran un peldaño por debajo del AEEI y dos con respecto a las aproximaciones que trabajan con las series univariantes de forma independiente. Por su parte, el iMLP y el k-NN para STI, aunque mejoran al método ingenuo, no son tan precisos como los métodos ya citados.

4.11. Conclusiones

En este capítulo se ha abordado el tema de la predicción de STI. Tal y como se indica en el apartado 4.3, este tipo de series son muy útiles para informar sobre la dispersión de un determinado fenómeno y, por tanto, son útiles en determinados contextos donde la variabilidad es crucial como, por ejemplo, las finanzas y la meteorología. Sin embargo, en dichos campos las STI normalmente no reciben un tratamiento que reconozca su naturaleza de intervalo.

El intervalo es un dato simbólico que implica una mayor complejidad que un dato clásico y que puede ser tratado de diferentes formas. Una aproximación que permite tratar los intervalos de forma sencilla consiste en descomponerlos en dos parejas de valores (extremo inferior y superior o centro y radio) de forma que el manejo de los intervalos se reduzca al manejo de una de esas parejas de valores. Otra aproximación para manejar datos de intervalos consiste en tratarlos empleando la aritmética de intervalos (Moore, 1966) que permite operar matemáticamente con ellos.

En el apartado 4.4 de este capítulo se han propuesto dos alternativas para medir el error en STI. Una de ellas mide el error de forma separada en cada uno de los componentes del intervalo y la otra utiliza la distancia como herramienta para representar la noción de diferencia entre un intervalo observado y su pronóstico. Cada una de las aproximaciones tiene su propio punto fuerte. La primera proporciona una mayor información sobre el error presentándolo desglosado por componentes y la segunda es más manejable, ya que muestra el error como un único valor que responde al criterio de la distancia empleada.

En este capítulo, también se han presentado distintas estrategias para predecir STI. En algunas de ellas, se utilizan métodos clásicos para predecir las series de los componentes de los intervalos. Mientras que en otras se han propuesto nuevos métodos que reconocen al intervalo como una entidad en sí misma y lo predicen como tal. Estos métodos son, en realidad, adaptaciones

de métodos clásicos ya existentes (e.g. alisados exponenciales, perceptrón multicapa y k-NN). La adaptación de estos métodos al contexto de las STI se ha realizado con ayuda de la aritmética de intervalos. Los métodos de predicción presentados en esta tesis se suman a los ya propuestos por Teles y Brito (2005) y por Maia et al. (2006a) y van conformando un cuerpo de técnicas para pronosticar STI.

El rendimiento de los métodos propuestos ha sido evaluado en el apartado 4.10. Las series que se han empleado proceden del ámbito económico (STI de la varición intradiaria de índices bursátiles y del cambio de divisas) y del ámbito meteorológico (STI de temperaturas). De los resultados obtenidos en los ejemplos pueden extraerse las siguientes conclusiones:

- Las aproximaciones propuestas para predecir STI han demostrado su capacidad de predicción en los ejemplos considerados. En todas las STI analizadas, se han conseguido mejores predicciones que las obtenidas por el método ingenuo con algunas, cuando no con todas, las aproximaciones propuestas. Esto ha sucedido también en las series de origen financiero donde es habitualmente complicado batir al método ingenuo. Dicho resultado demuestra la valía del enfoque planteado.

- No existe ninguna aproximación que haya resultado la mejor para todos los casos analizados. Sin embargo, la aproximación que trabaja de forma independiente con las series de los centros y de los radios y, en menor medida, los modelos VAR y VECM han obtenido muy buenos resultados en todos los casos, con lo que parecen buenas aproximaciones para predecir STI.

- En las STI financieras, la aproximación que predice la STI a partir de los modelos univariantes de los extremos de los intervalos no ha obtenido buenos resultados. La principal fuente de error en dichas STI estuvo en la serie temporal de los radios de los intervalos. Esto parece indicar que dicha aproximación, al trabajar con las series de forma independiente, no es capaz de predecir adecuadamente el radio en las series del contexto financiero. Sin embargo, el resto de aproximaciones propuestas si han obtenido predicciones del radio más precisas que las proporcionadas por el método ingenuo y por dicha aproximación.

- Los modelos propuestos en esta tesis que predicen la STI tratando al intervalo como tal, i.e., los alisados exponenciales, el iMLP y el k-NN, han obtenido buenos resultados en todos los casos analizados. Estos modelos son conceptualmente más correctos que los que trabajan con las series de los extremos de forma independiente, ya que estos, al trabajar con las series de los extremos de forma independiente, pueden dar lugar a predicciones sin sentido donde el extremo inferior del intervalo pronosticado esté por encima del extremo superior. Esta superioridad a

nivel conceptual se ha visto refrendada también en los resultados, porque, en los ejemplos analizados, los métodos de predicción que trabajan con el intervalo como tal han logrado obtener mejores predicciones para los extremos de los intervalos que la aproximación que predicen de forma independiente las series de los extremos.

- La diferenciación para STI propuesta en el apartado 4.9.1 ha demostrado su utilidad en las series del ámbito financiero analizadas. En dichas series existía una tendencia estocástica que afectaba a la posición del intervalo. Sin eliminar esa tendencia el iMLP y el k-NN para STI obtenían peores predicciones que el método ingenuo. Sin embargo, como era de esperar, una vez eliminada dicha tendencia, las predicciones de dichos modelos mejoraron considerablemente. Este comportamiento es similar al que se observo en el caso de las series temporales de los extremos y del centro el caso financiero, donde el MLP y el k-NN para series temporales clásicas necesitaban trabajar con las series diferenciadas para poder obtener buenas predicciones.

- En todos los casos analizados, las series temporales de los extremos de los intervalos eran series cointegradas. Sin embargo, al recoger dicha relación de cointegración en un modelo multivariante, i.e. al utilizar un modelo VECM, no se han obtenido predicciones sustancialmente mejores que las obtenidas por un modelo VAR que no recogiese explícitamente dicha relación.

- La aproximación planteada por Maia et al. (2006a) que consiste modelar de forma independiente las series de los centros y de los radios mediante un modelo híbrido ARMA+MLP no debe ser considerado como una regla de aplicación general, tal y como sugieren estos autores. Es cierto que la versatilidad de ambos modelos permite obtener predicciones aceptables para casi cualquier tipo de serie temporal, sin embargo, las buenas prácticas en predicción indican que cada serie debe ser pronosticada con el modelo que mejor se adapte a ella y que los modelos más parsimoniosos, i.e. con menos parámetros, suelen obtener mejores resultados. Por ello, dado que la serie temporal de los centros y la de los radios suelen tener un comportamiento muy diferente, lo adecuado es utilizar para cada una de ellas el modelo que mejor se adapte y no imponer un modelo con tantos parámetros como el híbrido sin un análisis previo. De hecho, en los ejemplos analizados el modelo híbrido no ha sido el más preciso en la predicción de las series de los centros en ninguno de los ejemplos planteados y sólo lo ha sido una vez en la predicción de la serie de los radios (en el ejemplo del apartado 4.10.4).

- El método de predicción de STI planteado por Teles y Brito (2005) no ha sido comparado en este capítulo. Dicho método es un modelo ARMA

para STI que trabaja con las series de los extremos, donde la ecuación que rige ambas series es la misma con excepción de la constante. A la vista de lo sucedido con la adaptación de otros métodos de predicción al contexto de las STI, es de esperar que este modelo ARMA hubiese obtenido unos resultados mejores que los obtenidos por los modelos ARIMA aplicados sobre las series de los extremos del intervalo. En cualquier caso, el modelo ARMA propuesto por Teles y Brito (2005) es más adecuado para tratar intervalos ya que asegura que en todo caso, la predicción obtenida será siempre un intervalo; lo cual no tiene por qué suceder al trabajar con las series de los extremos de forma independiente.

Como es natural, la evidencia empírica que proporcionan los ejemplos analizados es limitada y las conclusiones que se han obtenido no deben extrapolarse de forma directa a otros ámbitos, ni a otras STI, sino que deberán ser revisadas a medida que se obtengan nuevos resultados. Pese a ello, este trabajo constituye una referencia valiosa para futuros trabajos que pretendan abordar la predicción de STI.

Capítulo 5

Predicción de Series Temporales de Histogramas

Predecir es como conducir un coche con los ojos vendados siguiendo las indicaciones de alguien que mira por la luna trasera.

Anónimo

Las series temporales de histogramas son una herramienta que permite representar series temporales de distribuciones, es decir, series en las que cada instante se describe mediante una distribución. Este capítulo abordará diferentes aspectos sobre este nuevo tipo de series temporales como, por ejemplo, la idoneidad de los histogramas como método de representación de las distribuciones, la obtención de series temporales de histogramas o la definición de medidas de error para este tipo de series temporales. Para poder predecir estas series, en el capítulo se adaptarán dos métodos de predicción para series temporales clásicas. Los métodos en cuestión son los alisados exponenciales, que se adaptarán por medio de la aritmética de histogramas y por medio del uso de baricentros, y el método de los k vecinos más próximos que se adaptará empleando baricentros. La capacidad predictiva de estos métodos será evaluada por medio de una batería de ejemplos de diversos ámbitos.

5.1. Introducción

En el capítulo anterior, se abordaron las series temporales de intervalos. El intervalo representa la variabilidad de una observación en un determinado instante mediante un rango de valores (dicho rango puede ser el rango absoluto o, por ejemplo, el recorrido intercuartílico). Sin embargo, el intervalo no informa de lo que sucede entre los extremos considerados, es decir, no indica cómo se distribuyen las observaciones dentro del rango. Para ello,

es necesario un dato simbólico que permita representar de una forma más completa una distribución de datos cuantitativos. Los histogramas permiten cubrir dicha carencia. Este capítulo sienta las bases para trabajar y predecir series temporales de histogramas (STH). En primer lugar, se definirá este tipo de series temporales.

5.2. Definición de Serie Temporal de Histogramas

Definición. Una serie temporal de histogramas $\{h_{X_t}\}$ puede definirse como una secuencia de distribuciones observadas en instantes sucesivos en el tiempo denotados por $t = 1, ..., n$, donde cada distribución es representada mediante un histograma h_{X_t} que viene dado por

$$h_{X_t} = \{([I]_{t,1}, \pi_{t,1}), ..., ([I]_{t,p_t}, \pi_{t,p_t})\}, \text{ con } t = 1, ..., n, \tag{5.1}$$

donde $\{\pi_{t,i}\}$ con $i = 1, ..., p_t$ es una distribución de frecuencia o de probabilidad en el dominio considerado que cumple que $\pi_{t,i} \geq 0$ y que $\sum_{i=1}^{p_t} \pi_{t,i} = 1$; y donde $[I]_{t,i} \subseteq \mathcal{B}$, $\forall t, i$, es un intervalo (también llamado barra) definido como $[I]_{t,i} = [\underline{I}_{t,i}, \overline{I}_{t,i})$ con $-\infty < \underline{I}_{t,i} \leq \overline{I}_{t,i} < \infty$ e $\underline{I}_{t,i} \leq \overline{I}_{t,i-1}$ $\forall t, i$, con $i \geq 2$.

La notación que se ha propuesto para representar los histogramas que forman parte de una serie temporal no coincide con la notación de los histogramas mostada en el capítulo 2. Sin embargo, la notación que aquí se propone es la que se ha considerado más adecuada para trabajar en el contexto de las series temporales.

Desde la perspectiva del análisis de datos simbólicos, una STH puede definirse como una serie temporal donde las observaciones son realizaciones de variables aleatorias simbólicas de histograma. Cada histograma representa la densidad observada en cada instante temporal. Los histogramas son un tipo de estimador de densidad muy popular (Simonoff, 1996). Para poder trabajar con ellos, es necesario definir la función de densidad que llevan asociada. En esta tesis se asumirá que dentro de cada subintervalo del histograma las observaciones se encuentran distribuidas uniformemente. Esta hipótesis es la que se asume normalmente cuando se manejan histogramas en el contexto del análisis de datos simbólicos (Billard y Diday, 2003) y cuando se emplean los histogramas como estimadores de densidad. Dado el histograma $h = \{([I]_i, \pi_i)\}$ con $i = 1, ..., p$, sus funciones de densidad y de distribución se definen como se indica a continuación.

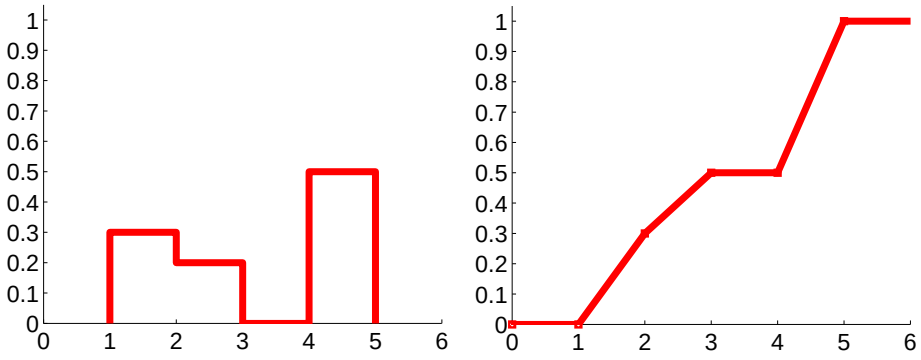


Figura 5.1: Función de densidad (izqda.) y de distribución (dcha.) del histograma $h = \{([1,2),.3),([2,3),.2),([3,4),.2),([4,5],.5)\}$.

Función de densidad de un histograma. La función de densidad del histograma h se define como

$$h(x) = \begin{cases} \frac{\pi_l}{\overline{I}_l - \underline{I}_l}, & \text{si } x \in [\underline{I}_l, \overline{I}_l), l \in \{1, ..., p\}. \\ 0, & \text{en otro caso} \end{cases} \tag{5.2}$$

Función de distribución de un histograma. La función de distribución del histograma h es su función de densidad acumulada de dicho histograma que se define como

$$H(x) = \int_{-\infty}^{x} h(x)dx = \begin{cases} 0, & \text{si } x \leq \underline{I}_1; \\ \sum_{j=1}^{l-1} \pi_j + \frac{x - \underline{I}_l}{\overline{I}_l - \underline{I}_l}\pi_l, & \text{si } x \in [\underline{I}_l, \overline{I}_l), l \in \{1, ..., p\} \\ 1, & \text{si } x \geq \overline{I}_p \ . \end{cases} \tag{5.3}$$

En la figura 5.1 se muestra la función de densidad y distribución de un histograma con el fin de ilustrar las definiciones anteriores.

5.3. ¿Por qué usar histogramas?

En esta tesis se propone el uso de los histogramas como una herramienta para representar distribuciones. Sin embargo, cabe preguntarse por qué razón usar histogramas en lugar de otros métodos que representen la densidad subyacente de una forma más suavizada y precisa, como, por ejemplo, los *kernels*. En este apartado se pretende responder a esa pregunta dando argumentos en favor del uso de los histogramas.

La versatilidad del histograma. En primer lugar, es interesante destacar que la definición de histograma dada en el apartado anterior es lo suficientemente flexible como para reflejar cualquier estimador de densidad basado en intervalos. Entre ellos se pueden destacar los siguientes

- Los histogramas equiespaciados. Son histogramas donde cada intervalo tiene exactamente la misma longitud. Este es el histograma clásico que se encuentra implementado en la mayoría de paquetes de software estadístico.

- Los histogramas equifrecuenciales, es decir, los histogramas donde la frecuencia asociada a cada intervalo es la misma; ver, por ejemplo, Burman (2002). Un caso particular de este tipo de histograma son los gráficos de caja (o *boxplots*) propuestos por Tukey (1977), que dividen una muestra en cuatro regiones de igual frecuencia. Tal y como indica Benjamini (1988), los gráficos de cajas ofrecen interesantes propiedades, una de ellas es que su sencilla representación gráfica permite conocer de un vistazo la forma de la distribución subyacente. Los *boxplots* son sólo un posible tipo de histograma equifrecuencial, pero pueden plantearse otros como, por ejemplo, aquellos construidos a partir de los deciles de la distribución.

- Los histogramas construidos sobre una partición. El analista divide el dominio de la variable considerada en los intervalos que considere relevantes para el problema que quiera analizar, sin necesidad de que los intervalos tengan la misma longitud. Un ejemplo de este tipo de histograma sería aquel que representa la distribución de ingresos de los españoles en determinadas bandas de distinta longitud, e.g. $[0-10000)$, $[10000-15000)$, $[15000-20000)$, $[20000-30000)$, $[30000-40000)$, etc.

- Los histogramas definidos sobre una una secuencia de cuantiles. En este caso, el analista determina la secuencia de cuantiles para el conjunto de datos que está estudiando. El histograma recogerá la frecuencia que se da entre cada par de cuantiles consecutivos. Este tipo de histograma permite poner un mayor énfasis en la parte de la distribución que se desee, e.g. la parte central de la distribución, alguno de sus extremos, etc. Un ejemplo de este tipo de histograma sería una variante del *boxplot* en el que se prestase una mayor atención a los extremos de la distribución mostrando también los cuantiles de .05, .1, .9 y .95.

Todos los tipos de histograma mencionados pueden resultar muy útiles en distintas situaciones y demuestran que la definición de histograma dada es lo suficientemente versátil y flexible como para adaptarse a distintos requerimientos. De hecho, los dos últimos son mencionados por Tay y Wallis (2000) como dos tipos de representación empleada en algunas aplicaciones a

la hora de representar las densidades de predicción. Normalmente, será responsabilidad del analista determinar qué tipo de representación le conviene más al problema que esté abordando en cada momento.

La precisión del histograma. Pese a su sencillez, los histogramas también pueden ofrecer representaciones precisas de la densidad subyacente si es necesario. La mayor parte de la investigación al respecto se centra en los histogramas equiespaciados, ya que éstos son, sin duda, los más utilizados en la práctica.

En un primer momento, el número de intervalos fue el parámetro que se utilizó para controlar la precisión del histograma equiespaciado como estimador de la distribución subyacente. Si el número de intervalos es demasiado bajo, se enmascara el aspecto de la distribución; por el contrario, si es demasiado alto, se puede estar representando la distribución con un nivel de detalle espurio.

Sturges (1926) propuso la primera regla para determinar el número de intervalos

$$n^{\text{o}} \textit{ de intervalos} = 1 + \log_2 n, \tag{5.4}$$

donde n es el tamaño de la muestra. Sin embargo, la regla de Sturges produce histogramas sobresuavizados, especialmente para valores grandes de n. Por ello, se desarrollaron otras aproximaciones para determinar el ancho del intervalo. Dichas aproximaciones miden la precisión con la que la densidad del histograma, $h(x)$, representa a la densidad subyacente, $f(x)$, por medio del error cuadrático integrado,

$$ECI = \int_{-\infty}^{\infty} [h(x) - f(x)]^2 du, \tag{5.5}$$

y de su valor esperado, el error cuadrático integrado medio (ECIM). Sin embargo, para ello es necesario conocer la distribución subyacente, lo cual no suele ser lo habitual. Scott (1979) propuso una regla para determinar el ancho cuando la densidad subyacente es Gaussiana

$$\textit{ancho del intervalo} = 3.491 \hat{\sigma} n^{-1/3}, \tag{5.6}$$

donde $\hat{\sigma}$ es una estimación de la desviación estándar. Scott (1992) desarrolló extensiones de esta regla para estimar el ancho permitiendo diferentes grados de asimetría y de curtosis.

Más interesante es la aproximación planteada por Wand (1997) que extiende las reglas de Scott para obtener el ancho de intervalo óptimo sea cual sea la densidad subyacente. Esta familia de reglas requiere de la definición de un parámetro l. Dicho parámetro indica el número de etapas de la estimación funcional que se realizan sobre el estimador empleado en la primera etapa (que suele ser una distribución normal). Por tanto, valores altos de l

implican un mayor número de etapas de estimación y dan lugar a una estimación con un sesgo menor, pero con una posible mayor varianza. Wand indica que $l = 2$ es un valor adecuado. Computacionalmente hablando, los requerimientos de este método son similares a los requeridos para estimar la densidad mediante un *kernel*.

Para el caso del histograma equifrecuencial, Burman (2002) propone un método para hallar el número adecuado de intervalos, tanto para el caso de un conjunto de datos univariante, como para el caso multivariante. El enfoque se basa en realizar validación cruzada con el objetivo de minimizar el error cuadrático integrado (5.5). Sin embargo, según el propio Burman, el estimador resultante no funciona bien si en la muestra existen regiones de baja densidad.

En teoría, tal y como indica Simonoff (1996), si no se impone la restricción de usar un ancho de intervalo fijo o de frecuencia fijo, se pueden obtener mejores representaciones de la densidad subyacente. El problema estriba en que es preciso conocer dicha distribución de antemano.

Otro parámetro que debe ser estimado a la hora de construir los histogramas equiespaciados es la posición de anclaje, i.e. la posición que toma el extremo izquierdo del primer intervalo del histograma. De acuerdo con Simonoff y Udina (1997), la posición de anclaje es determinante en la apariencia del histograma. Según estos autores, el ECIM no es muy efectivo a la hora de cuantificar lo bien que un estimador de densidad aproxima la apariencia de la densidad subyacente verdadera. Por ello, proponen un índice de estabilidad, G, que, dado un conjunto de datos y un ancho de intervalo, evalúa los cambios potenciales en la apariencia del histograma ante distintas posiciones de anclaje. Si $G >= 0.85$, el ancho de intervalo considerado produce histogramas estables. Para más detalles sobre el parámetro G se recomienda consultar el artículo de Simonoff y Udina (1997).

Facilidad de tratamiento computacional. Puede argumentarse que el uso de estimadores de densidad como los *kernels* o las mixturas de distribuciones Gaussianas permiten obtener representaciones más suaves de la densidad subyacente que las ofrecidas por el histograma. En otras palabras, puede considerarse que la representación escalonada que ofrece el histograma no es representativa de la densidad subyacente, ya que ésta rara vez será escalonada.

Sin embargo, ese inconveniente desde el punto de vista de la representación es su mayor virtud a la hora de tratar a los histogramas computacionalmente. La representación por medio de intervalos facilita el manejo computacional de los histogramas ya que permite de forma sencilla: realizar operaciones aritméticas con ellos, calcular distancias entre sus funciones de densidad y estimar el baricentro de un conjunto de histogramas. El hecho de manejar histogramas disminuye drásticamente el tiempo de computación

requerido por estas operaciones, que son, además, las operaciones en las que se van a basar los métodos de predicción que se desarrollan en este capítulo. El uso de histogramas facilita, por tanto, el uso de técnicas basadas en el cálculo intensivo como, por ejemplo, las búsquedas de parámetros óptimos para los métodos de predicción que se van a plantear.

Las virtudes del histograma como método de representación de datos pueden resumirse de la siguiente manera:

- Es una representación cercana a los datos originales y que no precisa de la imposición sobre los mismos de ninguna distribución *a priori*.
- Su versatilidad permite al analista centrarse en las características que más le interesen, e.g., en un conjunto de cuantiles o en una parte del rango de la variable.
- Permite describir las características esenciales de los datos con una precisión razonable.
- Su sencilla estructura simplifica su tratamiento computacional.

Dicho esto, aquellos que permanezcan escépticos ante las virtudes de los histogramas deben considerar si es realmente necesaria una herramienta más sofisticada. En muchos contextos prácticos es posible que baste con emplear histogramas y que los argumentos en favor de herramientas más sofisticadas no justifiquen su uso.

5.4. Medidas de Error para Series Temporales de Histogramas

En la predicción de series temporales clásicas, el error en el instante t se mide como la diferencia entre el valor pronosticado y el valor observado, i.e. $e_t = X_t - \hat{X}_t$. A la hora de desarrollar medidas de error para STH, hay que tener en cuenta que la complejidad del histograma es notablemente mayor que la del valor clásico, lo cual dificulta la medición del error. Por esta razón, para trabajar con histogramas es necesario redefinir el concepto de error.

En este apartado se analizarán una serie de posibles alternativas para medir el error en STH y se desarrollará a fondo la que se ha considerado adecuada.

5.4.1. Análisis de diferentes alternativas para elaborar medidas de error para STH

Una primera posibilidad para desarrollar medidas de error para STH consiste en definir el error en el instante t como la diferencia entre el histograma

observado y el histograma pronosticado, $h_{X_t} - \hat{h}_{X_t}$. Para ello se puede emplear la resta de la **aritmética de histogramas** propuesta por Colombo y Jaarsma (1980)[1]. Sin embargo, el resultado de esta operación no informa sobre la similitud entre los histogramas considerados. Para ilustrar este hecho supongamos el caso de una predicción perfecta tal que $h_{X_t} = \hat{h}_{X_t} = h_A$. Si utilizamos la operación de resta entre histogramas definida por Colombo y Jaarsma (1980), el error calculado como $h_{e_t} = h_{X_t} - \hat{h}_{X_t}$ tomará valor cero, i.e. $h_{e_t} = \{([0,0],1)\}$, si y sólo si $h_A = \{([a,a],1)\}$ con $a \in \Re$, es decir, si y solo si el histograma considerado es un valor clásico. Este hecho sucede porque la resta de histogramas propuesta por Colombo y Jaarsma (1980) tiene como objetivo representar la distribución de la resta entre cada par de valores posibles de cada uno de los histogramas considerados y no sirve para medir la diferencia que existe entre dos histogramas.

Descartado el uso de la aritmética de histogramas, puede proponerse el uso de **contrastes de bondad del ajuste** para determinar si un histograma pronosticado se corresponde o no con el histograma observado. En el área de las predicciones de densidad, Diebold, Gunther y Tay (1998) emplean los contrastes de bondad del ajuste para evaluar si una predicción de densidad se corresponde o no con la densidad verdadera. Sin embargo, este enfoque no sirve como base para desarrollar medidas de error en STH. La razón es que un contraste determina si ambas distribuciones son o no iguales de forma estadísticamente significativa, pero no ofrece información cuantitativa indicando lo similar o lo diferente que es una distribución de la otra. El concepto de error que buscamos debe ser mensurable, ya que de esa forma podrá usarse como referencia para ajustar los parámetros de un método de predicción o para elegir entre distintos métodos de predicción.

Otra posibilidad para medir el error en STH consiste en emplear **medidas de divergencia** para cuantificar el error existente en una predicción. Dentro también del área de la predicción de densidades, Hall y James (2007) utilizan el criterio de información de Kullback-Leibler para combinar predicciones de densidad y para cuantificar las diferencias entre la densidad pronosticada y la real. Dadas dos funciones de densidad $f(x)$ y $g(x)$ definidas sobre $\mathbb{R}$, el criterio de información o medida de divergencia de Kullback-Leibler se define como

$$D_{K-L}(f,g) = \int_{\Re} \log\{\frac{f(x)}{g(x)}\} f(x) dx. \tag{5.7}$$

Desafortunadamente, esta medida no es apropiada para las STH ya que requiere que el soporte de una de las densidades consideradas sea el mismo o esté contenido dentro del soporte de la otra densidad (ya que en caso contrario la medida toma el valor infinito). Esta condición es frecuentemente violada en STH donde el histograma observado y el pronosticado se suelen solapar pero sin llegar a ser tener el mismo soporte.

[1] Más detalles sobre la aritmética de histogramas en el apartado 5.5.1.1.

Sin embargo, pese a que el criterio de información de Kullback-Leibler no es válido para nuestro propósito, el uso de medidas de divergencia sí es un enfoque adecuado para proponer medidas de error para STH. Las medidas de divergencia proporcionan valores que representan de forma objetiva las diferencias entre la densidad verdadera y la pronosticada. Por tanto, basándonos en una medida de divergencia se puede desarrollar un concepto error entre densidades que sea cuantificable y definido en base a un criterio claro, como es la definición de la propia medida.

En el siguiente apartado, se abordará la definición de medidas de error basadas en medidas de divergencia.

5.4.2. Medidas de error para STH basadas en distancias

En este apartado, se va a analizar qué medidas de divergencia resultan adecuadas para reflejar el concepto de error entre histogramas. Como resultado del estudio, dos medidas de divergencia, que además cumplen las propiedades para ser consideradas distancias, serán consideradas como idóneas para proponer medidas de error para STH. Posteriormente se analizará la interpretación de las mismas y, por último, se propondrán una serie de medidas de error para STH basadas en ellas.

5.4.2.1. Análisis de medidas de divergencia para distribuciones

Gibbs y Su (2002) y Bock (2000) ofrecen sendas revisiones sobre las medidas de divergencia para distribuciones de probabilidad. En la tabla 5.1 muestran las más empleadas en la práctica, considerando que $f(x)$ y $g(x)$ son las funciones de densidad definidas para valores $x \in \Re$, que $F(x)$ y $G(x)$ son las funciones de distribución y que $F^{-1}(x)$ y $G^{-1}(x)$ son las inversas de dichas funciones de distribución.

Tabla 5.1: Medidas de divergencia para distribuciones

Divergencia de Kullback-Leibler	$D_{K-L}(f,g) = \int_{\Re} \log\{\frac{f(x)}{g(x)}\} f(x)dx$
Divergencia de Jeffrey	$D_J(f,g) = D_{K-L}(f,g) + D_{K-L}(g,f)$
Divergencia χ^2	$D_{\chi^2}(f,g) = \int_{\Re} \frac{\lvert f(x)-g(x)\rvert^2}{g(x)} dx$
Distancia de Hellinger	$D_H(f,g) = \left[\int_{\Re} (\sqrt{f(x)} - \sqrt{g(x)})^2 dx\right]^{\frac{1}{2}}$
Distancia de Variación Total	$D_{var}(f,g) = \int_{\Re} \lvert f(x) - g(x)\rvert dx$
Distancia de Wasserstein	$D_W(f,g) = \int_{\Re} \lvert F(x) - G(x)\rvert dx$
Distancia de Mallows	$D_M(f,g) = \sqrt{\int_0^1 (F^{-1}(t) - G^{-1}(t))^2 dt}$
Distancia de Kolmogorov	$D_K(f,g) = \text{máx}_{\Re} \lvert F(x) - G(x)\rvert$

Las divergencias de Jeffrey y de χ^2 sufren el mismo problema que el comentado en el apartado anterior para la divergencia de Kullback-Leibler.

Estas medidas toman un valor infinito si las funciones de densidad consideradas no tienen el mismo soporte, lo cual es habitual en STH. Por tanto, estas medidas no son adecuadas para desarrollar medidas de error para STH.

Tanto la distancia de Hellinger, como la distancia de Variación Total y la distancia de Kolmogorov toman valores en un intervalo acotado. La primera en el intervalo $[0, \sqrt{2}]$, la segunda en $[0, 2]$ y la tercera en $[0, 1]$. Estas distancias alcanzan su valor máximo cuando los soportes de las funciones de densidad no intersectan, sin importar lo alejados que estén los soportes. Esta característica no es adecuada para desarrollar medidas para STH ya que el error debe ser mayor cuanto más alejado se encuentre el histograma pronosticado del histograma observado. Por tanto, dichas medidas se pueden descartar.

Las distancias de Wasserstein y de Mallows (Mallows, 1972) sí resultan adecuadas ya que toman valores en el intervalo $[0, l]$, donde l es la longitud del dominio, por lo que cuanto más alejadas estén las funciones de densidad consideradas, mayor será la distancia entre ellas.

Es interesante reseñar que la distancia de Wasserstein puede representarse en función de las inversas de las funciones de distribución de la siguiente manera

$$D_W(f, g) = \int_{\Re} |F(x) - G(x)|dx = \int_0^1 |F^{-1}(t) - G^{-1}(t)|dt. \tag{5.8}$$

De esta forma, puede verse que las distancias de Wasserstein y de Mallows son casos particulares de la siguiente expresión

$$D(f, g) = \left(\int_0^1 (F^{-1}(t) - G^{-1}(t))^p dt \right)^{1/p} \tag{5.9}$$

con $p = 1$ y $p = 2$, respectivamente. Es decir, ambas distancias mantienen una relación similar a las que guardan la distancia Manhattan y la Euclídea, que son casos particulares de la métrica de Minkowski.

Puede verse un análisis riguroso sobre las propiedades de las distancias presentadas en Gibbs y Su (2002). A continuación, se muestra un ejemplo que ilustra cómo se comportan estas medidas de divergencia y que servirá para reforzar el argumento de que las distancias de Wasserstein y de Mallows resultan adecuadas para medir errores en STH.

Un ejemplo ilustrativo. Consideremos un histograma observado $h_A = \{([0, 1), .7), ([1, 2), .3)\}$ y dos predicciones o estimaciones de h_A como, por ejemplo, $h_B = \{([1, 2), .2), ([2, 3), .8)\}$ y $h_{B'} = \{([1, 2), .2), ([5, 6), .8)\}$. La figura 5.2 muestra las funciones de densidad y de distribución de estos histogramas. Si observamos dichas imágenes, podemos concluir que los histogramas h_A y h_B son más similares entre sí de lo que lo son los histogramas h_A y $h_{B'}$.

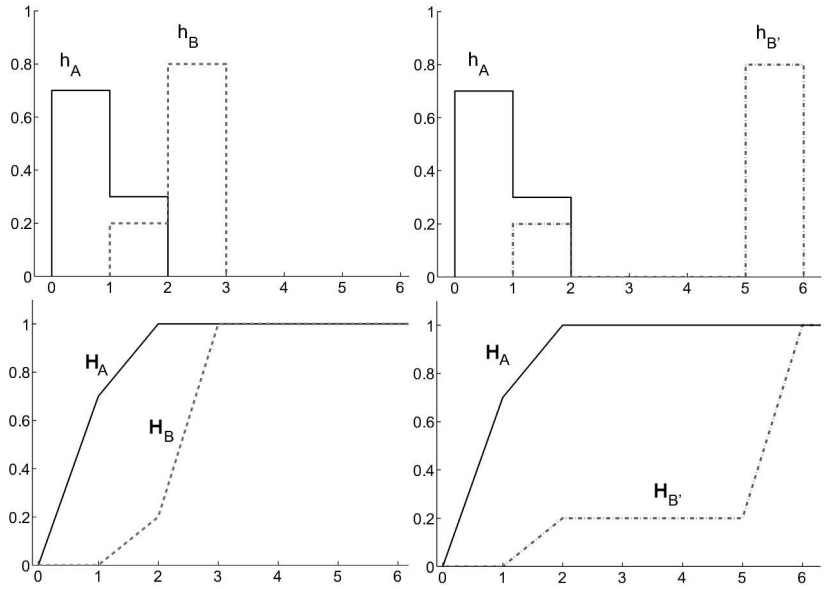


Figura 5.2: Arriba: Funciones de densidad de h_A y h_B (izqda.) y de h_A y $h_{B'}$ (dcha.). Abajo: Funciones de distribución acumulada de h_A y h_B (izqda.) y de h_A y $h_{B'}$ (dcha.).

Tal y como se puede ver en la tabla 5.2, de todas las medidas de divergencia consideradas, sólo las distancias de Wasserstein y de Mallows reflejan que el histograma h_B es más similar a h_A de lo que es $h_{B'}$. Es decir, son las únicas medidas que reflejan de forma adecuada el concepto de semejanza que capta el ojo humano.

Observando las fórmulas de la tabla 5.1 se puede ver que la distancia de Variación Total mide el área encerrada entre las funciones de densidad, mientras que la distancia de Wasserstein mide el área encerrada entre las funciones de distribución. Tal y como se puede ver en la figura 5.2, el área encerrada por las funciones de densidad entre h_A y h_B es idéntica a la encerrada entre h_A y $h_{B'}$. Mientras que no sucede así si comparamos sus funciones de distribución. Esto ilustra la idea de que las medidas de divergencia basadas en las funciones de distribución son más eficientes a la hora de reflejar la semejanza que aquellas basadas en las funciones de densidad. Esta aseveración es reforzada también por el hecho de que los contrastes de bondad del ajuste para distribuciones de probabilidad como el de Kolmogorov-Smirnov o el de Cramer-von Mises estén basados en funciones de distribución.

Sin embargo, no todas las medidas de divergencia basadas en funciones de distribución reflejan adecuadamente la semejanza que percibe el ojo humano. La distancia de Kolmogorov es un contraejemplo de esta teoría, porque emplea funciones de distribución y en el ejemplo obtiene unos resultados que contradicen el análisis visual de los histogramas. Esto es debido a

Tabla 5.2: Divergencia entre los histogramas h_A y h_B y los histogramas h_A y $h_{B'}$ según las diferentes medidas consideradas

Medida de divergencia	$D(h_A, h_B)$	$D(h_A, h_{B'})$
Divergencia de Kullback-Leibler	∞	∞
Divergencia de Jeffrey	∞	∞
Divergencia χ^2	∞	∞
Distancia de Hellinger	1.229	1.229
Distancia de Variación Total	1.6	1.6
Distancia de Wasserstein	1.5	3.9
Distancia de Mallows	1.52	4.11
Distancia de Kolmogorov	0.8	0.8

que esta distancia sólo tiene en cuenta el punto de la recta real para el cual la diferencia entre las dos funciones de distribución es máxima, no teniendo en cuenta el comportamiento de las funciones en el resto de la recta real.

El ejemplo mostrado prueba que las distancias de Wasserstein y de Mallows ofrecen una noción de divergencia que casa con la que se deriva de una inspección visual. No se trata de un caso aislado, pueden proponerse otros distintos y las conclusiones que se obtienen son similares.

A la luz del análisis realizado, se concluye que, de las distancias consideradas en este apartado, sólo las distancias de Wasserstein y de Mallows son adecuadas para reflejar el concepto de error en STH. En el apartado siguiente se profundizará sobre su interpretación.

5.4.2.2. Interpretación de las distancias de Wasserstein y de Mallows

Las distancias de Wasserstein y de Mallows tienen una interpretación clara e intuitiva, lo cual es una condición deseable en una medida de error. Dicha interpretación está relacionada con la de la *Earth Mover's Distance* o Distancia de los Transportistas de Arena (DTA). La DTA fue propuesta por Rubner et al. (2000) y es muy empleada en visión artificial para medir las diferencias entre histogramas que se emplean para representar las imágenes.

Los histogramas que se emplean en el área de visión artificial son distintos a los histogramas que se están considerando en este capítulo. Se trata de histogramas de una magnitud discreta cuyo rango va normalmente entre 0 y 255. Las magnitudes son, por ejemplo, el brillo y el color de la imagen. El histograma recoge el número de pixels que hay en la imagen analizada que toman cada uno de los valores entre 0 y 255 de la magnitud considerada.

La DTA es una solución al problema de transferencia de masas de Monge-Kantorovich (Rachev, 1984). Esto quiere decir que la DTA entre dos histo-

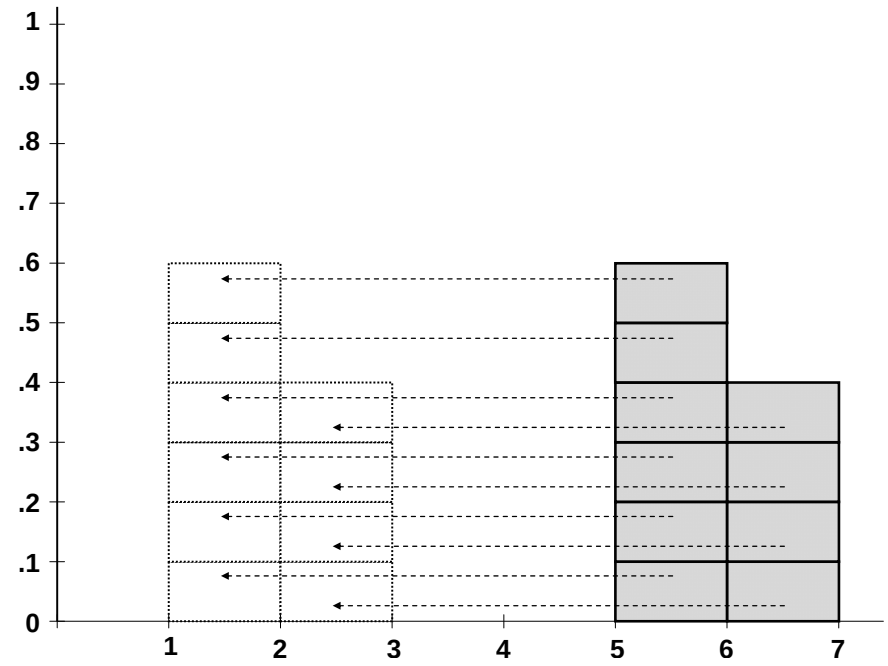


Figura 5.3: Representación gráfica de la Distancia de los Transportistas de Arena para los histogramas h_A y h_B.

gramas puede interpretarse como la cantidad mínima de trabajo requerida para transformar un histograma en otro por medio del transporte. Dado que Levina y Bickel (2001) demuestran que la DTA aplicada sobre distribuciones de probabilidad es idéntica a la distancia de Mallows y de Wasserstein, estas distancias pueden interpretarse de igual forma que la DTA. Por tanto, las distancias de Wasserstein y de Mallows pueden ser interpretadas como el trabajo requerido para transformar una distribución en otra transportándola hasta que ocupe su lugar. A continuación, se mostrará un ejemplo para explicar esta idea.

Consideraremos dos histogramas y consideraremos que uno de ellos es una masa de arena y el otro es un molde cuya capacidad es exactamente la de la masa de arena disponible. Los histogramas son $h_A = \{([1,2),.6),([2,3],.4)\}$ y $h_B = \{([5,6),.6),([6,7],.4)\}$. Estos histogramas son mostrados en la figura 5.3 en la que, para hacer más comprensible la idea, el histograma h_A es representado como una pila de bloques de longitud 1 y alto 0.1 en lugar de como una masa de arena.

Si únicamente consideramos el trabajo que hay que hacer para desplazar horizontalmente los bloques hacia el molde, el trabajo mínimo para mover los bloques del histograma h_A hasta completar el molde del histograma h_B consiste en realizar los movimientos que indican las flechas de la figura 5.3.

En física, el trabajo es una magnitud que depende de dos magnitudes vectoriales, la fuerza y la trayectoria. Si el módulo de la fuerza es constante y el ángulo que forma con la trayectoria también es constante, el trabajo será el producto escalar entre el vector de fuerza y el vector de distancia

$$W = \overrightarrow{F} \cdot \overrightarrow{d}. \tag{5.10}$$

En el caso que nos ocupa, hemos dicho que sólo se considerará el desplazamiento horizontal, por lo que la fuerza que hay que aplicar es paralela a la trayectoria de desplazamiento. Esto hace que el trabajo se calcule como el producto entre la magnitud de la fuerza y de la distancia

$$W = F \cdot d. \tag{5.11}$$

En nuestro caso, la distancia mínima que recorrerá cada bloque viene determinada por las flechas de la figura 5.3, cuya magnitud es $d_i = 4$ para los diez desplazamientos a realizar. Por su parte, la fuerza a aplicar para desplazar el bloque se considerará igual a la frecuencia que representa dicho bloque y dicha frecuencia es igual al producto entre el ancho y el largo del bloque. Como todos los bloques son de igual tamaño, la fuerza a aplicar para desplazar cada uno es $F_i = 0.1$. Por tanto, el trabajo realizado en cada uno de los desplazamientos es $W_i = F_i d_i = 0.1 \cdot 4 = 0.4$. Como en total hay que realizar diez desplazamientos idénticos, el trabajo para convertir un histograma en otro tiene un valor total de $W = \sum_i W_i = 4$.

