I

UNIVERSIDAD POLITÉCNICA DE VALENCIA

Escuela Técnica Superior de Ingenieros de Caminos, Canales y Puertos

Departamento de Ingeniería Hidráulica
y Medio Ambiente

TESIS DOCTORAL

PREDICCIÓN A CORTO PLAZO DE LA DEMANDA

DE AGUA URBANA EN ÁREAS DENSAMENTE POBLADAS

Presentado por:

César Alejandro Espinoza Rodríguez

Director: Juan B. Marco Segura

Enero del 2010

Dedicatoria

A mi mujer e hija. A mis padres y hermanas.

Agradecimientos

Como autor de esta tesis quiero agradecer:

A EMIVASA-Aguas de Valencia por el financiamiento e información facilitada para desarrollar este trabajo.

Al Consejo Nacional de Ciencia y Tecnología de México por el financiamiento brindado.

A mi director de tesis, Juan Bautista Marco Segura, por ver en un estudiante anónimo a un colaborador, por depositar en mí su confianza, por brindarme la oportunidad de colaborar en un proyecto tan interesante, por su orientación y sus siempre adecuados comentarios y ánimos.

Imprescindible agradecer al personal del Departamento de Ingeniería Hidráulica y medio ambiente, profesores y personal administrativo, todos fueron un apoyo para mí.

A mis compañeros estudiantes del departamento, fueron siempre una fuente de consejos y ejemplos. Sin su compañía y apoyo estos años habrían sido muy duros, en cambio han sido una grata experiencia, por tantos cafés, almuerzos y tertulias llenas de risas. No particularizo, todos han sido importantes.

A la colonia de compatriotas mexicanos en Valencia, han sido mi familia todo este tiempo.

Al Club Deportivo Runnersworld Valencia por trascender a lo académico y darme la oportunidad de disfrutar de grandes momentos y experiencias durante varios años.

A mi padre por su ejemplo, alma de minero. A mi madre por estar siempre al pendiente a pesar de la distancia.

A Mabel, mi mujer, por su apoyo en los buenos y malos ratos, por su paciencia y amor.

A la pequeña Andrea, por alegrar cada día.

Resumen

Ante un escenario donde el recurso agua es limitado y con una sociedad que la demanda cada vez con más garantías, la ingeniería es exigida a desarrollar técnicas y metodologías eficientes para asegurar que el vital líquido sea entregado en óptimas condiciones de calidad y cantidad a los usuarios domésticos, comerciales e industriales que conforman el conjunto de abonados de una ciudad. Cada tipo de usuario demanda el agua en diferentes escalas temporales y de cantidad, pero el conjunto de ellos consumiendo agua a la vez generan la demanda global de una ciudad.

Los operadores de los sistemas de abastecimiento y distribución de agua potable están obligados a gestionar sus operaciones de tal manera que el conjunto de abonados cuente con el servicio en el momento que lo demanden. La experiencia que acumula el personal de operación se vuelve fundamental para que este objetivo se cumpla ya que son capaces de predecir con gran precisión las demandas futuras.

En las búsqueda de predicciones con un fundamento matemático y estadístico sólido, hemos desarrollado este trabajo en el cual se han revisado las metodologías más destacadas que se han utilizando a lo largo de las últimas décadas para modelar y predecir la demanda de agua urbana en áreas densamente pobladas, encontrando que los modelos estocásticos del tipo ARIMA son la base de las principales metodologías. Sin embargo, encontramos también que los modelos existentes están desarrollados y pensados para ciudades en las cuales la demanda presenta un patrón con poca variabilidad a lo largo del ciclo anual y que solo es alterado principalmente por componentes climáticas y meteorológicas. Este no es el caso de muchas de las ciudades europeas, españolas y mediterráneas, que presentan una gran variabilidad derivada de patrones sociológicos y donde las componentes climáticas son poco relevantes. Esta variabilidad es generada por eventos puntuales que perturban el proceso de demanda y que cuando ocurren alteran los patrones repetitivos esperados. La ignorancia de este tipo de eventos en un escenario de predicción a corto plazo de la demanda mediante modelos estocásticos, resulta primero, en predicciones erradas para el momento de la ocurrencia de un evento perturbador puntual, y segundo, en predicciones distorsionadas hasta un determinado orden después de la ocurrencia del evento. Es esperable pues, que la ignorancia de este tipo de eventos disminuya la eficiencia de los modelos estocásticos.

A lo largo de este trabajo, se aplicó también la metodología de las redes neuronales para evaluar su eficiencia para modelar y predecir la demanda de agua. Su desempeño es comparado con los modelos antes menciona-

dos, encontrando que es posible obtener resultados muy similares con ambas metodologías, aún y cuando sus postulados parten desde puntos muy diferentes.

En este trabajo de tesis se propone una metodología para incorporar implícitamente en un modelo estocástico de predicción, el conjunto de eventos sociológicos perturbadores del proceso de demanda. La metodología es probada en un caso real para la ciudad de Valencia, España, encontrando que se consigue mejorar los resultados de las predicciones que obtienen tanto los modelos ARIMA como las redes neuronales. Se obtienen unos errores que están muy cercanos a un ruido blanco con una menor varianza residual, lo cual nos indica que la metodología propuesta capta tanto la variabilidad sistemática de la serie, así como la variabilidad irregular generada por los patrones sociológicos.

El esquema de predicción propuesto ha mostrado ser una buena herramienta para realizar predicciones de la demanda de agua a corto plazo para el caso analizado y que podría ser fácilmente implementable en un sistema que opere en tiempo real.

Abstract

In an scenario where water resources are limited and with a developing society demanding water with more and more guarantees every day, engineering is required to develop efficient techniques and methodologies to assure that the vital liquid is delivered in optimum conditions of quality and quantity for the domestic, commercial and industrial users that integrate the set of subscribers of a city. Every type of users demands water at different time and quantity scales, but the set of them consuming water jointly generates a city's global water demand.

Potable water demand and distribution facilities operators are obligated to manage its operations such that all subscribers receive the service at the moment they demand it. Experience accumulated in operations personnel becomes fundamental to achieve this objective, because they are capable to predict with great precision future water demands.

In the search for forecastings with strong mathematic and statistic foundations, we have developed this work where the more important methodologies that have been used in the last decades to model and predict urban water demand in densely populated areas were reviewed. It was found that stochastic models of the type of ARIMA are the bases for the more important reviewed methodologies. However, we also found that existing models were developed and thought for cities where water demand shows a cyclic pattern with little variability along the annual cycle, and is only affected mainly by climatic and meteorological components. This is not the case for many European, Spanish and Mediterranean cities, that show great variability derived from sociologic patterns and where climatic components are little relevant. This variability is generated for punctual events that disturb the water demand process and that when these occur the expected repetitive patterns change. The ignorance of these events in a short-term water demand forecasting scenario with stochastic models, first result, in wrong predictions at the moment of occurrence of a punctual disturbing event, and second in, distorted predictions until a determined order after the occurrence of the event. It is expected that the ignorance of these types of events diminishes the efficiency of the stochastic models.

Along this work, neural networks methodology was also applied to evaluate its efficiency to model and predict water demand. This performance is compared with the aforementioned models, finding that it is possible to obtain similar results with both methodologies even when their origins are from different points.

In this thesis work a methodology is proposed to incorporate implicitly in a stochastic prediction model the group of sociologic events that disturb the water demand process. The methodology is tested in a real case for the city of Valencia, Spain, finding predictions performances that overcome the results obtained not only by ARIMA models but also by neural networks. The errors obtained are very close to a white noise with a minor residual variance, which tells us that the proposed methodology captures not only the systematic variability of the series, but also the irregular variability generated by sociological patterns.

The proposed forecasting prediction outline has shown to be a good tool to predict short-term water demand for the analyzed case, and could easily be implemented in a system operating in real time.

Resum

Davant d'un escenari on el recurs aigua es limitat i en una societat que la sol·licita cada vegada amb més garanties, l'enginyeria és exigida a desenvolupar tècniques i metodologies eficients per assegurar que el vital líquid s'entregui en òptimes condicions de qualitat i quantitat als usuaris domèstics, industrials i comercials que conformen el conjunt d'abonats d'una ciutat. Cada tipus d'usuari sol·licita l'aigua en diferents escales temporals i de quantitat, però tots aquests consumint aigua conjuntament genera la demanda global d'una ciutat.

Els operadors dels sistemes d'abastiment i distribució d'aigua potable estan obligats a gestionar les seves operacions de tal forma que el conjunt d'abonats pugui contar amb el servei en el moment que el sol·liciten. L'experiència que acumula el personal d'operació es fonamental per a que aquest objectiu es compleixi ja que son capaços de preveure amb gran precisió les demandes futures.

En la recerca de prediccions amb un fonament matemàtic i estadístic sòlid, hem desenvolupat el següent treball en el qual s'han revisat les metodologies més destacades que s'han vingut utilitzant al llarg de les últimes dècades per a modelar i preveure la demanda d'aigua urbana en àrees densament poblades, trobant que els models estocàstics del tipus ARIMA son la base de les principals metodologies. No obstant, trobem també que els models existents estan desenvolupats i pensats per a ciutats en les quals la demanda presenta un patró amb poca variabilitat al llarg del cicle anual i que solament és alterat principalment per components climàtiques i meteorològiques. Aquest no és el cas de moltes ciutats europees, espanyoles i mediterrànies, que presenten una gran variabilitat derivada de patrons sociològics i on les components climàtiques son poc rellevants. Aquesta variabilitat es generada per esdeveniments puntuals que pertorben el procés de demanda i que quan ocorren alteren els patrons repetitius esperats. La ignorància d'aquests tipus d' esdeveniments en un escenari de predicció a curt termini de la demanda per mitjà de models estocàstics, resulta primer, en prediccions errònies per al moment de l'ocurrència d'un esdeveniments pertorbador puntual, i segon, en prediccions distorsionades fins a un determinat ordre després de l'ocurrència del esdeveniment. Es d'esperar per tant, que la ignorància d'aquest tipus d'esdeveniments redueixi l'eficiència dels models estocàstics.

Al llarg del següent treball, s'ha aplicat també la metodologia de les xarxes neuronals per a avaluar la seva eficiència per a modelar i preveure la demanda d'aigua. El seu acompliment és comparat amb els models abans

mencionats, trobant que és possible obtenir resultats molt similars amb ambdues metodologies, fins i tot quan els seus postulats parteixen des de punts molt diferents.

En aquest treball de tesis es proposa una metodologia per a incorporar implícitament en un model estocàstic de predicció, el conjunt d'esdeveniments sociològics pertorbadors del procés de demanda. La metodologia és comprovada en un cas real per a la ciutat de Valencia, Espanya, concloent que s'aconsegueix millorar els resultats de les prediccions que s'obtenen tant amb els models ARIMA com en les xarxes neuronals. S'obtenen uns errors que estan molt propers a un soroll blanc amb una menor variància residual, el qual ens indica que la metodologia proposta capta tant la variabilitat sistemàtica de la sèrie com la variabilitat irregular generada pels patrons sociològics.

L'esquema de predicció proposat ha mostrat ser un bona eina per a realitzar prediccions de la demanda d'aigua a curt termini per al cas analitzat i que podria ser fàcilment implementat en un sistema que opera en temps real.

X

Índice general

Índice de figuras

Índice de cuadros

Parte I

Introducción

Capítulo 1

Introducción

If you can't describe what you are doing as a process,
you don't know what you're doing.
W. Edwards Deming
1900-1993
Estadístico estadounidense, profesor universitario.

The distinction between the past, present and future
is only a stubbornly persistent illusion.
Albert Einstein
1879-1955
Físico Alemán.

En este capítulo se describe la motivación de la que surge la idea de desarrollar este trabajo de tesis, se enumeran los objetivos que se persiguen y finalmente se detalla la estructura que se ha empleado en la tesis resumiendo el contenido de los diferentes capítulos que la componen, resaltando los puntos de unión de unos capítulos con otros.

1.1. Motivación de la investigación

En las últimas décadas las empresas gestoras de sistemas de agua potable han ido evolucionando para ofrecer a sus clientes un servicio de calidad a un mínimo coste, tanto medioambiental como económico. Esta evolución va de la mano de los cambios en las normativas que tienen como fin una gestión sostenible de un bien tan escaso como es el agua.
La problemática del agua no es un asunto nuevo en la historia de España

ni tampoco lo es en la comunidad Valenciana. Las constantes en esta problemática son siempre una desigual distribución temporal y espacial de los recursos hídricos disponibles para todos los usos. Hasta no hace mucho tiempo la solución se abordaba mediante la realización de obras que aseguraran un mayor caudal disponible, acorde con el aumento de la demanda que potenciara las zonas demandantes. En lo que respecta a la demanda de agua urbana, la concentración de población en núcleos urbanos cada vez más poblados, ha provocado que la demanda aumente a una mayor tasa que en otros sectores, requiriendo además una mayor calidad y eficiencia.

Uno de los principales factores que intervienen en la mejora de las funciones de operación y planificación de un sistema de abastecimiento, es la disponibilidad de modelos de demanda que cuenten con suficiente precisión. La modelación de la demanda puede ser abordada desde distintos enfoques. Uno de ellos es el de la modelación estocástica que utiliza los datos históricos de la demanda global del sistema de forma univariada, y que en determinados momentos se complementan con registros de variables climáticas y/o determinísticas que aporten información relevante del comportamiento de la demanda. Mediante estos modelos es posible obtener predicciones de la demanda (horaria, diaria, mensual, anual) de agua. En la parte correspondiente al diseño, la predicción se aplica para proyectar y planear nuevos desarrollos o expansiones. En la operación y gestión, la predicción cobra especial relevancia en sistemas con déficit en sus suministros o en sistemas con variabilidades máximas y mínimas que someten al sistema a un estrés mucha veces evitable.