El valor obtenido coincide con el que resulta de calcular la distancia de Wasserstein con la ecuación (5.8). Como se mencionó anteriormente, dicho valor mide el área encerrada entre las funciones de distribución acumulada de los histogramas considerados. La figura 5.4 muestra las funciones de distribución para los histogramas h_A y h_B. En este caso, resulta trivial calcular la distancia de Wasserstein gráficamente porque corresponde a la suma del área de los dos paralelogramos comprendidos entre ambas funciones de distribución.

La relación entre el concepto de trabajo y la distancia de Mallows para funciones de distribuciones discretas viene expresada según la siguiente fórmula

$$W = (\sum_i F_i d_i^p)^{\frac{1}{p}}, \tag{5.12}$$

es decir, que se calcularía de manera similar a una métrica de Minkowski de orden $p = 2$. Mientras que en el caso de la distancia de Wasserstein la métrica sería de orden $p = 1$.

Por último, otro aspecto importante sobra la interpretación de las distancias de Wasserstein y de Mallows es que ambas se calculan como diferencias entre las funciones de los cuantiles de los histogramas que se estén considerando. Más concretamente, ambas distancias se calculan a partir de los

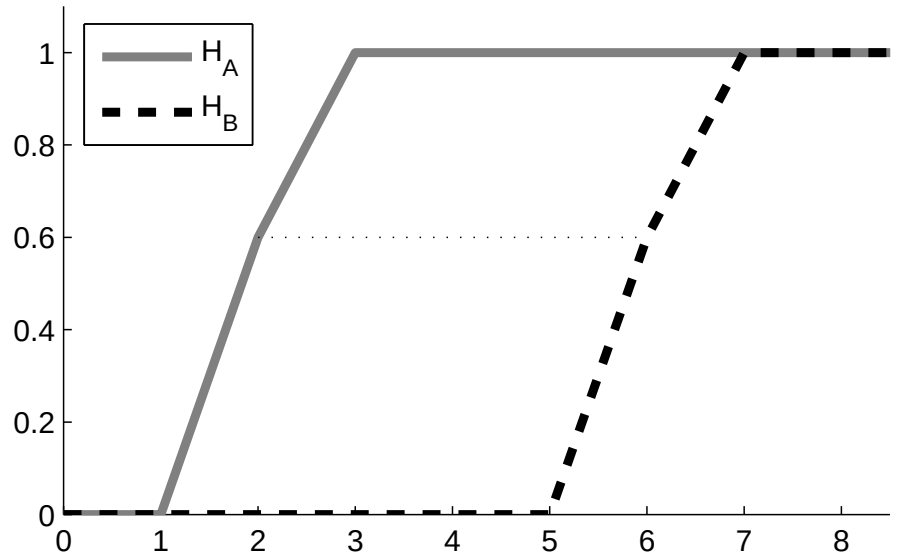


Figura 5.4: Funciones de distribución de los histogramas h_A y h_B.

valores $\delta_t = |F^{-1}(t) - G^{-1}(t)|$, $\forall t \in [0,1]$. La inversa de la función de distribución se conoce como la función de los cuantiles, i.e. el t-cuantil de f viene dado por el valor de la inversa de su función de distribución, $F^{-1}(t)$. Por ello, tal y como muestra la figura 5.5, cada valor δ_t puede considerarse como la distancia L_1 entre los cuantiles en t de las dos funciones de distribución. En la distancia de Mallows dichos valores están elevados al cuadrado. Por lo tanto, puede trazarse un paralelismo entre la distancia de Wasserstein y la distancia Manhattan, y entre la distancia de Mallows y la distancia Euclídea.

5.4.2.3. Definición del Error Medio basado en una Distancia

En los apartados anteriores se ha estimado adecuado reflejar los errores por medio de medidas de divergencia para distribuciones. De entre las medidas más habituales, sólo la de Wasserstein y la de Mallows se han considerado adecuadas. Ambas medidas se basan en las funciones de distribución acumuladas ya que, como se ha mostrado, este hecho hace que reflejen mejor las diferencias entre las distribuciones consideradas. Es interesante mencionar que cuando Diebold et al. (1998) plantean evaluar si las predicciones de densidad son adecuadas o no, proponen el uso de contrastes de bondad del ajuste sobre las funciones de distribución acumuladas.

Las medidas de divergencia consideradas son distancias, ya que cumplen los siguientes requisitos: son medidas definidas positivas, son simétricas y satisfacen la propiedad de la desigualdad triangular. Estas propiedades dotan

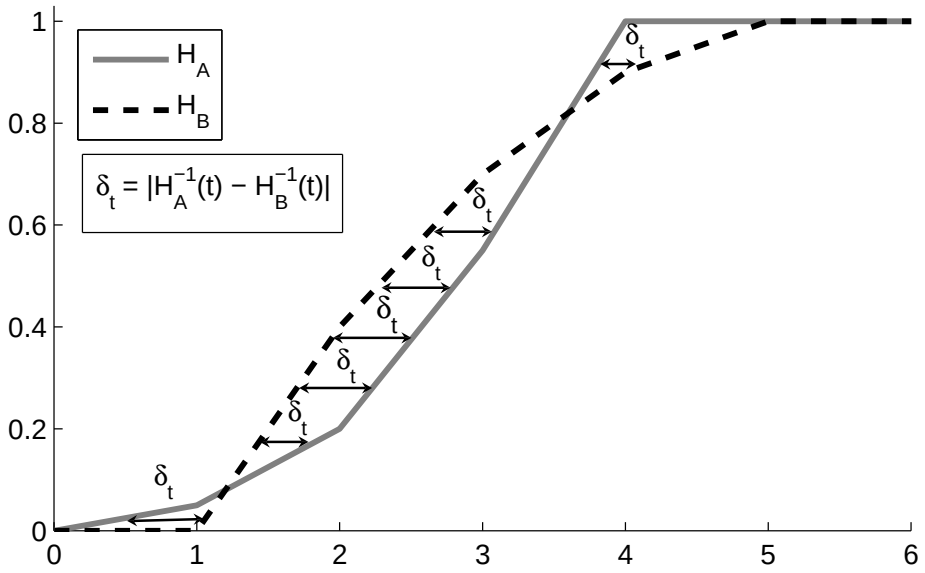


Figura 5.5: Representación de los valores $\delta_t = |H_A^{-1}(t) - H_B^{-1}(t)|$ sobre las funciones de distribución de dos histogramas cualesquiera h_A y h_B.

de mayor potencia significativa a esas medidas de divergencia. Además, el hecho de que las distancias sean definidas positivas hace que no sea necesario usar el valor absoluto ni el cuadrado para hacer que los valores resultante sean positivos. Al contrario de lo que sucede en las series temporales clásicas, donde se emplea el Error Absoluto Medio y el Error Cuadrático Medio para evitar la compensación de errores positivos y negativos, en STH los errores podrán ser agregados simplemente mediante la suma, sin necesidad de requerir a ninguna operación adicional. No obstante, también se puede plantear una medida que sea del estilo de la Raíz Cuadrada del Error Cuadrático Medio.

Sea $\{h_{X_t}\}$ la STH observada, y $\{\hat{h}_{X_t}\}$ la predicción de dicha STH, con $t = 1, ..., n$ el **Error Medio basado en la Distancia de Wasserstein o de Mallows** se define de la siguiente forma

$$EMD^q(\{h_{X_t}\}, \{\hat{h}_{X_t}\}) = \left(\frac{\sum_{t=1}^{n} \left(D(h_{X_t}, \hat{h}_{X_t}) \right)^q}{n} \right)^{1/q}, \tag{5.13}$$

donde $D(h_{X_t}, \hat{h}_{X_t})$ es la distancia de Wasserstein o de Mallows que aparece en la tabla 5.1 y q es el orden de la medida de error, de forma que si consideramos $q = 1$, agregamos las distancias como en el Error Absoluto Medio, y si $q = 2$, agregamos las distancias como en la Raíz Cuadrada del Error Cuadrático Medio.

Una medida de error escalada. Hyndman y Koehler (2006) proponen una nueva aproximación para el desarrollo de medidas de error robustas que no se vean afectadas por la unidad medida. Siguiendo las ideas que ellos proponen, puede plantearse una versión del EMD independiente de las unidades de medida. La medida que se va proponer escalará el EMD mostrado en (5.13) para $q = 1$.

En primer lugar, el error (i.e. la distancia) en t debe ser escalado por el EMD cometido por el método ingenuo en un periodo de tiempo que consideremos de referencia, es decir,

$$q_t = \frac{D(h_{X_t}, \hat{h}_{X_t})}{EMD_m}, \tag{5.14}$$

donde EMD_m es el error cometido por el método ingenuo en el periodo de referencia que se calcula como

$$EMD_m = \frac{1}{m-1} \sum_{i=2}^{m} D(h_{X_i}, h_{X_{i-1}}), \tag{5.15}$$

donde m es la longitud del periodo de referencia. Normalmente, el periodo de referencia será el periodo muestral disponible y el método de referencia será el método ingenuo, aunque podría ser empleado otro método al que se quiera considerar como el método a batir. Dicho esto, el **Error Medio Escalado basado en una Distancia** se define como

$$EMED(\{h_{X_t}\}, \{\hat{h}_{X_t}\}) = \frac{1}{n} \sum_{t=1}^{n} q_t = \frac{\frac{1}{n} \sum_{t=1}^{n} D(h_{X_t}, \hat{h}_{X_t})}{EMD_m}, \tag{5.16}$$

donde la distancia aplicada tanto en el numerador como en el denominador es la misma y puede ser la distancia de Mallows o la de Wasserstein. Si el EMED del periodo analizado es menor que 1, el método que se está empleando obtiene mejores predicciones en media que las conseguidas por el método de referencia para el periodo muestral. Por contra, si el resultado es mayor que 1, el método considerado predice peor.

5.5. Predicción mediante alisados exponenciales

En las series temporales clásicas, los métodos de alisado se emplean para eliminar las fluctuaciones de la serie en el corto plazo, resaltando la tendencia a largo plazo y el comportamiento cíclico. Dichos métodos obtienen las predicciones como un promedio de n valores pasados consecutivos de la serie temporal.

Dentro de los métodos de alisado, los más sofisticados son los métodos de alisado exponencial. Estos métodos surgen en la década de los 50 con los trabajos de Brown (1959) y de Holt (1957). Desde entonces se han realizado

avances muy significativos en la materia y se han propuesto múltiples variantes. Gardner (2006) ofrece una excelente revisión sobre la evolución de esta familia de métodos.

Los métodos de alisado exponencial, pese a partir de un concepto muy sencillo, como es el de generar las predicciones como un promedio, y pese a tener más de 50 años de antigüedad, siguen siendo un método de referencia en predicción debido al buen rendimiento que ofrecen en los estudios empíricos (Gardner, 2006).

En este apartado se adaptarán los métodos de alisado a las STH, prestando una mayor atención a los alisados exponenciales. Para realizar la adaptación se propondrán dos posibles caminos: uno basado en la aritmética de histogramas y otro basado en el método de los baricentros. En el punto A.2 de los apéndices se realiza un breve repaso a los conceptos básicos de los métodos de alisado en las series temporales clásicas.

5.5.1. El alisado de STH empleando la aritmética de histogramas

El alisado se realiza, o bien como un promedio, o bien como una suma ponderada de dos términos: el valor real y la predicción en el instante pasado. Por tanto, se requieren únicamente unas operaciones aritméticas muy simples para realizarlos: suma y división, o suma y multiplicación.

Una forma de adaptar los métodos de alisado para trabajar con datos de histograma es reemplazar la aritmética tradicional que opera sobre valores clásicos por una aritmética que permita operar histogramas que represente distribuciones de probabilidad. A continuación, se van a presentar los principios básicos de la aritmética de histogramas.

5.5.1.1. La aritmética de histogramas

Colombo y Jaarsma (1980) proponen un método para realizar operaciones aritméticas con variables aleatorias representadas por medio de histogramas. En su artículo, Colombo y Jaarsma simplemente esbozan el método. Por ello, para ver una explicación del mismo con mayor detalle se recomienda acudir a la tesis doctoral de Williamson (1989).

Según Colombo y Jaarsma, el histograma que resulta al realizar una operación aritmética entre dos histogramas se obtiene considerando todas las posibles parejas de intervalos de uno y otro histograma y operando con ellas según la aritmética de intervalos (Moore, 1979). Las operaciones básicas de la aritmética de intervalos fueron descritas en el apartado 2.7.1.1.

El método de Colombo y Jaarsma se describe como sigue. Dados dos histogramas, $h_A = \{([I]_{A,i}, \pi_{A,i})\}$ con $i = 1, ..., n$, y $h_B = \{([I]_{B,j}, \pi_{B,j})\}$ con $j = 1, ..., m$, que representan a dos variables aleatorias A y B, respectivamente, y siendo $\square$ un operador aritmético del conjunto $\{+, -, *, /\}$, enton-

ces $C = A\Box B$ puede ser representado mediante el histograma desordenado $h_C = \{([I]_{C,k}, \pi_{C,k})\}$ con $k = 1, ..., n \cdot m$, donde

$$\begin{aligned}
\underline{I}_{C,(i-1)m+j} &= \text{mín}\{\underline{I}_{A,i}\Box\underline{I}_{B,j}, \overline{I}_{A,i}\Box\underline{I}_{B,j}, \underline{I}_{A,i}\Box\overline{I}_{B,j}, \overline{I}_{A,i}\Box\overline{I}_{B,j}\},\\
\overline{I}_{C,(i-1)m+j} &= \text{máx}\{\underline{I}_{A,i}\Box\underline{I}_{B,j}, \overline{I}_{A,i}\Box\underline{I}_{B,j}, \underline{I}_{A,i}\Box\overline{I}_{B,j}, \overline{I}_{A,i}\Box\overline{I}_{B,j}\},\\
\pi_{C,(i-1)m+j} &= \pi_{A,i}\cdot\pi_{B,j}.
\end{aligned} \tag{5.17}$$

El número de intervalos del histograma resultante h_C es $k = n \cdot m$ y puede llegar a ser un valor alto. Además, los intervalos de dicho histograma, $[I]_{C,k}$, pueden estar desordenados y solaparse unos con otros. Estos hechos hacen que sea complicado interpretar el histograma resultado h_C, y que, si hay utilizarlo a continuación como operando, el número de intervalos del histograma resultadante se dispare. Por tanto, para evitar estos problemas es conveniente representar h_C de una forma más manejable. Esto puede realizarse reajustando h_C en un histograma de l intervalos disjuntos. Para ello, en primer lugar se deben ordenar los intervalos que constituyen el histograma h_C de forma que se considera que $[I]_{C,p} \leq [I]_{C,q}$ si

$$\underline{I}_{C,p} < \underline{I}_{C,q}, \text{ o si } (\underline{I}_{C,p} = \underline{I}_{C,q} \text{ e } \overline{I}_{C,p} < \overline{I}_{C,q}). \tag{5.18}$$

A partir del histograma donde los intervalos están ordenados, se debe construir un histograma disjunto, $h_D = \{([I]_{D,k}, \pi_{D,k})\}$ con $k = 1, ..., n \cdot m$. Para simplificar la notación consideraremos que el histograma ordenado no disjunto es $h_C = \{([z]_k, r_k)\}$, mientras que el histograma ordenado y disjunto es $h_D = \{([w]_k, s_k)\}$ con $k = 1, ..., n \cdot m$, el pseudocódigo para hallar el histograma disjunto es el siguiente

```
for (i := 1; i ≤ nm; i++){
    w_i := z_i;
    w̄_i := z_{i+1};
    s_i := 0;
    for (j := i; j > 0; j--){
        if (z̄_j > w_i and z_j < w̄_i){
            if (w̄_i > z̄_j and w_i ≥ z_j)
                s_i := s_i + r_j (z̄_j − w_i)/(z̄_j − z_j) ;
            else if (w̄_i ≤ z̄_j and w_i ≥ z_j)
                s_i := s_i + r_j (w̄_i − w_i)/(z̄_j − z_j) ;
        }
    }
}
```

Una vez que se tiene el histograma disjunto lo normal es representarlo como un histograma de menos intervalos y del mismo tipo que los histogramas operando. Por ejemplo, si el histograma disjunto tiene $n \cdot m$ intervalos,

puede reconstruirse como un histograma de l intervalos donde $l = \lfloor\sqrt{n \cdot m}\rfloor$, siendo $\lfloor x \rfloor$ la parte entera de x. Si los histogramas operando eran equiprobables, el histograma solución, al que por claridad denotaremos como $h_S = \{([v]_i, t_i)\}$ con $i = 1, ..., l$, tendrá todos los intervalos con la misma frecuencia $t_i = 1/(l-1)$ y se halla de la siguiente manera

```
u:=0;
for (i := 1, j := 1; i ≤ nm; i++){
      u := u + s_i;
      while (u ≥ 1/(l − 1)){
           j++;
           v_j := w_i + [(1/(l − 1) − (u − w_i))/w_i] ∗ (w̄_i − w_i);
           t_{j−1} := 1/(l − 1);
           u := u − 1/(l − 1);
      }
}
```

Es importante tener en cuenta que para realizar operaciones aritméticas entre histogramas e intervalos o entre histogramas y números reales, basta con considerar que todo intervalo $[a, b]$ es en realidad un histograma $h = \{([a, b], 1)\}$ y que todo número real a es un histograma $h = \{([a, a], 1)\}$. De hecho, la aritmética de histogramas subsume a la aritmética de intervalo y ésta a su vez, tal y como se dijo en el apartado 2.7.1.1, subsume a la aritmética clásica. Esto quiere decir que, si consideramos los histogramas $h_A = \{([a, a], 1)\}$ y $h_B = \{([b, b], 1)\}$ con $a, b \in \Re$ y realizamos operaciones con ellos de acuerdo a las leyes de la aritmética de histogramas, obtenemos los mismos resultados que operando con los intervalos $[a, a]$ y$[b, b]$, o con los números reales a y b.

En su artículo, Colombo y Jaarsma aplican su método sobre histogramas equiprobables, es decir, sobre histogramas donde la probabilidad asociada a todos los intervalos es la misma. Sin embargo, el método funciona igualmente para cualquier tipo de representación basada en intervalos de las mencionadas en el apartado 5.3.

El método de Colombo y Jaarsma asume que los operandos son independientes, es decir que las funciones de distribución de los histogramas considerados son independientes. Sin embargo, puede que exista dependencia entre las funciones de distribución consideradas y que, por ejemplo, valores altos de una de ellas se correspondan con valores bajos de otra. De suceder esto, los resultados de las operaciones aritméticas que se realizarían con estos histogramas serían distintos a los obtenidos si son independientes.

Por ello, en los últimos años, se han propuesto métodos para recoger la dependencia entre los operandos. Los métodos desarrollados por Li y Hyman (2004) y por Berleant y Zhang (2004) son los más sofisticados, aunque también implican una mayor complejidad y requieren de formas de representaciones más elaboradas que el histograma para caracterizar la distribución

de una variable aleatoria. En el ámbito de esta tesis, donde se va a realizar una primera aproximación al alisado de STH, se asumirá que los operandos son independientes, esto quiere decir que, si por ejemplo estamos sumando los valores dos histogramas h_A y h_B, los valores de h_A no ofrecen información de los valores de h_B y viceversa. Puede alegarse que la independencia es una hipótesis fuerte. Sin embargo, estimar la dependencia entre dos histogramas para el caso de una media móvil o de una ecuación de alisado, no es trivial y, dependiendo de cómo se hayan obtenido los histogramas, puede no ser posible por falta de información. Además, el coste computacional de acarrear la dependencia entre los operadores es muy alto, especialmente, en el caso del alisado exponencial donde los promedios tienen en cuenta todos los valores disponibles de la serie.

5.5.1.2. La adaptación del alisado mediante la aritmética de histogramas

Tanto las medias móviles, como las ecuaciones de alisado exponencial en forma recursiva mostradas en el apartado A.2 del apéndice A pueden representarse como una combinación lineal convexa de números reales. Para adaptar dichos métodos para trabajar con STH basta con reemplazar los valores reales por histogramas y la aritmética clásica por la aritmética de histogramas descrita en el apartado anterior.

Adaptación de la media móvil. Dada un serie temporal de histogramas $\{h_{X_t}\}$, la predicción para el instante $t+1$ producida por una media móvil de orden q es una suma ponderada de los últimos q valores de la serie

$$\hat{h}_{X_{t+1}} = \omega_1 h_{X_t} + \omega_2 h_{X_{t-1}} + \ldots + \omega_q h_{X_{t-(q-1)}}, \tag{5.19}$$

de manera que $\sum_{i=1}^{q} \omega_i = 1$ y $\omega_i \geq 0, \forall i$.

En el caso de la media móvil simple de orden n, los pesos asignados a cada valor serán $\omega_i = \frac{1}{q}$. En el caso de la media móvil con pesos aritméticamente decrecientes, los pesos serán $\omega_i = \frac{q-i+1}{\sum_{i=1}^{q} i}$ y en el caso de la media móvil con pesos exponencialmente decrecientes, los pesos serán $\omega_i = \alpha(1-\alpha)^{i-1}$ donde $\alpha = \frac{2}{q+1}$. Todos estos alisados pueden adaptarse sin problemas empleando la aritmética de histogramas.

Adaptación del alisado exponencial. La adaptación mediante la aritmética de histogramas del alisado exponencial empleando la fórmula en forma de corrección del error no es equivalente a la de la forma recurrente. Esto ya sucedía en la adaptación de los alisados empleando aritmética de intervalos (ver apartado 4.6.1). Este hecho es natural ya que la aritmética de histogramas está basada en la aritmética de intervalos y, por tanto, tampoco cumple

la propiedad distributiva, sino la subdistributiva. Debido a este hecho, la mejor forma de generar predicciones para una STH empleando el alisado exponencial es hacerlo mediante la fórmula recurrente

$$\hat{h}_{X_{t+1}} = \alpha h_{X_t} + (1-\alpha)\hat{h}_{X_t}, \tag{5.20}$$

con $\alpha \in [0,1]$. Resulta trivial comprobar que, al igual que en las series temporales clásicas, esta ecuación es equivalente a la media móvil con pesos exponencialmente decrecientes si el número de términos que se consideran en la serie es lo suficientemente grande. Para comprobarlo, basta con sustituir en (5.20) el término $\hat{h}_{X_t}$ por la ecuación de alisado que lo ha originado y repetir el paso sucesivamente. Las propiedades de la aritmética de histogramas garantizan que la igualdad entre ambas expresiones se cumplirá.

Tal y como se indicó en el apartado 5.5.1.1, la aritmética de histogramas subsume a la aritmética clásica. Por tanto, los métodos de alisado basados en aritmética de histogramas, son una generalización de los métodos de alisado sobre números reales para poder tratar con STH. Esto es indudablemente un buen aval para esta adaptación.

A continuación, se va a analizar cómo se comportan los alisados realizados empleando la aritmética de histogramas.

5.5.1.3. Análisis del efecto del alisado basado en aritmética de histogramas

El alisado de una serie temporal clásica permite suavizar la serie y eliminar su comportamiento extremo, obteniendo como resultado una serie que caracteriza mejor el fenómeno. Para entender en qué consiste el alisado de una STH empleando aritmética de histogramas es importante comprender cómo se comportan, utilizando dicha aritmética, el promedio de un conjunto de histogramas y la ecuación recursiva del alisado exponencial.

Promedio de histogramas. Dado un conjunto de histogramas $h_X(i)$ con $i = 1, ..., q$, a los que por simplificar denotaremos como h_i, el histograma promedio de dicho conjunto se calcula como

$$h_{\bar{X}} = \frac{h_1 + ... + h_q}{q}, \tag{5.21}$$

donde las operaciones aritméticas requeridas, i.e. suma y cociente, se realizan conforme al método descrito en (5.17).

El promedio de un conjunto de histogramas presenta las siguientes características:

1. El rango de $h_{\bar{X}}$ es el promedio de los rangos de los q histogramas considerados.

2. La posición de anclaje de $h_{\bar{X}}$, i.e. la posición que toma el extremo izquierdo de su primer intervalo, es el promedio de las posiciones de anclaje de los n histogramas considerados

3. Sea el conjunto de histogramas $h_1 = h_2 = ... = h_q$, su promedio no satisface $h_{\bar{X}} = h_i \ \forall i$, a menos que $h_i = \{([c, c), 1)\}$ con $c \in \mathbb{R}$, $\forall i$. Es decir, que el promedio de un conjunto de histogramas idénticos no es igual a dichos histogramas, a menos que los histogramas sean números reales idénticos.

La característica 3 denota un comportamiento que no se corresponde con lo que cabría esperar de un promedio. Sin embargo, la explicación a este comportamiento se encuentra en el propósito mismo de la aritmética de histogramas. Dicho propósito es el de proporcionar una herramienta que permita realizar operaciones aritméticas con variables aleatorias representadas en forma de histograma. Por ello, al operar con histogramas obtenemos el mismo comportamiento que al operar con variables aleatorias.

Para entenderlo mejor, consideraremos el siguiente ejemplo en el que se quiere obtener el promedio de un conjunto de variables aleatorias idénticas, $X_1, X_2, ..., X_{10}$, donde $X_i = N(2, 4), \forall i$. En ese caso, el promedio no es una variable aleatoria idéntica a las originales, sino que es $\bar{X} = N(2, 0.4)$. Es decir, es una variable aleatoria que sigue una distribución normal con menos desviación típica que las variables originales.

A continuación, se muestran dos ejemplos para ilustrar el comportamiento del promedio de un conjunto de histogramas.

En primer lugar, se considera un conjunto de histogramas idénticos tal que $h_i = \{([1, 2), 0.1), ([2, 3), 0.15), ([3, 4), 0.25), ([4, 5], 0.5)\}$, con $i = 1, .., 5$. Su promedio representado como un histograma de 4 intervalos de igual longitud es $h_{prom1} = \{([1, 2), 0.002), ([2, 3), 0.115), ([3, 4), 0.615), ([4, 5], 0.268)\}$. La parte izquierda de la figura 5.6 muestra dicho promedio, donde puede verse claramente que el promedio es menos asimétrico y más centrado que los histogramas que le han dado lugar. Esto es debido al hecho que se acaba de comentar.

A continuación, se considera el conjunto de histogramas formado por los histogramas $h_6 = \{([19, 20), 0.1), ([20, 21), 0.2), ([21, 22], 0.7)\}$ y $h_7 = \{([0, 3), 0.35), ([3, 6), 0.3), ([6, 9], 0.35)\}$. El promedio de dicho conjunto de histogramas representado mediante un histograma con cinco intervalos equiespaciados es $h_{prom2} = \{([9.5, 10.7), 0.07), ([10.7, 11.9), 0.23), ([11.9, 13.1), 0.26), ([13.1, 14.3), 0.27), ([14.3, 15.5], 0.17)\}$. En la parte derecha de la figura 5.6 se muestra dicho promedio. Su comportamiento se corresponde al comportamiento de un promedio basado en aritmética de histogramas.

Ambos ejemplos muestran que tanto la amplitud del histograma promedio, como su posición son el promedio de las amplitudes y las posiciones, respectivamente, de los histogramas considerados. Por su parte, la forma del

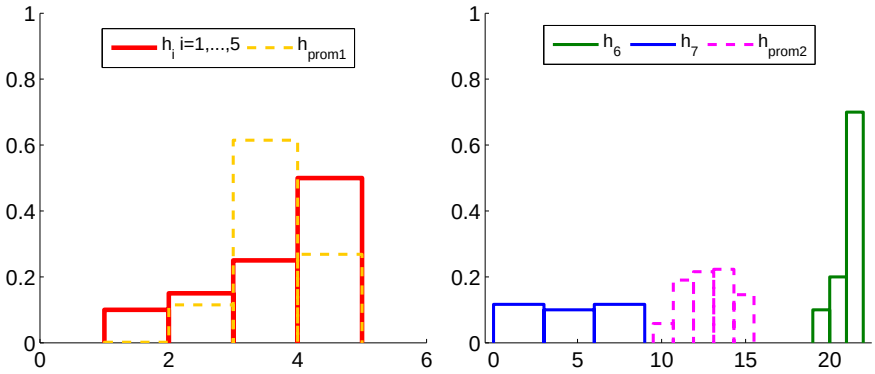


Figura 5.6: Promedio de los histogramas h_i con $i = 1, .., 5$ (izqda.) y de los histogramas h_6 y h_7 (dcha.)

histograma promedio puede verse como un promedio de las formas de los histogramas considerados, pero donde el resultado es ligeramente más simétrico de lo que cabría esperar.

La ecuación recursiva de alisado para STH. A continuación, se analizará el efecto de realizar una suma ponderada de dos términos tal y como se realiza en la ecuación recursiva del alisado exponencial (5.20).

Para ello, se considerará $h_{X_t} = h_6$ y $\hat{h}_{X_t} = h_7$. Si en la ecuación se toma $\alpha = 0.9$, se da un mayor peso al valor actual y un menor peso a la predicción. En ese caso, el resultado representado como un histograma con cinco intervalos de igual longitud es $h_{exp1} = \{([17.1, 17.82), 0.03), ([17.82, 18.54), 0.10), ([18.54, 19.26), 0.21), ([19.26, 19.98), 0.41), ([19.98, 20.7], 0.25)\}$. Por el contrario, si $\alpha = 0.1$, se da más peso a la predicción que al valor actual y el histograma resultante viene dado por $h_{exp2} = \{([1.9, 3.58), 0.19), ([3.58, 5.26), 0.21), ([5.26, 6.94), 0.18), ([6.94, 8.62), 0.21), ([8.62, 10.3], 0.21)\}$. Los resultados se muestran en la figura 5.7, donde puede verse que la predicción resultante con valores del parámetro α se comporta según lo esperado. Si $\alpha = 0.5$, se obtiene el mismo resultado que realizando el promedio, es decir, el resultado mostrado en la parte derecha de la figura 5.6.

El efecto de alisado en una STH. Por último, se va a mostrar el resultado de realizar el suavizado de una STH con la ecuación recursiva del alisado exponencial (5.20). En la figura 5.8 se muestra la STH obtenida tras aplicar el alisado exponencial basado en la aritmética de histogramas con $\alpha = 0.4$. De acuerdo con el propósito del alisado exponencial, la serie resultante elimina las fluctuaciones de la serie y resalta su comportamiento a largo plazo, mostrando también histogramas con menor variabilidad.

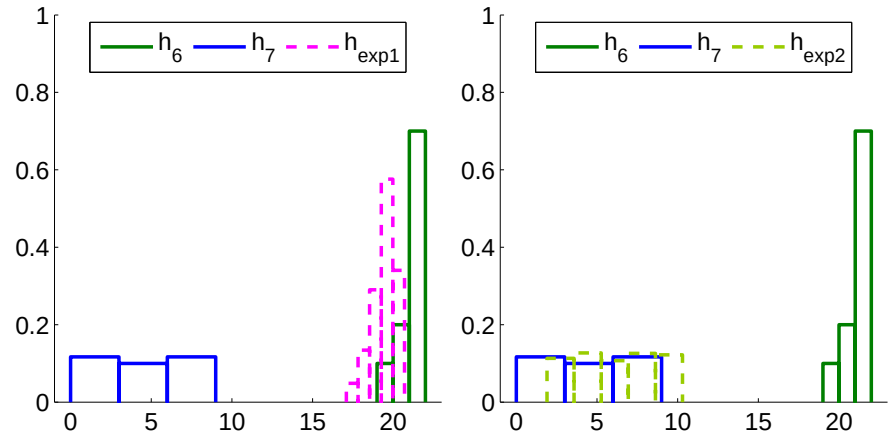


Figura 5.7: Resultado de la ecuación recursiva de alisado considerando que $h_{X_t} = h_6$ y que $\hat{h}_{X_t} = h_7$ y que $\alpha = 0.9$ (izqda.) y que $\alpha = 0.1$ (dcha.)

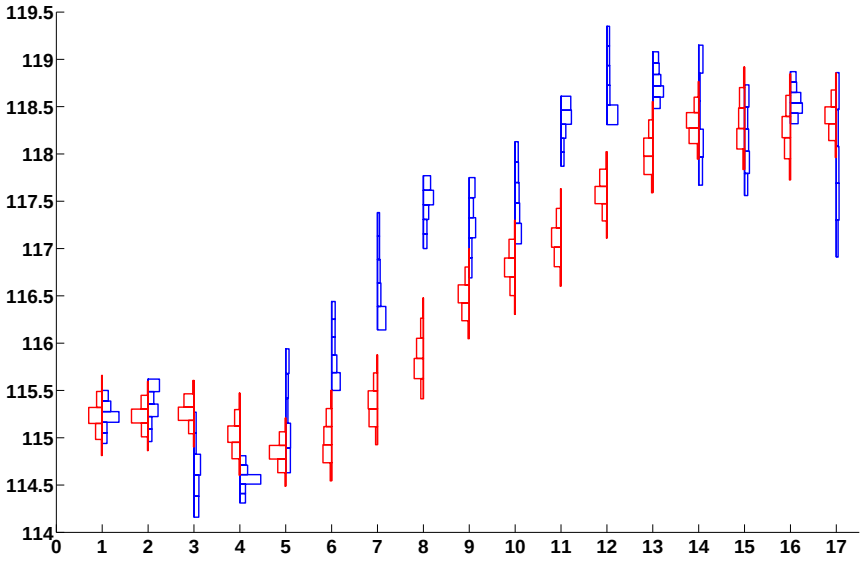


Figura 5.8: STH real (azul) y suavizada (rojo) empleando el alisado exponencial basado en aritmética de histogramas con $\alpha = 0.4$

5.5.2. El alisado de STH empleando baricentros

El alisado de una serie temporal se puede realizar, o bien mediante una media móvil, como el promedio ponderado de valores pasados y consecutivos de la serie, o bien mediante el alisado exponencial en forma recursiva, como el promedio ponderado del valor real de la serie y su predicción. En ambos casos, la operación que se realiza es un promedio.

Para adaptar los métodos de alisado para trabajar sobre STH, además del uso de la aritmética de histogramas mostrado en el apartado anterior, se pueden reemplazar los promedios por otra técnica que obtenga el mismo efecto. Una posibilidad es emplear baricentros como sustitutos de los promedios. Para ello se va a definir el concepto de histograma baricéntrico.

5.5.2.1. El histograma baricéntrico

En el apartado 4.8.2, de esta tesis se estudió la relación existente entre los conceptos de baricentro y promedio. Recapitulando brevemente lo mencionado en ese apartado, puede decirse que el baricentro de un sistema de partículas se obtiene como el promedio ponderado de las coordenadas de los partículas. De forma equivalente, dicho baricentro también puede obtenerse como el punto que minimiza la suma de las distancias euclídeas ponderadas entre sí mismo y el resto de partículas.

Tomando como punto de partida esta idea, puede desarrollarse el concepto de histograma baricéntrico de un conjunto de histogramas. El histograma baricéntrico será aquel que minimiza la distancia entre sí mismo y el resto de histogramas de un conjunto, utilizando para ello una distancia para histogramas. Dado un conjunto de histogramas h_{X_i} con $i = 1, ..., n$, el histograma baricéntrico h_{X_B} de dicho conjunto de histogramas se hallaría como

$$min_{h_{X_B}} \sum_{i=1}^{n} D(h_{X_i}, h_{X_B}), \tag{5.22}$$

donde $D(h_{X_i}, h_{X_B})$ es una distancia que se consideré adecuada. Tal y como muestran Verde y Irpino (2007), la elección de la distancia condiciona en gran medida las propiedades del baricentro. Por ello, más adelante se abordará qué distancia resulta adecuada para el caso del alisado.

De acuerdo a su definición, el histograma baricéntrico debe ser un representante apropiado del conjunto de histogramas que lo ha originado, de forma que elimine su comportamiento extremo y realce la tendencia central del conjunto. Esta es una propiedad que le hace adecuado para ser empleado como herramienta para realizar alisados en STH.

5.5.2.2. El histograma baricéntrico como herramienta de alisado

Debido a la conexión mostrada en el apartado 4.8.2 entre el concepto de baricentro y promedio, el histograma baricéntrico se utilizará para reemplazar el promedio en las ecuaciones de alisado y realizar de esa forma el suavizado de una STH. A continuación, se va a analizar el uso de los baricentros para adaptar tanto la media móvil, como la ecuación recursiva del alisado exponencial.

Adaptación de la media móvil. Sea una STH observada $\{h_{X_t}\}$ con $t = 1, ..., n$, la predicción $\hat{h}_{X_{t+1}}$ de dicha serie obtenida mediante una media móvil es el resultado de esta expresión

$$\hat{h}_{X_{t+1}} = \omega_1 h_{X_t} + \omega_2 h_{X_{t-1}} + \ldots + \omega_q h_{X_{t-(q-1)}}. \tag{5.23}$$

Esta expresión puede representarse como el cálculo del baricentro $\hat{h}_{X_{t+1}}$

$$\arg\min_{\hat{h}_{X_{t+1}}} \hat{h}_{X_{t+1}} = \sum_{i=1}^{q} \omega_i D(\hat{h}_{X_{t+1}}, h_{X_{t-(i-1)}}), \tag{5.24}$$

donde q es el número de periodos promediados, $D(\cdot, \cdot)$ es una distancia para distribuciones y ω_i es el peso asignado a $h_{X_{t-(i-1)}}$.

Adaptación del alisado exponencial. La adaptación de la fórmula del alisado exponencial expresado en forma recursiva (A.9), también puede realizarse empleando el concepto baricentro. Esto es debido a que en ese caso la predicción se calcula como una media ponderada del valor pasado y de la predicción pasada

$$\hat{h}_{X_{t+1}} = \alpha h_{X_t} + (1 - \alpha)\hat{h}_{X_t}. \tag{5.25}$$

Dicha media ponderada puede representarse como un baricentro, de forma que la predicción $\hat{h}_{X_{t+1}}$ sea el baricentro solución al siguiente problema de minimización

$$\arg\min_{\hat{h}_{X_{t+1}}} \left(\alpha D(\hat{h}_{X_{t+1}}, h_{X_t}) + (1 - \alpha) D(\hat{h}_{X_{t+1}}, \hat{h}_{X_t}) \right), \tag{5.26}$$

donde $\alpha \in [0, 1]$.

Sin embargo, no es posible adaptar la fórmula del alisado exponencial en forma de corrección del error (A.10) por medio de los baricentros ya que en ella desaparece la media ponderada. Los baricentros nos brindan una manera de realizar promedios (ponderados o no) de histogramas, pero no permiten realizar otras operaciones aritméticas básicas como la resta.

En el siguiente apartado se determinará qué distancias son adecuadas para adaptar el alisado exponencial para trabajar sobre STH.

5.5.2.3. Análisis de las posibles distancias a utilizar para realizar alisados

Las ecuaciones de alisado mediante baricentros (5.24) y (5.26) requieren de una distancia. En el apartado 5.4.2.1 de esta tesis, se analizaron una serie de medidas de divergencia con el fin de determinar si eran adecuadas o no para representar el concepto de error en una STH. El resultado fue que sólo las distancias de Mallows y Wasserstein servían para dicho propósito.

En este caso, el objetivo es determinar qué distancias sirven para estimar el histograma baricéntrico, teniendo en cuenta que dicho histograma debe ser un buen representante del conjunto de histogramas que combine las características del conjunto.

Verde y Irpino (2007) analizan los baricentros que se obtienen empleando distintas medidas de divergencia. El objetivo de su análisis es determinar con qué medidas se obtienen baricentros que se comporten de forma adecuada para representar las particiones producidas por un algoritmo de clustering dinámico sobre datos de histograma. Los autores concluyen que los baricentros que se obtienen empleando la distancia de Mallows presentan propiedades adecuadas. En su análisis, descartan distancias como las de Variación Total, Hellinger, Kolmogorov y Prokhorov-Lévy. Los histogramas baricéntricos que se obtienen con las métricas de Kolmogorov y de Prokhorov-Lévy no son únicos, lo cual no es adecuado para un representante de un *cluster* de histogramas. Mientras que en el caso de las distancias de Hellinger y de Variación Total, los histogramas baricéntricos se comportan como mixturas. Esto quiere decir que, dado un conjunto de histogramas unimodales, el histograma baricéntrico que se obtiene con estas distancias es multimodal, lo cual tampoco es un comportamiento adecuado para un representante.

Si el objetivo es emplear el histograma baricéntrico como el histograma pronosticado por un método de alisado aplicado sobre una STH, tampoco resulta deseable que dicho baricentro no sea único, ni que su comportamiento sea el de una mixtura. La no-unicidad del baricentro obligaría a optar por un determinado baricentro entre un conjunto potencialmente infinito de ellos. Mientras que el comportamiento en forma de mixtura tenderá a generar como baricentro un histograma cuyo rango sea tan amplio como la unión de los histogramas considerados y que posiblemente sea multimodal con tantas modas como tenga el conjunto de histogramas original. Ninguna de estas características es buena para una predicción en forma de histograma. Por ello, es mejor considerar distancias cuyos baricentros no presenten estas características, como la distancia de Mallows.

Para ilustrar la idea se va a calcular el histograma baricéntrico empleando las distancias de Variación Total y Mallows para los los histogramas $h_1 = \{([1,2),.3),([2,3),.5),([3,4),.2)\}$ y $h_2 = \{([3,4),.3),([4,5),.5),([5,6),.2)\}$. Tal y como indican Verde y Irpino (2007), el histograma baricéntrico basado en la distancia de Variación Total tiene como soporte la unión de los sopor-

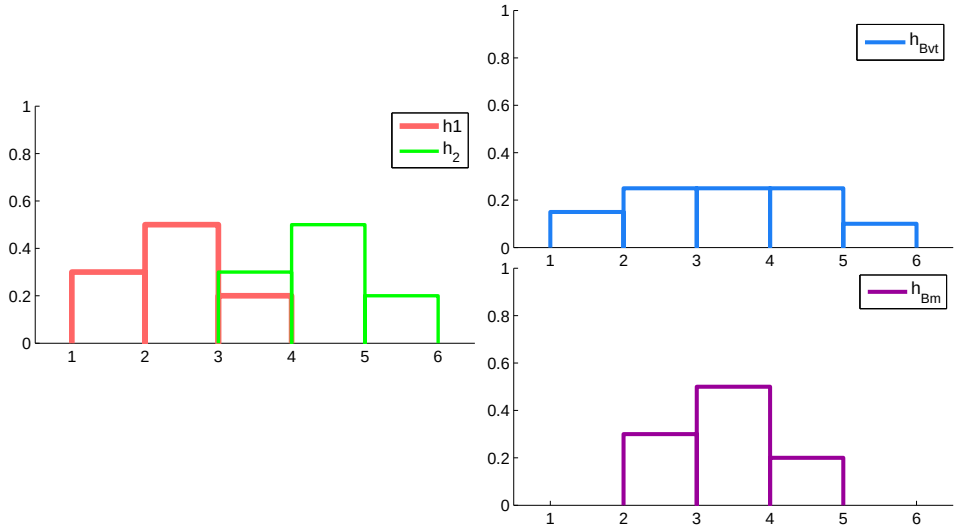


Figura 5.9: Histogramas h_1 y h_2 (izqda.). Histograma baricéntrico de h_1 y h_2 obtenido utilizando la distancia de Variación Total (dcha. arriba) y la distancia de Mallows (dcha. abajo)

tes de los histogramas considerados y el peso asociado a cada intervalo es la mediana del peso a dicho intervalo en los histogramas considerados. En este caso es $h_{Bvt} = \{([1,2),.15),([2,3),.25),([3,4),.25),([4,5),.25),([5,6),.1)\}$. El cálculo del histograma baricéntrico utilizando la distancia de Mallows es más complejo (más detalles en el apartado B.1 de los apéndices). En este caso el histograma que se obtiene es $h_{Bm} = \{([2,3),.3),([3,4),.4),([4,5),.2)\}$. La figura 5.9 muestra los histogramas considerados y los histogramas baricéntricos que se obtienen con una y otra distancia. En ella se puede ver que el histograma baricéntrico con la distancia de Variación Total tiene es una mixtura de los histogramas considerados. Por su parte, el histograma baricéntrico con la distancia de Mallows es, en este caso que los dos histogramas originales son idénticos, un histograma igual que los histogramas originales.