El sistema de abastecimiento de agua potable y distribución típico, consiste de una o varias fuentes de abastecimiento desde donde el agua es conducida por gravedad, aunque muchas veces lo es también mediante bombeo. Posteriormente su calidad es mejorada mediante potabilización y es almacenada cortos periodos de tiempo antes de ser entregada a la red de distribución. De nueva cuenta, en ocasiones tiene que ser bombeada para asegurar que los usuarios reciben el agua para su uso en adecuadas condiciones de presión y calidad, con tiempos de residencia mínimos en los tanques de almacenamiento y en la red de distribución. En cada una de las etapas se consumen energía eléctrica y productos químicos. Es evidente por lo tanto que al contar con buenas predicciones de la demanda se consigue que la operación y gestión sean más eficientes, porque aporta argumentos para una más efectiva toma de decisiones, por consecuencia un menor coste económico y medioambiental.

1.2. Objetivo

El presente trabajo de tesis tiene como principal objetivo proponer, analizar y comparar el desempeño con otras técnicas, de un modelo estocástico para la estimación a corto plazo de la demanda global de agua potable en sistemas de abastecimiento y distribución. Una vez que se hayan analizado los modelos de demanda más frecuentemente utilizados, el objetivo general de la tesis se subdivide en los siguiente objetivos particulares:

1. Proponer mejoras a los modelos de demanda actualmente más en uso con fundamento en las metodologías presentadas en Maidment and Parzen (1984a), Maidment et~al. (1985).

2. Evaluar los resultados obtenidos de los modelos de demanda con fundamento en técnicas estocásticas.

3. Demostrar la validez de técnicas de series temporales mediante la comparación de resultados con técnicas más recientes como son las redes neuronales artificiales.

4. Proponer un modelo de demanda que se ajuste a las características de ciudades con alta variabilidad estacional y fuertemente afectadas por variables deterministas.

1.3. Contenido y estructura de la tesis

El contenido de la tesis como tal inicia en el Capítulo 2, donde se presentan el conjunto de conceptos matemáticos y estadísticos en los que se fundamentan las distintas técnicas que se han utilizado en la modelación y predicción de la demanda de agua potable. Los conceptos se presentan con un enfoque más profundo al que se abordará en las secciones posteriores.

Una vez definidos los conceptos desde los que partiremos en nuestra investigación, el Capítulo 3 contiene una extensa revisión del estado del arte de la modelación y predicción de la demanda de agua potable. Se presenta a modo de resumen, la evolución de las distintas técnicas utilizadas para las diferentes escalas, ya sea mensual, diaria u horaria. Se destacan de cada una de ellas, tanto sus principales aportaciones como sus limitaciones más evidentes.

La línea de investigación que desarrolla esta tesis surge de la revisión del estado actual de los modelos de demanda que se presentan en el Capítulo 3. Al concluir el desarrollo de este capítulo se encuentra que los principales modelos presentados tienen un mejor desempeño en ciudades de los Estados Unidos de Norteamérica –donde fueron desarrollados y probados–, ya que el patrón que sigue la demanda a lo largo del año difiere del de las ciudades del tipo de Valencia, mediterráneas, europeas. Los errores en la modelación y predicción de la demanda de agua se magnifican durante la ocurrencia de eventos puntuales como son las festividades locales y nacionales. Al producirse estos eventos, la estimación de los parámetros de los modelos –de series temporales y de redes neuronales– se ven afectados y producen peores ajustes y predicciones. Los capítulos 4, 5 y 6 se dedicarán al desarrollo y explicación de la metodología propuesta para incorporar implícitamente en modelos estocásticos de predicción y análisis de intervención, las variables deterministas que alteran el proceso de demanda.

El Capítulo 7 es una aplicación para una serie temporal de la ciudad de Valencia, España, de las metodologías que se han utilizado más frecuentemente en la modelación de la demanda de agua potable y que fueron presentadas en el capítulo que precede a este. Se ha ajustado un modelo de series temporales tipo ARIMA con variables exógenas y sin ellas para evaluar su desempeño.

Ya que en los últimos años han sido extensivamente utilizadas, en el Capítulo 8, se construyeron un conjunto de redes neuronales con diferentes arquitecturas del tipo *feedforward* en busca de un mejor desempeño. Los resultados de las metodologías de los Capítulos 7 y 8 son comparadas al final de éste último capítulo y se evalúa el impacto para ambas de las variables deterministas que alteran el proceso de demanda.

En el Capítulo 9 se aplica la metodología desarrollada en los capítulos 4, 5 y 6 con la cual se trata la alta variabilidad de las series de demanda de agua incorporándola implícitamente mediante el análisis de intervención.

Finalmente en el Capítulo 10 resume el trabajo realizado a lo largo de la tesis y concluye las más importantes aportaciones a la metodología de modelación de la demanda de agua. Se proponen también un conjunto de líneas de investigación que podrían aportar mejoras interesantes.

Parte II

Marco Teórico

Capítulo 2

Marco Teórico

Predecir es barato, predecir mal cuesta caro
Proverbio Chino

A lo largo de este capítulo presentaremos el conjunto de teorías y metodologías en las que se fundamentan las distintas técnicas que se han utilizado en la modelación y predicción de la demanda de agua potable. Intentaremos presentar la información con un sentido más profundo al que se abordará en las secciones posteriores. Este conjunto de teorías acotan el área de trabajo de nuestra investigación posterior, por lo que trataremos de definir un conjunto de conceptos y proposiciones que nos permitan abordar el problema desde una base metodológica sólida.

2.1. Introducción a las series temporales

2.1.1. *Descripción y análisis preliminar*

Consiste en describir las principales características de la serie utilizando estadísticos resumen y/ó métodos gráficos. Cuando obtenemos una serie temporal, el primer paso en el análisis es usualmente graficar las observaciones a lo largo del tiempo para obtener medidas descriptivas simples de las principales características de la serie. Este gráfico mostrará rasgos de la serie tales como tendencia, estacionalidad, outliers, discontinuidades y cambios graduales o bruscos de las propiedades de la serie. Por todo lo anterior, el graficado de los datos es vital para describir los datos y como ayuda en la formulación de un modelo sensible. El graficado de los datos puede ayudar a

decidir si los datos necesitan ser transformados antes del análisis. Por ejemplo, si se observa una tendencia ascendente en el gráfico de los datos y la varianza aumenta conforme lo hace el valor medio. Entonces una transformación puede ser recomendable para estabilizar la varianza. La transformación de los datos también se recomienda en el caso de que las observaciones presenten algún sesgo. En este caso se recomienda una normalización de los datos. Finalmente, si los efectos estacionales se observan multiplicativos, entonces sería deseable transformar los datos para convertir en aditivo el efecto estacional. Una transformación también es recomendable cuando los datos presentan asimetría y/ó cuando los cambios en la varianza son severos.