En su trabajo, Verde y Irpino (2007) no consideran la distancia de Wasserstein que sí fue considerada en el apartado 5.4.2.1 de esta tesis para medir errores. En esta tesis se ha analizado si resulta adecuado su uso o no para realizar alisados y se ha descartado.

En el apéndice B, se explica cómo calcular los histogramas baricéntricos que se obtienen con la distancia de Wasserstein y con la distancia de Mallows y se estudian sus propiedades[2]. En el punto B.2 del apéndice se muestra cómo el baricentro de Mallows puede considerarse como la media de los histogramas

[2]Nota del autor: Los baricentros de Wasserstein y de Mallows también se emplean en el punto 5.6 para adaptar el método de k-NN para predecir STH. Por ello, el cálculo de dichos baricentros y al análisis de su comportamiento figura en un apéndice.

considerados, mientras que el baricentro de Wasserstein puede considerarse como la mediana de los mismos.

En el punto B.3 del apéndice, se analiza la idoneidad de los baricentros de Wasserstein y de Mallows para ser empleados como sustitutos de los promedios en los métodos de alisado[3]. La conclusión que se obtiene en dicho punto es que ambos baricentros permiten adaptar las medias móviles, pero sólo el baricentro de Mallows permiten adaptar la fórmula del alisado exponencial (A.9) de forma que sea equivalente con la media móvil de pesos aritméticamente decrecientes (A.8), como sucede en el caso de las series temporales clásicas.

En principio, sería posible realizar alisados exponenciales con el baricentro de Wasserstein utilizando la fórmula de la media móvil, pero si se hiciese de esa forma, el resultado de dicho alisado exponencial para valores de $\alpha > 0.5$ sería igual al del método ingenuo, lo cual no parece adecuado para un método de alisado. Más detalles en el punto B.3 de los apéndices.

5.5.2.4. Traslación de un histograma

Ya hemos visto que el promedio que se realiza en una media móvil y en la ecuación recursiva del alisado exponencial puede ser reemplazado por el cálculo del histograma baricéntrico. Sin embargo, la adaptación de algunos métodos de alisado exponencial más sofisticados requieren, además, otras operaciones. Por ello, se va a definir cómo se realiza el desplazamiento o traslación de un histograma a lo largo de la recta real.

Esta operación se requiere, como se verá más adelante, en el alisado exponencial con tendencia (ver el apartado 5.5.3.2) y en el alisado exponencial estacional donde la estacionalidad se modela como un número real (ver el apartado 5.5.3.3). En dichos alisados, hay una componente de la serie que se modela como un número real y que deberá ser sumada o restada a un histograma. Para realizar estas operaciones de suma y resta entre un histograma y un número real se ha optado por considerarlas como la traslación del histograma sobre la recta real en la que se le esté representando. De esta forma la suma entre el histograma $h_X = \{([I]_i, \pi_i)\}$ con $i = 1, ..., p$ y el entero d, $h_X + d$, es un histograma h'_X que se calcula de la siguiente forma

$$h'_X = \{([\underline{I}_i + d, \overline{I}_i + d], \pi_i)\} \text{ con } i = 1, ..., p. \tag{5.27}$$

De manera análoga, la resta entre el histograma h_X y el entero d, $h_X - d$, es un histograma h'_X que se calcula de la siguiente forma

$$h'_X = \{([\underline{I}_i - d, \overline{I}_i - d], \pi_i)\} \text{ con } i = 1, ..., p. \tag{5.28}$$

[3]Nota del autor: El punto B.3 de los apéndices es específico de los alisados exponenciales y debería aparecer en este capítulo. Sin embargo, está tan ligado al contenido del apéndice B que se ha considerado más adecuado incluirlo en él.

Estas operaciones de traslación obtienen el mismo resultado que la suma de un histograma y un número real y a la resta entre un histograma y un número real que se dan en la aritmética de histogramas (ver el apartado 5.5.1.1) al representar un número real a como un histograma $h = \{([a, a], 1)\}$.

5.5.2.5. Análisis del efecto del alisado empleando baricentros

A continuación, se va a mostrar qué efecto se consigue al alisar una STH empleando el baricentro de Mallows como herramienta para realizar el suavizado. El cálculo del baricentro de Mallows se explica en el apartado B.1 de los apéndices.

Los ejemplos que se van a mostrar en este apartado son los mismos que se mostraron en el apartado 5.5.1.3 donde se utilizaba la aritmética de histogramas. De esta forma, se puede comparar el comportamiento que se obtiene con una y otra técnica.

Promedio de histogramas. Dado un conjunto de histogramas $h_X(i)$ con $i = 1, ..., q$, a los que, por simplificar, denotaremos como h_i, el histograma promedio de dicho conjunto se calcula como

$$\arg\min_{h_{\bar{X}}} \sum_{i=1}^{q} \frac{1}{q} D(h_{\bar{X}}, h_i). \tag{5.29}$$

Para conocer cómo se realiza el cálculo del baricentro se recomienda acudir al apartado B.1 de los apéndices.

El promedio de un conjunto de histogramas utilizando baricentros presenta dos características que también tenía el promedio que se obtenía con la aritmética de histogramas (ver el apartado 5.5.1.3). Dichas características son:

1. El rango de $h_{\bar{X}}$ es el promedio de los rangos de los q histogramas considerados.

2. La posición de anclaje de $h_{\bar{X}}$, i.e. la posición que toma el extremo izquierdo de su primer intervalo, es el promedio de las posiciones de anclaje de los n histogramas considerados

Estas condiciones son las que cabe esperar en un promedio.

Respecto a la forma de la distribución del histograma promedio, existen notables diferencias entre el resultado que se obtiene utilizando el baricentro de Mallows y el que se obtiene con la aritmética de histogramas. El ejemplo más claro puede verse con un conjunto de histogramas idénticos $h_1 = h_2 = ... = h_q$, ya que el promedio utilizando el baricentro de Mallows sí satisface $h_{\bar{X}} = h_i$, $\forall i$. Si, por ejemplo, consideramos un conjunto de histogramas idénticos como $h_i = \{([1, 2), 0.1), ([2, 3), 0.15), ([3, 4), 0.25), ([4, 5], 0.5)\}$, con

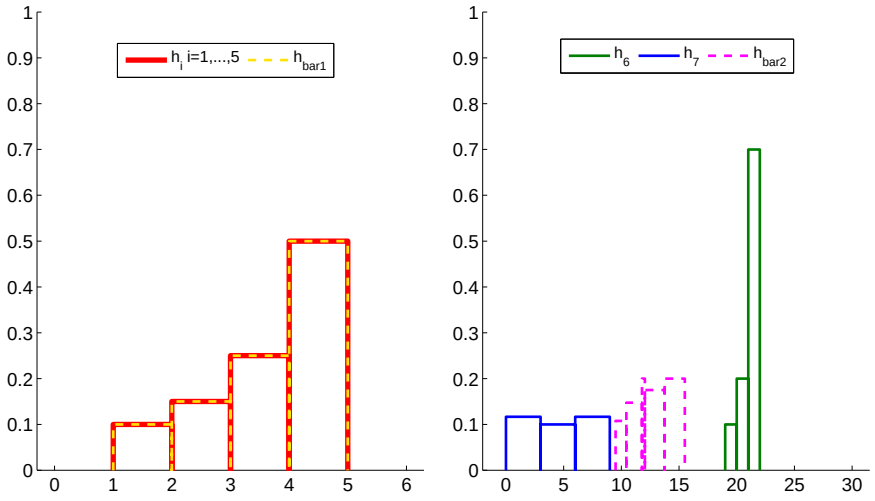


Figura 5.10: Baricentro de los histogramas h_i con $i = 1, .., 5$ (izqda.) y de los histogramas h_6 y h_7 (dcha.)

$i = 1, .., 5$, el promedio resultante será $h_{prom1} = h_i, \forall i$, tal y como se muestra en la parte izquierda de la figura 5.10. Esto es así porque el histograma que minimiza la distancia entre sí mismo y un conjunto de histogramas idénticos es un histograma idéntico a los del conjunto. El resultado que se obtenía con la aritmética de histogramas no cumplía esta propiedad (ver figura 5.6).

Por otra parte, el histograma que se obtiene al realizar el promedio de la pareja de histogramas $h_6 = \{([19, 20), 0.1), ([20, 21), 0.2), ([21, 22], 0.7)\}$ y $h_7 = \{([0, 3), 0.35), ([3, 6), 0.3), ([6, 9], 0.35)\}$ se muestra en la parte derecha de la figura 5.10. La forma del histograma promedio resultante es un combinación de la de los histogramas h_6 y h_7. Puede apreciarse, por ejemplo, cierta asimetría izquierda por la influencia de h_7. Dicha asimetría no quedaba reflejada en el promedio que se obtenía utilizando la aritmética de histogramas (ver figura 5.6).

La ecuación recursiva de alisado para STH. Para estudiar el comportamiento de la ecuación del alisado exponencial empleando el baricentro de Mallows, ecuación (5.26), se considerará que $h_{X_t} = h_6$ y que $\hat{h}_{X_t} = h_7$. Si se toma $\alpha = 0.9$, se da un mayor peso al valor actual y un menor peso a la predicción. Por el contrario, si $\alpha = 0.1$, se da más peso a la predicción que al valor actual. En la figura 5.11 se muestra el histograma resultante en ambos casos. En ella se puede ver como el histograma alisado es más similar a h_6 o a h_7, según el valor de α. La similitud afecta a la posición del histograma, a su amplitud y a su forma. Si $\alpha = 0.5$, se obtiene el mismo resultado que el mostrado en la parte derecha de la figura 5.10.

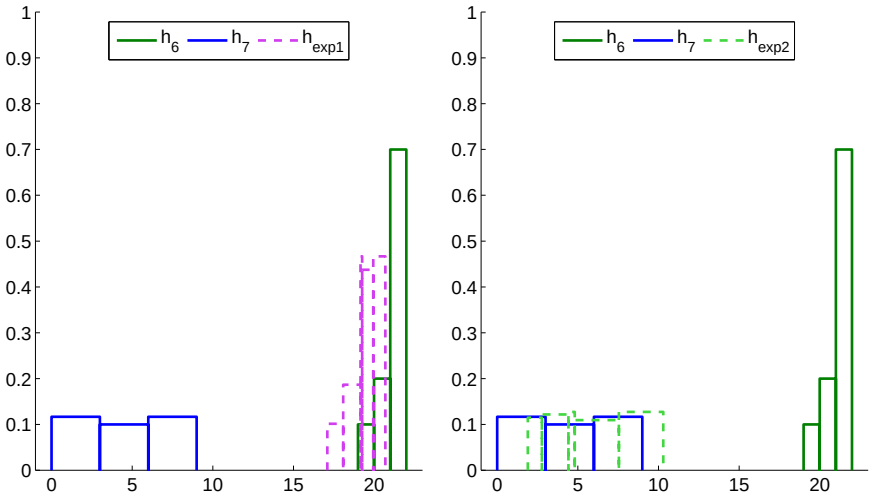


Figura 5.11: Resultado de la ecuación recursiva de alisado considerando que $h_{X_t} = h_6$ y que $\hat{h}_{X_t} = h_7$ y que $\alpha = 0.9$ (izqda.) y que $\alpha = 0.1$ (dcha.)

La diferencia existente con los resultados que obtenía el alisado que utiliza la aritmética de histogramas y que se mostraban en la figura 5.7, es que, en aquel caso, la forma del histograma alisado tendía a ser más centrada, no reflejando adecuadamente la forma de los histogramas considerados. Esto no sucede al utilizar el baricentro de Mallows.

El efecto de alisado en una STH. En la figura 5.12 se muestra una STH suavizada con la ecuación recursiva del alisado exponencial (5.26) empleando el baricentro de Mallows. Puede verse que la STH obtenida tras aplicar el alisado exponencial con $\alpha = 0.4$ presenta un comportamiento con menos variabilidad que la serie original.

Las diferencias con la figura 5.8 que mostraba el mismo alisado sobre la misma STH pero empleando aritmética de histogramas, simplemente atañen a la forma de los histogramas (es decir, a la distribución de sus pesos), no a su posición, ni a su rango. En el alisado basado en la aritmética de histogramas los histogramas suavizados son menos asimétricos y tienen un mayor peso en su parte central. Mientras que en el alisado empleando baricentros, la forma del histograma alisado guarda una mayor semejanza con la del histograma inmediatamente anterior.

5.5.3. Métodos de alisado exponencial

A continuación, se van a formular los métodos de alisado para una STH tal que $\{h_{X_t}\}$ con $t = 1, ..., n$. Se propondrá un alisado simple para los casos en los que no haya ni tendencia, ni estacionalidad. Para manejar la tendencia

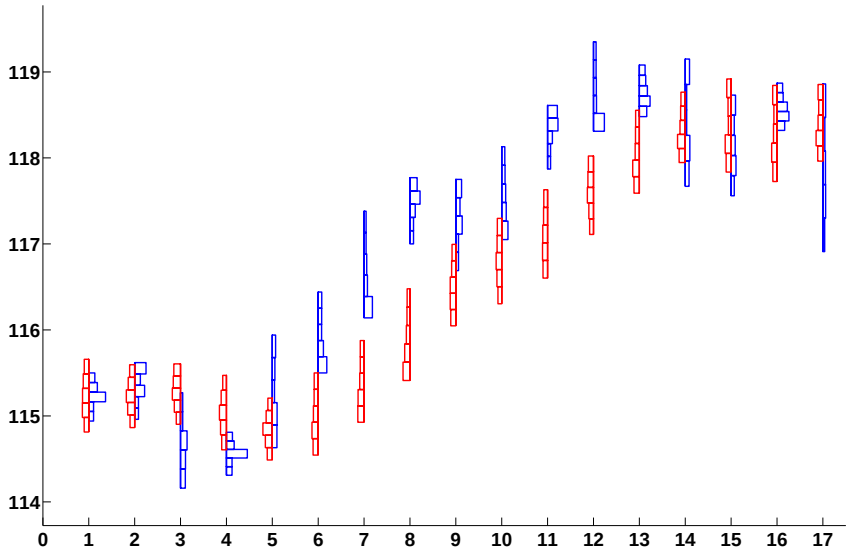


Figura 5.12: STH real (azul) y suavizada (rojo) utilizando el alisado exponencial empleando baricentros con $\alpha = 0.4$.)

se propondrán dos alisados, uno con tendencia y otro que permite atenuar la tendencia realizando estimaciones más conservadoras de la misma. Por su parte, para recoger la estacionalidad se proponen dos alisados, uno que refleja la estacionalidad en forma de histograma y otro que lo hace considerando que la estacionalidad sólo afecta a la posición de los histogramas. En todos los casos, cada una de las componentes de la serie se agrega de forma aditiva para obtener la predicción.

Los métodos que se van a mostrar pueden emplearse tanto con la aritmética de histogramas (Secc. 5.5.1) como con el baricentro de Mallows (Secc. 5.5.2). Dichos métodos siguen la notación descrita en Gardner (2006). En el apéndice A.2 se muestra que los alisados exponenciales pueden expresarse tanto en forma de corrección del error, como en forma recursiva. Sin embargo, tal y como se ha mencionado con anterioridad, ni en la adaptación basada en aritmética de histogramas, ni en la basada en los baricentros, resulta adecuado utilizar la forma de corrección del error. Por ello, todos los métodos que se van a presentar estarán expresados en forma recursiva.

5.5.3.1. Alisado exponencial simple

El alisado exponencial simple (AES) se define como

$$h_{X_{t+1}} = \alpha h_{X_t} + (1-\alpha)\hat{h}_{X_t}, \tag{5.30}$$

donde $\alpha \in [0,1]$. Para inicializar el método se puede optar por tomar $\hat{h}_{X_1} = h_{X_1}$. Sin embargo, existen procedimiento más sofisticados como, por ejemplo, tomar como valor inicial el promedio de los tres o cuatro primeros valores. Otra opción es realizar *backcasting* que consiste en invertir la serie original, de forma que $\{h_{X_{t'}}\}$ con $t' = n, ..., 1$, y predecirla hasta obtener el valor pronosticado para $t' = 1$ que se empleará como valor inicial.

5.5.3.2. Alisado exponencial con tendencia

En el contexto de las series temporales clásicas, Holt (1957) propuso un método de alisado exponencial para predecir series que muestran un comportamiento creciente o decreciente a largo plazo, es decir, series con tendencia. Este método consta de dos ecuaciones de alisado una para el nivel de la serie y otra para la tendencia, dichas ecuaciones pueden verse en la tabla que se muestra en la figura A.1 del apéndice A.

Para adaptar este método al contexto de las STH es preciso definir cómo se va a representar cada uno de los dos componentes de la serie. Resulta razonable que el nivel de la serie, es decir, su componente básica, se represente mediante un histograma. Por otro lado, la tendencia debe modelar los cambios en la posición de los histogramas a largo plazo y la posición tiene sentido que se represente como mediante un número real. Este número real podría ser el valor mínimo o máximo del histograma. Sin embargo, estos valores presentan el inconveniente de que son sensibles a la presencia de valores extremos por lo que pueden ofrecer una información errónea sobre la posición de la mayor parte de la distribución. Para evitar este problema se ha optado por representar la posición del histograma mediante su centro de gravedad.

Dado un histograma $h = \{([I_i], \pi_i)\}$ con $i = 1, ..., p$, y asumiendo que dentro de cada intervalo $[I_i]$ las observaciones se distribuyen uniformemente, el centro de gravedad de h viene dado por

$$c(h) = \sum_{i=1}^{p} \frac{(\underline{I}_i + \overline{I}_i)}{2} \pi_i. \tag{5.31}$$

El centro de gravedad es una medida de tendencia central que representa la media del histograma, asumiendo una distribución uniforme dentro de cada intervalo. Al ser una medida de tendencia central es más robusto a la hora de indicar la posición del histograma que valores como el mínimo y el máximo.

En el método de alisado exponencial con tendencia para STH que se va a proponer, el nivel de la serie en t será representado mediante un histograma al que se denotará como h_{S_t}, mientras que la tendencia de la serie en t se representará mediante un valor real T_t. El valor de T_t modelará la tendencia que existe en las posiciones de los histogramas de la serie, siendo el centro de gravedad del histograma el indicador de su posición. Cada una de las dos componentes, nivel y tendencia, se alisarán de forma independiente.

El alisado exponencial con tendencia (AET) viene dado por

$$h_{S_t} = \alpha h_{X_t} + (1-\alpha)(h_{S_{t-1}} + T_{t-1}), \quad (5.32)$$

$$T_t = \gamma(c(h_{S_t}) - c(h_{S_{t-1}})) + (1-\gamma)T_{t-1}, \quad (5.33)$$

$$\hat{h}_{X_{t+m}} = h_{S_t} + mT_t, \quad (5.34)$$

donde $\alpha, \gamma \in [0,1]$, $c(h_A)$ es el centro de gravedad del histograma h_A tal y como se define en (5.31), y m es un factor que multiplica el valor de la tendencia para obtener la predicción de la serie para el periodo futuro $t+m$.

Los valores de inicio del método, i.e. los valores para $t = 1$, pueden ser obtenidos mediante *backcasting*, es decir invirtiendo la serie y prediciéndola hasta obtener los valores iniciales del nivel y de la tendencia. La inicialización del proceso de *backcasting* puede realizarse utilizando los siguientes valores $T_n = c(h_{X_{n-1}}) - c(h_{X_n})$ y $h_{S_n} = h_{X_n}$.

Alisado exponencial con tendencia atenuada. Gardner y McKenzie (1985) afirman que el método de Holt con tendencia lineal tiende a sobreestimar el valor real de la serie debido a la extrapolación que hace de la tendencia. Para paliar este hecho proponen el uso de un parámetro ϕ que permita reducir el efecto de la tendencia y obtener así una predicción más conservadora.

Para adaptar este modelo al contexto de las STH, se tomará como punto de partida el alisado exponencial con tendencia presentado en el apartado anterior, donde h_{S_t} es el histograma que representa el nivel de la serie en t y T_t es el número real que representa la tendencia de la serie en t.

El alisado exponencial con tendencia atenuada aditiva (AETA) se define como

$$h_{S_t} = \alpha h_{X_t} + (1-\alpha)(h_{S_{t-1}} + \phi T_{t-1}), \quad (5.35)$$

$$T_t = \gamma(c(h_{S_t}) - c(h_{S_{t-1}})) + (1-\gamma)\phi T_{t-1}, \quad (5.36)$$

$$\hat{h}_{X_{t+m}} = h_{S_t} + \sum_{i=1}^{m} \phi^i T_t, \quad (5.37)$$

donde $\alpha, \gamma \in [0,1]$, $\phi \geq 0$, $c(h)$ es el centro de gravedad del histograma h, y m es el horizonte de predicción. Si $\phi = 0$, el método es el alisado exponencial simple. Si $\phi \in (0,1)$, el valor de la tendencia se atenúa. Si $\phi = 1$ el método es igual al alisado exponencial con tendencia, y si $\phi > 1$, la tendencia de la serie pronosticada será exponencial.

5.5.3.3. Alisado exponencial con estacionalidad

A la hora de modelar la estacionalidad se va a considerar que ésta puede aparecer como una variación que, o bien afecta al histograma en su conjunto,

o bien afecta sólo a su posición. Cada una de estas aproximaciones requiere proponer un método de alisado distinto. El tipo de estacionalidad de la STH puede ser determinada mediante la inspección visual de la representación gráfica de la serie.

Estacionalidad que sólo afecta a la posición del histograma. Si la estacionalidad sólo afecta a la localización del histograma, la componente estacional, I_t, puede ser representada por una serie temporal clásica que recoja los cambios en la posición del histograma debidos al efecto estacional. Al igual que en el caso de la tendencia, la localización de un histograma vendrá dada por su centro de gravedad (5.31). Por su parte, el nivel en t será representado como el histograma desestacionalizado h_{S_t}, es decir, será un histograma al que se le ha eliminado el efecto estacional. Las predicciones se compondrán mediante la suma del histograma del nivel y de la componente estacional, que será un número real.

El alisado exponencial con estacionalidad clásica aditiva (AEEc) viene dado por

$$\begin{aligned} h_{S_t} &= \alpha(h_{X_t} - I_{t-p}) + (1-\alpha)h_{S_{t-1}}, && (5.38)\\ I_t &= \delta(c(h_{X_t}) - c(h_{S_t})) + (1-\delta)I_{t-p}, && (5.39)\\ \hat{h}_{X_{t+1}} &= h_{S_t} + I_{t-p+1}, && (5.40) \end{aligned}$$

donde $\alpha, \delta \in [0,1]$, $c(h_A)$ es el centro de gravedad de un histograma h_A tal y como se define en (5.31), y p es la longitud de la estacionalidad. Al tratarse de un método con estacionalidad, los p primeros valores de la serie son necesarios para inicializar el método.

Los valores iniciales para el nivel y para las componentes estacionales de los p primeros periodos pueden ser obtenidos mediante un proceso de *backcasting* cuyos valores iniciales serían

$$\begin{aligned} h_{S_{n-(p-1)}} &= \frac{h_{X_n} + h_{X_{n-1}} + \ldots + h_{X_{n-(p-1)}}}{p},\\ I_n &= c(h_{X_n}) - c(h_{S_{n-(p-1)}}),\\ I_{n-1} &= c(h_{X_{n-1}}) - c(h_{S_{n-(p-1)}}), \ldots,\\ I_{n-(p-1)} &= c(h_{X_{n-(p-1)}}) - c(h_{S_{n-(p-1)}}). \end{aligned}$$

Estacionalidad que afecta al histograma en su conjunto. En el segundo método, la componente estacional será representada mediante un histograma, h_{I_t}, y el nivel será representado mediante un valor clásico, S_t. Este enfoque permite predecir adecuadamente las situaciones en las que la estacionalidad no sólo afecte a la localización del histograma, sino también a su rango o a su forma.

El alisado exponencial con estacionalidad aditiva en forma de histograma (AEEh) viene dado por

$$S_t = \alpha(c(h_{X_t}) - c(h_{S_{t-p}})) + (1-\alpha)S_{t-1}, \quad (5.41)$$

$$h_{I_t} = \delta(h_{X_t} - S_t) + (1-\delta)h_{I_{t-p}}, \quad (5.42)$$

$$\hat{h}_{X_{t+1}} = S_t + h_{I_{t-p+1}}, \quad (5.43)$$

donde $\alpha, \delta \in [0,1]$, y p es la longitud del periodo estacional. Por ser un método con estacionalidad, los p primeros valores de la serie son necesarios para inicializar el método.

El valor inicial del nivel y de los p primeros valores de la componente estacional pueden determinarse mediante un proceso de *backcasting* que puede ser inicializado según estos valores

$$\begin{aligned} S_{n-(p-1)} &= \frac{c(h_{X_n}) + c(h_{X_{n-1}}) + ... + c(h_{X_{n-(p-1)}})}{p}, \text{ y} \\ h_{I_n} &= h_{X_n} - S_{n-(p-1)}, \\ h_{I_{n-1}} &= h_{X_{n-1}} - S_{n-(p-1)}, \ldots, \\ h_{I_{n-(p-1)}} &= h_{X_{n-(p-1)}} - S_{n-(p-1)}. \end{aligned}$$

5.6. Predicción mediante el Método de los k Vecinos Más Cercanos

En este apartado se adaptará el método de los k vecinos más próximos para emplearlo en la predicción de STH. Por comodidad, se hará referencia a él mediante su abreviatura inglesa k-NN (*k-Nearest Neighbours*). La sencillez del método de k-NN y los buenos resultados que obtiene cuando es aplicado a la predicción de series temporales le convierten en un método atractivo para ser adaptado al contexto de las STH. En el punto A.3 del apéndice se puede encontrar una sencilla introducción al k-NN como técnica de predicción de series temporales clásicas.

En realidad, la adaptación del método de k-NN que se va a proponer no sólo permite predecir STH, sino que, de forma más general, permite utilizar el k-NN como herramienta de aprendizaje estadístico cuando los datos de entrada y de salida son representados como histogramas. La extensión de esta propuesta para el caso en el que la variable de salida es categórica resulta directa.

5.6.1. El papel de las distancias en la adaptación el k-NN para STH

La adaptación del algoritmo básico de k-NN, mostrado en el apéndice A.3, a la predicción de STH depende en gran medida del uso de una medida de distancia para datos de histograma. Dicha medida será usada para

1. Calcular los vecinos más próximos entre vectores de histogramas.
2. Obtener las predicciones en forma de histograma como un histograma baricéntrico que minimice la distancia entre sí mismo y los histogramas considerados.

En el siguiente punto se analizarán qué distancias son apropiadas para adaptar el k-NN para trabajar con datos de histograma. A continuación, se detallará el método del k-NN para la predicción de STH, para ello se explicará cómo medir distancias entre secuencias de histogramas y cómo obtener predicciones.

5.6.2. Análisis de las posibles distancias a utilizar en el k-NN sobre datos de histograma

Antes de analizar las distintas distancias, hay que plantearse cómo se mide la distancia entre dos histogramas. Seguramente existan distintas respuestas a esta cuestión, pero la que parece más adecuada en el marco de esta tesis es la de considerar que los histogramas son representaciones de una distribución subyacente y que, por tanto, resulta adecuado medir la distancia entre ellos empleando alguna de las distancias ya existentes para distribuciones.

En el apartado 5.4.2.1, se analizaron una serie de medidas de divergencia entre distribuciones con el fin de determinar cuáles eran las más adecuadas para representar el concepto de error en las STH. Tras el análisis, se consideró que las distancias de Wasserstein y de Mallows eran adecuadas para dicho propósito. La medición del error en cometido al predecir una STH consiste en estimar lo similar que son los histogramas pronosticados a los histogramas reales. Esta idea es la misma que se aplica a la hora de buscar la secuencia de histogramas más similar a la secuencia actual. Por tanto, las mismas razones que se expusieron en favor de las distancias de Wasserstein y de Mallows como distancias para medir el error, pueden ser esgrimidas para defender su aplicación en la búsqueda de los vecinos más próximos en la adaptación del k-NN a las STH.

Rubner, Puzicha, Tomasi y Buhmann (2001) analizan de forma empírica la efectividad de una serie de distancias para histogramas discretos a la hora de realizar tareas relacionadas con la visión artificial: clasificación, segmentación y recuperación de imágenes. En dicho artículo, la distancia de los transportistas de arena o *Earth Mover's Distance* (EMD) obtiene muy buenos resultados y uno de los clasificadores que se utilizan es un k-NN en el que la variable objetivo es cualitativa. Como ya se comentó en el apartado 5.4.2.1, la EMD es una distancia para histogramas multivariantes discretos que guarda relación con las distancias de Mallows y de Wasserstein para funciones de distribución univariantes. Por tanto, se puede considerar que la solvencia de la EMD en el área de visión artificial es otro argumento a favor

del uso de las distancias de Wasserstein y de Mallows en la adaptación del k-NN a las STH.

En esta tesis, se propone un k-NN donde la variable objetivo es también de histograma. Para ello, hay que definir cómo obtener el histograma que generará como resultado el k-NN. Dicho histograma debe comportarse como un promedio de los histogramas objetivo de cada uno de los k vecinos. En el k-NN para datos de histograma que propone esta tesis, dicho histograma será calculado, en lugar de como un promedio, como el histograma baricéntrico de los histogramas considerados.

La sustitución del concepto de promedio de un conjunto de histogramas por el del baricentro de dicho conjunto fue abordada con anterioridad en el punto 5.5.2.1 de esta tesis. En el apartado 5.5.2 se analizaba el uso del histograma baricéntrico para sustituir a las medias móviles y a la media ponderada de la ecuación recursiva de alisado. De igual forma que el histograma baricéntrico sirvió para aquellos propósitos, también puede servir para realizar un promedio ponderado de un conjunto de histogramas que requiere el k-NN. En ambos casos el objetivo es obtener una combinación lineal convexa de un conjunto de histogramas y el resultado de dicha combinación debe calcularse a partir de los histogramas originales y ser un adecuado representante de los mismos.

A la hora de analizar qué distancias emplear para calcular el histograma baricéntrico en el k-NN es importante considerar las conclusiones de los trabajos de Irpino y Verde. Irpino y Verde (2006a) e Irpino y Verde (2006b) realizan clustering de datos de histograma empleando la distancia de Mallows y acuñan el concepto de baricentro de datos de histograma. Posteriormente, Verde y Irpino (2007) analizan las propiedades de los histogramas baricéntricos que se producen con distintas medidas de divergencia. En este análisis descartan los histogramas baricéntricos que se obtienen con muchas medidas y sólo consideran como adecuados aquellos que se obtienen con la distancia de Mallows, alegando como razones su unicidad y que no se comportan como una mixtura del conjunto de histogramas considerado. En el punto B.2 de los apéndices se muestra que se comporta como una media de las funciones de distribución de los histogramas considerados. Estos argumentos también son válidos para justificar el empleo de los histogramas baricéntricos obtenidos mediante la distancia de Mallows como las predicciones que genera el k-NN para datos de histograma.

Entre las distancias analizadas por estos autores, no se encuentra la distancia de Wasserstein. Los histogramas baricéntricos que se obtienen con esta distancia sí son analizados en el punto B.2 de los apéndices, donde se prueba que su comportamiento no es del tipo mixtura, y que se comportan como la mediana de las funciones de distribución del conjunto de histogramas considerados. El uso de un histograma mediano como predicción del k-NN de STH, puede resultar adecuado en determinados contextos.

En conclusión, se estima que tanto la distancia de Mallows, como la de Wasserstein resultan adecuadas para adaptar el k-NN para trabajar con datos de histograma y, más concretamente, para la predicción de STH. La diferencia de comportamiento de los baricentros que generan ambas distancias, los de una se comportan como una media y los de otra como una mediana, permite al analista elegir el uso de una distancia u otra, según le interese obtener las predicciones de una forma o de otra.

Los procesos de determinación de los vecinos más próximos y de obtención de las predicciones son detallados a continuación.

5.6.3. Determinación de los vecinos más próximos

La STH $\{h_{X_t}\}$ con $t = 1, ..., n$ se transforma en una serie de vectores de histograma d-dimensionales tales como

$$h_{X_t^d} = (h_{X_t}, h_{X_{t-1}}, ..., h_{X_{t-d+1}}) \tag{5.44}$$

con $t = d, ...n$. A continuación, se calcula la distancia entre el último vector de la serie $h_{X_n^d}$ y el resto de vectores $h_{X_t^d}$ con $t = d, ..., n-1$ de la siguiente manera

$$D^q(h_{X_n^d}, h_{X_t^d}) = \left(\frac{\sum_{i=1}^{d} (D(h_{X_{n-i+1}}, h_{X_{t-i+1}}))^q}{d} \right)^{\frac{1}{q}}, \tag{5.45}$$

donde $D(h_{X_{n-i+1}}, h_{X_{t-i+1}})$ puede ser la distancia de Mallows o la distancia de Wasserstein, mostradas en la tabla 5.1. El parámetro q permite medir la discrepancia entre los histogramas de forma similar a la raíz cuadrada del error cuadrático medio ($q = 2$) o al error absoluto medio ($q = 1$).

Una vez que se han calculado las $n - d$ distancias, se determinan los k vectores más próximos al vector $h_{X_n^d}$. Dichos vectores se denotarán como $h_{X_{T_p}^d}$ con $p = 1, ..., k$.

5.6.4. Obtención de predicciones

En el método de k-NN para la predicción de series temporales explicado en el apartado A.3 del apéndice, las predicciones se calculan como la media (ponderada o no) de los valores siguientes de cada una de las k secuencias vecinas. Para adaptar el k-NN para manejar histogramas se sustituirá el promedio por la estimación de un baricentro.

Tal y como se indica en los apartados apartado 5.5.2.1 y 4.8.2 de esta tesis, el concepto físico del baricentro de un conjunto de partículas está relacionado con el concepto de media porque las coordenadas del baricentro se obtienen como la media ponderada de las coordenadas de las partículas, donde la ponderación es proporcional a la masa de cada una de las partículas. Teniendo en cuenta este hecho, resulta adecuado utilizar el concepto

de baricentro para reemplazar el promedio que se emplea en el k-NN para generar las predicciones.

En el k-NN para STH, la predicción $\hat{h}_{X_{n+1}}$ se obtendrá como el histograma baricéntrico que minimiza la suma de las distancias entre sí mismo y cada uno de los histogramas siguientes de sus k vecinos más próximos. La predicción $\hat{h}_{X_{n+1}}$ será la solución al siguiente problema de minimización

$$\min_{\hat{h}_{X_{n+1}}} \sum_{p=1}^{k} \omega_p D(\hat{h}_{X_{n+1}}, h_{X_{T_p+1}}), \tag{5.46}$$

donde $D(\hat{h}_{X_{n+1}}, h_{X_{T_p+1}})$ es la distancia de Mallows o de Wasserstein, $h_{X_{T_p+1}}$ es el siguiente histograma de la secuencia $h_{X^d_{T_p}}$, y ω_p es el peso asignado al vecino p tal que satisface $\omega_p \geq 0$ y $\sum_{p=1}^{k} \omega_p = 1$.

Si todos los vecinos tienen el mismo peso asignado, entonces $\omega_p = 1/k$, $\forall p$. Sin embargo, pueden aplicarse una distribución de pesos más sofisticada. Por ejemplo, el peso asignado al valor p puede ser inversamente proporcional a la distancia entre la secuencia actual y la secuencia vecina p tal que

$$\omega_p = \frac{\psi_p}{\sum_{l=1}^{k} \psi_l}, \tag{5.47}$$

con $\psi_p = (D(h_{X^d_n}, h_{X^d_{T_p}}) + \xi)^{-1}$ siendo $p = 1, ..., k$, y donde $D(h_{X^d_n}, h_{X^d_{T_p}})$ viene dado por la ecuación (5.45). El objetivo de la constante $\xi = 10^{-8}$ es evitar que el peso tome el valor infinito en el caso en el que la distancia entre las secuencias consideradas sea cero.

El apéndice B muestra cómo estimar de manera directa el histograma baricéntrico basado en las distancias de Mallows y de Wasserstein. Pese a que el baricentro es el resultado de un proceso de minimización, el método que se muestra en el apéndice no requiere el uso de técnicas de optimización. Esto es debido a que la representación del histograma en base a intervalos facilita enormemente el cálculo y reduce drásticamente el esfuerzo computacional que se requiere. En dicho apéndice también se muestra por qué el baricentro de Mallows de un conjunto de histogramas puede considerarse como el histograma medio, mientras que el baricentro de Wasserstein puede considerarse como el histograma mediano.

5.7. Elaboración y predicción de una STH

En este apartado se describirá el proceso a seguir para construir una STH y obtener predicciones de ella. Este proceso consta de los siguientes pasos.

1. Selección de los datos iniciales. Para poder construir una STH es necesario un conjunto de datos temporales de alguno de estos dos tipos:

1. Se dispone de los valores de una variable medida a lo largo del tiempo en los individuos de un conjunto. Si los individuos del conjunto son los mismos a lo largo de toda la serie, en realidad se dispone de un conjunto de series temporales, pero esto no es un requisito indispensable para construir la STH. En este caso, para cada instante temporal se construirá un histograma con los datos del conjunto de individuos.

2. Se dispone de una serie temporal de una frecuencia mayor a la que interesa analizar. Por ejemplo, se dispone una serie temporal con valores tomados cada minuto y el interés reside en analizar dichos datos con una frecuencia diaria. En este caso se construirá la STH con la frecuencia mayor y para construir cada histograma se utilizarán todos los datos que se han dado entre dos instantes consecutivos de la frecuencia mayor. En el ejemplo dado, se construirá un histograma para cada día con los datos tomados cada minuto.

La primera de las circunstancias muestra un caso prototípico de agregación contemporánea que puede resultar útil en áreas como, por ejemplo, la estadística oficial o los datos de panel, donde se recogen los valores de una o varias variables a lo largo del tiempo en un conjunto de individuos. La segunda circunstancia muestra un caso de agregación temporal que puede surgir, por ejemplo, cuando la serie es generada por un sensor que monitoriza el valor de una variable de forma continua o prácticamente continua.

2. Construcción de la STH. Una vez determinado el conjunto de datos inicial, se han de construir los histogramas que darán lugar a la STH. Para ello, se debe elegir el tipo de histograma que más interesa para representar la STH de acuerdo con el objetivo del análisis que se quiera realizar. Los tipos de histograma fueron comentados en el apartado 5.3 y son:

1. Histogramas equiespaciados: son los histogramas más populares. Se debe usar un número de intervalos que permitan describir los datos de una forma precisa sin enmascarar las características de la distribución.

2. Histogramas equifrecuenciales: permiten construir histogramas del tipo *boxplot* o definidos a partir de los deciles de la distribución.

3. Histogramas construidos sobre una partición del dominio de la variable considerada. Permiten estudiar el comportamiento de la variable en unos determinados intervalos que el analista considere interesantes.

4. Histogramas definidos a partir de una secuencia de cuantiles. Permiten estudiar el comportamiento de la distribución en los cuantiles que se desee, pudiéndose hacer énfasis en aquellas partes de la distribución que sean de mayor interés.

Es posible que se generen STH donde el número de intervalos de cada histograma a lo largo de la STH no sea constante. Esto sucede, por ejemplo, cuando se determina el ancho óptimo del intervalo de cada histograma de acuerdo con una regla como la de Wand (1997), de la que ya se habló en la sección 5.3. Si el analista decide utilizar la regla de Wand obtendrá una representación muy precisa de la densidad subyacente, pero, a cambio, deberá tratar con una STH donde el número de histogramas no es constante.

3. Análisis de la STH. Una vez construida la STH, ésta debe ser representada gráficamente y analizada para determinar cómo se ha comportado a lo largo del tiempo, si tiene tendencia, estacionalidad, valores extremos, patrones que se repitan en el tiempo, etc. Tras este análisis se determinará qué método o métodos resultan adecuados para predecir la STH.

4. Predicción de la STH. Consiste en generar predicciones de la STH con los métodos que fueron considerados apropiados en el paso anterior. En primer lugar, se aconseja dividir la serie en tres periodos:

1. inicialización: constará de tantos periodos como requiera el método de predicción que se va a utilizar.

2. entrenamiento: se empleará para ajustar los parámetros de los métodos de predicción que se estén empleado. Los parámetros de los métodos de predicción proporcionados se pueden determinar mediante una búsqueda en rejilla en el espacio s-dimensional, donde s es el número de parámetros, que tenga como objetivo encontrar la combinación de parámetros que produzca el menor error en el conjunto de entrenamiento.

3. prueba: permite comprobar el rendimiento de los métodos estimados en el entrenamiento y determinar cuál es el que mejor predice la STH.

Es importante comentar un aspecto práctico sobre la generación de predicciones de una STH. Los histogramas de la STH original serán de un tipo determinado (equiespaciados, equifrecuenciales, etc...). Sin embargo, el histograma pronosticado por los métodos de predicción no tiene por qué ser del mismo tipo (y normalmente no lo estará). Por ejemplo, si se aplica el método de k-NN, la predicción se generará como un baricentro y este baricentro normalmente no será un histograma del mismo tipo que los que le han dado lugar. Lo mismo sucede en el caso de los alisados, ya sea con el método de los baricentros o con la aritmética de histogramas. Por eso, parece razonable que las predicciones que generen estos métodos sean reconstruidas en un histograma del mismo tipo que la serie original. Para hacerlo basta con transformar el histograma pronosticado para convertirlo en un histograma del tipo deseado, considerando que dentro de cada intervalo de un histograma las observaciones se distribuyen según una uniforme.

En el caso de los alisados basados en la aritmética de histogramas este problema es más patente porque el histograma resultante de una operación aritmética, como ya se mencionó en el apartado 5.5.1.1 tendrá un gran número de intervalos que, además, pueden estar solapados entre sí. El problema se agrava si se utiliza el histograma resultado de una operación, como operando en otra. Por ello, los histogramas resultantes de las ecuaciones de alisado deben ser expresados como histogramas del tipo que se esté empleando en la serie temporal. En el apartado 5.5.1.1 se explica cómo transformar un histograma no disjunto en otro disjunto equiprobable de un número menor de intervalos. Resulta trivial adaptar dicho procedimiento para generar histogramas equiespaciados o de alguno de los otros tipos considerados en la tesis.