2.1.2. *Modelación*

Un modelo es una representación matemática de la realidad y puede ser usado para varios propósitos, incluyendo los siguientes:

1. Dar una descripción que nos ayude a modelar la variación sistemática y la variación no explicada (o componente error).
2. Al describir la variación sistemática, un modelo puede ayudar a confirmar o rechazar relaciones teóricas sugeridas a priori, o puede ayudar también para darle un sentido físico al proceso subyacente generador de datos.
3. La parte sistemática del modelo facilita la computación de buenas predicciones puntuales, mientras que la descripción de la variación no explicada ayudará a computar los intervalos de predicción.

Consiste en encontrar un modelo estadístico adecuado para describir el proceso generador de los datos de la serie, por lo que tiene como meta fundamental descubrir el patrón en la serie de datos históricos y extrapolar ese patrón hacia el futuro. Un modelo univariado se basa solamente en valores pasados de esa variable, mientras que un modelo multivariado se basa en valores pasados y presentes de otras variables (predictoras). En el último caso, la variación en una serie puede ayudar a explicar la variación en otra serie.
Debe tenerse claro que un modelo ajustado es una aproximación a los datos, por lo que existirán desviaciones del modelo en mayor o menor medida dependiendo de la complejidad del fenómeno que está siendo modelado y en la complejidad y la precisión del modelo.
La construcción de un modelo estadístico usualmente tiene tres etapas principales:

1. Especificación del modelo (o identificación del modelo)

2. Ajuste del modelo (o estimación del modelo)

3. Validación del modelo

Una vez definido, el modelo puede ser modificado, mejorado, ampliado o simplificado como respuesta a los resultados obtenidos, por lo que la construcción de un modelo es un proceso interactivo e iterativo. Existe una gran variedad de métodos disponibles que varían en precisión, alcance, horizonte de predicción, etc. Algunos puntos claves a la hora de decidirse por un método a aplicar en cada situación son, cuánto confiamos en el método, cuánto peso darle al método y cuánta modificación se requiere para incorporar el juicio personal antes que las predicciones sean usadas como una base para planear acciones futuras. Es importante saber también que se requieren distintos tipos de modelos para cada situación y para los fines deseados, es decir que el mejor modelo para predicción (out of sample) puede no ser el mejor modelo para describir la evolución de una serie temporal.

2.1.3. *Predicción*

Un problema muy importante en muchas áreas es el de estimar valores futuros de una serie temporal. Existen una gran variedad de procedimientos de predicción y es importante tener claro que ningún método es universalmente aplicable. En cambio el analista debe elegir el procedimiento que sea más apropiado para un grupo de condiciones. Es importante tener en mente también que la predicción es una forma de extrapolación, con todos los peligros que implica.
Los métodos de predicción en términos generales se pueden clasificar en tres grupos.

Subjetivos

Las predicciones se hacen con una base subjetiva usando el juicio, la intuición, el conocimiento de la serie, etc. Estos métodos no serán descritos ni utilizados aquí ya que estamos buscando métodos más objetivos.

Univariados

Las predicciones de una variable se basan en un modelo ajustado solamente a observaciones pasadas y presentes de una serie temporal, así que $\hat{x}_N(h)$ depende solo de valores de x_N, x_{N-1},..., posiblemente aumentado por una función simple en el tiempo, tal como una tendencia lineal global. Esto significa, por ejemplo, que las predicciones univariadas del nivel de una presa se basarán enteramente en los niveles pasados de la misma y no tomarían en cuenta otros factores hidrológicos o climáticos como podrían ser el caudal en el río que la abastece o la precipitación. Entre estos podemos mencionar los de suavizado exponencial (ES, Exponential Smoothing) que son un grupo general de procedimientos de predicción que se basa en la simple actualización de ecuaciones que implica disminuir exponencialmente los pesos que aporta cada observación pasada para calcular las predicciones. La forma más básica se llama simple exponential smoothing o SES (Brown, 1963) y puede ser utilizado en series que no muestren estacionalidad ni tendencia o deben ser removidas antes de utilizar este método. Las series temporales que contengan tendencia y no variación estacional pueden ser tratadas con el método de suavizado exponencial de Holt´s de dos parámetros, mientras que el método con el que usualmente se trata a las series con tendencia y variación estacional es el de suavizado exponencial Holt-Winters de tres parámetros. El planteamiento de este método es el de generalizar la ecuación de los SES introduciendo términos para la tendencia y la variación estacional, lo cuales también son actualizados por suavizado exponencial.

La familia de modelos ARIMA (Box et~al., 1976), también llamados modelos Box-Jenkins es una herramienta de predicción muy importante y la base de muchas ideas fundamentales en el análisis de series temporales. La autocovarianza y la función de autocorrelación juegan un rol importante en la selección de modelos ARIMA. La abreviación ARIMA significa por sus siglas en inglés, *Autoregressive Integrated Moving Average* o modelo autoregresivo integrado de media móvil. La palabra integrado produce confusión pero se refiere a la diferenciación de la serie de datos. Se puede definir un modelo como autoregresivo si la variable endógena de un periodo t es explicada por las observaciones de ella misma correspondientes a periodos anteriores añadiéndose un término de error. Los modelos autorregresivos se abrevian con la palabra AR tras la que se indica el orden del modelo: AR(1), AR(2),...etc. El orden del modelo expresa el número de observaciones retrasadas de la serie temporal analizada. Los tres números que se escriben después de ARIMA se refieren al orden del proceso AR o autoregresivo, al orden de diferenciación y al orden del proceso MA o de media móvil. Un modelo denominado de media móvil es aquel que explica el valor de una determinada variable en un periodo t en función de un término independiente y una sucesión de errores

correspondientes a períodos precedentes, ponderados convenientemente. Al igual que en el caso de los modelos autorregresivos, el orden de los modelos de medias móviles se indica como MA(1), MA(2),...etc. La combinación de modelos AR, MA y ARMA son casos especiales de los modelos ARIMA.

Multivariados

En muchas ocasiones las observaciones se registran simultáneamente para varias series temporales. Por citar un ejemplo, en meteorología se registran la temperatura, presión barométrica y lluvia en el mismo sitio para la misma secuencia de puntos en el tiempo. En economía, muchas medidas diferentes de actividades económicas son típicamente registradas a intervalos regulares. Las variables podrían ser el Índice de Precios al Consumo, el nivel de desempleo, etc. Dados estos datos se podría intentar desarrollar un modelo multivariado para describir las interrelaciones entre las series, y usar este modelo para hacer predicciones. Con datos de series temporales, el proceso de modelación se complica por la necesidad no solo de modelar la interdependencia entre las series, sino también la dependencia dentro de cada serie.