Por otro lado, si el número de intervalos de los histogramas de la STH varía a lo largo de la serie (como sucederá si se utiliza la regla de Wand (1997)), también se debe determinar con cuántos intervalos se construye cada uno de los histogramas pronosticados. Como *a priori* es imposible conocer el número de intervalos o el ancho óptimo de los intervalos del histograma futuro, lo más razonable es construir todas las predicciones con el mismo número de intervalos. Para elegir este número se puede emplear la moda del número de intervalos de los histogramas de la serie o cualquier otro estadísticos de tendencia central.

5.8. Ejemplos ilustrativos de la predicción de STH

El objetivo de esta sección es aplicar las técnicas de predicción de STH sobre conjuntos de datos reales. Esto permitirá ilustrar los conceptos que han sido enunciados a lo largo de todo el capítulo.

Los conjuntos de datos elegidos provienen de distintos ámbitos como son la meteorología, el medio ambiente y las finanzas. Esto demuestra que las STH tienen una gran aplicabilidad y que no se circunscriben a un ámbito concreto. Además, son posibles otros ámbitos de aplicación no recogidos en esta tesis. Uno de los más interesantes que no aparece en esta tesis debido a la imposibilidad de conseguir datos es el de la estadística oficial. En dicha área se puede estudiar la evolución de temporal de variables medidas sobre los individuos de un país (o de una región) utilizando como estadístico resumen el histograma, el cual ofrece más información sobre los individuos que otros estadísticos empleados habitualmente como la media. Otros posibles campos de aplicación incluyen el análisis de datos de panel, el control de calidad y el sector energético.

En algunos de los ejemplos que se van a mostrar, las STH han sido generadas mediante agregación contemporánea, mientras que en otros se ha empleado la agregación temporal, según fuese necesario en cada caso en cada caso. Las características de las STH resultantes son notablemente dis-

tintas, por ejemplo, se va a trabajar con series con estacionalidad clara, con estacionalidad cambiante y con series que, aunque originalmente no son estacionarias, al transformarse se convierten en estacionarias.

En los ejemplos también se utilizarán los distintos tipos de histogramas presentados en el apartado 5.3: equiespaciados, equifrecuenciales, definidos sobre una partición de la variable del dominio y definidos como una secuencia de cuantiles. Cuando se utilicen histogramas equiespaciados se empleará la regla propuesta por Wand (1997) para estimar de forma precisa la densidad subyacente. Los histogramas equifrecuenciales que se emplearán serán *box-plots*, i.e. histogramas equifrecuenciales de cuatro intervalos, por la claridad de su representación gráfica y por su fácil interpretación. Los histogramas definidos sobre una partición de la variable del dominio y los definidos sobre una secuencia de cuantiles están más ligados al problema que se está abordando, por lo que se darán más detalles sobre ellos en el ejemplo en el que se apliquen.

5.8.1. Descripción de la metodología seguida en la predicción de cada STH

En cada una de las STH se han aplicado las tres técnicas de predicción que se han presentado en este capítulo: los alisados exponenciales basados en la aritmética de histogramas, los alisados exponenciales basados en el método de los baricentros y el k-NN basado en el método de los baricentros con los distintos esquemas de ponderación: igual peso para todos los elementos (en las tablas de resultados aparecerá como *cte.* de constante) o un peso inversamente proporcional a la distancia (al que se hará referencia en las tablas como *inv.* de inverso).

Cada una de las STH se ha divido en tres partes: inicialización, entrenamiento y prueba. La longitud de los tres periodos varía según la serie, pero es aproximadamente un tercio de la longitud total.

El periodo de inicialización se ha incluido porque la técnica de k-NN necesita que la serie tenga un periodo histórico suficientemente grande como para poder realizar la búsqueda de los vecinos más cercanos y encontrar vecinos suficientemente similares a la secuencia actual. Los alisados con estacionalidad también necesitan de p valores de inicialización, siendo p la longitud del ciclo estacional. Para evitar que el periodo de la serie para los cuales se mide el error de entrenamiento sea distinto en cada uno de los métodos, se ha optado por fijar un periodo de inicialización para cada serie que será igual para todos los métodos. La longitud de dicho periodo será la que se considere adecuada para el método de k-NN (ya que es el que necesita de un periodo de inicialización más largo). Esto se hace para evitar confusiones a la hora de leer el error medio cometido en el conjunto de entrenamiento de la serie por los distintos métodos. En el periodo de inicialización no se medirá el error.

El resto de la serie se divide en el periodo de entrenamiento y el de prueba. Los parámetros óptimos de cada uno de los métodos se estimarán mediante una búsqueda en rejilla sobre un espacio s-dimensional, donde s es el número de parámetros que requiere cada uno de los métodos. La búsqueda en rejilla tendrá como objetivo minimizar el error cometido por el método durante el periodo de entrenamiento.

Para cada una de las series consideradas se mostrará el error de entrenamiento y de prueba cometido por cada una de las técnicas consideradas. Las medidas de error que se mostrarán serán el error medio basado en la distancia de Mallows o (EMD_M) y el error medio escalado basado también en la distancia de Mallows ($EMED_M$), en este último se escalará utilizando como método de referencia el método ingenuo con o sin estacionalidad, según la serie tenga o no componente estacional. El orden de estas medidas será $q = 1$, por lo que en ellas el error, medido como una distacia, entre la predicción y el valor observado será agregado sin elevarse al cuadrado, es decir, como en el error absoluto medio. Estas medidas fueron presentadas en el apartado 5.4.2.3. Se ha optado por mostrar los errores basados en las distancias de Mallows por su semejanza con el error cuadrático medio que se emplea habitualmente para mostrar el error en las series temporales clásicas. La búsqueda de parámetros óptimos se realizará también minimizando el EMD_M en el periodo de entrenamiento.

Es importante reseñar que se han realizado también las pruebas empleando el error medio basado en la distancia de Wasserstein y que pese al cambio de la distancia no se producen diferencias significativas en las conclusiones del análisis. Por ello, y para no resultar prolijos, se ha optado por sólo mostrar los resultados de los errores basado en la distancia de Mallows.

5.8.2. Predicción de datos meteorológicos con agregación contemporánea

El conjunto de datos considerado ha sido obtenido de la base de datos *Long-Term Instrumental Climatic Database of the People's Republic of China*[4]. El archivo contiene 14 variables climatológicas registradas con una frecuencia mensual por 60 estaciones meteorológicas. Cada estación es representativa de un región climática de China y el conjunto de las 60 estaciones forma una red con una distribución espacial aproximadamente uniforme.

Estas características hacen de este conjunto de datos un conjunto apropiado para emplear agregación espacial, ya que el histograma permite resumir los valores obtenidos por las 60 estaciones. La STH resultante describirá la evolución temporal de la distribución mensual de la variable considerada a lo largo y ancho de China. El hecho de representar las variables climátoló-

[4]Disponible en `http://dss.ucar.edu/datasets/ds578.5/data/` para usuarios registrados (el registro es gratuito)

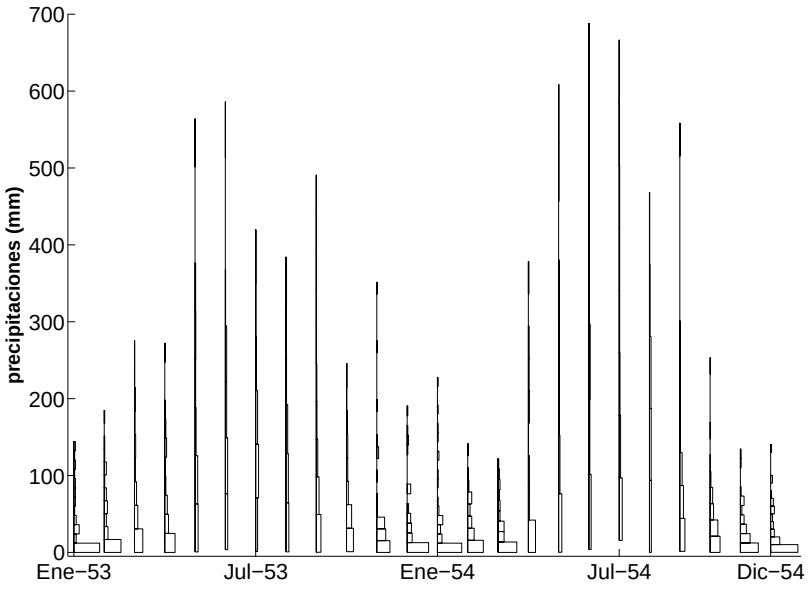


Figura 5.13: Extracto de la STH de la distribución de precipitaciones mensuales en China

gicas mediante distribuciones agregadas permite obtener un punto de vista diferente sobre algunos fenómenos interesantes como el aumento de las temperaturas producido en los últimos años.

Se ha considerado el periodo entre 1951 y 1988, ambos años incluidos. Por tanto, se cuenta con 60 series mensuales de 456 valores. Como periodo de inicialización se usarán los primeros 144 periodos de la serie, siendo los 156 siguientes el periodo de entrenamiento y los últimos 156 el de prueba.

5.8.2.1. Precipitaciones en la República Popular China

En este caso se abordará la distribución de la precipitación mensual en China medida en *mm*. Los histogramas que se emplearán para resumir la información serán histogramas equiespaciados, donde el ancho del intervalo de cada histograma ha sido determinado por la regla de Wand (ver sección 5.3). De esta forma, se obtendrán histogramas equiespacidados que resulten óptimos a la hora de representar de forma precisa la distribución subyacente, pero el número de intervalos de cada histograma variará a lo largo de la STH.

La figura 5.13 muestra un extracto de la serie resultante. En ella puede apreciarse que la STH de las precipitaciones tiene una componente estacional muy clara que afecta al histograma en su conjunto, ya que tanto su forma como su rango sufren una variación periódica a lo largo del año. La estacio-

Tabla 5.3: Errores de predicción cometidos por los diferentes métodos en la STH de las precipitaciones en China

	Entrenamiento		Prueba	
Método	EMD_M	$EMED_M$	EMD_M	$EMED_M$
Ingenuo	41.75	1.25	39.2	1.17
Ingenuo estacional	33.5	1	31.77	0.95
k-NN M. cte. ($k = 14$ $d = 18$)	24.51	0.73	23.3	0.7
k-NN W. cte. ($k = 14$ $d = 23$)	25.04	0.75	**23**	**0.69**
k-NN M. inv. ($k = d =$)	24.61	0.73	23.28	0.69
k-NN W. inv. ($k = 14$ $d = 23$)	25.39	0.76	23.14	0.69
AES Arit. ($\alpha = .94$)	41.4	1.24	38.95	1.16
AES Baric. ($\alpha = 1$)	41.36	1.23	38.61	1.15
AEEc Arit. ($\alpha = .81$ $\delta = .17$)	35.76	1.07	33.06	0.99
AEEc Baric. ($\alpha = .81$ $\delta = .19$)	36.19	1.08	33.45	1
AEEh Arit. ($\alpha = .02$ $\delta = .01$)	29.54	0.88	28.46	0.85
AEEh Baric. ($\alpha = .02$ $\delta = .05$)	24.56	0.73	23.27	0.69

nalidad es debida al efecto del Monzón que tiene lugar entre los meses de mayo y octubre y que ocasiona grandes lluvias en China. Una distribución de las precipitaciones en China puede ayudar al gobierno a planificar asuntos relacionados con las reservas hídricas como, por ejemplo, el suministro de agua para consumo y para regadío, la producción de energía, etc.

Al usar la regla de Wand, el número de intervalos de los histogramas varía a lo largo de la serie. Por ello, a la hora de generar los histogramas pronosticados hay que determinar con cuántos intervalos se representará la predicción. En este caso, se ha decidido que sea 10 porque la mediana del número de intervalos en la STH es 10 y porque su media es 10.65.

La serie que se está analizando es de frecuencia mensual y su ciclo estacional es anual. Por ello, como método de referencia se empleará el método ingenuo con estacionalidad, $\hat{h}_{X_{t+1}} = h_{X_{t+1-p}}$, siendo $p = 12$ la longitud del ciclo estacional. Los alisados que se emplearán para predecir la serie serán los alisados estacionales. El alisado con estacionalidad en forma de histograma (AEEh) es el que, a juzgar por la figura 5.13, parece más apropiado. Sin embargo, también se empleará el alisado en el que la estacionalidad se representa como un valor clásico (AEEc) y los alisados sin estacionalidad para observar la mejora que se produce al tener en cuenta esta componente. Como ya se ha dicho anteriormente, se emplearán los alisado basados en aritmética de histogramas y los basados en los baricentros.

Respecto al método de k-NN, se empleará tanto el que emplea la distancia de Wasserstein como el que emplea la distancia de Mallows y usando los dos esquemas de ponderación: el mismo peso para todos (cte.) y el peso inversamente proporcional a la distancia (inv.).

La tabla 5.3 muestra los errores de predicción cometidos tanto en el periodo de entrenamiento como en el periodo de prueba por los métodos considerados. Al tratarse de una serie con estacionalidad, se ha empleado como método de referencia para calcular el $EMED_M$ el método ingenuo con estacionalidad. Los métodos que no tienen en cuenta la estacionalidad, como el método ingenuo o como los alisados exponenciales simples (AES), no obtienen buenos resultados. Tampoco predicen bien los alisados que consideran que la estacionalidad sólo afecta a la posición del histograma (AEEc), ya que, como muestra la figura 5.13, la estacionalidad afecta a todo el histograma. Sin embargo, los alisados que modelan la estacionalidad en forma de histograma (AEEh) obtienen una mejora notable respecto al método ingenuo estacional. Si se observa el $EMED_M$ para el caso del AEEh que emplea los baricentros, se puede concluir que su rendimiento mejora en media el rendimiento del método ingenuo en el conjunto de entrenamiento en torno a un 30 %, lo cual es una gran mejoría. El rendimiento obtenido por los métodos de k-NN es también muy bueno y comparable al del AEEh basado en baricentros. No existen diferencias significativas según se emplee la distancia de Mallows o de Wasserstein, ni según se emplee un esquema de ponderación u otro.

5.8.2.2. Temperatura Media en la República Popular China

En este otro ejemplo se analizará la distribución de la temperatura media mensual en China medida en grados centígrados. En este caso también se ha optado por representar la distribución mediante histogramas equiespaciados donde el ancho del intervalo se ha estimado mediante la regla de Wand (1997). Por ello, como en el ejemplo anterior, el número de intervalos a lo largo de la STH no es constante. Esto hace que haya que fijar *a priori* el número de intervalos con el que se generarán las predicciones. En este caso, se ha decidido que dicho valor sea 5 porque la mediana y la moda del número de intervalos en la STH es 5 y porque su media es 5.6.

La figura 5.14 muestra un extracto de la STH que representa la distribución de las temperaturas medias mensuales en China. En esta serie la componente estacional es evidente: los meses de invierno muestran una gran variación térmica a lo largo de China, mientras que en los meses de verano hay menor variación. También se puede apreciar como en los meses de verano la distribución de las temperaturas es asimétrica a la derecha, mientras que en los meses de invierno existe una mayor simetría. Por tanto, la estacionalidad no sólo afecta a la posición del histograma, sino a todo él.

El análisis de la STH de la temperatura media en China puede ser útil para analizar fenómenos de actualidad como el calentamiento global. A modo de curiosidad cabe comentar que, al margen del trabajo que se realiza en este punto, se han analizado los centros de gravedad de los histogramas por meses y se ha observado que en los meses de invierno sí se aprecia cierta

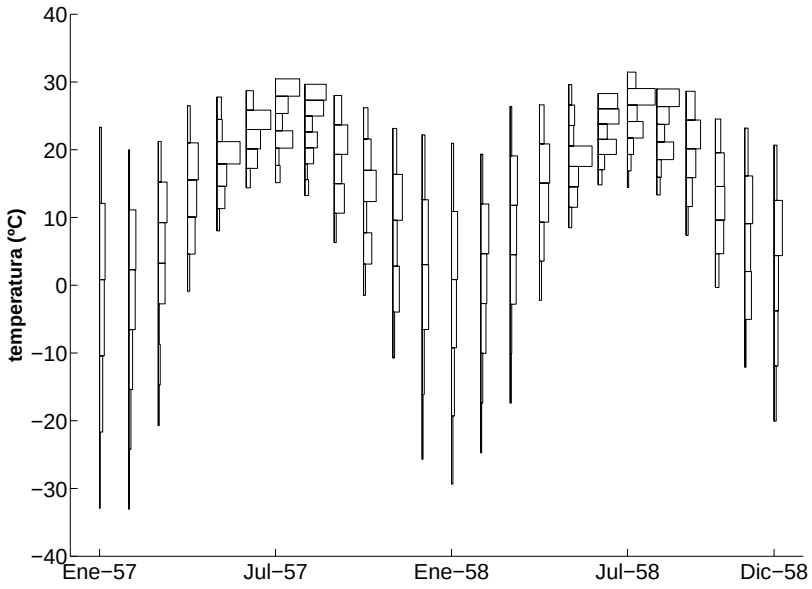


Figura 5.14: Extracto de la STH de la distribución de las temperaturas medias mensuales en China

tendencia ascendente a lo largo de los años. Esto indicaría que aparentemente los inviernos son algo más calurosos, especialmente a partir de los años 80.

Al igual que en el caso anterior, se empleará como método de referencia el método ingenuo con estacionalidad, $\hat{h}_{X_{t+1}} = h_{X_{t+1-p}}$, con $p = 12$. También se empleará el k-NN de Wasserstein y de Mallows con los dos esquemas de ponderación propuestos. Respecto a los alisados exponenciales, se probarán tanto los que no tienen estacionalidad como los que sí, aunque el análisis visual de la serie indica que funcionarán mejor los que recogen la estacionalidad y, de esos, mejor los que la modelan mediante un histograma (AEEh).

En la tabla 5.4 se muestran los resultados de los distintos métodos de predicción aplicados. La tabla recoge el error medio basado en la distancia de Mallows. El $EMED_M$ se ha escalado empleando como error de referencia el cometido por el método ingenuo con estacionalidad en el conjunto de entrenamiento. Como era de esperar, los métodos que no tienen en cuenta la componente estacional, como los alisados exponenciales simples (AES) o el método ingenuo, obtienen los peores resultados. Los alisados que manejan la estacionalidad como un número real (AEEc) obtienen peores resultados que el método ingenuo con estacionalidad, tal y como muestra el $EMED_M$. El alisado exponencial basado en aritmética de histogramas y que maneja la estacionalidad en forma de histograma (AEEh Arit.) sólo obtiene unos resultados ligeramente mejores que el método ingenuo con estacionalidad.

Tabla 5.4: Errores de predicción cometidos por los diferentes métodos en la STH de las temperaturas medias en China

	Entrenamiento		Prueba	
Método	EMD_M	$EMED_M$	EMD_M	$EMED_M$
Ingenuo	4.79	3.44	4.7	3.37
Ingenuo estacional	1.39	1	1.44	1.04
k-NN M. cte. ($k = 12$ $d = 5$)	1.09	0.79	1.11	0.8
k-NN W. cte. ($k = 8$ $d = 19$)	1.11	0.8	1.11	0.8
k-NN M. inv. ($k = 13$ $d = 4$)	1.09	0.79	**1.09**	**0.79**
k-NN W. inv. ($k = 13$ $d = 3$)	1.12	0.81	**1.09**	**0.79**
AES Arit. ($\alpha = 1$)	4.81	3.45	4.7	3.37
AES Baric. ($\alpha = 1$)	4.81	3.45	4.7	3.37
AEEc Arit. ($\alpha = .89$ $\delta = .41$)	2	1.43	1.93	1.39
AEEc Baric. ($\alpha = .91$ $\delta = .3$)	2.08	1.49	2.06	1.48
AEEh Arit. ($\alpha = .01$ $\delta = .93$)	1.37	0.98	1.39	1
AEEh Baric. ($\alpha = .01$ $\delta = .03$)	1.09	0.79	1.14	0.81

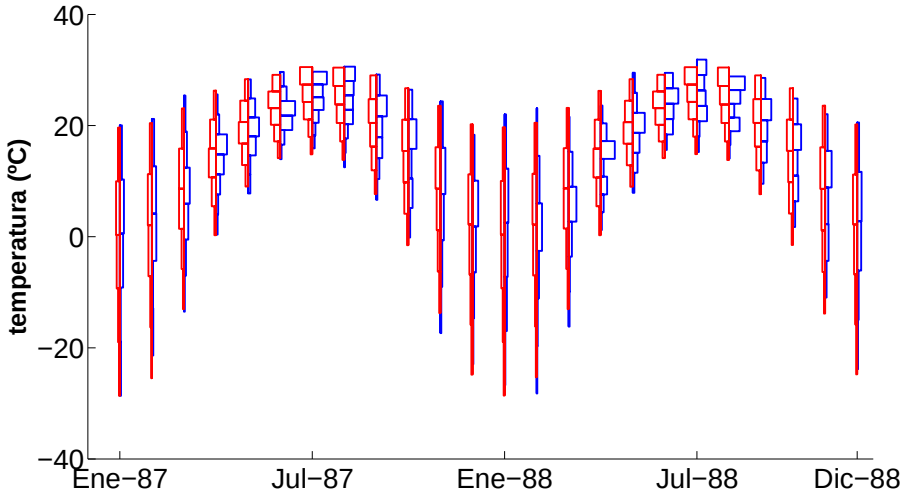


Figura 5.15: STH real (azul) y pronosticada (rojo) en una parte del periodo de prueba

Sin embargo, el AEEh basado en baricentros funciona notablemente mejor y reduce el error cometido por el método de referencia en torno al 20 %, lo cual es una mejora considerable. Los métodos de k-NN obtienen unos resultados muy similares al AEEh basado en baricentros por lo que también resultan métodos adecuados para predecir esta serie. Cabe reseñar que no se aprecian grandes diferencias en los métodos de k-NN según la distancia o el tipo de ponderación que se emplee, todos ellos funcionan con similar solvencia.

La figura 5.15 muestra la STH real y la pronosticada empleando el alisado exponencial con estacionalidad en forma de histograma que emplea el método de los baricentros. En ella se puede ver como la predicción se ajusta muy bien a la serie real. Las series pronosticadas por los métodos de k-NN también obtienen un muy buen ajuste, pero no serán reproducidas por cuestiones de espacio.

5.8.3. Predicción de datos medioambientales empleando agregación contemporánea

Las estaciones medioambientales monitorizan los niveles de contaminantes atmosféricos a lo largo del tiempo. Dichas estaciones suelen estar situadas en puntos estratégicos, de tal manera que forman una red espacial que describe de forma precisa el comportamiento de la región o ciudad en la que se encuentran. Como consecuencia de la monitorización continua, estas estaciones producen una gran cantidad de datos que deben ser resumidos o agregados para poder ser manejados. Los datos medioambientales son, por tanto, un área propicia para aplicar la metodología simbólica.

En este caso se han considerado los datos registrados por una red de 27 estaciones medioambientales situadas en la ciudad de Madrid. Los datos de los que se dispone son las medias mensuales de una serie de contaminantes atmosféricos registrados entre 1994 y 2006. Estos datos son públicos y pueden ser descargados de la web de la Concejalía de Medioambiente del Ayuntamiento de Madrid[5].

Cada una de las estaciones es representativa de un zona de Madrid, por ello parece adecuado agregar las 27 medias mensuales mediante un histograma que represente la distribución de los contaminantes en Madrid capital. En otras palabras, cada histograma representará la distribución de la media mensual de un determinado contaminante en la ciudad de Madrid. La STH de histogramas resultantes permite analizar la evolución de la distribución de cada contaminante a lo largo del tiempo para observar patrones de comportamiento, tendencias, estacionalidad, etc.

Para representar esta distribución se han usado histogramas definidos sobre una partición en el espacio de la variable de interés que, en este caso, es el nivel de un contaminante. Los niveles de contaminante suelen representarse

[5] `http://www.mambiente.munimadrid.es`

mediante índices de calidad del aire (ICAs). Los indices de calidad del aire se diseñan para informar de forma sencilla a la población de las condiciones del aire que se dan en el día y, en el caso de que alcancen niveles peligrosos para la salud, para lanzar las correspondientes alertas. Para ello, se toman como referencia los valores límite que marca la legislación vigente y que dependen de la frecuencia con la que se registren los datos (típicamente se manejan valores horarios, octohorarios y diarios).

En nuestro caso, los datos que se manejan son las medias mensuales. Obviamente, al tratarse de una serie de frecuencia mensual su predicción no sirve para lanzar alertas, ni informar en el corto plazo. Sin embargo, sí permite estudiar la evolución a más largo plazo y de forma conjunta de los contaminantes en la ciudad. Para elaborar los índices de frecuencia mensual se ha contado con la ayuda de Manuel Vellón, un profesional con una amplia experiencia en proyectos de investigación medioambiental y cofundador del portal web sobre contaminación atmosférica TROPOSFERA[6].

Para elaborar un ICA a partir de los promedios mensuales se tomarán como referencia los valores límite anuales existentes en la legislación para la protección de salud humana. Los datos brutos del contaminante se transforman para representarse como un porcentaje calculado con respecto al valor límite

$$ICA_{mes} = concentración_{mes} \cdot (100/(VL + MdT)), \tag{5.48}$$

donde $VL + MdT$ es el acrónimo que recibe el Valor Límite y Margen de Tolerancia de un determinado contaminante marcado por la ley. El valor del ICA mensual se clasifica conforme a un rango que indica el nivel de calidad

$$\begin{aligned} BUENA \quad & si \quad ICA_{mes} \in [0, 50] \\ ADMISIBLE \quad & si \quad ICA_{mes} \in (50, 100], \\ MALA \quad & si \quad ICA_{mes} \in (100, 150], \\ MUYMALA \quad & si \quad ICA_{mes} > 150. \end{aligned} \tag{5.49}$$

Este rango será el que se utilice como partición del espacio de la variable. Es decir, no se utilizará una partición sobre los niveles del contaminante en bruto, sino sobre su porcentaje con respecto al valor límite.

Como ya se ha mencionado, se dispone las medias mensuales de varios contaminantes recogidas en 27 estaciones medioambientales de Madrid entre 1994 y 2006. Para cada contaminante, se representarán los 27 datos como un histograma construido sobre la partición del ICA mostrada en la fórmula (5.49), dando lugar a una STH de 156 periodos mensuales. En cada STH, los primeros 48 periodos se emplearán para la inicialización, los siguientes 56 como conjunto de entrenamiento y los últimos 52 como conjunto de prueba.

[6] `http://www.troposfera.org`

5.8.3.1. Los niveles de dióxido de nitrógeno en el aire en la ciudad de Madrid

Una de las variables que se analiza en las estaciones medioambientales es el nivel de dióxido de nitrógeno, NO_2, medido en $\mu g/m^3$. El NO_2 es un contaminante tóxico por inhalación. Una exposición prolongada a altas concentraciones de NO_2, $40-100\mu g/m^3$, puede causar problemas de salud, como la disminución de la función pulmonar o el aumento de la probabilidad de padecer problemas respiratorios, especialmente en los niños. La exposición a niveles altos de NO_2 puede incrementar las reacciones alérgicas respiratorias. Puesto que la presencia del NO_2 está muy relacionada con la formación o presencia de otros contaminantes del aire, no se sabe con certeza si la exposición a largo plazo a concentraciones relativamente bajas de NO_2 puede afectar, por sí sola, a la mortalidad o al agravamiento de enfermedades.

Las fuentes principales de NO_2 son los motores de combustión interna (como los del tráfico rodado) y, en general, la quema de combustibles fósiles. Su presencia en el aire contribuye a la formación y modificación de otros contaminantes del aire tales como el ozono y las partículas en suspensión, así como a la aparición de la lluvia ácida. La ausencia de lluvia y viento en épocas calurosas favorece que la concentración de NO_2 aumente.

En los datos de los que se dispone se refleja el nivel medio mensual de NO_2 medido en $\mu g/m^3$. Sin embargo, en lugar de considerar los valores brutos del nivel de NO_2, se considerarán los valores en porcentaje dividiendo el valor bruto entre el valor límite anual establecido por ley en 2006[7], que corresponde a $48\mu g/m^3$. Este valor refleja el Índice de Calidad del Aire mensual del NO_2

$$ICA_{NO_2} = concentración_{NO_2\ mensual} \cdot (100/48). \tag{5.50}$$

Con los valores del ICA_{NO_2} mensuales de cada una de las estaciones se construirá un histograma definido sobre una partición del ICA_{NO_2} similar a la mostrada en la ecuación (5.49). La diferencia es que la partición en (5.49) no tiene una cota superior y los histogramas con los que se trabaja en esta tesis deben estar perfectamente acotados (ver la definición del histograma en el apartado 5.2). Por ello, para recoger los valores altos que se producen excepcionalmente en el ICA_{NO_2} de algunas estaciones se amplia dicha partición para recoger los intervalos $(150, 200]$, $(200, 250]$ y $(250, 300]$.

La STH resultante consta de 156 histogramas mensuales correspondientes a los 13 años comprendidos entre 1994 y 2006. Cada histograma representa la distribución de la media mensual del nivel de NO_2 en Madrid. La figura 5.16 muestra una parte de la serie resultante. En ella, además de comprobar que la calidad del aire en Madrid no es buena, es complicado determinar si existe algún patrón de estacionalidad.

[7] Los valores límite establecidos por ley son distintos para cada año y siguen una tendencia decreciente. Para 2002 el valor era $56\mu g/m^3$ y para 2010 el valor límite está fijado en $40\mu g/m^3$ disminuyendo a raíz de $2\mu g/m^3$ al año.

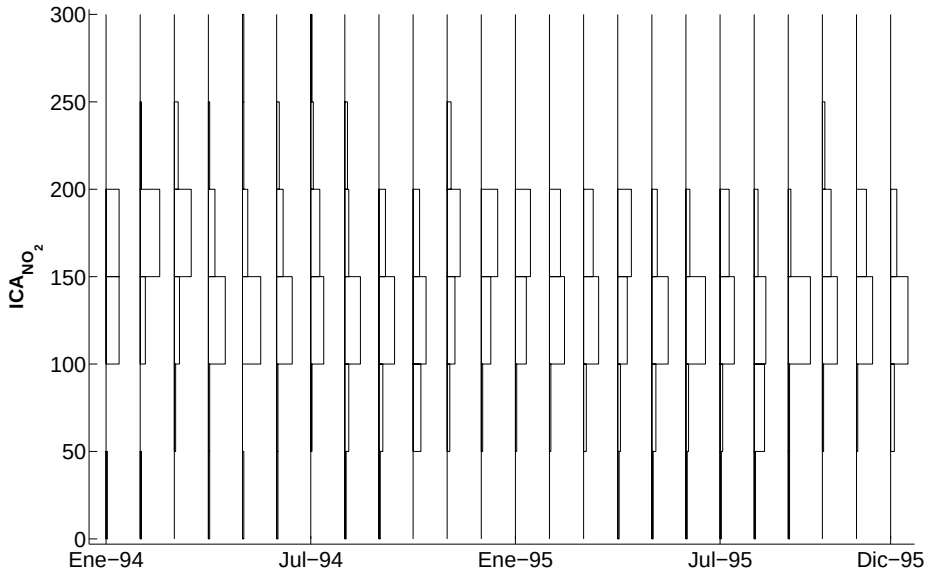


Figura 5.16: Extracto de la STH de la distribución del ICA_{NO_2} en la ciudad de Madrid

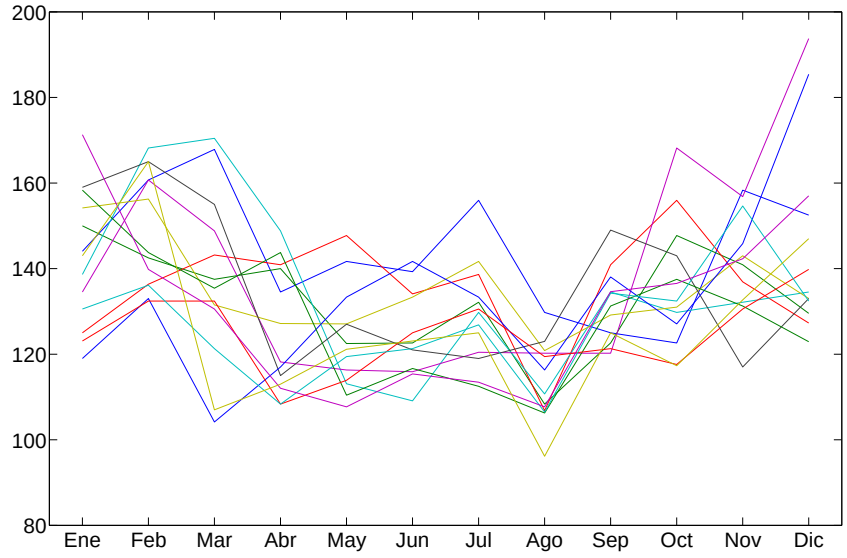


Figura 5.17: Representación de las series temporales anuales de los centros de gravedad de la STH del ICA_{NO_2} en Madrid en cada uno de los años considerados

Tabla 5.5: Errores de predicción cometidos por los diferentes métodos en la STH del nivel de ICA_{NO_2} en Madrid

	Entrenamiento		Prueba	
Método	EMD_M	$EMED_M$	EMD_M	$EMED_M$
Ingenuo	17.86	1	16.22	0.91
Ingenuo estacional	23.39	1.31	17.11	0.96
k-NN M. cte. ($k = 6$ $d = 19$)	17.75	0.99	16.59	0.93
k-NN W. cte. ($k = 5$ $d = 19$)	17.14	0.96	14.39	0.81
k-NN M. inv. ($k = 6$ $d = 19$)	17.77	0.99	16.49	0.92
k-NN W. inv. ($k = 5$ $d = 19$)	17.14	0.96	14.39	0.81
AES Arit. ($\alpha = .75$)	16.78	0.94	15.6	0.87
AES Baric. ($\alpha = .85$)	18.03	1.01	16	0.9
AEEc Arit. ($\alpha = .71$ $\delta = .37$)	16.78	0.94	**13.44**	**0.75**
AEEc Baric. ($\alpha = .76$ $\delta = .16$)	21.02	1.18	19.28	1.08
AEEh Arit. ($\alpha = .69$ $\delta = .17$)	16.8	0.94	15.13	0.85
AEEh Baric.($\alpha = .64$ $\delta = .06$)	17.26	0.97	15.81	0.89

Para determinar si existe estacionalidad, en la figura 5.17 se muestran los centros de gravedad de los doce histogramas mensuales para cada uno de los años considerados. En esta imagen puede verse cierto patrón estacional entre los meses de julio y septiembre. El bajón que se produce en agosto puede ser debido a la disminución del tráfico rodado en Madrid durante este mes. Sin embargo, la estacionalidad no es lo suficientemente clara. La razón es que los niveles de NO_2 no sólo dependen del mes del año en el que nos encontremos, sino de otros factores cuya repercusión no es despreciable.

Al no quedar lo suficientemente claro si existe o no componente estacional, se han probado los alisados con y sin estacionalidad. Como métodos de referencia se usarán los métodos ingenuos con estacionalidad ($h_{\hat{X}_{t+1}} = h_{X_{t-11}}$) y sin ella ($h_{\hat{X}_{t+1}} = h_{X_t}$). En la tabla 5.5 se muestran los resultados obtenidos por los métodos considerados. Como el método ingenuo sencillo funciona mejor que el ingenuo estacional, se ha utilizado para escalar el error y obtener el $EMED_M$. Del análisis del error cometido por los métodos ingenuos se pueden concluir que la serie se predice mejor con el valor inmediatamente anterior que con el valor obtenido el mes pasado. Sin embargo, el error cometido por el método ingenuo estacional en el entrenamiento se ve reducido enormemente en el periodo de prueba, lo que indica que la componente estacional es más clara en dicho periodo. Otra conclusión es que la serie es más estable en el periodo de prueba porque los métodos ingenuos funcionan mejor en ella.

Los resultados mostrados en la tabla indican que el k-NN basado en la distancia de Mallows funciona de forma similar que el método ingenuo simple. Por su parte, el k-NN basado en la distancia de Wasserstein sí me-

jora los resultados del ingenuo. El hecho de que funcione mejor el k-NN con la distancia de Mallows que con la distancia de Wasserstein indica que las predicciones son más precisas si se generan como una mediana de valores pasados, en lugar de como una media de los mismos. La mediana no es sensible al comportamiento más extremo, mientras que la media sí lo es. Los métodos de alisado exponencial con estacionalidad también han obtenido, casi todos, mejores resultados que el método ingenuo. De entre ellos, el método que mejor ha funcionado es el alisado exponencial que modela la estacionalidad como un valor real y que está basado en aritmética de histogramas.

5.8.3.2. Los niveles de partículas en suspensión en el aire en la ciudad de Madrid

En este caso, la variable estudiada será el nivel de partículas en suspensión de diámetro inferior a 10 micras, PM_{10}, medido en $\mu g/m^3$. Las partículas en suspensión son aquellas partículas tanto sólidas como líquidas que se encuentran suspendidas en el aire entre las que se encuentran partículas de polvo, humo de tabaco, cenizas volantes, hollín, polen y esporas. Su composición y su tamaño son muy variables, lo que influye en la manera en que afectan a la salud humana. La exposición prolongada a niveles altos de partículas en suspensión puede afectar a los pulmones y reducir la esperanza de vida en unos cuantos meses, especialmente en el caso de personas con enfermedades cardíacas y pulmonares.

Las partículas pueden ser emitidas al aire de forma directa cuando provienen de fuentes como los procesos de combustión o el polvo arrastrado por el viento; o bien formarse en la atmósfera por la transformación de gases emitidos como el SO_2. En Europa, los sulfatos y la materia orgánica son los principales componentes del conjunto de partículas en suspensión que contaminan el aire. El polvo mineral, los nitratos y el hollín también pueden llegar a ser componentes mayoritarios en determinadas condiciones.

Los datos de los que se disponen son las medias mensuales de partículas en suspensión en 27 puntos de Madrid. Al igual que en el caso del NO_2, en lugar de considerar los niveles medidos en $\mu g/m^3$, se trabajará con el valor en porcentaje que se obtiene al dividir el valor bruto entre el valor límite anual establecido por ley, que corresponde a $40\mu g/m^3$. El Índice de Calidad del Aire mensual de partículas en suspensión ($ICA_{PM_{10}}$) resultante es

$$ICA_{PM_{10}} = concentración_{PM_{10}\ mensual} \cdot (100/40). \tag{5.51}$$

A partir del $ICA_{PM_{10}}$ mensual de cada una de las estaciones se construirá un histograma definido sobre una partición del $ICA_{PM_{10}}$ similar a la mostrada en la ecuación (5.49), pero cuyo último intervalo debe ser finito. Para ello y para recoger en la partición los valores anormalmente altos que se registran en algunas estaciones, se añaden a la partición mostrada en (5.49) los intervalos $(150, 200]$ y $(200, 250]$.

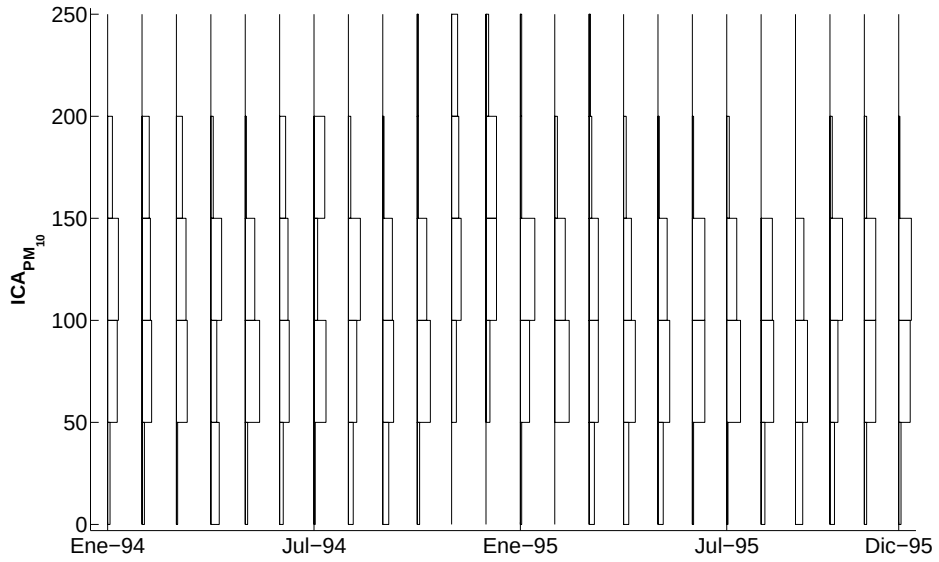


Figura 5.18: Extracto de la STH de la distribución del $ICA_{PM_{10}}$ en la ciudad de Madrid

La STH resultante consta de 156 histogramas mensuales (correspondientes al periodo 1994-2006) donde cada histograma representa la distribución de la media mensual del $ICA_{PM_{10}}$ en Madrid. Tal y como puede verse en la figura 5.18, la STH no muestra un patrón estacional lo suficientemente claro.

Por su parte, la representación de los centros de gravedad de los histogramas a lo largo del año que muestra la figura 5.19 tampoco revela una componente estacional clara. Sí parece que en abril los niveles de partículas en suspensión descienden habitualmente, para repuntar en el mes de junio. Sin embargo, en los meses de invierno la variabilidad es mucho mayor y no se aprecian indicios de estacionalidad.

Al no quedar claro si la serie posee estacionalidad o no, se emplearán métodos que tengan en cuenta la estacionalidad y métodos que no la consideren. En la tabla 5.6 se muestran los resultados obtenidos. De los métodos ingenuos, el sencillo funciona mejor que el estacional. Por tanto, será el error cometido por el método ingenuo sencillo en el periodo de entrenamiento el que se emplee para calcular los errores escalados ($EMED_M$).

La conclusión principal que puede obtenerse al analizar los resultados es que resulta complicado batir al método ingenuo. Sólo el alisado exponencial simple basado en la aritmética de histogramas mejora notablemente los resultados obtenidos por el método ingenuo. El resto de métodos o funcionan aproximadamente igual que el método ingenuo o empeoran sus resultados.