Ajustar un modelo multivariado a una serie temporal no es fácil, aun con los grandes avances en recursos computacionales que se han conseguido en los últimos años. Los modelos univariados pueden ser útiles para muchos fines y es obvio que los modelos multivariados ofrecen mucho en la parte de la comprensión de la estructura subyacente de un sistema dado y también en ocasiones en la obtención de mejores predicciones. Sin embargo no todo el tiempo esta afirmación es cierta. Los modelos multivariados normalmente consiguen un mejor ajuste que los modelos univariados, pero hay varias razones por las cuales la obtención de mejores ajustes no necesariamente se refleja –aunque en ocasiones lo hacen– en mejores predicciones. Algunas razones se mencionan a continuación:

1. Al existir más parámetros por estimar hay más oportunidades de variación muestral que incrementa la incertidumbre y afecta las predicciones.

2. Al existir más variables a medir hay más oportunidades para que errores y/ó outliers pasen desapercibidos.

3. No todo el tiempo las series observadas multivariadas son adecuadas para ajustar un modelo multivariado.

4. El cálculo de predicciones de una variable dependiente puede requerir valores futuros de variables explicativas que no estarán disponibles al momento en que la predicción tenga que hacerse. Las variables explicativas entonces deben ser predichas de alguna manera antes de que pueda ser hecha la predicción de la variable dependiente y esto inevitablemente conlleva una reducción en la precisión. Si las predicciones de las variables explicativas tienen una pobre precisión, entonces muy probablemente las predicciones resultantes de la variable dependiente tendrán peor precisión que las realizadas con algún modelo univariado.

2.1.4. *Control*

Si se consigue construir un modelo que entregue buenas predicciones, se puede permitir al analista pasar a controlar un proceso dado. Es decir, las predicciones aportan elementos para tomar decisiones de control óptimo de un proceso. Por ejemplo, para el caso de una serie temporal que mide las presiones registradas en una red de abastecimiento de un sector, el propósito del análisis podría ser mantener las presiones operando dentro de un rango adecuado para las tuberías existentes en el sector. Los problemas de control están muy relacionados con la predicción en muchas situaciones. Por ejemplo, si se predice que las presiones del sector van a alejarse del objetivo, entonces se deben tomar medidas correctivas.

2.2. Definición y conceptos básicos de Series Temporales

2.2.1. *MODELOS ARIMA*

Proceso estocástico y sus propiedades

Tanto los modelos autorregresivos como los de medias móviles que más adelante serán analizados necesitan para su comprensión la introducción del concepto de proceso estocástico.

Un proceso estocástico es una sucesión de variables aleatorias $X(t)$ ordenadas a lo largo de una dimensión física, pudiendo tomar t cualquier valor entre -∞ y ∞. El subíndice t representa el paso del tiempo.

Cada una de las variables $X(t)$ que configuran un proceso estocástico tendrán su propia función de distribución con sus correspondientes momentos. Así mismo, cada par de esas variables tendrán su correspondiente función de distribución conjunta y sus funciones de distribución marginales. Esto mismo ocurrirá, ya no para cada par de variables, sino para conjuntos más amplios de las mismas. De esta forma, para caracterizar un proceso estocástico deberíamos especificar las funciones de distribución conjunta de cualquier conjunto de variables. El análisis de series temporales es diferente de otros problemas estadísticos por el hecho de que la serie temporal observada es usualmente la única que podremos observar. En otras palabras, $x(t)$ es usualmente la única observación que tendremos para $X(t)$. Sin embargo el análisis de series temporales se ocupa de evaluar las propiedades subyacentes del modelo probabilístico desde esta serie temporal observada, aunque sea la única realización que se podrá observar.

Habitualmente resulta complejo llegar a conocer esas funciones de distribución, de forma que para caracterizar un proceso estocástico, basta con especificar la media y la varianza para cada $x(t)$ y la covarianza para variables referidas a distintos valores de t:

Media. La media $\mu(t)$ se define para todo t por

$$\mu(t) = E[X(t)]$$

Varianza. La varianza $\sigma^2(t)$ se define para todo t por

$$\sigma^2(t) = Var[X(t)]$$

Autocovarianza. La varianza sola no es suficiente para especificar el segundo momento de una secuencia de variables aleatorias. Habitualmente se define la función de autocovarianza (acv.f.) $\gamma(t_1, t_2)$ como la covarianza de $X(t_1)$ con $X(t_2)$,

$$\gamma(t_1, t_2) = E\{[X(t_1) - \mu(t_1)][X(t_2) - \mu(t_2)]\}$$

Es fácil ver que la varianza es un caso especial de la acv.f. cuando $t_1 = t_2$.

Proceso Estacionario

Una clase importante de procesos estocásticos son aquellos que son estacionarios. Una serie temporal se denomina estacionaria definida en sentido estricto (Chatfield, 2004), si la distribución conjunta de $X(t_1), ..., X(t_k)$ es la misma distribución conjunta de $X(T_1 + \tau), ..., X(T_k + \tau)$ para todo $t_1, ..., t_k, \tau$. En otras palabras, desplazar el origen en una cantidad τ no tiene efecto en la distribución conjunta, la cual dependerá solamente de los intervalos entre $t_1, t_2, ..., t_k$. Esta definición es válida para cualquier valor de k. En particular si $k = 1$, la estacionaridad en sentido estricto implica que la distribución de $X(t)$ es la misma para todo t, siempre que los dos primeros momentos sean finitos. Suponemos que

$$\mu(t) = \mu$$

lo que significa que es constante y no depende del valor de t.

$$\sigma^2(t) = E[(x(t) - \mu)^2] = \sigma^2$$

La segunda suposición de estacionaridad es que la varianza del proceso sea constante. La varianza de la serie expresa el grado de variación alrededor del valor medio asumido constante y proporciona una medida de la incertidumbre alrededor de esta media.
Además, si $k = 2$ la distribución conjunta de $X(t_1)$ y $X(t_2)$ depende solamente de la distancia entre estos puntos en el tiempo $(t_2 - t_1) = \tau$, que se denomina retardo. De esta manera la acv.f. $\gamma(t_1, t_2)$ también depende solamente de $(t_2 - t_1)$ y se puede escribir como $\gamma(\tau)$, donde

$$\begin{aligned}\gamma(\tau) &= E\{[X(t) - \mu][X(t + \tau) - \mu]\} \\ &= Cov[X(t), X(t + \tau)]\end{aligned}$$

se denomina coeficiente de autocovarianza al retardo τ.
El valor de la autocovarianza depende de las unidades en las cuales $X(t)$ sea

medida. Por eso, para propósitos interpretativos, es de gran ayuda estandarizar la acv.f. para producir una función denominada como de autocorrelación (ac.f.), que se define como

$$\rho(\tau) = \frac{\gamma(\tau)}{\gamma(0)}$$

Esta cantidad mide la correlación entre $X(t)$ y $X(t+\tau)$. Se aclara que el argumento τ de $\gamma(\tau)$ y $\rho(\tau)$ es discreto si el tiempo es discreto, pero continuo si el tiempo es continuo. Normalmente se utiliza $\gamma(k)$ y $\rho(k)$ para denotar estas funciones cuando son discretas en el tiempo.

Estacionareidad de Segundo Orden En la práctica es de gran ayuda definir estacionareidad en un sentido menos estricto que el descrito anteriormente. Un proceso se define como estacionario de segundo orden (o estacionario débil) si su media es constante y su acv.f. depende solamente del retardo, así que

$$E[X(t)] = \mu$$

y

$$Cov[X(t), X(t+\tau)] = \gamma(\tau)$$

De esta última condición se desprende que si un fenómeno es estacionario, sus variables pueden estar relacionadas linealmente entre sí, pero de forma que la relación entre dos variables solo depende de la distancia temporal k transcurrida entre ellas.
Dejando $\tau = 0$, notamos que la forma de una acv.f. estacionaria implica que la varianza junto con la media sean constantes. La definición también implica que la varianza y la media sean finitas. Esta definición de proceso estacionario débil será la utilizada en adelante, ya que muchas de las propiedades de un proceso estacionario dependen solamente de la estructura del proceso definido por su primer y segundo momento. Una clase importante de procesos donde esto es particularmente cierto son los procesos *normales* donde la distribución conjunta de $X(t_1), ..., X(t_k)$ es multivariada normal para todo $t_1, ..., t_k$. La distribución multivariada normal se caracteriza completamente por su primer y segundo momento y por lo tanto por $\mu(t)$ y $\gamma(t_1, t_2)$. Se cumple que la estacionareidad de segundo orden implica estacionareidad estricta para procesos normales. No obstante, μ y $\gamma(\tau)$ podrían no describir adecuadamente procesos estacionarios que sean muy distintos del normal.
Una vez introducido el concepto de proceso estocástico puede decirse que una serie temporal cualquiera es en realidad una muestra, una realización concreta con unos valores concretos de un proceso estocástico teórico, real.

El análisis de series que vamos a estudiar tratará, a partir de los datos de una serie temporal, inferir las características de la estructura probabilística subyacente del verdadero proceso estocástico.

2.2.2. *Modelos de Series Temporales más comunes*

Proceso aleatorio puro o Ruido Blanco

Un proceso que es discreto en el tiempo se denomina proceso aleatorio puro si consiste de una secuencia de variables aleatorias $\{Z_t\}$, las cuales son mutuamente independientes e idénticamente distribuidas. Normalmente se asume que las variables aleatorias siguen una distribución normal con media cero y varianza σ_Z^2. De la definición anterior se entiende que el proceso tiene varianza y media constantes. Además si asumimos independencia de los datos significa que:

$$\gamma(k) = Cov(Z_t, Z_{t+k}) = \begin{cases} \sigma_Z^2 & k = 0 \\ 0 & k = \pm 1, \pm 2, ... \end{cases}$$

esto significa que los diferentes valores no están correlacionados. La ac.f. es:

$$\sigma(k) = \begin{cases} 1 & k = 0 \\ 0 & k = \pm 1, \pm 2, ... \end{cases}$$

Ya que la media y la acv.f. no dependen del tiempo, el proceso es estacionario de segundo orden. De hecho, ya que se asume independencia de los datos, esto implica que el proceso es también estacionario de primer orden o estacionario en sentido estricto.
Los procesos aleatorios son de gran utilidad en muchas situaciones, particularmente como piedra angular para procesos más complicados tales como los de media móvil que más adelante describiremos.

Paseo Aleatorio

Es un modelo de mucho interés práctico y se define como:

$$X_t = X_{t-1} + Z_t \tag{2.1}$$

donde $\{Z_t\}$ denota un proceso de ruido blanco. El proceso es habitualmente inicializado a 0 cuando $t = 0$, así que

$$X_1 = Z_1$$

y

$$X_t = \sum_{i=1}^{t} Z_i$$

Luego se encuentra que $E(X_t) = t\mu$ y que $Var(X_t) = t\sigma_Z^2$. Como la media y la varianza cambian con t, el proceso no es estacionario. De cualquier forma, es interesante destacar que la primera diferencia de un paseo aleatorio se define como

$$\nabla X_t = X_t - X_{t-1} = Z_t$$

que forma un proceso de ruido blanco, que es por tanto estacionario.

Procesos Autorregresivos (AR)

Un proceso $\{X_t\}$ se dice que es un proceso autoregresivo de orden p (se abrevia AR(p)) si:

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + Z_t \tag{2.2}$$

El orden del modelo expresa el número de observaciones retrasadas de la serie temporal analizada que intervienen en la ecuación. El valor al instante t depende linealmente de los últimos p valores retrasados.
Si se usa el operador de retardo B de tal forma que $BX_t = X_{t-1}$, el modelo AR(p) se puede escribir de forma resumida como:

$$\phi(B)X_t = Z_t \tag{2.3}$$

donde

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$$

es un polinomio en B de orden p.
El ejemplo más simple de un proceso AR es el de primer orden, AR(1)

$$X_t = \phi X_{t-1} + Z_t \tag{2.4}$$

Claramente, si $\phi = 1$, entonces el modelo se reduce a un paseo aleatorio como el de la ecuación (2.1), cuando el modelo no es estacionario. Si $|\phi| > 1$, entonces es obvio que la serie será explosiva y por tanto no estacionaria. En cambio, si $|\phi| < 1$, entonces resulta que el proceso es estacionario, con ac.f. dada por $\rho_k = \phi^k$ para $k = 0, 1, 2, ...$ La ac.f. de este proceso disminuye exponencialmente, pero para procesos estacionarios AR de mayor orden, la

ac.f. será típicamente una mezcla de términos que disminuyen exponencialmente o una superposición de ondas de seno y coseno.