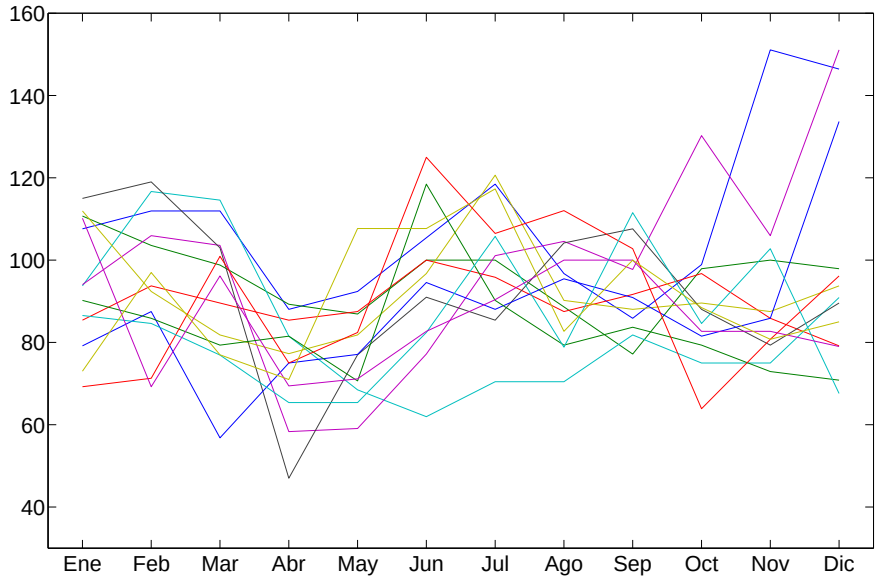


Figura 5.19: Representación de las series temporales anuales que representan los centros de gravedad de la STH del $ICA_{PM_{10}}$ en Madrid en cada uno de los años considerados

Tabla 5.6: Errores de predicción cometidos por los diferentes métodos en la STH del nivel de partículas en suspensión en Madrid

	Entrenamiento		Prueba	
Método	EMD_M	$EMED_M$	EMD_M	$EMED_M$
Ingenuo	18.6	1	17.92	0.96
Ingenuo estacional	22.56	1.21	20.26	1.09
k-NN M. cte. ($k = 6$ $d = 1$)	20.5	1.1	19.89	1.07
k-NN W. cte. ($k = 5$ $d = 1$)	16.51	0.89	19.55	1.05
k-NN M. inv. ($k = 6$ $d = 2$)	20.49	1.1	18.57	1
k-NN W. inv. ($k = 5$ $d = 7$)	17.8	0.95	17.4	0.94
AES Arit. ($\alpha = .08$)	17.64	0.95	**16.2**	**0.87**
AES Baric. ($\alpha = 1$)	19.04	1.02	18.69	1.01
AEEc Arit. ($\alpha = .51$ $\delta = .05$)	17.06	.92	19.27	1.04
AEEc Baric. ($\alpha = .99$ $\delta = .62$)	19.94	1.07	22.94	1.23
AEEh Arit. ($\alpha = .03$ $\delta = .01$)	16.94	.91	17.69	0.95
AEEh Baric. ($\alpha = .03$ $\delta = .01$)	17.47	0.94	17.99	0.97

5.8.4. Predicción de datos financieros intradiarios con agregación temporal

En los mercados financieros, las transacciones de compra-venta de acciones o de divisas marcan los precios de las mismas conforme a las leyes de la oferta y la demanda. Las series temporales resultantes muestran una serie de características que hace que sea complicado predecirlas empleando métodos estándar. Según Engle y Russell (2009), estas características son: espaciado temporal irregular, patrones de comportamiento diarios, precios discretos y dependencia compleja entre los valores de la serie. Engle y Russell (2009) proponen una serie de métodos para modelar datos intra-diarios.

Sin embargo, en lugar de trabajar con estas series financieras de alta frecuencia, lo habitual suele ser considerar únicamente los valores de cierre de cada sesión diaria o de cada semana. En esos casos, se está ignorando una gran cantidad de información que podría aprovecharse empleando STH. En una STH, los histogramas permiten realizar la agregación temporal de los valores de la serie intradiaria, presentando un resumen que informa de la distribución de los valores de la serie en cada uno de los periodos considerados (normalmente serán días, pero también podrían ser semanas o meses).

Puede argumentarse que el resumen que proporciona la STH ignora información contenida en la serie de alta frecuencia original, lo cual es cierto, ya que al agregar los valores intradiarios se pierde información sobre la secuenciación de los mismos. Sin embargo, se evitan también los problemas de las series intradiarias mencionados por Engle y Russell (2009).

En realidad, incluso asumiendo que se cuenta con métodos de predicción robustos, el pronóstico de toda la serie de valores intradiarios completa para el día siguiente sigue siendo una tarea titánica o, más bien, utópica. Por el contrario, el trabajar con una serie de valores agregados y generar predicciones para el día siguiente es notablemente más sencillo. El valor de cierre puede considerarse una magnitud agregada con la que resulta sencillo trabajar, pero no ofrece información sobre el comportamiento intradiario de la serie. Esta información sí es ofrecida por los histogramas, que describen la volatilidad de la acción en cada una de las sesiones.

En lugar de agregar las series temporales con los precios, se van a agregar las series temporales de los rendimientos. A continuación, se explica qué son los rendimientos y por qué son más adecuados que los precios.

Las series temporales de los rendimientos. En las series temporales de precios financieros los valores consecutivos suelen estar muy correlacionados y su varianza aumenta con el tiempo. Esto complica su estudio y hace que sea más conveniente analizar y predecir los cambios en los precios que trabajar con los propios precios (Aas y Dimakos, 2004). Para ello, la serie original se transforma en la serie de los rendimientos aritméticos o geométricos.

Los rendimientos aritméticos se obtienen como

$$q_t = \frac{y_t - y_{t-1}}{y_{t-1}} \tag{5.52}$$

donde y_t es el valor actual de la serie original y donde y_{t-1} es el valor inmediatamente anterior. Los rendimientos aritméticos cumplen lo siguiente:

- $q_t = +1.00 = +100\,\%$ cuando el valor de y_t es el doble que el de y_{t-1}
- $q_t > 0$ cuando se han obtenido beneficios respecto a y_{t-1}
- $q_t < 0$ cuando se han obtenido pérdidas respecto a y_{t-1}
- $q_t = -1.00 = -100\,\%$ cuando el valor y_t es cero, es decir, el activo que se está analizando ha perdido su valor por completo (se trata de un caso límite que no sucede en la práctica).

Por su parte, los rendimientos geométricos se definen como

$$g_t = \log(\frac{y_t}{y_{t-1}}) = \log(y_t) - \log(y_{t-1}). \tag{5.53}$$

Los rendimientos geométricos cumplen las siguientes propiedades.

- $g_t > 0$ cuando se han obtenido beneficios respecto a y_{t-1}
- $g_t < 0$ cuando se han obtenido pérdidas respecto a y_{t-1}

Al emplearse logaritmos, los rendimientos geométricos son más difíciles de interpretar, pero presentan otras propiedades que los hacen interesantes:

- Son simétricos, es decir, los rendimientos positivos y negativos de igual magnitud indican un cambio de igual proporción en la serie original. Esta propiedad no se cumple con los rendimientos aritméticos. Por ejemplo, supongamos que tenemos un euro invertido y que por él se obtiene un rendimiento geométrico del 0.50 y en el siguiente periodo otro del -0.50, tras ambos resultados el dinero invertido seguirá siendo un euro. Sin embargo, si los rendimientos son aritméticos el resultado final será 0.75 euros. La asimetría de los rendimientos aritméticos se hace más evidente cuando la magnitud de los rendimientos es grande.
- La suma de los rendimientos geométricos durante una serie de periodos sucesivos es igual al rendimiento geométrico obtenido considerando únicamente los periodos inicial y final considerados. Siguiendo con el ejemplo anterior, tras un rendimiento geométrico de 0.50 y otro de -0.50 obtenemos un rendimiento geométrico de 0. Sin embargo, no es posible expresar el rendimiento aritmético durante un periodo como función de los rendimientos aritméticos obtenidos durante los subperiodos que componen dicho periodo.

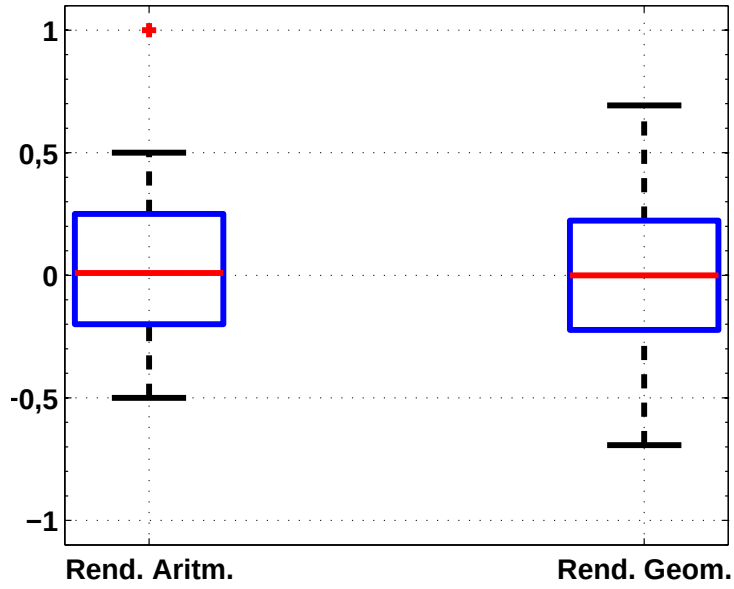


Figura 5.20: Rendimientos aritméticos (izqda.) y geométricos (dcha.) obtenidos a partir de la serie $\{X_t\}$.

A continuación se muestra un ejemplo para ilustrar el uso de los rendimientos empleando la agregación temporal. Sea la siguiente serie temporal que primero decrece de 50 en 50 y luego crece con el mismo ritmo hasta acabar en el mismo valor con que se inició $\{X_t\} = \{400, 350, 300, 250, 200, 150, 100, 50, 100, 150, 200, 250, 300, 350, 400\}$. Consideremos los rendimientos aritméticos y geométricos de esta serie y agreguémoslos por medio de un gráfico de cajas. El resultado que se muestra en la figura 5.20 indica que el *boxplot* de los rendimientos geométricos es perfectamente simétrico. Por su parte, el *boxplot* de los rendimientos aritméticos es asimétrico a la derecha, ya que tiene un valor extremo en el rango superior. Si se observa con mayor detenimiento se puede ver como la parte superior de la caja es ligeramente mayor que la parte inferior, cuando en el caso del *boxplot* de los rendimientos geométricos ambas partes son idénticas. Tal y como se ha indicado, para valores del rendimiento pequeños los rendimientos aritméticos son simétricos, pero para valores grandes no lo son. En los rendimientos financieros diarios o intradiarios las variaciones son normalmente pequeñas, pese a ello, parece adecuado evitar el sesgo que introduce el signo de la variación empleando los rendimientos geométricos. Por esta razón, para manejar datos agregados, tanto temporalmente como de forma contemporánea, se van a emplear rendimientos geométricos.

A continuación, se mostrarán dos ejemplos en los que se trabajará con las series temporales de rendimientos geométricos intradiarios agregados me-

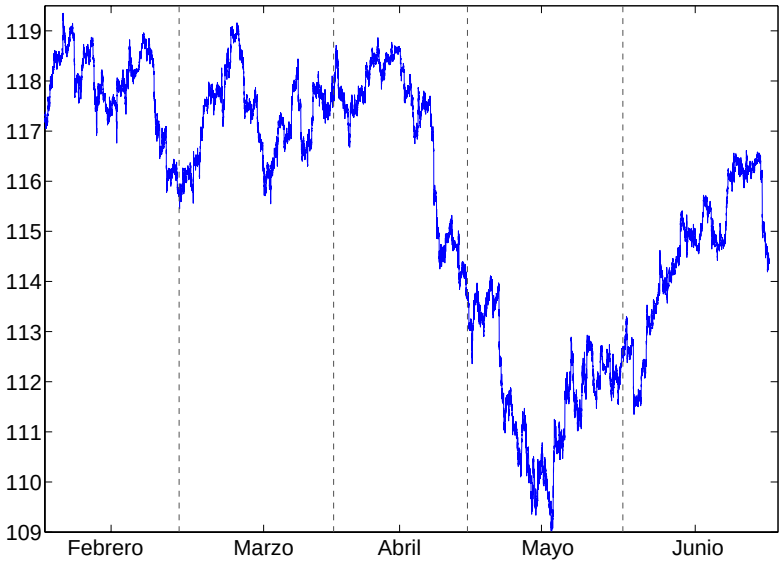


Figura 5.21: Series temporal de los precios intradiarios del cambio de divisa $\$ - ¥$ entre el 1-2-2006 y el 30-6-2006

diante histogramas. Las series temporales originales representan los precios intradiarios del cambio de divisas Dólar-Yen y Euro-Dólar. Para ambas series temporales se estudiará el periodo comprendido entre el 1 de Febrero de 2006 hasta el 30 de Junio de 2006. En dicho periodo hubo 107 días de negociación. Para cada día de negociación se cuentan con 288 valores intradiarios. De los 107 periodos de la serie, los 41 primeros periodos se utilizarán para inicialización, entre el 42 y el 85, ambos inclusive, para el entrenamiento y los periodos entre el 86 y el 107 se han empleado para la prueba.

5.8.4.1. La distribución del cambio Dólar-Yen intradiario en 2006

En primer lugar se analizará visualmente la serie de precios del cambio de divisas Dólar-Yen. La figura 5.21 muestra la serie de valores intradiarios durante dicho periodo. Se puede apreciar que desde febrero hasta mediados de abril la serie sufre una serie de oscilaciones en torno al 115 y al 120. Sin embargo, desde mediados de abril hasta mediados de mayo la serie sufre una caída de la que se recupera parcialmente durante el mes de junio.

Para eliminar los saltos que se producen en la serie y trabajar con una serie más sencilla de predecir, se ha optado por transformar la serie original en la serie de los rendimientos geométricos mediante la transformación mostrada en la ecuación (5.53). De esta forma, se elimina la tendencia de la serie.

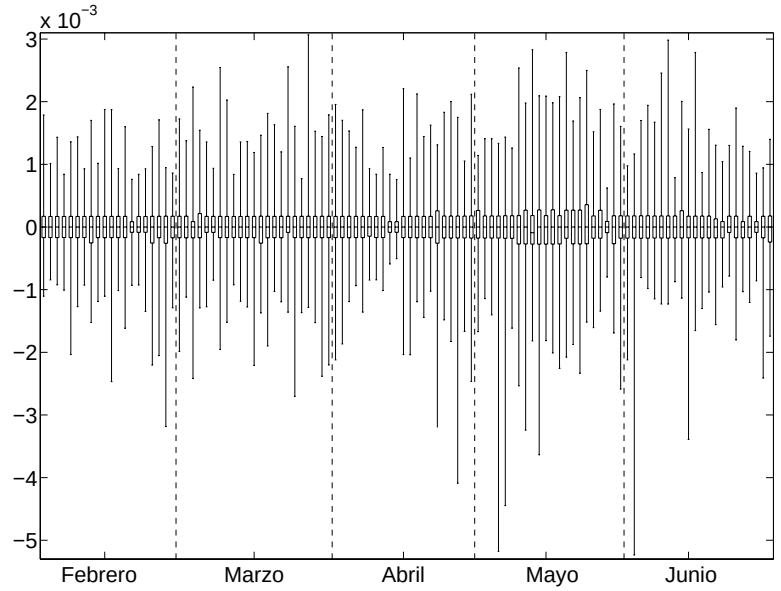


Figura 5.22: Serie temporal de *boxplots* de los rendimientos geométricos diarios del cambio de divisas $\$ - ¥$ entre el 1-2-2006 y el 30-6-2006

Es fácil entender que si la serie de precios presenta tendencia y cambios de nivel bruscos, como es el caso, el método de k-NN no obtendrá buenos resultados ya que no tendrá precedentes en el pasado de la serie para encontrar vecinos similares. Sin embargo, al utilizar la serie transformada tomando los rendimientos geométricos, la tendencia es eliminada y se obtiene una serie de media cero.

La serie de los rendimientos geométricos será agregada por medio de *boxplots* que resumirán los rendimientos de cada uno de los días. La STH resultante se muestra en la figura 5.22. Como ya se ha mencionado, los *boxplots* son histogramas equifrecuenciales de cuatro intervalos, donde cada intervalo tiene una frecuencia relativa asociada de 0.25. Tal y como muestra la figura 5.22, los *boxplots* dividen el conjunto de rendimientos en cuatro regiones, ofreciendo una representación visual muy intuitiva que describe la volatilidad de cada uno de los días de negociación.

En la STH resultante no hay atisbos de estacionalidad, ni de tendencia. Por ello, los métodos que se emplearán serán sin estacionalidad. Los resultados de los métodos empleados se muestran en la tabla 5.7. El error cometido por el método ingenuo en el entrenamiento se ha empleado como método de referencia. En el conjunto de entrenamiento, todos los métodos de predicción mejoran ampliamente al método ingenuo. En el conjunto de prueba las diferencias se estrechan, aunque todos los métodos mejoran en más de un 10 %

Tabla 5.7: Errores de predicción cometidos por los diferentes métodos en la STH de los rendimientos geométricos del cambio $\$ - ¥$

	Entrenamiento		Prueba	
Método	EMD_M	$EMED_M$	EMD_M	$EMED_M$
Ingenuo	$3.68{\cdot}10^{-4}$	1	$3.65{\cdot}10^{-4}$	0.99
k-NN M. cte. ($k = 8$ $d = 12$)	$2.48{\cdot}10^{-4}$	0.67	$3.15{\cdot}10^{-4}$	0.86
k-NN W. cte. ($k = 6$ $d = 3$)	$2.57{\cdot}10^{-4}$	0.7	$3.24{\cdot}10^{-4}$	0.88
k-NN M. inv. ($k = 8$ $d = 12$)	$2.48{\cdot}10^{-4}$	0.67	$3.15{\cdot}10^{-4}$	0.86
k-NN W. inv. ($k = 6$ $d = 12$)	$2.59{\cdot}10^{-4}$	0.7	$3.21{\cdot}10^{-4}$	0.87
AES Arit. ($\alpha = .06$)	$2.87{\cdot}10^{-4}$	0.78	$3.63{\cdot}10^{-4}$	0.99
AES Baric. ($\alpha = .04$)	$2.47{\cdot}10^{-4}$	0.67	$\mathbf{3.07{\cdot}10^{-4}}$	**0.83**

al método ingenuo. La única excepción es el alisado exponencial basado en la aritmética de histogramas que obtiene unos resultados muy similares al método ingenuo. Curiosamente, de entre el resto de métodos el que mejor predice es el alisado exponencial basado en el método de los baricentros.

5.8.4.2. La distribución del cambio Euro-Dólar intradiario en 2006

En este ejemplo se aborda la predicción de la serie de precios del cambio de divisas Euro-Dólar. En la figura 5.23 se muestra la serie de valores intradiarios durante dicho periodo. Durante los meses de marzo a mediados de mayo, el Euro muestra gran fortaleza respecto al Dólar y el cambio Euro-Dólar inicia una tendencia ascendente que se estabiliza a mediados de mayo. Por otra parte, a principios de junio el cambio de estas divisas sufre una acusada caída que es corregida en parte a finales de mes.

Con el fin de hacer que la serie sea estacionaria en media, la serie de los precios será transformada en la serie de los rendimientos geométricos. Sobre la serie de los rendimientos geométricos se realizará la agregación temporal para dar lugar a la STH. En este caso, los histogramas que se emplearán son los *boxplots*. En la figura 5.24 se muestra la STH resultante, donde puede puede apreciarse que la serie es estacionaria en media. Los *boxplots* permiten identificar muy claramente los días con mayor volatilidad.

En la tabla 5.8 se muestran los errores de predicción cometidos por los distintos métodos empleados. Para obtener el error escalado $EMED_M$ se utiliza como error de referencia el cometido en el entrenamiento por el método ingenuo. Todos los métodos empleados obtienen mejores resultados que el método ingenuo en el periodo de entrenamiento y en el de test. De ellos, el que peor funciona es el alisado exponencial simple basado en aritmética de histogramas y los que mejor los k-NNs basados en la distancia de Mallows. Tal y como muestra la figura 5.24, la variabilidad de la serie aumenta en el periodo de prueba. Esto hace que el método ingenuo empeore notablemente su rendimiento en dicho periodo. Pese a ello, los k-NNs mejoran en similar proporción los resultados del método ingenuo en dicho periodo.

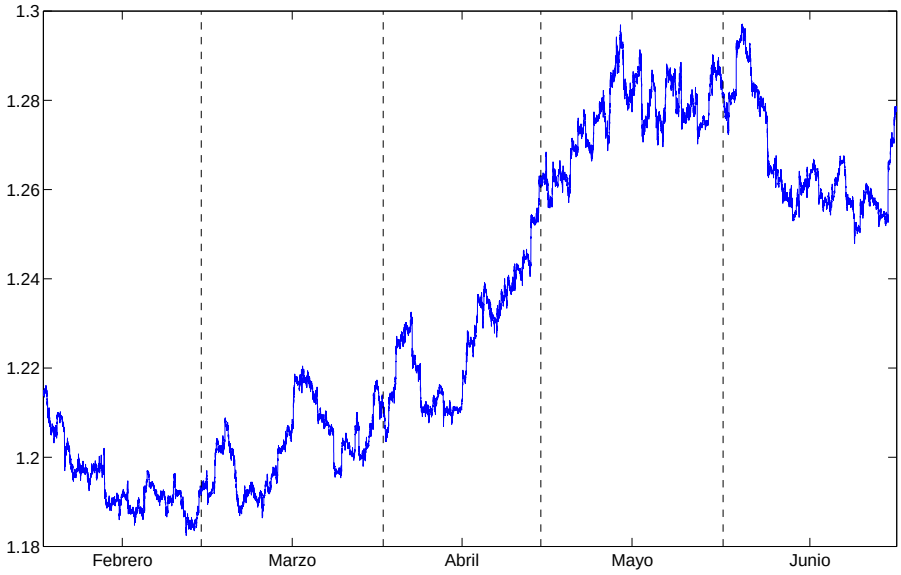


Figura 5.23: Series temporal de los precios intradiarios del cambio de divisa € – $ desde el 1-2-2006 y el 30-6-2006

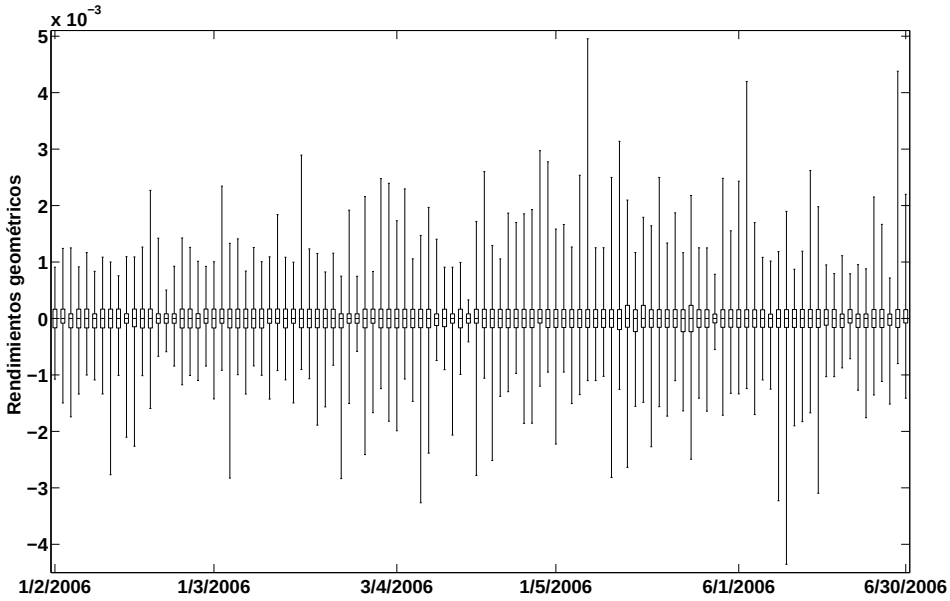


Figura 5.24: Serie temporal de *boxplots* de los rendimientos geométricos diarios del cambio de divisas € – $ entre el 1-2-2006 y el 30-6-2006

Tabla 5.8: Errores de predicción cometidos por los diferentes métodos en la STH de los rendimientos geométricos del cambio $€ - \$$

	Entrenamiento		Prueba	
Método	EMD_M	$EMED_M$	EMD_M	$EMED_M$
Ingenuo	$3.33{\cdot}10^{-4}$	1	$3.79{\cdot}10^{-4}$	1.14
k-NN M. cte. ($k = 10$ $d = 17$)	$2.52{\cdot}10^{-4}$	0.76	$\mathbf{2.88{\cdot}10^{-4}}$	**0.87**
k-NN W. cte. ($k = 6$ $d = 1$)	$2.43{\cdot}10^{-4}$	0.73	$3.09{\cdot}10^{-4}$	0.93
k-NN M. inv. ($k = 10$ $d = 17$)	$2.52{\cdot}10^{-4}$	0.76	$\mathbf{2.88{\cdot}10^{-4}}$	**0.87**
k-NN W. inv. ($k = 7$ $d = 1$)	$2.42{\cdot}10^{-4}$	0.73	$3.18{\cdot}10^{-4}$	0.96
AES Arit. ($\alpha = .12$)	$2.67{\cdot}10^{-4}$	0.8	$3.53{\cdot}10^{-4}$	1.06
AES Baric. ($\alpha = .06$)	$2.44{\cdot}10^{-4}$	0.73	$2.98{\cdot}10^{-4}$	0.9

5.8.5. Predicción de datos financieros resumidos mediante agregación contemporánea

Los datos financieros ofrecen una gran cantidad de oportunidades de aplicar metodologías de predicción simbólicas. La agregación temporal de datos intradiarios empleando histogramas es una posibilidad que ha sido explorada en el apartado anterior. Otra opción interesante consiste en realizar la agregación contemporánea de los valores de las cotizaciones de las acciones que constituyen un determinado índice bursátil a lo largo del tiempo. De esta forma se obtiene una representación original del comportamiento del índice en cuestión que sirve para conocer cómo ha evolucionado la distribución de las acciones en el tiempo.

Esta enfoque es planteado en el trabajo realizado por González-Rivera et al. (2008) donde se muestra un STH que representa la distribución de los rendimientos semanales de las 500 acciones que integran el índice bursátil del Standard & Poor's. En dicho trabajo, el interés se centra en clasificar las 500 acciones del índice en un ranking de acuerdo a sus rendimientos semanales y en analizar los saltos que dan las acciones entre los rankings de dos semanas consecutivas. Tomando como punto de partida dicho trabajo, a continuación se plantea la predicción de la STH resultante de agregar los rendimientos geométricos semanales de las acciones del IBEX-35.

5.8.5.1. Predicción de la distribución de los rendimientos de las acciones del IBEX-35

El periodo que se analizará es el comprendido entre el 1 de septiembre y el 31 de diciembre de 2006. Las acciones que integraban el IBEX durante aquel periodo son mostradas en la tabla 5.9. Como los precios de estas acciones tienen magnitudes muy dispares, no es conveniente agregarlos en bruto, sino que es mejor transformar las series y agregar las series transformadas. Para ello, se han convertido las series de precios diarios en las series de los rendimientos geométricos diarios.

Tabla 5.9: Acciones que constituían el IBEX-35 entre septiembre y diciembre de 2006 con su respectiva capitalización bursatil en miles de millones de euros

Acción	Capitalización	Acción	Capitalización
BSCH	83.370	FCC	8.983
Telefónica	74.949	Acciona	8.884
BBVA	63.985	Cintra	5.992
Endesa	37.406	Telecinco	5.103
Iberdrola	32.546	Acerinox	4.783
Repsol	32.536	Bankinter	4.609
Inditex	23.313	Enagas	4.519
Banco Popular	16.591	Gamesa	4.333
Gas Natural	14.083	Mapfre	4.157
ACS	13.956	REE	4.143
Metrovacesa	12.316	Fadesa	3.983
Abertis	12.150	A3 TV	3.744
Union Fenosa	12.141	Sogecable	3.439
Sacyr Vallehermoso	11.755	Prisa	2.816
Banesto	11.186	Indra	2.534
Ferrovial	10.387	Iberia	2.275
Altadis	9.930	NH Hoteles	1.987
Banco Sabadell	9.269	-	-

El conjunto de las 35 series de los rendimientos geométricos son agregadas para dar lugar a una STH. En dicha STH, los histogramas se construirán a partir de una partición del espacio de frecuencias. Los cuantiles que han sido elegidos para formar la partición son los siguientes $\{0, 10, 40, 60, 90, 100\}$. De esta forma, se separan los valores que se encuentran en los extremos, y el 20 % central de la distribución. La imagen 5.25 muestra la STH resultante.

En la tabla se muestran los resultados obtenidos por los métodos de predicción analizados 5.10. El método ingenuo se ha empleado como método de referencia. El resto de métodos utilizados superan al ingenuo ampliamente tanto en el periodo de entrenamiento, como en el de prueba. Los resultados del método ingenuo indican que la serie se vuelve más difícil de predecir en el conjunto de prueba. De hecho, aunque no quede reflejado en la figura 5.25, en el mes de diciembre la variabilidad de la serie es mayor. El método que mejor funciona es el k-NN basado en la distancia de Wasserstein con pesos inversamente proporcionales a la distancia, aunque el esquema de ponderación que asigna el mismo peso a todos los vecinos obtiene prácticamente el mismo resultado. Resulta interesante comprobar que el k-NN basado en la distancia de Mallows, aunque supera al ingenuo, obtiene peores resultados que el k-NN basado en Wasserstein. Una posible explicación es que el k-NN basado en Mallows al obtener las predicciones como una media de los histo-

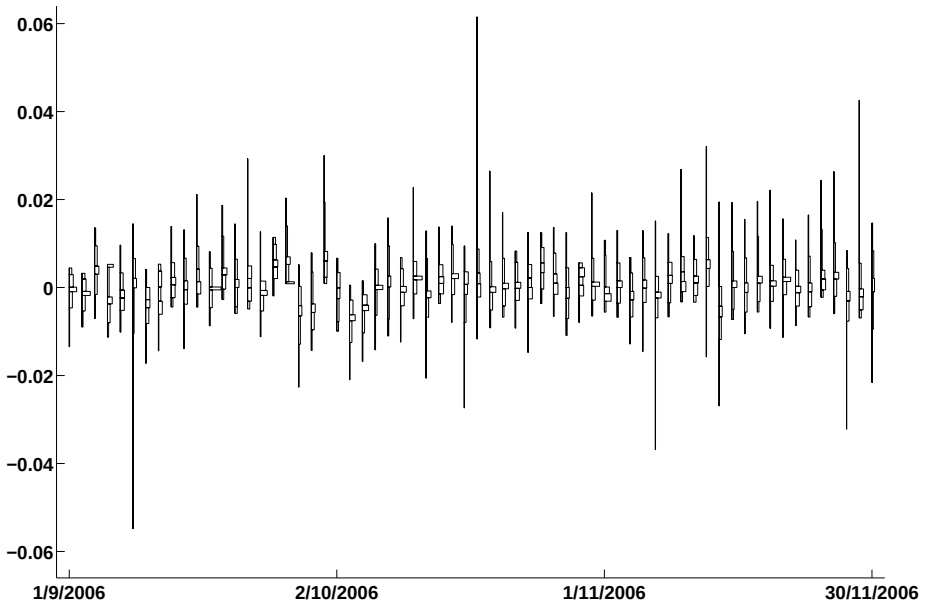


Figura 5.25: Serie temporal de histogramas que representan los rendimientos geométricos diarios obtenidos por las acciones constituyentes del IBEX-35 entre el 1-9-2006 y el 30-11-2006

Tabla 5.10: Errores de predicción cometidos por los diferentes métodos en la STH de los rendimientos geométricos de las acciones del IBEX-35

	Entrenamiento		Prueba	
Método	EMD_M	$EMED_M$	EMD_M	$EMED_M$
Ingenuo	$0.42{\cdot}10^{-2}$	1	$0.51{\cdot}10^{-2}$	1.23
k-NN M. cte. ($k = 10$ $d = 10$)	$0.29{\cdot}10^{-2}$	0.68	$0.45{\cdot}10^{-2}$	1.07
k-NN W. cte. ($k = 10$ $d = 1$)	$0.28{\cdot}10^{-2}$	0.67	$0.41{\cdot}10^{-2}$	0.98
k-NN M. inv. ($k = 10$ $d = 10$)	$0.29{\cdot}10^{-2}$	0.68	$0.45{\cdot}10^{-2}$	1.07
k-NN W. inv. ($k = 9$ $d = 1$)	$0.29 \cdot 10^{-2}$	0.68	$\mathbf{0.41{\cdot}10^{-2}}$	**0.97**
AES Arit. ($\alpha = .12$)	$0.28{\cdot}10^{-2}$	0.68	$0.42{\cdot}10^{-2}$	1
AES Baric. ($\alpha = .02$)	$0.29{\cdot}10^{-2}$	0.68	$0.42{\cdot}10^{-2}$	1

gramas se ve más afectado por los histogramas de comportamiento extremo que se dan en la última parte del conjunto de prueba. Por el contrario, el k-NN basado en Wasserstein, al generar las predicciones como una mediana de los histogramas, no se ve afectado por este problema. Por otro lado, los alisados exponenciales simples obtienen unos resultados similares a los que obtiene el k-NN basado en Wasserstein.

5.9. Conclusiones

En este capítulo se ha propuesto el uso de las STH como una herramienta para representar series temporales de distribuciones y se han propuesto diferentes métodos de predicción para este tipo de series. El trabajo presentado en el capítulo es pionero en el área de predicción, ya que no se tiene conocimiento de otros desarrollos que permitan predecir series temporales de distribuciones, ni en forma de histograma, ni utilizando otro estimador.

En el apartado 5.3, se ha defendido el histograma como herramienta de representación de una distribución. La representación como conjunto de intervalos con pesos asociados que caracteriza al histograma es lo suficientemente versátil como para permitir un abanico de representaciones que se adaptan a las distintas necesidades que se le pueden plantear a un analista ante distintos conjuntos de datos. En el apartado 5.8 se han mostrado ejemplos que han ilustrado el uso de los distintos tipos de histogramas: equiespaciados, equifrecuenciales, construidos sobre una partición en el espacio de la variable y construidos sobre una partición del espacio de frecuencias.

La complejidad que implica el manejo de histogramas con respecto a la que requería el manejo de intervalos es notable. Esto hace que sea más complicado definir conceptos y métodos para trabajar con las STH. Para definir medidas de error se ha decidido utilizar el concepto de distancia, tal y como se muestra en el apartado 5.4. Para desarrollar métodos de predicción se han propuesto dos aproximaciones la aritmética de histogramas y el método de los baricentros. Los métodos de predicción propuestos han sido los alisados exponenciales (en el apartado 5.5) y el método de los k vecinos más cercanos (en el apartado 5.6).

Todos los métodos propuestos, aunque están basados en ideas sencillas, obtienen habitualmente buenos resultados cuando se aplican a la predicción de series temporales clásicas. En el apartado 5.8, se ha mostrado que sus homólogos para STH también son capaces de obtener buenos resultados ante series de distinto tipo y origen, batiendo en todos los casos al método ingenuo que ha sido tomado como referencia. Más complicado resulta extraer conclusiones sobre la superioridad o inferioridad de uno u otro modelo, porque todos han sido capaces de demostrar su buen comportamiento en al menos una serie temporal de las analizadas. Por ello, todas estas técnicas pueden ser consideradas como posibles candidatas para la predicción de nuevas STH.

Capítulo 6

Conclusiones y Trabajo Futuro

Una conclusión es el punto
donde te cansas de pensar.

Arthur Bloch, La ley de Murphy

En este capítulo se resumirán las contribuciones de esta tesis en el campo de la predicción de series temporales de intervalos e histogramas. En el caso de las series temporales de histogramas, el propio tipo de serie temporal es en sí mismo una aportación novedosa. En el capítulo se expondrán brevemente las conclusiones obtenidas en la aplicación práctica de los métodos propuestos. Por último, se enunciarán posibles líneas de trabajo futuro, no sólo a nivel teórico, sino también a nivel práctico.

6.1. Conclusiones

Esta tesis supone una aportación novedosa dentro de las áreas del análisis de datos simbólicos y de la predicción de series temporales. En el año 2004, año en el que se planteó esta tesis, no existía ningún método que abordase la predicción de series temporales de intervalo o de histograma. Cuatro años después, la situación ha empezado a cambiar con la aparición de los trabajos de Teles y Brito (2005) y de Maia et al. (2006a) en series temporales de intervalos. Esta tesis supone una contribución más que sirve para consolidar el área.

La tesis ha supuesto un esfuerzo multidisciplinar, no sólo por combinar dos áreas como la predicción de series temporales y el análisis de datos simbólicos, sino también porque en su desarrollo se han tocado con mayor o menor profundidad temas relativos a ámbitos tan variados como son el estudio de las distancias (tanto de intervalos como de histogramas), la visión artificial (al emplear la *Earth Mover's Distance* para trabajar con histogramas), las series temporales caóticas (al trabajar con el k-NN), la estimación

de densidades (al trabajar con histogramas), la aritmética de intervalos y de funciones de densidad (para realizar operaciones), la econometría (al utilizar modelos econométricos como el VAR o el VECM y al trabajar con series financieras), la meteorología y el medioambiente (como ámbitos de aplicación), etc.

A continuación, se resumirán las aportaciones más relevantes de esta tesis, los artículos y contribuciones a congresos a los que ha dado lugar y otros frutos originados durante el desarrollo de la misma.

6.1.1. Aportaciones de la tesis

6.1.1.1. Series temporales de intervalos

Las series temporales de intervalos (STI) aparecen habitualmente en contextos tales como la meteorología y las finanzas, para representar los rangos de las temperaturas y de las cotizaciones, respectivamente. El interés por el análisis de los intervalos en el ámbito financiero ha despertado en los últimos años, tal y como ha sido documentado en los apartados 3.7.1 y 4.5.2.2. Sin embargo, las STI no han sido identificadas o tratadas como tales, salvo en Teles y Brito (2005), Maia et al. (2006a) y en esta tesis. El trabajo aquí presentado ha abordado el tema en profundidad y ha servido para consolidar las STI. Sus aportaciones más relevantes han sido las siguientes:

- Propuesta de dos enfoques para medir el error en las STI.
  - En los componentes de la STI (valores mínimo, máximo, centro y radio) utilizando medidas escaladas.
  - En el intervalo como un todo utilizando distancias.
- Desarrollo de distintos métodos de predicción de STI.
  - Alisados exponenciales utilizando aritmética de intervalos.
  - Método de k-NN.
  - Perceptrón multicapa (iMLP) basado en aritmética de intervalos.
  - Predecir una STI a partir de las series temporales de sus componentes (mínimo, máximo, centro y radio) aplicando para ello métodos de predicción (univariantes o multivariantes) para series temporales clásicas. Más concretamente, ha propuesto predecir las series temporales de los mínimos y de los máximos o, alternativamente de los centros y los radios.

Las conclusiones más relevantes obtenidas a lo largo de la investigación son las siguientes

- El intervalo permite representar la variabilidad de un conjunto de observaciones en forma de rango o de rango intercuartílico ofreciendo información complementaria a la que se obtiene con otros valores agregados como, por ejemplo, la media.
- El manejo de intervalos supone una mayor complejidad que el manejo del valores reales. Sin embargo, un intervalo puede descomponerse cómodamente en una pareja de reales (mínimo y máximo, o centro y radio), lo que facilita enormemente su manejo.
- El concepto de error en las STI no puede ser representado mediante la aritmética de intervalos. Para medir el error, en la tesis se propone el uso de distancias para intervalos o medir el error escalado en cada una de las series de los componentes de los intervalos. El primer enfoque permite medir el error con un único valor, mientras que el segundo permite conocer con más detalle en qué componente se está cometiendo más error.
- Se ha demostrado que las predicciones que se obtienen al modelar la STI como un VAR de orden p sobre las series del mínimo y del máximo o sobre las del centro y del radio son equivalentes.
- En los ejemplos planteados, las series temporales de los mínimos y de los máximos estaban cointegradas. Sin embargo, al recoger dicha relación de cointegración mediante un modelo VECM no se han obtenido mejoras con respecto a la predicción que se obtenía con un modelo VAR que ignorase dicha relación.
- En los ejemplos analizados, todas las aproximaciones para predecir STI presentadas en esta tesis mejoran los resultados obtenidos por el método ingenuo. De entre ellas, el enfoque de predicción univariante que predice de forma independiente el centro y el radio ha resultado ser de las que mejores resultados han obtenido en todos los ejemplos.
- En los ejemplos analizados, los métodos que trabajan con el intervalo como un todo, i.e., los alisados, el iMLP y el k-NN para STI mejoran las predicciones que obtiene el método ingenuo. Además, se ha observado que las predicciones que obtienen estos métodos en los extremos de los intervalos son mejores que las que se obtienen utilizando sus homólogos univariantes para pronosticar las series de los extremos de forma independiente.

6.1.1.2. Series temporales de histogramas

En el campo de las series temporales de histogramas (STH), esta tesis ha explorado un camino muy innovador, ya que no se conocía ningún artículo

previo que tratase este tipo de series. Por tanto, la tesis ha abierto camino definiendo las STH y motivando debidamente su uso para resolver situaciones que no pueden ser abordadas de otra manera, si no es perdiendo información.

Las contribuciones más relevantes relativas a las STH han sido:

- Propuesta de una aproximación para medir el error en las STH basada en el uso de distancias para funciones de densidad que reflejen adecuadamente las diferencias entre histogramas, como las distancias de Wasserstein y de Mallows.
- Desarrollo de métodos de predicción de STH.
  - Alisados exponenciales utilizando aritmética de histogramas.
  - Alisados exponenciales utilizando el baricentro de Mallows.
  - Método de k-NN utilizando el baricentro de Wasserstein y de Mallows.

Las principales conclusiones que se han obtenido en el desarrollo de la investigación son las siguientes:

- El histograma, en sus diferentes versiones, es una herramienta muy versátil a la hora de representar distribuciones ya que es capaz de adaptarse a las distintas necesidades del analista.
- La complejidad que supone el manejo de histogramas es mayor que la que supone el manejo de intervalos, pero es mucho menor que la que implicaría el manejo de otras herramientas de representación de distribuciones.
- Las distancias de Wasserstein y de Mallows reflejan la semejanza entre histogramas de una forma bastante similar a como la hace el ojo humano y, por tanto, son adecuadas para medir errores en las STH.
- El histograma baricéntrico permiten obtener, con determinadas distancias, un histograma que represente la tendencia central de un conjunto de histogramas. Si se emplea la distancia de Mallows se obtiene un histograma que se comporta como el promedio del conjunto de histogramas resultantes, mientras que si se emplea la distancia de Wasserstein se obtiene un histograma que se comporta como la mediana de dicho conjunto.
- Los alisados exponenciales basados en la aritmética de histogramas presentan el problema de que el promedio con dicha aritmética no se comporta exactamente como requieren los métodos de alisado. Además, normalmente no se podrá asumir la independencia de los operandos y tener en cuenta la dependencia entre los operandos convierte el problema en intratable.

- Los alisados exponenciales basados en el baricentro que emplea la distancia de Mallows no presentan los inconvenientes del alisado basado en aritmética de histogramas y reflejan el concepto de promedio tal y como cabe esperar. Por su parte, los baricentros que se obtienen con la distancia de Wasserstein no son adecuados para realizar alisados exponenciales, porque su comportamiento es similar al de una mediana y eso desvirtúa el proceso de alisado.

- El método de k-NN puede adaptarse correctamente a la predicción de STH utilizando las distancias de Mallows o de Wasserstein para buscar los vecinos más próximos y para obtener las predicciones como histogramas baricéntricos. El uso de una distancia u otra hace que el método tenga un comportamiento ligeramente diferente. La principal diferencia se da a la hora de componer la predicción, ya que con Mallows se obtienen histogramas promedios y con Wasserstein histogramas medianos.

- Tal y como se ha mostrado en los ejemplos, la capacidad predictiva de los métodos presentados es notable. En todos los ejemplos analizados, al menos uno de los métodos propuestos ha obtenido mejores resultados que el método de referencia (método ingenuo). El orden de dichas mejoras varía entre el 10 % y el 25 %, según los ejemplos.