Si hacemos sustituciones sucesivas en la ecuación (2.4) se escribiría como:

$$\begin{aligned} X_t &= \phi(\phi X_{t-2} + Z_{t-1}) + Z_t \\ &= \phi^2(\phi X_{t-3} + Z_{t-2}) + \phi Z_{t-1} + Z_t \end{aligned}$$

y eventualmente encontraríamos que X_t podría ser expresado como un proceso MA de orden infinito en la forma

$$X_t = Z_t + \phi Z_{t-1} + \phi^2 Z_{t-2} + \cdots$$

siempre que $-1 < \phi < +1$, para que la suma converja.
La posibilidad de que un proceso AR pueda ser escrito en forma MA y viceversa, significa que hay una dualidad entre los procesos AR y MA que es muy útil para una variedad de fines. En vez de hacer sustituciones sucesivas para explorar esta dualidad, es más simple utilizar el operador de retardo B. La ecuación (2.4) se escribiría

$$(1 - \phi B)X_t = Z_t \tag{2.5}$$

y entonces:

$$\begin{aligned} X_t &= Z_t/(1 - \phi B) \\ &= (1 + \phi B + \phi^2 B^2 + \cdots)Z_t \\ &= Z_t + \phi Z_{t-1} + \phi^2 Z_{t-2} + \cdots \end{aligned}$$

Expresado de esta forma es claro que:

$$E(X_t) = 0$$
$$Var(X_t) = \sigma_Z^2(1 + \phi^2 + \phi^4 + \cdots)$$

Por lo tanto la varianza es finita siempre y cuando $|\phi| < 1$ en cuyo caso:

$$Var(X_t) = \sigma_X^2 = \sigma_Z^2/(1-\phi^2)$$

La acv.f. se define como:

$$\begin{aligned}
\gamma(k) &= E[X_t, X_{t+k}] \\
&= E[(\sum \phi^i Z_{t-i})(\sum \phi^j Z_{t+k-j})] \\
&= \sigma_Z^2 \sum_{i=0}^{\infty} \phi^i \phi^{k+i} \quad para\, k \geq 0 \\
&= \phi^k \sigma_Z^2/(1-\phi^2) \quad siempre\, que\, |\phi| < 1 \\
&= \phi^2 \sigma_X^2
\end{aligned}$$

Para $k < 0$, encontramos que $\gamma(k) = \gamma(-k)$. Ya que $\gamma(k)$ no depende de t, un proceso AR de orden 1 es estacionario de segundo orden siempre que $|\phi| < 1$, y la ac.f. se define entonces como ya fue comentado anteriormente:

$$\rho(k) = \phi^k \quad k = 0, 1, 2...$$

Una propiedad muy útil de un proceso AR(p) es que se puede demostrar que la ac.f. parcial es cero para todos los retardos mayores que p. Esto significa que la ac.f. parcial muestral puede ser utilizada para determinar el orden de un proceso AR por medio de observar el valor de retardo en el cual la ac.f. parcial muestral es aproximadamente cero, o al menos no significativamente diferente de cero para retardos mayores.

Proceso de Media Móvil (MA)

Una serie $\{X_t\}$ es un proceso de media móvil de orden q (se abrevia MA(q)) si

$$X_t = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q} \tag{2.6}$$

donde $\{Z_t\}$ representa los errores y es un proceso de ruido blanco con media cero y varianza constante σ_z^2. Se denomina a θ_p parámetro de media

móvil y representa el efecto de los errores pasados en X_t y deben ser estimados. La ecuación (2.6) se puede escribir también de la forma:

$$X_t = \theta(B)Z_t \tag{2.7}$$

donde

$$\theta(B) = 1 + \theta_1 B + \cdots + \theta_q B^q$$

es un polinomio en B de orden q.

De la ecuación fundamental de procesos MA se puede deducir que:

$$E(X_t) = 0$$
$$Var(X_t) = \sigma_Z^2 \sum_{i=0}^{q} \theta_i^2$$

ya que los Z_s son independientes. También tenemos que:

$$\begin{aligned}
\gamma(k) &= Cov(X_t, X_{t+k}) \\
&= Cov(\theta_0 Z_t + \cdots + \theta_q Z_{t-q}, \theta_q Z_{t+k} + \cdots + \theta_q Z_{t+k-q}) \\
&= \begin{cases} 0 & k > q \\ \sigma_Z^2 \sum_{i=0}^{q-k} \theta_i \theta_{i+k} & k = 0, 1, ..., q \\ \gamma(-k) & k < 0 \end{cases}
\end{aligned}$$

ya que

$$Cov(Z_s, Z_t) = \begin{cases} \sigma_Z^2 & s = t \\ 0 & s \neq t \end{cases}$$

Ya que $\gamma(k)$ no depende de t, y la media es constante, el proceso es estacionario de segundo orden para todos los valores de $\{\theta_i\}$. Además, si los Z_s están normalmente distribuidos, entonces también lo estarán los X_s, y tendremos un proceso normal estacionario en sentido estricto.
La ac.f. del proceso MA(q) esta dada por:

$$\rho(k) = \begin{cases} 1 & k = 0 \\ \sum_{i=0}^{q-k} \theta_i \theta_{i+k} \diagup \sum_{i=0}^{q} \theta_i^2 & k = 1, ..., q \\ 0 & k > q \\ \rho(-k) & k < 0 \end{cases}$$

La ac.f. se anula en el retardo q. Esta es una propiedad especial de los procesos MA y se utiliza para determinar el orden del proceso por medio de la observación del retardo en el cual la ac.f. muestral no es significativamente distinta de cero.

Un proceso MA(q) es siempre estacionario por ser la suma de procesos estacionarios. Diremos que el proceso es invertible si las raíces del operador $\theta_q(B) = 0$ son en módulo, mayores que la unidad. Para el caso de un MA(1) la explicación es la siguiente: la condición de estacionareidad no impone restricción alguna en el valor de θ_1. Por ejemplo, si se permitiera que $|\theta_1| > 1$, entonces tendríamos que aceptar una solución ilógica que resultaría en que la influencia de valores pasados (retardos de $Z_t's$) se incrementaría al ir más atrás en el tiempo. Es muy irreal pensar que un evento que ocurrió muchos años o días atrás tenga más influencia en la situación actual de un proceso que uno que ocurrió recientemente. Por lo tanto, se debe restringir θ_1 para satisfacer que cumpla con que $|\theta_1| < 1$. Esto es lo que se denomina condición de invertibilidad de un modelo MA(1). Para el caso de un MA(2) las restricciones a los parámetros necesarias para cumplir la condición de invertibilidad son:

$$-1 < \theta_2 < 1 \qquad \theta_2 + \theta 1 < 1 \qquad \theta_2 - \theta_1 < 1$$

Existen condiciones más complicadas para cuando $q \geq 3$. Las mismas restricciones aplicadas a los parámetros con el fin de cumplir con la condición de invertibilidad y de tener soluciones únicas aquí comentadas para procesos MA(q) se aplican para procesos AR(p). En todo caso la correlación entre una variable y su pasado va reduciéndose a medida que nos alejamos más en el tiempo del momento para el que estamos considerando dicha correlación (proceso ergódico). La explicación intuitiva de esta situación derivaría de que si le especificáramos una variable en función de ciertos coeficientes que nos determinen su correlación con los valores pasados de ella misma, los valores de dichos coeficientes deberían ser necesariamente inferiores a uno, porque sino el proceso de infinitos números sería explosivo.