6.1.2. Artículos publicados

El trabajo desarrollado a lo largo de estos años ha sido presentado en diversos foros relacionados con la predicción, la estadística y la inteligencia computacional. A continuación, se citan las publicaciones en congresos:

- *Forecasting time series of observed distributions with smoothing methods based on the barycentric histogram.* Escrito con Carlos Maté y publicado en el libro de actas de la *International FLINS Conference on Computational Intelligence in Decision and Control* editado por *World Scientific* en 2008.

- *Forecasting histogram time series with k-Nearest Neighbours methods.* Presentado en el *International Symposium on Forecasting* en 2007 y galardonado con un premio de una cuantía de 1000$ como uno de los mejores trabajos de estudiantes de doctorado.

- *Exponential smoothing methods for interval time series.* Escrito con Antonio Muñoz San Roque, Carlos Maté y Ángel Sarabia, y publicado en el libro de actas del *European Symposium on Time Series Prediction* editado por *Helsinki University of Technology* en 2007.

- *Introducing interval time series: accuracy measures.* Escrito con Carlos Maté y publicado en el libro de actas de la conferencia de la *International Association for Statistical Computing (COMPSTAT)* y editado por *Physica-Verlag* en 2006.

- *Smoothing methods for histogram-valued time series.* Escrito con Carlos Maté, Antonio Muñoz San Roque, y Ángel Sarabia, y presentado en el *International Symposium on Forecasting* celebrado en 2006.

Además, se ha publicado algunas parte del contenido de la tesis en revistas científicas internacionales:

- *Forecasting histogram time series with k-nearest neighbours methods.* Escrito por Javier Arroyo y Carlos Maté y pendiente de publicación en la revista *International Journal of Forecasting* (Volumen 24, número 4) que tiene un factor de impacto de 1.409 (*Journal Citation Report* 2007).

- *iMLP: Applying Multi-Layer Perceptrons to Interval-Valued Data.* Escrito (por orden de firma) por Antonio Muñoz San Roque, Carlos Maté, Javier Arroyo y Ángel Sarabia y publicado en la revista *Neural Processing Letters* (Volumen 25, número 2, pags. 157-169) que tiene un factor de impacto de 0.580 (*Journal Citation Report* 2007).

Actualmente, se está trabajando en la publicación de algunas aportaciones de la tesis en revistas internacionales, a ser posible con índice de impacto JCR (*Journal Citation Report*). El objetivo es publicar un artículo que resuma las contribuciones del capítulo de series temporales de intervalos y dos artículos que aborden los alisados exponenciales en series temporales de histogramas (uno mediante la aritmética de histogramas y otro mediante los baricentros).

6.1.3. Otros aspectos a mencionar

La tesis se ha enmarcado dentro del proyecto PRESIM (*Modelos de PREdicción para Datos SIMbólicos*) dirigido por el profesor Carlos Maté, que también ha dirigido esta tesis doctoral. El proyecto PRESIM es un proyecto de investigación financiado por la Universidad Pontificia Comillas en el que colaboran profesores de dicha universidad, como Antonio Muñoz San Roque y Ángel Sarabia, que comenzó en octubre de 2005 y que finalizará en septiembre de 2009. El objetivo del proyecto era el desarrollo de métodos de predicción para datos simbólicos.

Tanto el proyecto, como la tesis tenían también entre sus objetivos el de la difusión de la predicción con datos simbólicos. A ese respecto, además de la publicación de los artículos de investigación, se ha establecido contacto con

profesores de otras universidades que han manifestado, de una u otra manera, su interés por el trabajo desarrollado en el proyecto y que, en algunos casos, han llegado a colaborar con nosotros. Entre los profesores con los que se ha contactado se incluyen Gloria González Rivera (Universidad de California Riverside), Paula Brito (Universidad de Oporto), Robert Fildes (Universidad de Lancaster) y Kenneth F. Wallis (Universidad de Warwick).

Por último, mencionar que los métodos de predicción específicos de STI y de STH han sido implementados en MATLAB y que pueden ser solicitados para su uso académico.

6.2. Líneas de trabajo futuro

Las series temporales de intervalos y de histogramas acaban de nacer, por tanto, las posibilidades de crecimiento del área son enormes. Para ello, uno de los aspectos fundamentales reside en difundir su potencial y su utilidad, y en atraer hacia ella gente con capacidad investigadora e interés por explotar dicho potencial. Si eso se consigue, la supervivencia de este área de investigación está asegurada.

A continuación, se presentarán algunas líneas de trabajo futuro que sirven de continuación a la investigación realizada en esta tesis. Se distinguirá entre las líneas de trabajo teóricas y las que tienen un carácter más aplicado.

6.2.1. Líneas de trabajo a nivel teórico y metodológico

De cara a proponer nuevos métodos se debe tener como referencia el trabajo ya realizado en las áreas de la predicción de series temporales y en el análisis de datos simbólicos e intentar tender puentes entre ambas. Al ser las STI y las STH un campo novedoso, las posibilidades de desarrollo en el plano teórico son enormes. La siguiente lista muestra sólo algunos de los posibles caminos:

- Es necesaria la creación de conceptos que permitan describir el comportamiento de las STI y de las STH. En las series temporales clásicas existen conceptos tales como, por ejemplo, la tendencia, la estacionalidad y la componente cíclica, que caracterizan dichas series. En esta tesis se han definido algunos de estos conceptos al proponer los métodos de alisado; por ejemplo, la tendencia se definió como el cambio a largo plazo en el centro de gravedad del intervalo o del histograma. Sin embargo, es posible plantear otras definiciones de tendencia que, por ejemplo, tengan en cuenta la evolución a largo plazo de otras característica como el ancho de los intervalos en las STI o los cambios de forma en la distribución de los histogramas en las STH. Por ello, es necesario realizar un estudio más profundo sobre el tema.

- También es necesario desarrollar una teoría sobre los procesos estocásticos simbólicos, definir conceptos relacionados con dichos procesos como la estacionariedad y analizar qué relación existe entre procesos estacionarios y el modelado de los mismos con determinados métodos. En el caso de las STI se puede trabajar con la estacionariedad de las series temporales individuales, pero hay muchas preguntas por responder al respecto. Por ejemplo, ¿qué efecto tiene dicha estacionariedad a la hora de modelar la serie mediante mediante modelos como el modelo ARIMA propuesto por Teles y Brito (2005)? ¿qué relación existe entre la estacionariedad de las series del centro y del radio y la de las series del mínimo y del máximo?, etc.

- En la tesis se ha planteado el uso de distancias para medir el error en STI y en STH. En los dos tipos de series, la aritmética no era una altertiva adecuada. Sin embargo, es posible que existan otras aproximaciones para medir el error en dichas series temporales sin necesidad de usar las distancias. De hecho, para las STI también se ha propuesto medir el error escalado en las componentes individuales. Otra línea de trabajo consiste, por tanto, en estudiar otras formas de medir el error en las series.

- Es obvio que el catálogo de métodos de predicción para STI y STH debe ser ampliado. Hay tres vías de ampliación:

  - Seguir refinando los métodos de predicción planteados en esta tesis para hacerlos más efectivos. Los métodos planteados admiten mejoras, por ejemplo, los alisados pueden ser extendidos o replanteados para modelar otras componentes de las series consideradas, es posible plantear variantes del k-NN que usen otros esquemas de ponderación o nuevas formas de medir la similitud para hacerlos efectivos a la hora de recoger la tendencia, etc. También cabe la posibilidad de plantear otras formas de realizar los alisados o el k-NN distintas a las presentados en esta tesis.
  - Adaptar al contexto simbólico métodos ya existentes en el contexto clásico y no adaptados en esta tesis. Un candidato obvio debido a su popularidad en el contexto clásico son los modelos autorregresivos.
    - Autorregresión de STI: Actualmente, existen dos propuestas de autorregresión para STI (Teles y Brito, 2005; Maia et al., 2006a). En esta tesis se han explorado otras alternativas para realizar la autorregresión: la que consiste en estimar un modelo autorregresivo para cada una de las series de los extremos y la que emplea modelos vectoriales autorregresivos (VAR y VECM). Sin embargo, pueden plantearse nuevos modelos.

Por ejemplo, los métodos de regresión revisados en el apartado 2.5.1 pueden ser adaptados al contexto temporal y puede compararse su eficacia con los ya desarrollados. También sería interesante definir un concepto de correlación para este tipo de series que permitiese orientar al analista para determinar cómo debe ser el modelo autorregresivo, de forma similar a como sucede en la metodología Box-Jenkins.

- Autorregresión de STH: La adaptación del enfoque autorregresivo a la predicción de STH es más compleja ya que, por el momento, no existe un modelo de regresión que permita relacionar variables de entrada y de salida en forma de histograma.

- Desarrollar métodos de predicción innovadores que no tengan un equivalente en el contexto clásico y que exploten las posibilidades y las peculiaridades propias de las STI y de las STH. Indudablemente, esta vía es la que más imaginación requiere ya que implica crear métodos partiendo desde cero.

- La combinación de predicciones es otro posible campo de investigación. En esta tesis se han planteado distintos métodos de predicción, pero no se ha estudiado si la combinación de las predicciones de dichos modelos produce o no una mejora en la precisión. En las series temporales clásicas, existe evidencia empírica de que a menudo la combinación de predicciones es más precisa que cada una de las predicciones sin combinar. De hecho, algo tan simple como el promedio de predicciones a menudo se presenta como una buena estrategia de combinación (Winkler y Makridakis, 1983). Un estudio en esa línea para STI y STH también tendría interés y sería de esperar que la combinación obtuviese buenos resultados, ya que puede explotar la forma notablemente distinta de obtener predicciones que tienen métodos como, por ejemplo, el k-NN y el alisado exponencial.

- Otra línea de trabajo más alejada del enfoque de las STI y de las STH, consiste en utilizar la aritmética de intervalos y de histogramas revisada en la tesis, para incorporar intervalos y densidades de predicción como variables externas en modelos de predicción clásicos como la función de transferencia o el perceptrón multicapa. Esto permitiría comprobar el efecto de la incertidumbre en la salida del modelo.

6.2.2. Líneas de trabajo a nivel aplicado

Los métodos de predicción propuestos en esta tesis han sido probados sobre ejemplos de diferentes ámbitos y en todos ellos se han obtenido resultados satisfactorios. La aplicación de los métodos se ha hecho con el objetivo

principal de demostrar la capacidad predictiva de éstos y no con el fin de resolver un problema concreto o de obtener un conocimiento más profundo de un determinado fenómeno.

Sin embargo, hay ámbitos donde la aplicación de las series temporales de intervalo o de histogramas pueden ser de gran utilidad. Algunos de ellos han sido trabajados en la tesis, pero merecen una mayor atención. En el caso de las series temporales de intervalos, los ámbitos de aplicación donde es más interesante incidir son los siguientes:

- En meteorología, un posible campo de investigación consiste en comparar las predicciones del intervalo de temperaturas obtenidas a partir de la STI y de los modelos de predicción meteorológica, y en estudiar si las combinaciones de dichas predicciones mejoran los modelos originales o no.

- En el apartado 4.10, se predicen los valores futuros de STI que reflejan los valores mínimos y máximos diarios del cambio de divisas. Los resultados que se obtienen son satisfactorios, pero el objetivo en ese caso ha sido batir al método ingenuo. Sería interesante profundizar en este área para determinar si es posible obtener beneficios empleando dichas predicciones como base para generar reglas de negociación (*trading rules*) de compra-venta de valores y comparar los resultados con otras estrategias basadas únicamente en las predicciones de los valores de cierre, como las mostradas en Andrada-Félix, Fernadez-Rodriguez, Garcia-Artiles y Sosvilla-Rivero (2003). Otra alternativa es ver la utilidad de la predicción del intervalo o de la distribución de los precios intradiarios de una acción o de una divisa en el área de la gestión de riesgos.

- Siguiendo con el ámbito de las finanzas, otra posible aplicación de las STI reside en la predicción de gráficos de velas (o *candlesticks*). Los *candlesticks* ofrecen una buena síntesis de la información intradiaria (o intraperiodo) mediante dos intervalos: apertura-cierre y mínimo-máximo. Por ello, pueden ser pronosticados con los métodos planteados en esta tesis para STI o con extensiones de los mismos y comparar sus resultados con el de los métodos ya planteados (ver apartado 3.5.2.2). La información proporcionada por una predicción *candlestick* fiable sería de gran utilidad para orientar a los inversores. Los resultados obtenidos por Fiess y MacDonald (2002), según los cuales se mejora la predicción de los valores de cierre al incorporar los valores mínimo y máximo al modelo, auguran buenos resultados en esta línea.

Por otro lado, sería interesante profundizar en la aplicación de las series temporales de histograma a los siguientes contextos:

- El caso del instituto de estadística ha sido mencionado repetidas veces a lo largo de la tesis porque es uno de los que mayor poder explicativo tiene. Sin embargo, no se ha podido poner en práctica por falta de datos. Una aplicación interesante sería explorar la predicción de STH que muestren la evolución temporal de una variable medida en el conjunto de los habitantes de una región. Para ello, es necesario contactar con el Instituto Nacional de Estadística.

- Las STH también pueden ser aplicadas en el ámbito del control de calidad. En ese caso, el histograma permite resumir los valores de un indicador de calidad en todos los productos de un lote. La STH resultante mostraría la evolución de dicho indicador en lotes sucesivos. Una línea de investigación interesante consiste en crear procedimientos que permitan detectar situaciones anómalas en los lotes antes de que éstas se agraven.

- El estudio de los flujos de datos (*stream data*) desde la perspectiva de las STH también puede ser interesante. Los flujos de datos son generados de forma continua por sensores con una frecuencia que dificulta su almacenamiento y que favorece el hecho de que sean analizados a medida que se generan. Hébrail y Lechevallier (2007) muestran una primera aproximación a los flujos de datos desde la perspectiva del análisis de datos simbólicos. Por otro lado, Guha, Koudas y Shim (2006) abordan los problemas que supone la construcción de histogramas para resumir flujos de datos. El desarrollo de métodos para analizar STH originadas a partir de flujos de datos y su predicción constituye otra línea de trabajo interesante.

- En el apartado 5.8.3, se mostró una aplicación de las STH en el ámbito medioambiental, donde los histogramas reflejaban la distribución mensual de un determinado contaminante en la ciudad de Madrid. Tal y como se indicó en dicho apartado, los índices de calidad del aire se construyen normalmente con datos horarios u octohorarios para poder controlar que no se rebasen los límites que perjudican la salud humana y para poder alertar a la población en caso de ser necesario. Una aplicación interesante de las STH en este ámbito es el uso de histogramas para agregar temporalmente los niveles de contaminante y obtener STH de frecuencia horaria u octohoraria sobre las que desarrollar procedimientos de control preventivo.

La enumeración de aplicaciones de las STI y de las STH no pretende ser exhaustiva, sino mostrar áreas de desarrollo que sería interesante explorar. Evidentemente, hay otras muchas aplicaciones que no han sido mencionadas y que aguardan ser descubiertas. De ahí la importancia de difundir este nuevo tipo de series temporales.

Apéndice A

Métodos de predicción clásicos adaptados en esta tesis

Nunca profetices:
Si lo haces bien,
nadie se acordará,
y si lo haces mal,
nadie va a dejar que lo olvides.

Mark Twain

Con el fin de hacer esta tesis lo más autocontenida posible, en este apéndice se introducen brevemente los métodos de predicción clásicos que son trabajados en la tesis desde la perspectiva simbólica. Estos métodos son los modelos de vectores autorregresivos, los alisados y el k-NN. El primero se empleará en la predicción de series temporales de intervalos porque permite predecir estas series sin realizar ninguna adaptación previa. Los dos siguientes son adaptados en la tesis de forma que permitan predecir series temporales de intervalos y series temporales de histogramas.

A.1. Los modelos vectoriales autorregresivos

Los modelos vectoriales autorregresivos (VAR) son una generalización de los modelos autorregresivos al contexto multivariante. Los VAR son empleados en econometría para capturar la evolución de las interdependencias entre un conjunto de series temporales sin necesidad de determinar *a priori* la relación de dependencia entre las variables que entran en juego.

En un modelo VAR todas las variables son tratadas de igual forma, ya que todas ellas tienen su propia ecuación en la que se modela su comportamiento incluyendo sus propios retardos y los retardos del resto de variables consideradas.

El modelo VAR de orden p para k variables viene dado por

$$\boldsymbol{Y}_t = \boldsymbol{C} + \boldsymbol{A}_1\boldsymbol{Y}_{t-1} + \boldsymbol{A}_2\boldsymbol{Y}_{t-2} + \cdots + \boldsymbol{A}_p\boldsymbol{Y}_{t-p} + \boldsymbol{E}_t \tag{A.1}$$

donde $\boldsymbol{Y}_t$ es el vector de $k \times 1$ variables, $\boldsymbol{C}$ es un vector de $k \times 1$ constantes, $\boldsymbol{A}_i$ es una matriz de coeficientes de $k \times k$ (con $i = 1, ..., p$) y $\boldsymbol{E}_t$ es el vector $k \times 1$ con los términos de error, los cuales satisfacen

1. $\mathrm{E}(\boldsymbol{E}_t) = 0$, los términos de error tienen media cero.

2. $\mathrm{E}(\boldsymbol{E}_t\boldsymbol{E}_t') = \Omega$, la matriz de covarianzas entre los errores es constante.

3. $\mathrm{E}(\boldsymbol{E}_t\boldsymbol{E}_{t-k}') = 0$, no hay correlación serial entre los términos de error.

Como puede verse, los modelos VAR presentan una estructura sencilla, pero pese a esta sencillez son capaces de recoger dinámicas muy ricas entre múltiples series temporales, de una forma relativamente fácil de usar y de interpretar. Por otro lado, los VAR precisan de un número alto de observaciones para poder tener grados de libertad suficientes para estimar el modelo, o, en su defecto, necesitan que se restrinjan las variables o el número de retardos.

Los modelos VAR también requieren que el vector de series consideradas se rija por una normalidad multivariante y que sus covarianzas sean invariantes a lo largo del tiempo. Sin embargo, en el terreno económico la hipótesis de normalidad multivariante no suele cumplirse. Esto supone un serio problema, ya que la capacidad de un VAR para representar un proceso generador de datos se basa, en gran media, en dicha hipótesis y la inferencia estadística sólo es válida si las hipótesis que se consideran son ciertas. Los estudios de simulación demuestran que la inferencia estadística es sensible a la validez de algunas hipótesis como son la constancia de los parámetros, la ausencia de correlación serial de los residuos y la simetría de los residuos; mientras que es moderadamente robusta frente a problemas de exceso de kurtosis (distribuciones con colas grandes) y de heterocedasticidad residual. Por ello, se aconseja que para garantizar el éxito de un modelo VAR se cumplan al menos las tres primeras hipótesis. La inspección visual de los residuos así como el cálculo de algunos estadísticos descriptivos pueden ayudar a esta tarea.

En el caso en el que las variables que integran el modelo VAR no sean estacionarias debe analizarse si existe una relación de cointegración entre alguna de ellas. Si es el caso, dicha relación debe ser recogida e incorporada al modelo. A continuación, se explica brevemente qué es la cointegración y los modelos vectoriales de corrección del error.

A.1.1. Cointegración y modelos vectoriales de corrección del error

El concepto de cointegración es introducido en Granger (1981) y en Engle y Granger (1987) para recoger el comportamiento de series temporales que evolucionan de forma similar a lo largo del tiempo, y que aunque se puedan separar a corto plazo, tienden a no divergir demasiado a largo plazo. En el plano económico, la cointegración sucede a menudo entre variables como importaciones y exportaciones, ventas y costes de producción, precios y salarios, etc.

Murray (1994) explica la cointegración entre dos series temporales con gran sentido del humor mediante el símil de un borracho y su perro. Cada uno de estos dos personajes, mientras camina, da sus propios rodeos, sin embargo, ambos se preocupan de no alejarse mucho el uno del otro y si esto sucede intentan reducir dicha distancia. Un observador que se fije en la ruta que sigue cada uno de ellos por separado puede llegar a la conclusión de que vagan sin rumbo, es decir, que se mueven según un proceso no estacionario. Sin embargo, si el observador se fija en la distancia que separa a ambos llegará a la conclusión de que es estacionaria.

Más formalmente, diremos que dos series temporales univariantes $\{A_t\}$ y $\{B_t\}$ están cointegradas si ambas series son integradas de orden d,[1] $I(d)$, y existe una combinación lineal entre ellas que es de un orden de integración $d_1 < d$. Es decir, si podemos construir una serie $\{C_t\}$ que sea $I(d_1)$ y que se obtenga como

$$C_t = \alpha_1 B_t + \alpha_2 A_t. \tag{A.2}$$

A la combinación (α_1, α_2) se le llama relación de cointegración. En la práctica, el objetivo es hallar una combinación lineal cuyo orden de integración resultante sea $I(0)$.

Cuando la relación de cointegración es muy importante para el modelo, esta debe aparecer explícitamente en él. Para ello, se utilizan los modelos vectoriales de corrección del error

$$\Delta \boldsymbol{Y}_t = \boldsymbol{B} + \boldsymbol{\Pi} \boldsymbol{Y}_{t-1} + \boldsymbol{\Gamma}_1 \Delta \boldsymbol{Y}_{t-1} + \boldsymbol{\Gamma}_2 \Delta \boldsymbol{y}_{t-2} + \cdots + \boldsymbol{\Gamma}_{p-1} \Delta \boldsymbol{Y}_{t-p+1} + \boldsymbol{E}_t, \tag{A.3}$$

que se obtiene restando $\boldsymbol{Y}_{t-1}$ a ambos lados de la ecuación (A.1) y reajustando los términos. En dicha ecuación, el operador Δ se utiliza para referirse a la diferenciación $\Delta \boldsymbol{Y}_t = \boldsymbol{Y}_t - \boldsymbol{Y}_{t-1}$, $\boldsymbol{B}$ es un vector de $k \times 1$ constantes, y las matrices de coeficientes se hallan de la siguiente forma $\Pi = -(I_k - A_1 - \cdots - A_p)$ y $\Gamma_j = -(A_{j+1} + \cdots + A_p)$ con $j = 1, ..., p-1$.

Si las variables consideradas en el modelo son como mucho $I(1)$, entonces $\Delta \boldsymbol{Y}_t$ no tendrá tendencia estocástica (será estacionario en media) y, por tanto, sólo el término $\boldsymbol{\Pi} \boldsymbol{Y}_{t-1}$ tendrá variables $I(1)$. Sin embargo, como $\Delta \boldsymbol{Y}_t$

[1] Una serie integrada de orden d es una serie a la que hay que diferenciar d veces para convertirla en estacionaria.

no tiene tendencia estocástica, $\mathbf{\Pi} \boldsymbol{Y}_{t-1}$ tampoco la tiene, luego es $I(0)$. Este término es el que contiene las relaciones de cointegración del modelo.

Es importante darse cuenta de que el término de cointegración $\mathbf{\Pi} \boldsymbol{Y}_{t-1}$ representa una relación a largo plazo entre las variables del modelo, mientras que el resto de términos $\mathbf{\Gamma}_j$ recogen las dinámicas a corto plazo del modelo. Los modelos vectoriales de corrección del error son una herramienta adecuada para separar ambos componentes.

A.1.2. Estrategia para especificar un modelo VAR

Allen y Fildes (2001) propone, como estrategia general para predecir con un modelo VAR, usar como especificación de partida un modelo VAR con un número suficientemente alto de retardos. Según afirman estos autores, algunos estudios comparativos indican que se obtienen mejores resultados si se reduce el número de parámetros. La estrategia a seguir consiste en ir probando sucesivamente con un modelo de menor orden. Para ello, Allen y Fildes (2001) proponen usar contrastes de errores de especificación (no-normalidad, outliers, heterocedasticidad, autocorrelación, constancia de los parámetros), para detectar si el modelo es correcto. Como existen muchos tipos de contrastes, lo normal es que siempre falle alguno. Si un modelo simplificado falla los contrastes de errores en la especificación es una prueba de que la simplificación es inapropiada. Sin embargo, según estos autores no existe evidencia para relacionar la capacidad predictiva del modelo y los fallos en los contrastes de errores en la especificación. Según Allen y Fildes (2001), simplificar el modelo adecuadamente suele dar mejores resultados.

Además de intentar reducir el número de retardos de cada variable, también se debe analizar si existen raíces unidad en las series, o realizar un contraste de cointegración, ya que si se da la cointegración un modelo vectorial de corrección del error suele ofrecer mejores predicciones. La relación entre la cointegración y los modelos de corrección del error es estudiada por Engle y Granger (1987) que demuestran que si dos variables están cointegradas, entonces pueden ser representadas mediante un modelo de corrección del error y viceversa.

En este apartado se ha pretendido ofrecer una visión general sobre el tema, para obtener más información sobre los modelos VAR y modelos de corrección del error se recomienda acudir al manual de Lütkepohl (2005) o a una referencia más sintética como Lütkepohl (2006).

A.2. Los métodos de alisado en las series temporales clásicas

Por lo general, los valores de una serie temporal acostumbran a tener una acusada variabilidad a lo largo de toda la serie. Dicha variabilidad se

refleja en la multitud de "picos"que aparecen en su representación gráfica. El comportamiento global de la serie se aprecia mejor cuando estas fluctuaciones son rebajadas. El proceso que permite rebajar las fluctuaciones de una serie recibe el nombre de alisado. Los métodos más básicos para alisar una serie temporal son las medias móviles.

Sin embargo, los métodos de alisado también pueden obtenerse para generar predicciones de una serie temporal. Robert G. Brown propuso el primer método de alisado exponencial (Brown, 1959). Por su parte, Charles C. Holt desarrolló un método de alisado que manejase tendencias aditivas y que permitiese alisar datos estacionales (Holt, 1957). Posteriormente, Winters (1960) prueba los métodos propuestos por Holt sobre datos empíricos aumentando espectacularmente la difusión y la repercusión de los mismos.

Pese a la sencillez de los métodos de alisado exponencial, estos métodos son muy utilizados en la práctica, especialmente cuando el objetivo es pronosticar una gran cantidad de series de forma automática, como en el caso del control de inventario. Además, estos métodos permiten predecir distintos tipos de series temporales con notables resultados, tal y como muestran los análisis empíricos (Gardner, 2006). Por esta razón y por el principio de parsimonia, resulta aconsejable usar los métodos de alisado exponencial como referencia a batir por otros métodos más sofisticados y complejos.

Gardner (1985) y Gardner (2006) presentan dos revisiones sobre el estado del arte de los alisados exponenciales, las cuales demuestran que la técnica sigue en constante desarrollo. Uno de los avances más significativos en los últimos tiempos ha sido el desarrollo de unos modelos espacio-estado con una única fuente de error que dan un soporte estadístico riguroso a los métodos de alisado exponencial desarrollados por Hyndman, Koehler, Snyder y Grose (2002).

A continuación, se resumirán los conceptos básicos de las técnicas de alisado.

A.2.1. Las medias móviles

La media móvil es el método de alisado más sencillo. Dada un serie temporal observada $\{X_t\}$, la predicción $\hat{X}_{t+1}$ en el instante $t+1$ obtenida como una media móvil de orden q, MM(q), es la media de los últimos q valores de la serie.

$$\hat{X}_{t+1} = \frac{X_t + X_{t-1} + \ldots + X_{t-(q-1)}}{q}. \tag{A.4}$$

El valor de q es el orden de la media móvil y permite realizar un alisado a corto, medio y largo plazo. En las series temporales financieras de frecuencia diaria, los valores típicos de q son $q = \{10, 40, 100\}$.

Si el objetivo de la media móvil no es la predicción, sino el eliminar las fluctuaciones de la serie considerada, suele ser interesante eliminar el sesgo

introducido al usar sólo datos pasados. Para ello, se emplean medias móviles centradas que se calculan usando datos pasados y futuros de la serie en torno al instante t de la siguiente forma

$$\hat{X}_t = \frac{X_{t-\frac{q}{2}} + X_{t-\frac{q-1}{2}} + \ldots + X_t + \ldots + X_{t-\frac{q-1}{2}} + X_{t+\frac{q}{2}}}{q}, \tag{A.5}$$

donde q es un valor impar. Sin embargo, en el contexto de esta tesis, el objetivo es la predicción por lo que no se emplearán medias móviles centradas.

Las medias móviles ponderadas se basan en el supuesto de que los valores más recientes de la serie son más relevantes de cara a elaborar la predicción y, consecuentemente, asignan un mayor peso a los valores cuanto más próximos son al instante actual.

La media móvil ponderada con pesos aritméticamente decrecientes, MMPA (q), es

$$\hat{X}_{t+1} = \frac{q \cdot X_t + (q-1)X_{t-1} + \ldots + X_{t-(q-1)}}{q + (q-1) + \ldots + 1}. \tag{A.6}$$

En la MMPA los pesos decrecen aritméticamente, pero puede ser interesante utilizar otro tipo de ponderación que le dé aún más importancia a las observaciones recientes sin llegar a descartar del todo a las observaciones más antiguas. Éste es el propósito de la media móvil ponderada con pesos exponencialmente decrecientes, MMPE(q), que se representa como

$$\hat{X}_{t+1} = \frac{X_t + (1-\alpha)X_{t-1} + \ldots + (1-\alpha)^{q-1}X_{t-(q-1)}}{1 + (1-\alpha) + \ldots + (1-\alpha)^{q-1}}, \tag{A.7}$$

con $\alpha = \frac{2}{q+1}$. Si el número de periodos considerados, q, es lo suficientemente grande, entonces $1 + (1-\alpha) + \ldots + (1-\alpha)^{q-1} \simeq \alpha^{-1}$ y la ecuación (A.7) puede ser reescrita como

$$\hat{X}_{t+1} = \sum_{j=1}^{t} \alpha(1-\alpha)^{j-1}X_{t-(j-1)}. \tag{A.8}$$

Esta ecuación suele presentarse de forma abreviada como

$$\hat{X}_{t+1} = \alpha X_t + (1-\alpha)\hat{X}_t, \tag{A.9}$$

donde $\alpha \in [0,1]$. Esta ecuación representa el método conocido como alisado exponencial simple. Dicha ecuación representa el método de una forma recurrente (o recursiva) y es equivalente a esta otra representada en forma de corrección de error

$$\hat{X}_{t+1} = \hat{X}_t + \alpha(X_t - \hat{X}_t). \tag{A.10}$$

De esta forma, la predicción en el instante $t+1$ es la predicción generada para el instante en t más el error cometido en el instante t rebajado mediante el parámetro $\alpha \in [0,1]$.

Tabla A.1: Notación de los métodos de alisado exponencial mostrados en la figura A.1

Símbolo	Definición
α	Parámetro de alisado para el nivel de la serie
γ	Parámetro de alisado para la tendencia
δ	Parámetro de alisado para la estacionalidad
ϕ	Parámetro para la atenuación
S_t	Nivel alisado de la serie en el instante t
T_t	Tendencia aditiva alisada en el instante t
R_t	Tendencia multiplicativa alisada en el instante t
I_t	Índice estacional alisado en el instante t
X_t	Valor observado de la serie en el instante t
p	Número de periodos en el ciclo estacional
m	Número de periodos adicionales de la predicción
$\hat{X}_t(m)$	Predicción de la serie en el instante $t+m$
$e_t = X_t - \hat{X}_{t-1}(1)$	Error de predicción en el instante t

El alisado exponencial simple es el método de alisado más sencillo. Existen otros métodos más avanzados y que permiten predecir series con distintas características.

A.2.2. Los métodos de alisado exponencial

Pegels (1969) propuso la primera clasificación de los métodos de alisado exponencial. Esta taxonomía ha sido ampliada a lo largo del tiempo para recoger los nuevos métodos de alisado que iban surgiendo. La última versión de la misma es recogida en el artículo de Gardner (2006) y puede verse en la figura A.1.

En la esquina superior izquierda de la tabla de la figura A.1, se muestra la versión más sencilla del alisado exponencial, la cual permite modelar una serie temporal sin tendencia ni estacionalidad. Dicho método es el mismo que muestra la ecuación (A.9). En la tabla se muestran otros métodos de alisado que permite recoger la estacionalidad y la tendencia de la serie, tanto en forma multiplicativa, como en forma aditiva. Además, la tendencia, tanto la aditiva como la multiplicativa, puede presentarse también en forma atenuada (en inglés *damped*). La combinación de las distintas estacionalidades y tendencias da lugar al catálogo de métodos de alisado básicos que muestra la tabla.

Con el catálogo de métodos mostrados se pueden predecir una gran cantidad de tipos de series temporales. Los métodos con tendencia atenuada obtienen un gran rendimiento en la predicción de series temporales, tal y como indica Taylor (2003). Según Gardner (2006), el buen rendimiento de

Tendencia	Estacionalidad: N (No)	A (Aditiva)	M (Multiplicativa)
N (No)	$S_t=\alpha X_t+(1-\alpha)S_{t-1}$ $\hat{X}_t(m)=S_t$	$S_t=\alpha(X_t-I_{t-p})+(1-\alpha)S_{t-1}$ $I_t=\delta(X_t-S_t)+(1-\delta)I_{t-p}$ $\hat{X}_t(m)=S_t+I_{t-p+m}$	$S_t=\alpha(X_t/I_{t-p})+(1-\alpha)S_{t-1}$ $I_t=\delta(X_t/S_t)+(1-\delta)I_{t-p}$ $\hat{X}_t(m)=S_tI_{t-p+m}$
	$S_t=S_{t-1}+\alpha e_t$ $\hat{X}_t(m)=S_t$	$S_t=S_{t-1}+\alpha e_t$ $I_t=I_{t-p}+\delta(1-\alpha)e_t$ $\hat{X}_t(m)=S_t+I_{t-p+m}$	$S_t=S_{t-1}+\alpha e_t/I_{t-p}$ $I_t=I_{t-p}+\delta(1-\alpha)e_t/S_t$ $\hat{X}_t(m)=S_tI_{t-p+m}$
A (Aditiva)	$S_t=\alpha X_t+(1-\alpha)(S_{t-1}+T_{t-1})$ $T_t=\gamma(S_t-S_{t-1})+(1-\gamma)T_{t-1}$ $\hat{X}_t(m)=S_t+mT_t$	$S_t=\alpha(X_t-I_{t-p})+(1-\alpha)(S_{t-1}+T_{t-1})$ $T_t=\gamma(S_t-S_{t-1})+(1-\gamma)T_{t-1}$ $I_t=\delta(X_t-S_t)+(1-\delta)I_{t-p}$ $\hat{X}_t(m)=S_t+mT_t+I_{t-p+m}$	$S_t=\alpha(X_t/I_{t-p})+(1-\alpha)(S_{t-1}+T_{t-1})$ $T_t=\gamma(S_t-S_{t-1})+(1-\gamma)T_{t-1}$ $I_t=\delta(X_t/S_t)+(1-\delta)I_{t-p}$ $\hat{X}_t(m)=(S_t+mT_t)I_{t-p+m}$
	$S_t=S_{t-1}+T_{t-1}+\alpha e_t$ $T_t=T_{t-1}+\alpha\gamma e_t$ $\hat{X}_t(m)=S_t+mT_t$	$S_t=S_{t-1}+T_{t-1}+\alpha e_t$ $T_t=T_{t-1}+\alpha\gamma e_t$ $I_t=I_{t-p}+\delta(1-\alpha)e_t$ $\hat{X}_t(m)=S_t+mT_t+I_{t-p+m}$	$S_t=S_{t-1}+T_{t-1}+\alpha e_t/I_{t-p}$ $T_t=T_{t-1}+\alpha\gamma e_t/I_{t-p}$ $I_t=I_{t-p}+\delta(1-\alpha)e_t/S_t$ $\hat{X}_t(m)=(S_t+mT_t)I_{t-p+m}$
AA (Aditiva Atenuada)	$S_t=\alpha X_t+(1-\alpha)(S_{t-1}+\phi T_{t-1})$ $T_t=\gamma(S_t-S_{t-1})+(1-\gamma)\phi T_{t-1}$ $\hat{X}_t(m)=S_t+\sum_{i=1}^{m}\phi^iT_t$	$S_t=\alpha(X_t-I_{t-p})+(1-\alpha)(S_{t-1}+\phi T_{t-1})$ $T_t=\gamma(S_t-S_{t-1})+(1-\gamma)\phi T_{t-1}$ $I_t=\delta(X_t-S_t)+(1-\delta)I_{t-p}$ $\hat{X}_t(m)=S_t+\sum_{i=1}^{m}\phi^iT_t+I_{t-p+m}$	$S_t=\alpha(X_t/I_{t-p})+(1-\alpha)(S_{t-1}+\phi T_{t-1})$ $T_t=\gamma(S_t-S_{t-1})+(1-\gamma)\phi T_{t-1}$ $I_t=\delta(X_t/S_t)+(1-\delta)I_{t-p}$ $\hat{X}_t(m)=\left(S_t+\sum_{i=1}^{m}\phi^iT_t\right)I_{t-p+m}$
	$S_t=S_{t-1}+\phi T_{t-1}+\alpha e_t$ $T_t=\phi T_{t-1}+\alpha\gamma e_t$ $\hat{X}_t(m)=S_t+\sum_{i=1}^{m}\phi^iT_t$	$S_t=S_{t-1}+\phi T_{t-1}+\alpha e_t$ $T_t=\phi T_{t-1}+\alpha\gamma e_t$ $I_t=I_{t-p}+\delta(1-\alpha)e_t$ $\hat{X}_t(m)=S_t+\sum_{i=1}^{m}\phi^iT_t+I_{t-p+m}$	$S_t=S_{t-1}+\phi T_{t-1}+\alpha e_t/I_{t-p}$ $T_t=\phi T_{t-1}+\alpha\gamma e_t/I_{t-p}$ $I_t=I_{t-p}+\delta(1-\alpha)e_t/S_t$ $\hat{X}_t(m)=\left(S_t+\sum_{i=1}^{m}\phi^iT_t\right)I_{t-p+m}$
M (Multiplicativa)	$S_t=\alpha X_t+(1-\alpha)(S_{t-1}R_{t-1})$ $R_t=\gamma(S_t/S_{t-1})+(1-\gamma)R_{t-1}$ $\hat{X}_t(m)=S_t\ R_t^m$	$S_t=\alpha(X_t-I_{t-p})+(1-\alpha)S_{t-1}R_{t-1}$ $R_t=\gamma(S_t/S_{t-1})+(1-\gamma)R_{t-1}$ $I_t=\delta(X_t-S_t)+(1-\delta)I_{t-p}$ $\hat{X}_t(m)=S_tR_t^m+I_{t-p+m}$	$S_t=\alpha(X_t/I_{t-p})+(1-\alpha)S_{t-1}R_{t-1}$ $R_t=\gamma(S_t/S_{t-1})+(1-\gamma)R_{t-1}$ $I_t=\delta(X_t/S_t)+(1-\delta)I_{t-p}$ $\hat{X}_t(m)=(S_tR_t^m)I_{t-p+m}$
	$S_t=S_{t-1}R_{t-1}+\alpha e_t$ $R_t=R_{t-1}+\alpha\gamma e_t/S_{t-1}$ $\hat{X}_t(m)=S_tR_t^m$	$S_t=S_{t-1}R_{t-1}+\alpha e_t$ $R_t=R_{t-1}+\alpha\gamma e_t/S_{t-1}$ $I_t=I_{t-p}+\delta(1-\alpha)e_t$ $\hat{X}_t(m)=S_tR_t^m+I_{t-p+m}$	$S_t=S_{t-1}R_{t-1}+\alpha e_t/I_{t-p}$ $R_t=R_{t-1}+(\alpha\gamma e_t/S_{t-1})/I_{t-p}$ $I_t=I_{t-p}+\delta(1-\alpha)e_t/S_t$ $\hat{X}_t(m)=(S_tR_t^m)I_{t-p+m}$
MA (Multiplicativa Atenuada)	$S_t=\alpha X_t+(1-\alpha)(S_{t-1}R_{t-1}^{\phi})$ $R_t=\gamma(S_t/S_{t-1})+(1-\gamma)R_{t-1}^{\phi}$ $\hat{X}_t(m)=S_tR_t^{\sum_{i=1}^{m}\phi^i}$	$S_t=\alpha(X_t-I_{t-p})+(1-\alpha)S_{t-1}R_{t-1}^{\phi}$ $R_t=\gamma(S_t/S_{t-1})+(1-\gamma)R_{t-1}^{\phi}$ $I_t=\delta(X_t-S_t)+(1-\delta)I_{t-p}$ $\hat{X}_t(m)=S_tR_t^{\sum_{i=1}^{m}\phi^i}+I_{t-p+m}$	$S_t=\alpha(X_t/I_{t-p})+(1-\alpha)(S_{t-1}R_{t-1}^{\phi})$ $R_t=\gamma(S_t/S_{t-1})+(1-\gamma)R_{t-1}^{\phi}$ $I_t=\delta(X_t/S_t)+(1-\delta)I_{t-1}$ $\hat{X}_t(m)=\left(S_tR_t^{\sum_{i=1}^{m}\phi^i}\right)I_{t-p+m}$
	$S_t=S_{t-1}R_{t-1}^{\phi}+\alpha e_t$ $R_t=R_{t-1}^{\phi}+\alpha\gamma e_t/S_{t-1}$ $\hat{X}_t(m)=S_tR_t^{\sum_{i=1}^{m}\phi^i}$	$S_t=S_{t-1}R_{t-1}^{\phi}+\alpha e_t$ $R_t=R_{t-1}^{\phi}+\alpha\gamma e_t/S_{t-1}$ $I_t=I_{t-p}+\delta(1-\alpha)e_t$ $\hat{X}_t(m)=S_tR_t^{\sum_{i=1}^{m}\phi^i}+I_{t-p+m}$	$S_t=S_{t-1}R_{t-1}^{\phi}+\alpha e_t/I_{t-p}$ $R_t=R_{t-1}^{\phi}+(\alpha\gamma e_t/S_{t-1})/I_{t-p}$ $I_t=I_{t-p}+\delta(1-\alpha)e_t/S_t$ $\hat{X}_t(m)=\left(S_tR_t^{\sum_{i=1}^{m}\phi^i}\right)I_{t-p+m}$

Figura A.1: Taxonomía de los principales métodos de alisado exponencial tomada de Gardner (2006). En la parte superior de cada celda aparecen las fórmulas en forma recurrente y en la inferior en forma de error-corrección

los métodos de alisado en general y de los métodos con tendencia atenuada en particular es posiblemente una de las razones por las que no ha habido el suficiente progreso en la identificación y selección de métodos de alisado exponencial.

A.3. El método de k-NN

El método de los k vecinos más próximos (o k-NN según la abreviatura inglesa de *k-Nearest Neighbours*) es una técnica que se encuadra dentro del área del reconocimiento de patrones. Se trata de una técnica muy versátil que puede emplearse para tareas de clasificación, estimación de densidades, aproximación funcional y predicción de series temporales, entre otras.

La idea clave es que vectores de entrada similares deben tener también valores similares en la variable objetivo. El algoritmo básico es el siguiente:

1. Ante un nuevo vector, se buscan los k vectores más parecidos a él, es decir, sus vecinos. Los vecinos serán aquellos vectores más cercanos al nuevo vector en el espacio de representación. Para medir la distancia entre vectores se emplea habitualmente la distancia euclídea.

2. A continuación, se utiliza el valor de la variable objetivo en los k vecinos para determinar el valor de la variable objetivo del nuevo vector.

   a) Si la variable objetivo es categórica, al nuevo elemento se le asigna como valor de la variable objetivo aquella categoría a la que pertenezcan la mayoría de sus k vecinos.

   b) Si la variable objetivo es numérica, el valor que toma dicha variable en el nuevo elemento suele ser la media aritmética de los valores que toma en sus k vecinos.

Éste es el esquema básico del algoritmo de k-NN, pero existen numerosas modificaciones. Una de las más comunes consiste en ponderar la contribución de cada uno de los vecinos al resultado de forma que aquellos más próximos contribuyan más que aquellos más lejanos.

El método de k-NN es, por tanto, un método de aprendizaje perezoso, es decir, no aproxima la función por completo, sino que realiza una aproximación local de la función y difiere la realización de los cálculos hasta el momento de la asignación del resultado.

Cuando la variable objetivo es categórica, el k-NN se suele emplear como técnica de clasificación (Cover y Hart, 1967). Sin embargo, también se puede aplicar como técnica de regresión cuando la variable objetivo es cuantitativa. En ese caso el método de k-NN puede considerarse como un método de regresión no-paramétrica que realiza regresiones locales. El valor de la variable objetivo cuantitativa para el elemento considerado se suele calcular como la

media de los valores de los vecinos más próximos. Una de las aplicaciones del k-NN como técnica de regresión es la predicción de series temporales (Yakowitz, 1987). Sobre este tema trata el siguiente apartado.

A.3.1. El k-NN como método de predicción de series temporales

La predicción de series temporales con k-NN consiste en, dada una secuencia de valores, identificar las k secuencias pasadas que sean más similares a la actual y, una vez determinadas dichas k secuencias, combinar sus valores futuros para calcular la predicción de la secuencia actual.

El k-NN ha sido empleado en multitud de ocasiones en el ámbito de las finanzas con el fin de modelar la dinámica no-lineal de las series financieras. Sirvan de ejemplo los trabajos de Meade (2002), Fernández-Rodríguez et al. (1999) y Aparicio, Pozo y Saura (2002). Otros campos de aplicación incluyen la hidrología (Brath, Montari y Toth, 2002) y el sector energético (Sorjamaa, Reyhani y Lendasse, 2005).

El algoritmo del k-NN puede describirse brevemente de la siguiente forma

1. La serie temporal considerada, $\{X_t\}$ con $t = 1, ..., n$, es transformada en una serie de vectores d-dimensionales tal que

$$X_t^d = (X_t, X_{t-1}, ..., X_{t-(d-1)}), \tag{A.11}$$

   donde $d \in \mathbb{N}$ y d es el número de retardos. Cada vector X_t^d puede ser representado como un punto dentro de un espacio de d-dimensiones.

2. A continuación, se calcula la distancia entre el último vector $X_n^d = (X_n, X_{n-1}, ..., X_{n-d+1})$ y cada uno de los vectores de la serie temporal $\{X_t^d\}$. Una vez calculadas las distancias se identifican los k vectores más próximos a X_n^d. Dichos vectores se denotaran como $X_{T_1}^d, X_{T_2}^d ..., X_{T_k}^d$. Para calcular las distancias, se suele emplear la distancia euclídea.

3. Dados los k vectores vecinos, $X_{T_1}^d, X_{T_2}^d ..., X_{T_k}^d$, la predicción se calcula como el promedio ponderado de sus valores siguientes

$$\hat{X}_{n+1} = \frac{\sum_{i=1}^{k} \omega_i X_{T_i+1}}{\sum_{i=1}^{k} \omega_i}, \tag{A.12}$$

   donde ω_i es el peso que se le asigna al valor siguiente del vector vecino $X_{T_i}^d$ de forma que $\omega_i \geq 0$ y $\sum_{i=1}^{k} \omega_i = 1$. Si a los valores siguientes de todos los vecinos se les asigna el mismo peso, entonces $\omega_i = 1/k \ \forall i$ y se está realizando la media aritmética de los valores siguientes. Otra forma más sofisticada de asignar pesos consiste en hacer que ω_i sea inversamente proporcional a la distancia entre el vector $X_{T_i}^d$ y el vector actual. De esa forma, se consigue que tengan un mayor peso los valores siguientes de aquellos vectores más similares al actual.

Los parámetros del algoritmo. Este algoritmo consta de dos parámetros: el número de vecinos a considerar, k, y el número de observaciones pasadas a considerar para describir los vectores de elementos. El parámetro k es un parámetro de suavizado, valores grandes de k reducen el ruido, pero hacen que las predicciones sean más homogéneas. Normalmente se emplean técnicas heurísticas, como la validación cruzada para determinar un valor de k apropiado.

En el k-NN, como en cualquier otro método de aprendizaje estadístico, si las variables de entrada son irrelevantes, las predicciones que se obtienen son irremediablemente pobres. Por ello, para obtener un modelo preciso hay que determinar un valor de d adecuado. En realidad, tal y como argumentan Sorjamaa et al. (2005), no sólo se debe determinar el valor de d, sino que puede resultar conveniente determinar qué subconjunto de los d retardos considerados es realmente relevante en nuestro problema. Sorjamaa et al. (2005) proponen tres métodos para seleccionar el subconjunto de retardos adecuado de entre los 2^d posibles subconjuntos. Esto permite eliminar variables de entrada redundantes y reducir las dimensiones de la representación del problema. Por otro lado, esto supone aumentar el esfuerzo computacional, pero debido a la simplicidad del k-NN, los cálculos propuestos por Sorjamaa et al. (2005) pueden realizarse en un tiempo razonable.

Adicionalmente, el instante actual puede ser descrito no sólo por los valores retardados de la serie que se quiere predecir, sino también por los valores retardados de otras series que se encuentren relacionadas con él. Para no contaminar el modelo, conviene sólo añadir aquellas variables y aquellos retardos que sean verdaderamente relevantes.

A.3.2. Otras versiones del k-NN para la predicción de series temporales

Existen numerosas versiones alternativas para realizar el k-NN sobre una serie temporal. Por ejemplo, a la hora de medir distancias entre el vector actual y los vectores pasados, se puede emplear la distancia euclídea estandarizada donde las series son transformadas para tener media cero y varianza uno (Casdagli y Weigend, 1994), o la distancia euclídea ponderada, donde los pesos asignados a cada valor del vector de los retardos decrecen cuanto más alejado en el tiempo se encuentre el retardo más alejadas en el tiempo (Murray, 1993). Una alternativa curiosa a la distancia euclídea es utilizada por Fernández-Rodríguez et al. (1999) que emplean el coeficiente de correlación serial como herramienta para elegir los vecinos más cercanos, que serán aquellos con una mayor correlación con el vector actual.

A la hora de calcular la predicción, en lugar de obtenerla como una media, se pueden utilizar otras alternativas. Una de ellas consiste en calcular la predicción como una media ponderada en la que se asigna más peso a los valores futuros de los vecinos más cercanos. Otra forma de calcular la

predicción consiste en hacerlo mediante un modelo de regresión local ponderada estimado a partir de los valores futuros de los vecinos más cercanos (Atkeson, Moore y Schaal, 1997). Al estar tratando series temporales, lo que se realiza es en realidad una autorregresión local entre X_{T_i+1} y el vector de los valores pasados considerados $X^d_{T_i}$ con $i = 1, ..., k$. La ecuación de autorregresión estimada a partir de los k vectores es aplicada al último vector X^d_n para obtener la predicción $\hat{X}_{n+1}$. Pueden verse más detalles sobre este predictor en Casdagli y Weigend (1994).

Es importante reseñar que otra área donde se aplica el método de k-NN es la predicción de series temporales caóticas. Según este enfoque, algunas series temporales pueden considerarse como el resultado de un proceso determinista, en lugar de ser consideradas como el resultado de un proceso estocástico. Si es posible obtener el modelo determinista que rige la serie, entonces es posible obtener predicciones precisas en el corto plazo. Estos sistemas deterministas son en realidad caóticos ya que son sensibles a las condiciones iniciales. El objetivo consiste en encontrar la función que relaciona los valores pasados de la serie con el valor futuro. Con el k-NN se realiza una estimación local de la función en cada instante temporal.

La primera aproximación al k-NN desde el área de las series temporales caóticas la realizan Farmer y Sidorowich (1987). El enfoque consiste en embeber la serie temporal $\{X_t\}$ en un espacio de estados usando coordenadas retardadas. La principal diferencia con el enfoque básico mostrado anteriormente consiste en que los vectores de estado d-dimensionales vienen representados de la siguiente forma

$$X^{d,\tau}_t = (X_t, X_{t-\tau}, ..., X_{t-(d-1)\tau}), \tag{A.13}$$

donde $d, \tau \in \mathbb{N}$, d es el número de retardos y τ es el parámetro de retardo. Si $\tau = 1$, como se asume en muchos casos, (e.g. Fernández-Rodríguez et al. (1999), Meade (2002) y Sorjamaa et al. (2005)), estamos ante el caso del k-NN básico mostrado anteriormente. Jayawardena, Li y Xu (2002) apuntan algunas alternativas para determinar el valor apropiado de τ: tomar el valor en el que la autocorrelación baja un determinado valor umbral, que es típicamente $1/e$; o tomar como valor el retardo en el cual la autocorrelación vale 0 o cruza la línea del cero; o tomar el valor 1 por defecto. Estos autores apuntan que si el valor de τ es pequeño los datos no serán independientes y si es muy grande se está suavizando en exceso la serie, perdiendo mucha información. En su artículo, Jayawardena et al. (2002) proponen un método para determinar los tres parámetros, τ, k y d, que, según sus pruebas, permite mejorar la capacidad predictiva del modelo.

De cara a adaptar el k-NN a la predicción de series temporales simbólicas no se va a realizar una aproximación desde el punto de vista de las series temporales caóticas. La razón es que muchos conceptos de la teoría de los sistemas caóticos no tienen (al menos, por el momento) correspondencia

en el campo de los datos simbólicos. Por ello, en la tesis, se adapta el k-NN a la predicción de series temporales simbólicas desde la perspectiva del aprendizaje estadístico. Por ello, se asumirá que $\tau = 1$ y se determinarán el resto de parámetros mediante validación cruzada.

Apéndice B

Los histogramas baricéntricos basados en las distancias de Wasserstein y de Mallows

¡Hasta un niño de cinco años
sería capaz de entender esto!
Rápido, busque a un niño de cinco años.

Groucho Marx, Sopa de Ganso

Irpino y Verde (2006b) proponen un procedimiento para estimar de manera sencilla el histograma baricéntrico de un conjunto de histogramas usando la distancia de Mallows. En este apéndice se muestra este procedimiento y se extiende para estimar el histograma baricéntrico usando la distancia de Wasserstein. Además, también se comentarán algunos aspectos relativos al comportamiento de los baricentros de Wasserstein y de Mallows y a la interpretación que puede dársele a cada uno de ellos.

B.1. Estimación del histograma baricéntrico usando las distancias de Wasserstein y de Mallows

Para estimar el histograma baricéntrico es necesario en primer lugar reescribir las distancias de Wasserstein y de Mallows para trabajar con datos de histograma. Al reescribir las distancias, su cálculo se simplifica sobremanera. A continuación, se debe resolver el problema de minimización que supone la obtención del baricentro. Al haber reescrito las distancias, la minimización se resuelve directamente sin tener que recurrir a técnicas de optimización. A continuación, se muestran ambos puntos.

B.1.1. Adaptación de las distancias para tratar con datos de histogramas

A continuación, se muestra cómo estimar las distancias de Mallows y de Wasserstein para dos histogramas cualesquiera, h_X y h_Y, tal que $h_X = \{([I]_{l_X}, \pi_{l_X})\}$ con $l_X = 1, ..., p_X$ y $h_Y = \{([I]_{l_Y}, \pi_{l_Y})\}$ con $l_Y = 1, ..., p_Y$, y asumiendo que en cada intervalo $[I]$ los valores se distribuyen según una uniforme.

Para calcular estas distancias de forma directa, ambas distancias deben ser escritas en función de los intervalos de la recta real para los cuales ambos histogramas, h_X y h_Y, son uniformemente densos. Para determinar estos intervalos se realiza el siguiente proceso.

En primer lugar, se establece el conjunto de pesos acumulados de las funciones de densidad de los histogramas h_X y h_Y

$$w = \{w_{0X}, w_{1X}, ..., w_{p_X X}, w_{0Y}, w_{1Y}, ..., w_{p_Y Y}\}, \tag{B.1}$$

where $w_{0X} = w_{0Y} = 0$, $w_{l_X X} = \sum_{i=1}^{l_X} \pi_{iX}$ and $w_{l_Y Y} = \sum_{i=1}^{l_Y} \pi_{iY}$.

A continuación, hay que determinar el conjunto de pesos acumulados para los que las funciones de densidad acumuladas de h_X y de h_Y intersectan. Los elementos del conjunto son los puntos del intervalo $(0, 1)$ donde $H_X(x) = H_Y(x)$ con $x \in \Re$. A dicho conjunto se le denotará como v.

Dados los conjuntos v y w, el vector $\mathbf{z} = [z_0, ..., z_l, ..., z_m]$, con $z_0 = 0$ y $z_m = 1$ se obtiene como el resultado de

- ordenar los elementos de w sin repeticiones, para el caso de la distancia de Mallows
- ordenar los elementos de $w \cup v$ sin repeticiones, para el caso de la distancia de Wasserstein

Los elementos del vector $\mathbf{z}$ son pesos acumulados. Cada par de elementos consecutivos (z_{l-1}, z_l) de dicho vector puede hacerse corresponder con dos intervalos de la recta real, uno para cada histograma, donde ambos histogramas son uniformemente densos. Para realizar la correspondencia entre los pesos acumulados y los valores de la recta real, es necesario definir la inversa de la función de distribución.

Dado el histograma $h = \{([I]_l, \pi_l)\}$ con $l = 1, ..., p$, y la definición de distribución mostrada en la ecuación (5.3), y asumiendo que dentro de los intervalos $[I]_l$ del histograma los valores se distribuyen uniformemente, la inversa de la función de distribución de h se define como

$$H^{-1}(t) = \underline{I}_l + \frac{t - w_{l-1}}{w_l - w_{l-1}}(\overline{I}_l - \underline{I}_l) \text{ if } t \in [w_{l-1}, w_l), \tag{B.2}$$

donde $w_0 = 0$, $w_l = \sum_{i=1}^{l} \pi_i$ y $t \in [0, 1]$.

Dada la definición de la inversa de la función de distribución acumulada (B.2), cada pareja de elementos consecutivos (z_{l-1}, z_l) del vector $\mathbf{z}$ se corresponde con dos intervalos uniformemente densos, uno para h_X y otro para h_Y, que se hallan como

$$I_{lX} = [H_X^{-1}(z_{l-1}), H_X^{-1}(z_l)] \text{ e } I_{lY} = [H_Y^{-1}(z_{l-1}), H_Y^{-1}(z_l)]. \quad \text{(B.3)}$$

Como estos intervalos son uniformemente densos, cada intervalo $I_l = [\underline{I}_l, \overline{I}_l]$ puede ser expresado como una función de la variable t en términos de su centro y de su radio

$$I_l(t) = c_l + r_l(2t-1) \text{ para } 0 \leq t \leq 1 \text{ con } c_l = \frac{\underline{I}_l + \overline{I}_l}{2}, r_l = \frac{\overline{I}_l - \underline{I}_l}{2}, \quad \text{(B.4)}$$

con $t \in [0,1]$. El peso asociado al intervalo I_l es $\pi_l = z_l - z_{l-1}$, con $l = 1, ..., m$.

Dados los intervalos de la recta real para los cuales los histogramas h_X y h_Y son uniformemente densos, la distancia de Mallows entre los histogramas h_X y h_Y puede ser reescrita en los términos de dichos intervalos como

$$\begin{aligned} D_M^2(h_X, h_Y) &= \sum_{l=1}^{m} \int_{z_{l-1}}^{z_l} (H_X^{-1}(t) - H_Y^{-1}(t))^2 dt \\ &= \sum_{l=1}^{m} \pi_l \int_0^1 [(c_{lX} + r_{lX}(2t-1)) - (c_{lY} + r_{lY}(2t-1))]^2 dt \\ &= \sum_{l=1}^{m} \pi_l [(c_{lX} - c_{lY})^2 + \frac{1}{3}(r_{lX} - r_{lY})^2]. \end{aligned} \quad \text{(B.5)}$$

De manera similar, la distancia de Wasserstein puede ser reescrita como

$$\begin{aligned} D_W(h_X, h_Y) &= \sum_{l=1}^{m} \int_{z_{l-1}}^{z_l} |H_X^{-1}(t) - H_Y^{-1}(t)| dt \\ &= \sum_{l=1}^{m} \pi_l \int_0^1 |(c_{lX} + r_{lX}(2t-1)) - (c_{lY} + r_{lY}(2t-1))| dt \\ &= \sum_{l=1}^{m} \pi_l |c_{lX} - c_{lY}|. \end{aligned} \quad \text{(B.6)}$$

B.1.2. Formulación y resolución del problema de minimización

El histograma baricéntrico h_{X_B} de un conjunto de k histogramas h_{X_1}, $h_{X_2}, \ldots, h_{X_k}$ es la solución al siguiente problema de minimización

$$\min_{h_{X_B}} \sum_{p=1}^{k} \omega_p D(h_{X_B}, h_{X_p}), \quad \text{(B.7)}$$

donde $D(h_{X_B}, h_{X_p})$ es la distancia de Mallows o de Wasserstein y ω_p es el peso que se le asigna a cada histograma h_{X_p}, siendo $\omega_p \geq 0$ y $\sum_{p=1}^{k} \omega_p = 1$.

Si los pesos ω_p son todos iguales para todos los histogramas, entonces ω_p es constante y, por tanto, no tiene efecto en la minimización y puede ser ignorado. Si los pesos no son iguales para todos los histogramas, entonces el problema de minimización puede ser reformulado para que también se ignoren los pesos. Dicha reformulación consiste en repetir cada histograma h_{X_p} un número de veces proporcional a su peso ω_p. Una vez hecho esto, el baricentro h_{X_B} del conjunto original puede ser calculado como el baricentro de un nuevo conjunto de k' histogramas $h'_{X_1}, h'_{X_2}, ..., h'_{X_{k'}}$ donde los pesos son constantes y, por tanto, pueden ignorarse.

Por ejemplo, dado el conjunto de $p = 3$ histogramas $\{h_{X_1}, h_{X_2}, h_{X_3}\}$ cuyos pesos son $\omega_1 = 0.2$, $\omega_2 = 0.3$ y $\omega_3 = 0.5$, respectivamente. Su baricentro puede ser calculado como el baricentro de un conjunto de 10 histogramas donde $h'_{X_1} = h'_{X_2} = h_{X_1}$, $h'_{X_3} = h'_{X_4} = h'_{X_5} = h_{X_2}$ y $h'_{X_6} = ... = h'_{X_{10}} = h_{X_3}$ y donde los pesos no son tenidos en cuenta. Puede demostrarse que las soluciones que se obtienen usando este método y minimizando la ecuación (B.7) con pesos no constantes son iguales tanto para el caso de la distancia de Wasserstein, como para el de la distancia de Mallows.

Por tanto, el problema de minimización de la ecuación (B.7) tanto si los pesos son idénticos, como si no, puede reformularse en ambos casos para prescindir de los pesos, quedando reducido a

$$\min_{h_{X_B}} \sum_{p=1}^{k} D(h_{X_B}, h_{X_p}). \tag{B.8}$$

Si se aplica la distancia de Mallows (B.5), entonces la solución para el problema de minimización de (B.8) es el histograma baricéntrico, $h_{X_B} = \{([I]_{lB}, \pi_{lB})\}$, donde $[I]_{lB} = [c_{lB} - r_{lB}, c_{lB} + r_{lB}]$ con $l = 1, ..., m_k$, y que viene dado por

$$\min_{h_{X_B}} \sum_{p=1}^{k} D_M^2(h_{X_B}, h_{X_p}) = \min_{c_{lB}, r_{lB}} \sum_{p=1}^{k} \sum_{l=1}^{m_k} \pi_l [(c_{lp} - c_{lB})^2 + \frac{1}{3}(r_{lp} - r_{lB})^2], \tag{B.9}$$

donde m_k es la longitud del vector $\mathbf{z}_k$. Este vector es el resultado de ordenar sin repetición el vector w, siendo w el conjunto de pesos acumulados de las funciones de densidad de los k histogramas considerados, tal y como se mostró en la sección B.1.1.

En el caso de la distancia de Wasserstein (B.6), h_{X_B} se calcula como

$$\min_{h_{X_B}} \sum_{p=1}^{k} D_W(h_{X_B}, h_{X_p}) = \min_{c_{lB}} \sum_{p=1}^{k} \sum_{l=1}^{m_k} \pi_l |c_{lp} - c_{lB}|, \tag{B.10}$$

donde m_k es la longitud del vector $\mathbf{z}_k$. En este caso, $\mathbf{z}_k$ es el resultado de ordenar sin repetición $w \cup v$, siendo w el conjunto de pesos acumulados de

las funciones de densidad de los k histogramas considerados, y siendo v el conjunto de las intersecciones de las funciones de densidad acumuladas de los k histogramas, tal y como se indica en la sección B.1.1.

Es importante tener en cuenta que la minimización en (B.9) es un problema de mínimos cuadrados, mientras que la minimización en (B.10) es un problema de mínimas desviaciones en valor absoluto. Por esta razón, en el caso de la distancia de Mallows (B.9), el mínimo se obtiene como una media

$$c_{lB} = \frac{\sum_{p=1}^{k} c_{lp}}{k} \quad \text{y} \quad r_{lB} = \frac{\sum_{p=1}^{k} r_{lp}}{k}, \tag{B.11}$$

para cada $l = 1, ..., m_k$. Mientras que en el caso de la distancia de Wasserstein, se obtiene como la mediana

$$c_{lB} = \text{mediana}(c_{lp}), \text{ con } p = 1, ..., k \tag{B.12}$$

para cada $l = 1, ..., m_k$. Si c_{lq} con $q \in 1, ..., k$ es el centro del intervalo l del baricentro que se ha obtenido como la mediana de los k centros, el radio de dicho intervalo será $r_{lB} = r_{lq}$. Como la solución óptima es una mediana, si el número de histogramas k es par, entonces el valor de c_{lp} puede ser cualquier valor que se encuentre entre los dos centros de c_{lp} con $p = 1, ..., k$ que delimiten el 50 % central del conjunto. Normalmente, se tomará como c_{lp} la media de de dicho par de centros.

A partir de los valores de c_{lB} y de r_{lB}, el histograma baricéntrico $h_{X_B} = \{([I]_{lB}, \pi_{lB})\}$ se representa como

$$[I]_{lB} = [c_{lB} - r_{lB}, c_{lB} + r_{lB}], l = 1, ..., m_k, \tag{B.13}$$

donde los pesos asociados a cada intervalo son $\pi_{lB} = \pi_l$ con $l = 1, ..., m_k$.

B.2. Comportamiento de los histogramas baricéntricos de Wasserstein y de Mallows

En primer lugar, conviene aclarar que ni el baricentro de Wasserstein, ni el baricentro de Mallows se comportan como una mixtura de las distribuciones de los histogramas que los han originado.

Se dice que una variable aleatoria continua X es una mixtura de n variables aleatorias continuas Y_i, si la función de densidad de X, $f_X(x)$, es una suma ponderada de la siguiente forma

$$f_X(x) = \sum_{i=1}^{n} a_i f_{Y_i}(x). \tag{B.14}$$

Normalmente, se impone la condición de que la suma sea una combinación lineal convexa, i.e. $a_i \geq 0$ y $\sum_{i=1}^{n} a_i = 1$.

El hecho de que el baricentro que se obtiene con la distancia de Wasserstein y con la de Mallows no se comporte como una mixtura es adecuado para poderlo utilizar como predicción de una STH. Si el baricentro se comportase como una mixtura, tendría un soporte que sería igual a la unión de todos los soportes de los histogramas considerados y tendría tantas modas como modas hubiese en el conjunto de histogramas utilizado para estimar el baricentro. Indudablemente, una mixtura representa al conjunto de histogramas original, pero lo hace de una forma poco compacta que no resulta adecuada en el contexto de predicción de STH. Tal y como indican Verde y Irpino (2007), los baricentros que se obtienen utilizando distancias, como la distancia de Variación Total sí se comportan como una mixtura de distribuciones, por lo que no resultan adecuados para este propósito.

Ahora bien, si los baricentros de Mallows y de Wasserstein no se comportan como una mixtura de distribuciones, ¿cómo lo hacen? El hecho de que el baricentro de Mallows se calcule como una media, ver ecuación (B.11), hace que su comportamiento sea precisamente el de una media. De forma similar, como el baricentro de Wasserstein se calcula como una mediana, ver ecuación (B.12), su comportamiento es el de una mediana y únicamente tiene en cuenta aquellos histogramas cuyas funciones de distribución delimitan el 50 % central de las funciones de distribución de los k histogramas considerados. Para poder apreciar mejor estos comportamientos se van a emplear una serie de ejemplos.

B.2.1. Ejemplos de baricentros

Para cada uno de los ejemplos, se va a representar gráficamente el conjunto de histogramas original y su baricentro. En las representaciones gráficas no sólo se mostrarán las funciones de densidad, sino también las funciones de distribución acumuladas. Es precisamente en estas últimas donde se aprecia mejor el comportamiento de los baricentros. Esto es debido a que los baricentros se calculan empleando las funciones de distribución acumulada. Para comprender mejor cómo se calculan los baricentros, el gráfico de las funciones de distribución acumuladas mostrará unas líneas horizontales que representan los valores z del vector $\mathbf{z}$. Dichos valores z, particionan el rango de la inversa de la función de distribución acumulada, i.e. el intervalo $[0, 1]$, en las regiones en las que se calculará cada intervalo del baricentro. El cálculo del intervalo en cada una de esas regiones se realiza de acuerdo con las fórmulas mostradas en el apartado B.1.2.

Consideremos los histogramas $h_1 = \{([1, 2), 0.7), ([2, 3), 0.2), ([3, 4], 0.1)\}$ y $h_2 = \{([11, 12), 0.1), ([12, 13), 0.2), ([13, 14], 0.7)\}$. La figura B.1 muestra en su parte superior estos histogramas y el baricentro que se obtiene utilizando la distancia de Mallows. En ella se aprecia claramente como la posición, el ancho y la forma del histograma baricéntrico son el resultado de promediar las características de los histogramas considerados (o, mejor dicho, de sus funcio-

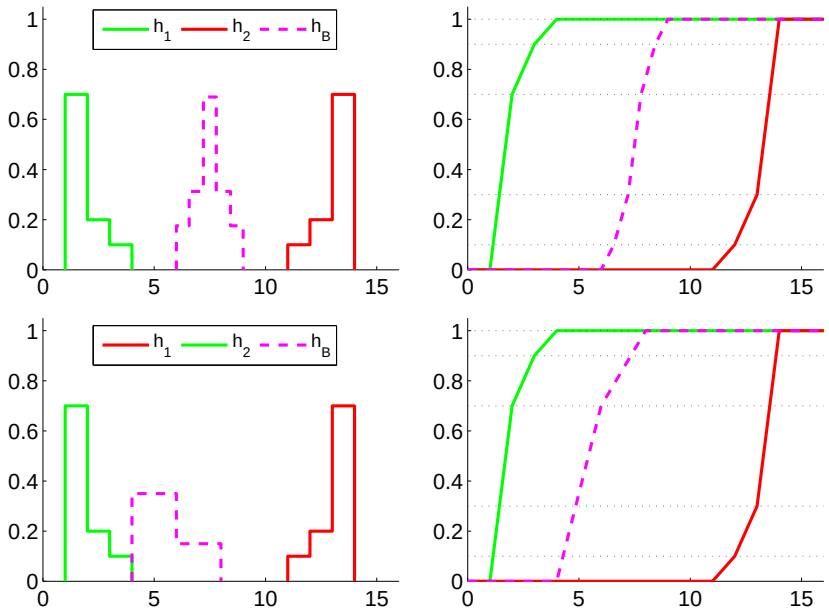


Figura B.1: Baricentro de Mallows (arriba) y un posible baricentro de Wasserstein (abajo) para los histogramas h_1 y h_2. Funciones de densidad (izqda.) y funciones de distribución acumulada (dcha.).

nes de distribución acumuladas). El baricentro de Mallows que se obtiene es $h_{B_W} = \{([6, 6.57), 0.1), ([6.57, 7.21), 0.2), ([7.21, 7.79), 0.4), ([7.79, 8.43), 0.2), ([8.43, 9], 0.1)\}$.

La parte inferior de la figura B.1 muestra los histogramas h_1 y h_2 y un posible baricentro de Wasserstein para los histogramas considerados. El baricentro de Wasserstein para los histogramas h_1 y h_2 no es único porque en este caso el número de histogramas considerados es par y, como se ha dicho, el baricentro de Wasserestein se calcula como una mediana. Los valores del baricentro representado son $h_{B_M} = \{([4, 6), 0.7), ([6, 8], 0.3)\}$. En realidad, cualquier función de distribución acumulada contenida dentro de las funciones de distribución acumuladas de h_1 y h_2 puede ser un baricentro de Wasserstein para h_1 y h_2. En estos casos, lo más adecuado es emplear la media de las dos funciones acumuladas de distribución que formen parte del 50 % central para cada una de las regiones del intervalo $[0, 1]$. En el caso que nos ocupa, con sólo dos histogramas, dicho histograma baricéntrico sería el mismo que el que se calcula utilizando la distancia de Mallows.

Consideremos ahora el siguiente conjunto de tres histogramas: $h_3 = \{([0, 1), 0.2), ([1, 2], 0.8)\}$, $h_4 = \{([5, 6), 0.4), ([7, 8], 0.6)\}$ y $h_5 = \{([6, 7), 0.7), ([7, 8], 0.3)\}$. En la parte inferior de la figura B.2 se muestra el baricentro de Wasserstein de este conjunto. Como el número de histogramas es impar, di-

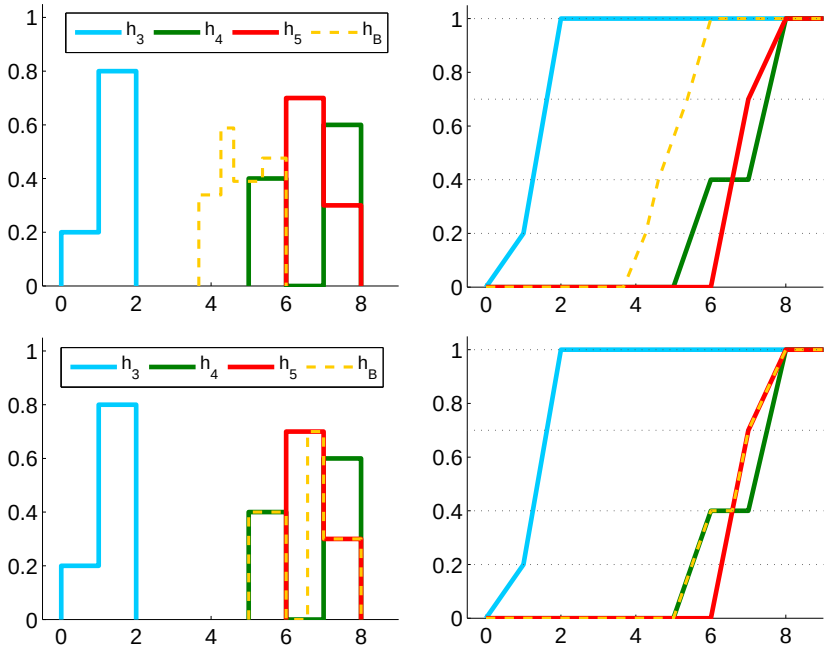


Figura B.2: Baricentro de Mallows (arriba) y un posible baricentro de Wasserstein (abajo) para los histogramas h_3, h_4 y h_5. Funciones de densidad (izqda.) y funciones de distribución acumulada (dcha.).

cho baricentro es único y viene dado por $h_{B_W} = \{([5, 5.5), 0.2), ([5.5, 6), 0.2), ([6, 6.57), 0), ([6.57, 7), 0.3), ([7, 8], 0.3)\}$. En el gráfico donde se muestran las funciones de distribución acumuladas, se puede apreciar muy claramente que el baricentro de Wasserstein se comporta como la mediana de dichas funciones. Tal y como puede verse, el baricentro coincide en todo momento con la función de distribución acumulada que ocupa la posición central del rango. El histograma h_3 no tiene efecto sobre el baricentro de Wasserstein, ya que su función de distribución acumulada se encuentra en un extremo del rango de valores en todas las regiones inducidas por los valores del vector $\mathbf{z}$.

Sin embargo, en el baricentro de Mallows todos los histogramas considerados tienen efecto porque dicho baricentro se calcula como una media. Para el conjunto de histogramas considerado, el baricentro de Mallows es $h_{B_M} = \{([3.67, 4.26), 0.2), ([4.26, 4.61), 0.2), ([4.61, 5.38), 0.3), ([5.38, 6], 0.3)\}$ y puede verse en la parte superior de la figura B.2. Este ejemplo sirve para ver que, en el caso del baricentro de Mallows, el histograma h_3 no sólo es tenido en cuenta en el cálculo del baricentro, sino que tiene mucho efecto sobre él por encontrarse bastante alejado de los histogramas h_4 y h_5. Este comportamiento es similar al que presenta la media aritmética de un conjunto de valores clásicos cuando uno de ellos se encuentra muy alejado del resto.

B.3. Análisis de la idoneidad de los histogramas baricéntricos de Wasserstein y de Mallows para ser empleados en métodos de alisado

Tal y como se indica en el apartado 5.5.2.2, la adaptación del alisado de una serie temporal utilizando baricentros consiste en reemplazar el cálculo del promedio de un conjunto de histogramas por la estimación de su baricentro. El promedio aparece tanto en la ecuación de la media móvil, como en la del alisado exponencial en forma recursiva, pero no en la ecuación del alisado en forma de corrección del error. A continuación, se estudia la adaptación de las fórmulas de alisado donde aparece el promedio sustiyéndolo por el baricentro de Wasserstein y el de Mallows.

B.3.1. Análisis de la adaptación de la media móvil

Dada una serie temporal $\{h_{X_t}\}$ con $t = 1, .., n$, la predicción mediante una media móvil de q términos puede obtenerse como

$$h_{X_{t+1}} = \omega_1 h_{X_t} + \omega_2 h_{X_{t-1}} + \ldots + \omega_q h_{X_{t-(q-1)}}, \tag{B.15}$$

donde los pesos ω_i con $i = 1, ..., q$ pueden ser iguales para todos los términos o decrecer aritméticamente o exponencialmente a medida que i aumenta.

La predicción de la media móvil mostrada en (B.15) puede obtenerse calculando el siguiente baricentro

$$\arg\min_{\hat{h}_{X_{t+1}}} \sum_{i=1}^{q} \omega_i D(\hat{h}_{X_{t+1}}, h_{X_{t-(i-1)}}), \tag{B.16}$$

donde $D(\cdot, \cdot)$ es la distancia de Wasserstein o de Mallows. En el punto B.1.2 se ha explicado cómo calcular este baricentro para ambas distancias y en el punto B.2 se ha comentado que el baricentro de Wasserstein es la función de distribución mediana de las funciones de distribución de los histogramas considerados, mientras que el del baricentro Mallows vendría a ser la función de distribución media. Esta diferencia de comportamiento entre ambas distancias resulta interesante ya que *a priori* permite realizar una media móvil con distintos matices. Por ejemplo, si se quiere eliminar el comportamiento extremo conviene emplear la mediana, si se busca un comportamiento que tenga en cuenta todos los términos de la media móvil.

Sin embargo, la media móvil empleando la distancia de Wasserstein presenta una serie de inconvenientes. Si el valor de $\omega_1 > 0.5$, entonces el resultado de la media móvil será exactamente el valor del último elemento de la la serie, h_{X_t}. En otras palabras, la predicción que obtendrá la media móvil o, mejor dicho, mediana móvil, sería la misma que ofrece el método ingenuo. Esto es así porque, tal y como se explica en el punto B.1.2, si el peso asignado

a cada histograma no es el mismo para todos, el problema de minimización del cálculo del baricentro se reformula de tal forma que los pesos asignados a todos los individuos sea el mismo. Según dicha reformulación, si el peso asignado al individuo u es, por ejemplo, $\omega_u = 0.51$, siendo $\sum_{i=1}^{q} \omega_i = 1$, entonces el 51 % de los individuos del problema reformulado serán idénticos al histograma cuyo peso asignado era ω_u. La mediana de un conjunto donde el 51 % de los datos son idénticos es dicho dato, ya que se repite más de la mitad de las veces. Por tanto, si $\omega_1 > 0.5$ la mediana móvil no es de gran utilidad.

Este problema no sucede cuando se calcula la media móvil empleando la distancia de Mallows porque dicho baricentro se comporta como una media y en una media todos los elementos que se usan para calcularla son tenidos en consideración. Si el peso asignado a un elemento es $\omega_u = 0.51$, siendo $\sum_{i=1}^{q} \omega_i = 1$, el baricentro se comporta como una media ponderada que obedece a dichas ponderaciones. Si el problema se reformula de acuerdo a lo explicado en el punto B.1.2, dicho elemento se repetirá el 51 % de las veces en el nuevo conjunto de individuos y, por tanto, tendrá más representación en la media. Es fácil comprobar que los resultados que se obtienen en ambos casos son idénticos.

B.3.2. Análisis de la adaptación del alisado exponencial

Dada una serie temporal $\{h_{X_t}\}$ con $t = 1, .., n$, la predicción para el instante $t + 1$ mediante el alisado exponencial simple en modo recursivo puede expresarse como

$$h_{X_{t+1}} = \alpha h_{X_t} + (1 - \alpha)\hat{h}_{X_t}, \tag{B.17}$$

donde $\alpha \in [0, 1]$.

La predicción que se obtiene con la ecuación del alisado en (B.17) puede obtenerse calculando el siguiente baricentro

$$\arg\min_{\hat{h}_{X_{t+1}}} \left(\alpha D(\hat{h}_{X_{t+1}}, h_{X_t}) + (1 - \alpha) D(\hat{h}_{X_{t+1}}, \hat{h}_{X_t})\right), \tag{B.18}$$

donde $D(\cdot, \cdot)$ es la distancia de Wasserstein o de Mallows.

En este caso, al igual que sucedía en el apartado anterior, la adaptación de la ecuación del alisado utilizando la distancia de Wasserstein no resulta adecuada porque sólo presenta tres resultados posibles:

- Si $\alpha > 0.5$, entonces $\hat{h}_{X_{t+1}} = h_{X_t}$.
- Si $\alpha < 0.5$, entonces $\hat{h}_{X_{t+1}} = \hat{h}_{X_t}$.
- Si $\alpha = 0.5$ el resultado es cualquier histograma cuya función de distribución acumulada se encuentre entre las funciones de distribución

acumuladas de h_{X_t} y $\hat{h}_{X_t}$. Lo razonable en este caso sería tomar como predicción la media de ambos histogramas, i.e. el baricentro de Mallows.

El problema no es sólo que existan tres resultados posibles es que, salvo en el caso de $\alpha = 0.5$, para el resto de valores de α la serie alisada sería constante para todo valor de t.

De nuevo, la explicación a este comportamiento se encuentra en la reformulación del problema de minimización que permite calcular el baricentro cuando los pesos asignados a cada histograma no son idénticos (ver el apartado B.1.2). La reformulación consiste en calcular el baricentro para un nuevo conjunto de histogramas donde los histogramas del problema original se encuentren repetidos un número de veces proporcional a su peso. Es decir, en este caso, si $\alpha = 0.6$, eso quiere decir que el baricentro de Wasserstein se obtendrá calculando el baricentro de un conjunto donde haya 6 individuos iguales h_{X_t} y 4 iguales $\hat{h}_{X_t}$. El baricentro de Wasserstein de dicho conjunto sería el histograma mediano, que en este caso sería h_{X_t} para todo valor de t.

Dicho problema no sucede para el caso en el que se adapten la ecuación de alisado (B.18) empleando la distancia de Mallows, porque en ese caso, el baricentro se comporta como una media ponderada. Cuanto más peso se asigne a α más similar será el baricentro resultante a h_{X_t} y cuanto menos peso se le asigne más se parecerá a $\hat{h}_{X_t}$.

B.3.2.1. Relación entre la fórmula de la media móvil y del alisado exponencial

Tal y como se explica en el punto A.2 de los apéndices, la ecuación de la media móvil de orden q con pesos exponencialmente decrecientes y la ecuación del alisado exponencial son equivalentes siempre y cuando el número de periodos q sea lo suficientemente grande. En este apartado se va a comprobar si dicha equivalencia se cumple también en la adaptación al contexto de las STH utilizando los baricentros de Wasserstein y de Mallows.

Dada una serie temporal $\{h_{X_t}\}$ con $t = 1, .., n$, la media móvil de orden q con pesos exponencialmente decrecientes puede reescribirse en función de α asumiendo que $\alpha = \frac{2}{q+1}$ y que q es lo suficientemente grande, de la siguiente forma

$$\arg\min_{\hat{h}_{X_{t+1}}} \sum_{j=1}^{t} \alpha(1-\alpha)^{j-1} D(h_{X_{t-(j-1)}}, \hat{h}_{X_{t+1}}). \tag{B.19}$$

Si la distancia considerada es la distancia de Mallows, esta ecuación es equivalente a la adaptación del alisado exponencial mostrada en (B.18). Sin embargo, si se considera la distancia de Wasserstein, la equivalencia no se cumple. La demostración de estos hecho sería extensa y por ello no se incluirá

aquí. Sin embargo, para entender por qué suceden basta con considerar el siguiente ejemplo en el contexto de las series temporales clásicas.

Consideremos que en un determinado instante se tienen los valores de la serie $X_t = 5$ y $X_{t-1} = 2$, y la predicción para el instante $t-1$ $\hat{X}_{t-1} = 1$. Supongamos que el valor de $\alpha = 0.4$. En ese caso, la predicción para los instantes t y $t+1$ será

$$\hat{X}_t = 0.4 \cdot X_{t-1} + 0.6 \cdot \hat{X}_{t-1} = 3.2 \quad \text{(B.20)}$$

$$\hat{X}_{t+1} = 0.4 \cdot X_t + 0.6 \cdot \hat{X}_t = 3.92 \quad \text{(B.21)}$$

La predicción que se obtiene para $t+1$ es la misma que la que se obtiene si en la ecuación (B.21) se sustituye $\hat{X}_t$ por la ecuación (B.20) y se calcula el resultado. En ese caso, la ecuación sería

$$\hat{X}_{t+1} = 0.4 \cdot X_t + 0.24X_{t-1} + 0.36 \cdot \hat{X}_{t-1} = 3.92 \quad \text{(B.22)}$$

La equivalencia en este caso resulta obvia por las propiedades de la aritmética básica. En este caso las predicciones se obtienen como promedios ponderados de dos términos. Por ello, este caso sería similar al cálculo de la predicción utilizando el baricentro de Mallows, que también se calcula como un promedio.

Por otro lado, si expresamos las predicciones como una mediana ponderada donde el peso asignado a cada valor indica el tanto por ciento de elementos de ese valor que existen en el conjunto para el que se calcula la mediana, en ese caso obtendríamos

$$\hat{X}_t = \text{mediana}(0.4X_{t-1}, 0.6\hat{X}_{t-1}) = 1 \quad \text{(B.23)}$$

$$\hat{X}_{t+1} = \text{mediana}(0.4X_t, 0.6\hat{X}_t) = 1 \quad \text{(B.24)}$$

Según esta adaptación la mediana en la ecuación (B.23) es el valor 1 porque se repite el 60 % de las veces. En la siguiente ecuación se sustituye su valor y sucede lo mismo. Y es obvio que ese valor se repetirá como predicción para toda la serie.

Sin embargo, si se calcula la predicción adaptando la ecuación en forma extendida (B.22) de forma que aparezcan todos los elementos de la serie en ella y asignándole a cada uno el peso que le corresponde, el resultado varía

$$\hat{X}_{t+1} = \text{mediana}(0.4X_t, 0.24X_{t-1}, 0.36\hat{X}_{t-1}) = 2. \quad \text{(B.25)}$$

La predicción en este caso no coincide con la obtenida con las ecuaciones de alisado en forma comprimida. Esto es debido a las propiedades de la mediana, la cual no se calcula mediante operaciones aritméticas, sino que necesita de todo el conjunto de datos ordenado para poder ser calculada. Por ello, no puede calcularse de manera fraccionada, es decir, no se puede

dividir el conjunto de datos en varios subconjuntos, calcular la mediana de cada uno de esos subconjuntos y, a continuación, calcular la mediana de las medianas de los subconjuntos. El resultado en ese caso no tiene por qué coincidir con la mediana original.

Este fenómeno que se produce en las series temporales clásica, también sucederá con las STH y es la razón de que adaptar las medias móviles con el baricentro que emplea la distancia de Wasserstein no sea adecuado, ya que dicho baricentro actúa como una mediana.

En conclusión, la equivalencia que se da en las series temporales clásicas entre la ecuación de la media móvil de orden q con pesos exponencialmente decrecientes y la ecuación del alisado exponencial, sólo se cumple en las STH utilizando el método de los baricentros con la distancia de Mallows y no empleando la distancia de Wasserstein.

Bibliografía

Teme al hombre de un solo libro.
Santo Tomás de Aquino

Aas, K. y Dimakos, X. K. Statistical modelling of financial time series: An introduction. Informe técnico, Norwegian Computing Center. Applied Research and Development, 2004.

Alizadeh, S., Brandt, M. W. y Diebold, F. X. Range-based estimation of stochastic volatility models. *The Journal of Finance*, vol. 57(3), páginas 1047–1091, 2002.

Allen, P. G. y Fildes, R. *Principles of forecasting: A handbook for researchers and practitioners*, capítulo Econometric Forecasting, páginas 303–362. Kluwer Academic, Norwell, MA, 2001.

Andrada-Félix, J., Fernadez-Rodriguez, F., Garcia-Artiles, M.-D. y Sosvilla-Rivero, S. An empirical evaluation of non-linear trading rules. *Studies in Nonlinear Dynamics & Econometrics*, vol. 7(3), páginas 1–30, 2003.

Angulo, C., Anguita, D., González-Abril, L. y Ortega, J. Support vector machines for interval discriminant analysis. *Neurocomputing*, vol. 71(7–9), páginas 1220–1229, 2008.

Aparicio, T., Pozo, E. y Saura, D. The nearest neighbour method as a test for detecting complex dynamics in financial series. An empirical application. *Applied Financial Economics*, vol. 12(7), páginas 517–525, 2002.

Appice, A., d'Amato, C., Esposito, F. y Malerba, D. Classification of symbolic objects: A lazy learning approach. *Intelligent Data Analysis*, vol. 10, páginas 301–324, 2006.

Arifin, A. Z. y Asano, A. Image segmentation by histogram thresholding using hierarchical cluster analysis. *Pattern Recognition Letters*, vol. 27(13), páginas 1515–1521, 2006.

Arroyo, J., Muñoz San Roque, A., Maté, C. y Sarabia, A. Exponential smoothing methods for interval time series. En *Proceedings of the 1st European Symposium on Time Series Prediction*, páginas 231–240. 2007.

Asharaf, S., Narasimha Murty, M. y Shevade, S. K. Rough set based incremental clustering of interval data. *Pattern Recognition Letters*, vol. 27(6), páginas 515–519, 2006.

Atkeson, C. G., Moore, A. W. y Schaal, S. Locally weighted learning. *Artificial Intelligence Review*, vol. 11(1–5), páginas 11–73, 1997.

Baltagi, B. H., editor. *Panel Data. Theory and Applications*. Springer, Berlin, 2004.

Beck, J. B., Kreinovich, V. y Wu, B. *Soft Methodology and Random Information Systems*, capítulo Interval-Valued and Fuzzy-Valued Random Variables: From Computing Sample Variances to Computing Sample Covariances, páginas 85–92. Springer-Verlag, Berlín, 2004.

Beckers, S. Variances of securityprice returns based on high, low, and closing prices. *The Journal of Business*, vol. 56(1), páginas 97–112, 1983.

Benjamini, Y. Opening the box of a boxplot. *The American Statistician*, vol. 42, páginas 257–262, 1988.

Berleant, D. y Zhang, J. Representation and problem solving with the distribution envelope determination (DEnv) method. *Reliability Engineering and System Safety*, vol. 85(1–3), páginas 153–168, 2004.

Bernstein, J. *The Compleat Day Trader: Trading Systems, Strategies, Timing Indicators and Analytical Methods*. McGraw-Hill, New York, 1995.

Bertrand, P. y Goupil, F. *Analysis of Symbolic Data. Exploratory Methods for Extracting Statistical Information from Complex Data*, capítulo Descriptive statistics for symbolic data, páginas 103–124. Springer, 2000.

Billard, L. Dependencies in bivariate interval-valued symbolic data. En *Classification, Clustering and Data Mining Applications: Proceedings of the 9th Conference of the IFCS, IFCS 2004*, páginas 319–324. Springer, Berlín, 2004.

Billard, L. y Diday, E. Regression analysis for interval-valued data. En *Data Analysis, Classification and Related Methods : Proceedings of the 7th Conference of the IFCS, IFCS 2002*, páginas 369–374. Springer, Berlín, 2000.

Billard, L. y Diday, E. Symbolic regression analysis. En *Classification, Clustering and Data Analysis: Proceedings of the 8th Conference of the IFCS, IFCS 2002*, páginas 281–288. Springer, Berlín, 2002.

Billard, L. y Diday, E. From the statistics of data to the statistics of knowledge: Symbolic data analysis. *Journal of the American Statistical Association*, vol. 98(462), páginas 470–487, 2003.

Billard, L. y Diday, E. Descriptive statistics for interval-valued observations in the presence of rules. *Computational Statistics*, vol. 21, páginas 187–210, 2006a.

Billard, L. y Diday, E. *Symbolic Data Analysis: Conceptual Statistics and Data Mining*. Wiley & Sons, Chichester, 1ª edición, 2006b.

Bishop, C. M., editor. *Neural Networks for Pattern Recognition Author*. Oxford University Press, Oxford, 1995.

Bock, H.-H. *Analysis of Symbolic Data. Exploratory Methods for Extracting Statistical Information from Complex Data*, capítulo Dissimilarity Measures for Probability Distributions, páginas 153–160. Springer, 2000.

Bock, H.-H. *Symbolic Data Analysis and the SODAS Software*, capítulo Visualizing symbolic data by Kohonen maps, páginas 205–234. John Wiley & Sons, Chichester, 2008.

Bock, H.-H. y Diday, E., editores. *Analysis of Symbolic Data. Exploratory Methods for Extracting Statistical Information from Complex Data*. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Berlín, 1ª edición, 2000.

Box, G. E. y Jenkins, G. M. *Time series analysis: forecasting and control*. Holden Day, San Francisco, 1970.

Brannas, K. y Ohlsson, H. Asymetric time series and temporal aggregation. *Review of Economics and Statistics*, vol. 81, páginas 341–344, 1999.

Brath, A., Montari, A. y Toth, E. Neural networks and non-parametric methods for improving real-time flood forecasting through conceptual hydrological models. *Hidrology and Earth Systems Sciences*, vol. 6(4), páginas 627–640, 2002.

Brida, J. G. y Punzo, L. F. Symbolic time series analysis and dynamic regimes. *Structural Change and Economic Dynamics*, vol. 14(2), páginas 159–183, 2003.

Brito, P. Modelling and analysing interval data. En *Proceedings of the 30th Annual Conference of GfKl*, páginas 197–208. Springer, 2007.

Brito, P. y Noirhomme-Fraiture, M. Editorial of the special issue on symbolic and spatial data analysis: Mining complex data structures. *Intelligent Data Analysis*, vol. 10, páginas 297–300, 2006.

Brown, R. G., editor. *Statistical forecasting for inventory control*. McGraw-Hill, New York, 1959.

Burman, P. Estimation of equifrequency histograms. *Statistics and Probability Letters*, vol. 56, páginas 227–238, 2002.

Canal, L. y Marques Pereira, R. Towards statistical indices for numeroid data. En *Proceedings of the Seminar on New Techniques and Technologies for Statistics*. Sorrento, 1998.

de Carvalho, F. d. A., de Souza, R. M. C. R., Chavent, M. y Lechevallier, Y. Adaptive Hausdorff distances and dynamic clustering of symbolic interval data. *Pattern Recognition Letters*, vol. 27(3), páginas 167–179, 2006a.

de Carvalho, F. d. A. T., Brito, P. y Bock, H.-H. Dynamic clustering for interval data based on l2 distance. *Computational Statistics*, vol. 21(2), páginas 231–250, 2006b.

de Carvalho, F. d. A. T., Lima Neto, E. d. A. y Tenorio, C. P. A new method to fit a linear regression model for interval-valued data. En *KI 2004: Advances in Artificial Intelligence, 27th Annual German Conference on AI*, páginas 295–306. Springer, 2004.

de Carvalho, F. d. A. T., Souza, R. M., Chavent, M. y Lechevallier, Y. Adaptative Hausdorff distances and dynamic clustering of symbolic interval data. *Pattern Recognition Letters*, vol. 27, páginas 167–179, 2006c.

Casdagli, M. C. y Weigend, A. S. *Time Series Prediction: Forecasting the Future and Understanding the Past*, capítulo Exploring the continuum between deterministic and stochastic modeling, páginas 347–366. Addison-Wesley, Reading, MA, 1994.

Chatfield, C. *Time-series Forecasting*, capítulo Calculating interval forecasts. Chapman and Hall/CRC Press, 1ª edición, 2001a.

Chatfield, C. *Time-series Forecasting*. Chapman and Hall/CRC Press, Boca Raton, 1ª edición, 2001b.

Chatfield, C. *Time-series Forecasting*, capítulo Multivariate forecasting methods. Chapman and Hall/CRC Press, 1ª edición, 2001c.

Chavent, M. y Saracco, J. On central tendency and dispersion measures for intervals and hypercubes. *Communications in Statistics - Theory and Methods*, vol. 37(9), páginas 1471–1482, 2008.

Cheung, Y.-W. An empirical model of daily highs and lows. *International Journal of Finance & Economics*, vol. 12(1), páginas 1–20, 2007.

Chouakria, A., Cazes, P. y Diday, E. *Analysis of Symbolic Data. Exploratory Methods for Extracting Statistical Information from Complex Data*, capítulo Symbolic principal component analysis, páginas 200–212. Springer, Berlín, 2000.

Chuang, C.-C. Extended support vector interval regression networks for interval input-output data. *Information Sciences*, vol. 178(3), páginas 871–891, 2008.

Colombo, A. y Jaarsma, R. A powerful numerical method to combine random variables. *IEEE Transactions on Reliability*, vol. 29(2), páginas 126–129, 1980.

Corrado, C. y Truong, C. Forecasting stock index volatility: Comparing implied volatility and the intraday high-low price range. *Journal of Financial Research*, vol. 30(2), páginas 201–215, 2007.

Corsaro, S. y Marino, M. Interval linear systems: the state of art. *Computational Statistics*, vol. 21(2), páginas 365–383, 2006.

Cover, T. y Hart, P. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, vol. 13(1), páginas 21–27, 1967.

Daw, C., Finney, C. y Tracy, E. A review of symbolic analysis of experimental data. *Review of Scientific Instruments*, vol. 74(2), páginas 916–930, 2003.

De Carvalho, F. d. A. T. Proximity coefficients between Boolean symbolic objects. En *Proceedings of the IFCS 1993*, páginas 387–394. 1994.

Deheuvels, P. La fonction de dependence empirique et ses proprietes. Un test non parametrique d'independence. *Academic Royale Belguique Bulletin de la Classe des Science*, vol. 65, páginas 274–292, 1979.

Denoeux, T. y Masson, M. Multidimensional scaling of interval-valued dissimilarity data. *Pattern Recognition Letters*, vol. 21(1), páginas 83–92, 2000.

Diamond, P. Least squares fitting of compact set-valued data. *Journal of Mathematical Analysis and Applications*, vol. 147(2), páginas 351–362, 1990.

Diday, E. *Digital Pattern Recognition*, capítulo Cluster Analysis, páginas 47–94. Springer-Verlag, Berlín, 1ª edición, 1976.

Diday, E. Introduction à l´approche symbolique en analyse des données. En *Premières Journées Symbolique-Numérique*, páginas 21–56. CEREMADE, Université Paris IX Dauphine, 1987.

Diday, E. y Esposito, F. An introduction to symbolic data analysis and the SODAS software. *Intelligent Data Analysis*, vol. 7, páginas 583–601, 2003.

Diday, E. y Noirhomme, M. *Symbolic Data and the SODAS Software*. Wiley & Sons, Chichester, 2008.

Diday, E. y Vrac, M. Mixture decomposition of distributions by copulas in the symbolic data analysis framework. *Discrete Applied Mathematics*, vol. 147, páginas 27–41, 2005.

Diebold, F. X., Gunther, T. A. y Tay, A. S. Evaluating density forecasts with applications to financial risk management. *International Economic Review*, vol. 39(4), páginas 863–883, 1998.

Do, T.-N. y Poulet, F. Kernel-based algorithms and visualization for interval data mining. En *Sixth IEEE International Conference on Data Mining - Workshops*, páginas 295 –299. IEEE, 2003.

Duarte Silva, A. P. y Brito, P. Linear discriminant analysis for interval data. *Computational Statistics*, vol. 21(2), páginas 289–308, 2006.

D'Urso, P. y Giordani, P. A least squares approach to principal component analysis for interval valued data. *Chemometrics and Intelligent Laboratory Systems*, vol. 70(2), páginas 179–192, 2004.

D'Urso, P. y Giordani, P. A robust fuzzy k-means clustering model for interval valued data. *Computational Statistics*, vol. 21(2), páginas 251–269, 2006.

El Golli, A., Conan-Guez, B. y Rossi, F. Self-organizing maps and symbolic data. *Journal of Symbolic Data Analysis*, vol. 2(1), páginas 1–8, 2004.

Engle, R. F. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, vol. 50, páginas 987–1007, 1982.

Engle, R. F. y Granger, C. W. J. Co-integration and error correction: representation, estimation, and testing. *Econometrica*, vol. 55, páginas 251–276, 1987.

Engle, R. F. y Russell, J. *Handbook of Financial Econometrics* (pendiente de publicación), capítulo Analysis of high frequency financial data. North-Holland, 2009.

Farmer, J. D. y Sidorowich, J. J. Predicting chaotic time series. *Physical Review Letters*, vol. 8(59), páginas 845–848, 1987.

Fernández-Rodríguez, F., Sosvilla-Rivero, S. y Andrada-Félix, J. Exchange-rate forecasts with simultaneous nearest-neighbour methods: evidence from the EMS. *International Journal of Forecasting*, vol. 15(4), páginas 383–392, 1999.

Ferson, S. What monte carlo methods cannot do. *Human and Ecological Risk Assessment*, vol. 2(4), páginas 990–1007, 1996.

Ferson, S., Ginzburg, L., Kreinovich, V., Longpré, L. y Aviles, M. Computing variance for interval data is np-hard. *ACM SIGACT News*, vol. 33(2), páginas 108–118, 2002.

Ferson, S., Ginzburg, L., Kreinovich, V., Longpré, L. y Aviles, M. Exact bounds on finite populations of interval data. *Reliable Computing*, vol. 11(3), páginas 207–233, 2005.

Ferson, S. y Kreinovich, V. Modeling correlation and dependence among intervals. En *Proceedings of the 2nd International Workshop on Reliable Engineering Computing*, páginas 697–704. Physica-Verlag, Heidelberg, 2006.

Fiess, N. M. y MacDonald, R. Technical analysis in the foreign exchange market: A cointegration-based approach. *Multinational Finance Journal*, vol. 3(3), páginas 147–172, 1999.

Fiess, N. M. y MacDonald, R. Towards the fundamentals of technical analysis: analysing the information content of high, low and close prices. *Economic Modelling*, vol. 19(3), páginas 353–374, 2002.

Gama, J. y Gaber, M. M., editores. *Learning from Data Streams: Processing Techniques in Sensor Networks*. Springer, New York, 2007.

Gardner, E. S. Exponential smoothing: The state of the art. *Journal of Forecasting*, vol. 4(1), páginas 1–28, 1985.

Gardner, E. S. Exponential smoothing: The state of the art. Part 2. *International Journal of Forecasting*, vol. 22(4), páginas 637–666, 2006.

Gardner, E. S. y McKenzie, E. Forecasting trends in time series. *Management Science*, vol. 31, páginas 1237–1246, 1985.

Garman, M. B. y Klass, M. J. On the estimation of security price volatilities from historical data. *The Journal of Business*, vol. 53(1), páginas 67–78, 1980.

Giacomini, R. y Granger, C. W. J. Aggregation of space-time processes. *Journal of Econometrics*, vol. 118, páginas 7–26, 2004.

Gibbs, A. L. y Su, F. E. On choosing and bounding probability metrics. *International Statistical Review*, vol. 70(3), páginas 419–435, 2002.

Gil, M. A., González-Rodríguez, G., Colubi, A. y Montenegro, M. Testing linear independence in linear models with interval-valued data. *Computational Statistics and Data Analysis*, vol. 51(6), páginas 3002–3015, 2007.

Gil, M. A., Lubiano, M. A., Montenegro, M. y López-García, M. T. Least squares fitting of an affine function and strength of association for interval-valued data. *Metrika*, vol. 56, páginas 97–111, 2002.

Gioia, F. y Lauro, C. N. Basic statistical methods for interval data. *Statistica applicata*, vol. 17(1), páginas 75–104, 2005.

Gioia, F. y Lauro, C. N. Principal component analysis on interval data. *Computational Statistics*, vol. 21(2), páginas 343–363, 2006.

Giordani, P. y Kiers, H. Three-way component analysis of interval valued data. *Journal of Chemometrics*, vol. 18(5), páginas 253–264, 2004.

Girosi, F., Jones, M. y Poggio, T. Regularization theory and neural networks architectures. *Neural Computation*, vol. 7(2), páginas 219–269, 1995.

González, L., Velasco, F., Angulo, C., Ortega, J. A. y Ruiz, F. Sobre núcleos, distancias y similitudes entre intervalos. *Inteligencia Artificial, Revista Iberoamericana de IA*, vol. 8(23), páginas 111–117, 2004.

González-Rivera, G., Lee, T.-H. y Mishra, S. Jumps in cross-sectional rank and expected returns: A mixture model. *Journal of Applied Econometrics*, vol. 23, páginas 585–606, 2008.

González-Rodríguez, G., Blanco, A., Corral, N. y Colubi, A. Least squares estimation of linear regression models for convex compact random sets. *Advances in Data Analysis and Classification*, vol. 1, páginas 67–81, 2007.

González-Rodríguez, G., Colubi, A., Coppi, R. y Giordani, P. On the estimation of linear models with interval-valued data. En *COMPSTAT 2006, Proceedings in Computational Statistics*, páginas 697–704. Physica-Verlag, Heidelberg, 2006.

Gowda, K. C. y Ravi, T. V. Divisive clustering of symbolic objects using the concepts of both similarity and dissimilarity. *Pattern Recognition*, vol. 28, páginas 1277–1282, 1995.

Granger, C. W. J. Some properties of time series data and their use in econometric model specification. *Journal of Econometrics*, vol. 16(1), páginas 121–130, 1981.

Granger, C. W. J. Implications of aggregation with common factors. *Econometric Theory*, vol. 3, páginas 208–222, 1987.

Granger, C. W. J. *Disaggregation in econometric modelling*, capítulo Aggregation of time-series variables: A survey, páginas 17–34. Routledger, London, 1990.

Grassberger, P. y Procaccia, I. Measuring the strangeness of strange attractors. *Physica D: Nonlinear Phenomena*, vol. 9(1–2), páginas 189–208, 1983.

Groenen, P. y Winsberg, S. Multidimensional scaling of histogram dissimilarities. En *Data Science and Classification, Proceedings of the IFCS 2006*, páginas 161–170. Springer, Berlín, 2006.

Groenen, P., Winsberg, S., Rodriguez, O. y Diday, E. I-Scal: Multidimensional scaling of interval dissimilarities. *Computational Statistics and Data Analsys*, vol. 51(1), páginas 360–378, 2006.

Guha, S., Koudas, N. y Shim, K. Approximation and streaming algorithms for histogram construction problems. *ACM Transactions on Database Systems*, vol. 31(1), páginas 396–438, 2006. ISSN 0362-5915.

Hall, S. G. y James, M. Combining density forecasts. *International Journal of Forecasting*, vol. 23(1), páginas 1–13, 2007.

Hébrail, G. y Hugueney, B. Symbolic representation of long time series. En *Conference on Applied Statistical Models and Data Analysis (ASMDA)*, páginas 537–542. 2001.

Hébrail, G. y Lechevallier, Y. *Selected Contributions in Data Analysis and Classification*, capítulo Building Symbolic Objects from Data Streams, páginas 83–94. Springer, 2007.

Hendry, D. F. y Hubrich, K. Forecasting economic aggregates by disaggregates. Informe de investigación 589, European Central Bank, Département des Sciences Économiques, 2006.

Holt, C. C. Forecasting trends and seasonals by exponentially weighted moving averages. Informe técnico, 1957. O.N.R. Memorandum, vol. 52, Carnegie Institute of Technology.

Hyndman, R. J. y De Gooijer, J. G. 25 years of time series forecasting. *International Journal of Forecasting*, vol. 22, páginas 443–473, 2006.

Hyndman, R. J. y Koehler, A. B. Another look at measures of forecast accuracy. *International Journal of Forecasting*, vol. 22(4), páginas 679–688, 2006.

Hyndman, R. J., Koehler, A. B., Snyder, R. D. y Grose, S. A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of Forecasting*, vol. 18(3), páginas 439–454, 2002.

Ichino, M. y Yaguchi, H. Generalized Minkowski metrics for mixed feature type data analysis. *IEEE Transactions on Systems, Man and Cybernetics*, vol. 24, páginas 698–708, 1994.

Irpino, A. "Spaghetti" PCA analysis: An extension of principal components analysis to time dependent interval data. *Pattern Recognition Letters*, vol. 27, páginas 504–513, 2006.

Irpino, A. y Verde, R. Dynamic clustering of histograms using wasserstein metric. En *COMPSTAT 2006, Proceedings in Computational Statistics*, páginas 869–876. Physica-Verlag, Heidelberg, 2006a.

Irpino, A. y Verde, R. A new Wasserstein based distance for the hierarchical clustering of histogram symbolic data. En *Data Science and Classification, Proceedings of the IFCS 2006*, páginas 185–192. Springer, Berlín, 2006b.

Ishibuchi, H. y Tanaka, H. An extension of the bp-algorithm to interval input vectors-learning from numerical data and expert's knowledge. En *Proceedings of the IEEE International Joint Conference on Neural Networks*, páginas 1588–1593. IEEE, 1991.

Ishibuchi, H., Tanaka, H. y Fukuoka, N. Discriminant analysis of multidimensional interval data and its application to chemical sensing. vol. 16(4), páginas 311–329, 1990.

Izumi, Y., Yamaguchi, T., Mabu, S., Hirasawa, K. y Hu, J. Trading rules on the stock markets using genetic network programming with candlestick chart. En *IEEE Congress on Evolutionary Computation, CEC 2006*, páginas 2362 – 2367. IEEE Computer Society, Washington, DC, USA, 2006.

Jayawardena, A. W., Li, W. K. y Xu, P. Neighbourhood selection for local modelling and prediction of hydrological time series. *Journal of Hydrology*, vol. 258(1–4), páginas 40–57, 2002.

Kalman, R. E. A new approach to linear filtering and prediction. *Journal of Basic Engineering*, vol. 83-D, páginas 95–108, 1960.

Kaplan, S. On the method of discrete probability distributions in risk and reliability calculations. Application to seismic risk assessment. *Risk Analysis*, vol. 1(3), páginas 189–195, 1981.

Kearfott, R. y Kreinovich, V., editores. *Applications of Interval Computation. Applied Optimization.* Kluwer, Dordrecht, 1ª edición, 1996.

Kearfott, R., Nakao, M., Neumaier, A., Rump, S., Shary, S. y van Hentenryck, P. Standardized notation in interval analysis. En *Proceedings of XIII Baikal International School-seminar "Optimization methods and their applications"*, vol. 4, páginas 106–113. Institute of Energy Systems SB RAS, 2005.

Keogh, E., Lonardi, S. y Chiu, B. Finding surprising patters in a time series database in linear time and space. En *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, páginas 550–556. ACM Press, Nueva York, 2002.

Kitchens, B. P. *Symbolic-Dynamics: One-Sided, Two-Sided and Countable State Markov Shifts.* Springer, Nueva York, 1ª edición, 1998.

Kohonen, T. *Self-Organizing Maps.* Springer, Berlín, 1ª edición, 1995.

Kreinovich, V., Xiang, G., Starks, S. A., Longpré, L., Ceberio, M., Araiza, R., Beck, J., Kandathi, R., Nayak, A., Torres, R. y Hajagos, J. G. Towards combining probabilistic and interval uncertainty in engineering calculations: Algorithms for computing statistics under interval uncertainty, and their computational complexity. *Reliable Computing*, vol. 12(6), páginas 471–501, 2006.

Laaksonen, S. *Symbolic Data Analysis and the SODAS Software*, capítulo People's life values and trust components in Europe: symbolic data analysis for 20-22 countries, páginas 405–419. John Wiley & Sons, Chichester, 2008.

Lee, C.-H. L., Liu, A. y Chen, W.-S. Pattern discovery of fuzzy time series for financial prediction. *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, páginas 613–625, 2006.

Lee, K. y Jo, G. Expert system for predicting stock market timing using a candlestick chart. *Expert Systems with Applications*, vol. 16(4), páginas 357–364, 1999.

Levina, E. y Bickel, P. J. The Earth Mover's Distance is the Mallows distance: Some insights from statistics. En *Proceedings of the 8th International Conference on Computer Vision*, páginas 251–256. 2001.

Li, S., Ogura, Y. y Kreinovich, Y. *Limit theorems and applications of set valued and fuzzy valued random variables*. Kluwer, Dordrecht, 2002.

Li, W. y Hyman, J. M. Computer arithmetic for probability distribution variables. *Reliability Engineering & System Safety*, vol. 85(1-3), páginas 191–209, 2004.

Li, W. y Mac Hyman, J. Representation and problem solving with the distribution envelope determination (denv) method. *Reliability Engineering and System Safety*, vol. 85(1–3), páginas 191–209, 2004.

Lima Neto, E. A. y de Carvalho, F. d. A. T. Centre and range method for fitting a linear regression model to symbolic interval data. *Computational Statistics and Data Analysis*, vol. 52, páginas 1500–1515, 2008.

Lima Neto, E. d. A., de Carvalho, F. d. A. T. y Freire, E. S. Applying constrained linear regression models to predict interval-valued data. En *KI 2005: Advances in Artificial Intelligence, 28th Annual German Conference on AI*, páginas 92–106. Springer, 2005.

Lima Neto, E. d. A., de Carvalho, F. d. A. T. y Tenorio, C. P. Univariate and multivariate linear regression methods to predict interval-valued features. En *AI 2004: Advances in Artificial Intelligence, 17th Australian Joint Conference on Artificial Intelligence*, páginas 526–537. Springer, 2004.

Limam, M. M., Diday, E. y Winsberg, S. Symbolic class description with interval data. *Journal of Symbolic Data Analysis*, vol. 1(1), páginas 1–10, 2003.

Lodwick, W. A. y Jamison, K. D. Estimating and validating the cumulative distribution of a function of random variables: Toward the development of distribution arithmetic. *Reliable Computing*, vol. 9(2), páginas 127–141, 2003.

Lütkepohl, H. *Forecasting aggregated vector ARMA processes*. Springer, Berlín, 1ª edición, 1987.

Lütkepohl, H. *New Introduction to Multiple Time Series Analysis*. Springer, Berlín, 1ª edición, 2005.

Lütkepohl, H. *Palgrave Handbook of Econometrics. Volume 1: Econometric Theory*, capítulo Vector Autoregressive Models, páginas 477–510. Palgrave Macmillan, Houndmills, 2006.

Luenberg, D. G., editor. *Linear and Nonlinear Programming*. Addison-Wesley, Reading, MA, 1984.

Maia, A. L. S., de Carvalho, F. d. A. T. y Ludermir, T. B. A hybrid model for symbolic interval time series forecasting. En *13th International Conference Neural Information Processing, ICONIP*, Lecture Notes in Computer Science, páginas 934–941. Springer, 2006a.

Maia, A. L. S., de Carvalho, F. d. A. T. y Ludermir, T. B. Symbolic interval time series forecasting using a hybrid model. En *SBRN '06: Proceedings of the Ninth Brazilian Symposium on Neural Networks*. IEEE Computer Society, Washington, DC, USA, 2006b.

Mallows, C. L. A note on asymptotic joint normality. *The Annals of Mathematical Statistics*, vol. 43(2), páginas 508–515, 1972.

Marino, M. y Palumbo, F. Interval arithmetic for the evaluation of imprecise data effects in least squares linear regression. *Statistica Applicata*, vol. 14, páginas 277–291, 2002.

Marino, M. y Palumbo, F. Interval linear regression: an application to soil permeability analysis. En *Analisi statistica multivariata per le Scienze Economico-Sociali, le Scienze Naturali e la Tecnologia, Convegno Intermedio della Società Italiana di Statistica*. 2003.

Mas, M. y Olaeta, H. *Symbolic Data Analysis and the SODAS Software*, capítulo Symbolic analysis of the time use survey in the Basque country, páginas 421–428. John Wiley & Sons, Chichester, 2008.

Maté, C. y González-Rivera, G. A PCA approach to forecast histogram-valued time series. Applications to expected returns in stock indices. En *Proceedings of the 27th International Symposium on Forecasting*. Nueva York, EEUU, 2007.

Mballo, C. y Diday, E. Kolmogorov-Smirnov for decision trees on interval and histogram variables. En *Classification, Clustering and Data Mining Applications: Proceedings of the 9th Conference of the IFCS, IFCS 2004*, páginas 341–350. Springer, Berlín, 2004.

Mballo, C. y Diday, E. The criterion of Kolmogorov-Smirnov for binary decision tree: Application to interval valued variables. *Intelligent Data Analysis*, vol. 10(4), páginas 325–341, 2006.

Meade, N. A comparison of the accuracy of short term foreign exchange forecasting methods. *International Journal of Forecasting*, vol. 18(1), páginas 67–83, 2002.

Meneses, E. y Rodríguez-Rojas, O. Using symbolic objects to cluster web documents. En *WWW '06: Proceedings of the 15th international conference on World Wide Web*, páginas 967–968. ACM, New York, NY, USA, 2006.

Michalski, R. S., Diday, E. y Step, R. *Progress in Pattern Recognition (vol 1)*, capítulo A recent advance in data analysis: Clustering objects into classes characterized by conjunctive concepts, páginas 33–56. North-Holland, Nueva York, 1ª edición, 1982.

Mok, D. M., Lam, K. y Li, W. Using daily high/low time to test for intraday random walk in two index futures markets. *Review of Quantitative Finance and Accounting*, vol. 14(4), páginas 381–397, 2000.

Moore, R. E. *Interval Analysis*. Prentice Hall, Englewood Cliffs, N.J., 1966.

Moore, R. E. *Methods and Applications of Interval Analysis*. SIAM, Philadelphia, PA, 1979.

Moore, R. E. Risk analysis without monte carlo methods. *Freiburger Intervall-Berichte*, vol. 84(1), páginas 1–48, 1984.

Morris, G. L. *Candlestick charting explained: timeless techniques for trading stocks and futures*. McGraw-Hill, Nueva York, 3ª edición, 2006.

Muñoz San Roque, A., Maté, C., Arroyo, J. y Sarabia, A. iMLP: Applying multi-layer perceptrons to interval-valued data. *Neural Processing Letters*, vol. 25(2), páginas 157–169, 2007.

Murray, D. Forecasting a chaotic time series using an improved metric for embedding space. *Physica D*, vol. 68, páginas 318–325, 1993.

Murray, M. P. A drunk and her dog: An illustration of cointegration and error correction. *The American Statistician*, vol. 48(1), páginas 37–39, 1994.

Nelsen, R. *An Introduction to Copulas*. Springer-Verlag, Nueva York, 1ª edición, 1999.

Nivlet, P., Fournier, F. y Royer, J. Interval discriminant analysis: an efficient method to integrate errors in supervised pattern recognition. En *2nd International Symposium on Imprecise Probability and their Application*, páginas 284–292. 2001.

Palumbo, F. Editorial of the special issue on interval data. *Computational Statistics*, vol. 21, páginas 183–185, 2006.

Palumbo, F. y Irpino, A. Multidimensional interval-data: Metrics and factorial analysis (*ponencia invitada*). En *Conference on Applied Statistical Models and Data Analysis (ASMDA)*, páginas 689–698. 2005.

Palumbo, F. y Lauro, C. N. A PCA for interval-valued data based on midpoints and radii. En *New Developments in Psychometrics*, páginas 641–648. Springer, Heidelberg, 2003.

Parkinson, M. The extreme value method for estimating the variance of the rate of return. *The Journal of Business*, vol. 53(1), página 61, 1980.

Pasley, A. y Austin, J. Distribution forecasting of high frequency time series. *Decision Support Systems*, vol. 37(4), páginas 501–513, 2004.

Patiño-Escarcina, R. E., Callejas Bedregal, B. R. y Lyra, A. Interval computing in neural networks: One layer interval neural networks. En *Intelligent Information Technology. Proceedings of the 7th International Conference on Information Technology*, páginas 68–75. Springer, 2004.

Pegels, C. C. Exponential forecasting: some new variations. *Management Science*, vol. 12(5), páginas 311–315, 1969.

Peña, D. *Análisis de Series Temporales*. Alianza Editorial, Madrid, 2005.

Perinel, E. Construire un arbre de discrimination binaire à partir de données imprécises. *Revue de Statistique Appliquée*, vol. 47(1), páginas 5–30, 1999.

Puzicha, J., Hofmann, T. y Buhmann, J. M. Histogram clustering for unsupervised segmentation and image retrieval. *Pattern Recognition Letters*, vol. 20(9), páginas 889–909, 1999.

Rachev, S. T. The Monge-Kantorovich mass transference problem and its stochastic applications. *Theory of Probability and its Applications*, vol. 29, páginas 647–676, 1984.

Rodríguez, O., Diday, E. y Winsberg, S. Generalization of the principal components analysis to histogram data. En *Workshop on Symbolic Data Analysis de la conferencia European Conference on Principles and Practice of Knowledge Discovery in Databases, PKDD 2004*. Pisa, Italia, 2000.

Rossana, R. J. y Seater, J. J. Temporal aggregation and economic time series. *Journal of Business and Economic Statistics*, vol. 13, páginas 441–451, 1995.

Rossi, F. y Conan-Guez, B. *Symbolic Data Analysis and the SODAS Software*, capítulo Multi-Layer Perceptrons and Symbolic Data, páginas 373–391. John Wiley & Sons, Chichester, 2008.

Rubner, Y., Puzicha, J., Tomasi, C. y Buhmann, J. M. Empirical evaluation of dissimilarity measures for color and texture. *Computer Vision and Image Understanding*, vol. 84(1), páginas 25–43, 2001.

Rubner, Y., Tomasi, C. y Guibas, L. J. The Earth Mover's Distance as a metric for image retrieval. *International Journal of Computer Vision*, vol. 40(2), páginas 99–121, 2000.

Rumelhart, D. E., Hinton, G. E. y Williams, R. J. *Parallel Distributed Processing, Volume 1: Foundations*, capítulo Learning Internal Representations by Error Propagation, páginas 318–362. MIT Press, Cambridge, 1987.

Scott, D. W. On optimal and data-based histograms. *Biometrika*, vol. 66, páginas 605–610, 1979.

Scott, D. W. *Multivariate Density Estimation*. John Wiley, Nueva York, 1992.

Silvestrini, A. y Veredas, D. Temporal aggregation of univariate linear time series model. Informe de investigación 44, Université catholique de Louvain, Département des Sciences Économiques, 2005.

Simonoff, J. S. *Smoothing methods in Statistics*. Springer-Verlag, New York, 1996.

Simonoff, J. S. y Udina, F. Measuring the stability of histogram appearance when the anchor position is changed. *Computational Statistics and Data Analysis*, vol. 23, páginas 335–353, 1997.

Sklar, A. Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, vol. 8, páginas 229–231, 1959.

Slutsky, E. The summation of random causes as the source of cyclic processes. *Econometrica*, vol. 5, páginas 105–146, 1937.

Song, Q. y Chissom, B. S. Forecasting enrollments with fuzzy time series - Part I. *Fuzzy Sets and Systems*, vol. 54(1), páginas 1–9, 1993a.

Song, Q. y Chissom, B. S. Fuzzy time series and its models. *Fuzzy Sets and Systems*, vol. 54(3), páginas 269–277, 1993b.

Song, Q., Leland, R. P. y Chissom, B. S. A new fuzzy time-series model of fuzzy number observations. *Fuzzy Sets and Systems*, vol. 73(3), páginas 341–348, 1995.

Sorjamaa, A., Reyhani, N. y Lendasse, A. Input and structure selection for k-NN approximator. En *Computational Intelligence and Bioinspired Systems, 8th International Work-Conference on Artificial Neural Networks, IWANN 2005*, Lecture Notes in Computer Science, páginas 985–992. Springer, 2005.

de Souza, R. M. C. R., de Carvalho, F. d. A. T. y Pizzato, D. F. A partitioning method for mixed feature-type symbolic data using a squared euclidean distance. En *KI 2006: Advances in Artificial Intelligence, 29th Annual German Conference on AI*, páginas 92–106. Springer, 2006.

Sturges, H. The choice of a class interval. *Journal of the American Stastical Association*, vol. 21, páginas 65–66, 1926.

Tay, A. S. y Wallis, K. F. Density forecasting: a survey. *Journal of Forecasting*, vol. 19(4), páginas 235–254, 2000.

Taylor, J. W. Exponential smoothing with a damped multiplicative trend. *International Journal of Forecasting*, vol. 19, páginas 715–725, 2003.

Teles, P. y Brito, M. P. Modelling interval time series data. En *Proceedings of the 3rd IASC World Conference on Computational Statistics & Data Analysis*. Limassol, Cyprus, 2005.

Theil, H., editor. *Applied economic forecasting*. Rand McNally, Chicago, IL, 1966.

Timmermann, A. Special issue on density forecasting in Economics and Finance (editorial). *Journal of Forecasting*, vol. 19(4), páginas 231–234, 2000.

Torkamani, M., Asgari, J. y Lucas, C. Estimating strange attractor's dimension in very noisy data, application to FOREX time series. En *International Conference on Information and Communication Technologies, ICTTA '06*, páginas 1944 – 1947. IEEE Computer Society, Washington, DC, USA, 2006.

Tukey, J. W., editor. *Exploratory Data Analysis*. Addison-Wesley, Reading, MA, 1977.

Tzanetakis, G. y Cook, P. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, vol. 10(5), páginas 293–302, 2002.

Tzanetakis, G., Ermolinskyi, A. y Cook, P. Pitch histograms in audio and symbolic music information retrieval. *Journal of New Music Research*, vol. 32(2), páginas 143–152, 2003.

Verde, R. y Irpino, A. *Selected Contributions in Data Analysis and Classification*, capítulo Dynamic clustering of histogram data: using the right metric, páginas 123–134. Springer, 2007.

Wand, M. P. Data-based choice of histogram bin width. *The American Statistician*, vol. 51(1), páginas 59–64, 1997.

Webby, R. y O'Connor, M. Judgemental and statistical time series forecasting: a review of the literature. *International Journal of Forecasting*, vol. 12(1), páginas 91–118, 1996.

Weigend, A. S. y Shi, S. Predicting daily probability distributions of S&P500 returns. *Journal of Forecasting*, vol. 19(4), páginas 375–392, 2000.

Williams, W. H. y Goodman, M. L. A simple method for the construction of empirical confidence limits for economic forecasts. *Journal of the American Statistical Association*, vol. 66(336), páginas 752–754, 1971.

Williamson, R. C. *Probabilistic Arithmetic*. Tesis Doctoral, University of Queensland (Australia), 1989.

Williamson, R. C. y Downs, T. Probabilistic arithmetic. I. Numerical methods for calculating convolutions and dependency bounds. *International Journal of Approximate Reasoning*, vol. 4(2), páginas 89–158, 1990.

Winkler, R. L. y Makridakis, S. The combination of forecasts. *Journal of the Royal Statistical Society: Series A*, vol. 146(2), páginas 150–157, 1983.

Winters, P. R. Forecasting sales by exponentially weighted moving averages. *Management Science*, vol. 6, páginas 324–342, 1960.

Xiang, G. Fast algorithm for computing the upper endpoint of sample variance for interval data: Case of sufficiently accurate measurements. *Reliable Computing*, vol. 12(1), páginas 59–64, 2006.

Yakowitz, S. Nearest-neighbour methods for time series analysis. *Journal of time series analysis*, vol. 8(2), páginas 235–247, 1987.

Yang, D. y Zhang, Q. Drift independent volatility estimation based on high, low, open, and close prices. *The Journal of Business*, vol. 73(3), páginas 477–492, 2000.

Yule, G. U. On a method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers. *Philosophical Transactions of the Royal Society of London A*, vol. 226, páginas 267–298, 1927.

Zadeh, L. Fuzzy sets. *Information and Control*, vol. 8, páginas 338–353, 1965.

Zellner, A. y Tobias, J. A note on aggregation, disaggregation and forecasting performance. *Journal of Forecasting*, vol. 19, páginas 457–469, 2000.

Zhang, G., Patuwo, B. E. y Hu, M. Y. Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting*, vol. 14(1), páginas 35–62, 1998.

Zhang, G. P. Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, vol. 50(1), páginas 159–175, 2003.

Zhao, Y., He, Q. y Chen, Q. An interval set classification based on support vector machines. En *Joint International Conference on Autonomic and Autonomous Systems and International Conference on Networking and Services*, páginas 81–87. IEEE, 2005.

Zhao, Y., Liu, Y. y He, Q. One kind of interval support vector regression algorithm and its application in web information mining. En *Proceedings of the 25th Chinese Control Conference*, páginas 825–829. IEEE, 2006.