2.2.3. *Procesos ARMA*

Una clase de modelos muy utilizados en series temporales es el formado por la combinación de un proceso AR y un proceso MA. La mezcla de procesos autorregresivos y de media móvil con p términos AR y q términos MA se denomina un proceso ARMA de orden (p, q). Se define por:

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q} \tag{2.8}$$

Usando la notación del operador de retardo la ecuación (2.8) se puede escribir de la forma

$$\phi(B)X_t = \theta(B)Z_t \tag{2.9}$$

donde $\phi(B)X_t = \theta(B)Z_t$ son polinomios en B de orden p, q respectivamente tal que

$$\phi(B) = 1 - \phi_1 B - \ldots - \phi_p B^p$$

y

$$\theta(B) = 1 - \theta_1 B - \ldots - \theta_q B^q$$

Esto combina las ecuaciones (2.2) y (2.6).

La ecuación (2.9) tiene solución causal estacionaria única siempre que las raíces de $\phi(x) = 0$ tengan valores fuera del círculo unidad. El proceso es invertible siempre que las raíces de $\theta(x) = 0$ tengan valores fuera del círculo unidad. En el caso estacionario, la ac.f. será generalmente una mezcla de ondas exponenciales o sinusoidales.
Es importante reconocer que con la introducción de modelos combinados AR y MA se gana en parsimonia a la hora de especificar un modelo. Es fácil demostrar que la ecuación (2.8) puede ser reescrita como un proceso MA(∞) por medio de sustituciones sucesivas de X_{t-1}, X_{t-2} etc. Como ya se comentó en la sección 2.2.2 en la página 19, de la misma forma la ecuación (2.8) puede ser reescrita como un proceso AR(∞) por medio de sustituciones sucesivas de $Z_{t-1}, Z_{t-2}, \ldots$, etc. El inconveniente de plantearlo de esta forma, es decir utilizando estas propiedades de los modelos AR y MA puros es que serán obtenidos a costa de muchos parámetros por estimar y un uso poco eficiente de los datos.

Procesos Integrados ARMA (ARIMA)

Los modelos ARIMA son en teoría, la clase de modelos más generalmente utilizados en predicción de series temporales que pueden convertirse en estacionarias por medio de transformaciones matemáticas tales como la diferenciación y/ó la aplicación de logaritmos. De hecho la forma más simple de pensar en los modelos ARIMA es como versiones ajustadas de modelos de paseo aleatorio y modelos con tendencia aleatoria: El ajuste consiste en

agregar retardos de la serie diferenciada y/ó retardos de los errores de predicción a la ecuación de predicción, tantos ordenes como sean necesarios para remover algún rastro de autocorrelación de los errores de predicción.

Los modelos de paseo aleatorio y tendencia aleatoria, los modelos autorregresivos y los modelos de suavizado exponencial son todos casos especiales de modelos ARIMA. En la práctica muchas series temporales no son estacionarias y por tanto no podemos aplicar procesos AR, MA ó ARMA directamente. Una posibilidad de manejar series no estacionarias es aplicar diferenciación para hacerlas estacionarias. Las primeras diferencias, $(X_t - X_{t-1}) = (1-B)X_t$, podrían ser diferenciadas de nuevo para generar las segundas diferencias y así sucesivamente. Las dth diferencias se escribirían como $(1-B)^d X_t$. Si la serie de datos original se diferencia d veces antes de ajustar un proceso ARMA(p, q), entonces el modelo de la serie original sin diferenciar se dice que es un proceso ARIMA(p, d, q) donde la letra "I" en el acrónimo significa *integrado* y d denota el número de diferencias realizadas.
La ecuación (2.9) se generaliza para dar:

$$\phi(B)(1-B)^d X_t = \theta(B) Z_t \tag{2.10}$$

Los operadores combinados AR son ahora $\phi(B)(1-B)^d$. Véase que cuando $\phi(B)$ y $\theta(B)$ son ambos casi de valor unidad (p y q son ambos cero) y d es igual a uno, entonces el modelo se reduce a un modelo ARIMA(0,1,0) dado por:

$$X_t - X_{t-1} = Z_t \tag{2.11}$$

Esta es obviamente la misma ecuación del modelo del paseo aleatorio (2.1) el cual puede ser considerado como un modelo ARIMA(0,1,0).

Cuando se ajustan modelos AR y MA, más que la dificultad de estimar los coeficientes, la principal dificultad es la de determinar el orden del proceso. Con los modelos ARIMA, surge un problema adicional, se debe determinar el orden de diferenciación requerido (valor de d). Existen varios procedimientos formales para determinar el orden de diferenciación, pero muchos analistas simplemente diferencian la serie hasta encontrar el valor mínimo de diferenciación que produzca una serie temporal que fluctúe alrededor de un valor medio bien definido y cuya ac.f. decaiga medianamente rápido hasta cero. La diferenciación de primer orden es usualmente adecuada para series no estacionales, aunque en ocasiones la diferenciación de segundo orden es necesaria.

Procesos ARIMA Estacionales (SARIMA)

La estacionalidad en una serie temporal provoca que la media de las observaciones no sea constante pero que evolucione de forma previsible de acuerdo con un patrón cíclico. Por ejemplo, en una serie de demandas de agua potable de una ciudad, la media de las demandas no es constante ya que varía con el mes, pero el mismo mes en distintos años es esperable que tenga un valor medio constante.

Diremos que una serie es estacional cuando su valor esperado no es constante, pero varía con una pauta cíclica (Peña, 2005). En concreto si:

$$E(x_t) = E(x_{t+s})$$

diremos que la serie tiene una estacionalidad de periodo s. El periodo estacional, s, define el número de observaciones que forman el ciclo estacional. Por ejemplo, $s = 12$ para series mensuales, $s = 4$ para series trimestrales, etc. Supondremos que el valor s es fijo en la serie, aunque esto puede no ser exactamente cierto, por ejemplo, si tenemos datos diarios y el periodo estacional es la longitud del mes, s será aproximadamente 30, pero variará de unos meses a otros. También puede existir más de un tipo de estacionalidad. Por ejemplo, con datos diarios podemos tener una estacionalidad semanal, con $s = 7$, otra mensual, con $s = 30$ y otra anual, con $s = 365$.

Si la serie es estacional, con s periodos por año, entonces un modelo ARIMA estacional (SARIMA) puede obtenerse como una generalización de (2.10). B^s denota el operador tal que $B^s X_t = X_{t-s}$. De esta forma la diferencia estacional puede escribirse como $(X_t - X_{t-s}) = (1 - B^s)X_t$. Un término autoregresivo estacional, por ejemplo, es uno donde X_t depende linealmente de X_{t-s}. Un modelo SARIMA con términos no estacionales de orden (p, d, q) y términos estacionales de orden (P, D, Q) se abrevia como modelo ARIMA(p, d, q)x$(P, D, Q)_s$ y puede ser escrito como

$$\phi_p(B)\Phi_P(B^s)(1 - B)^d(1 - B^s)^D X_t = \theta_q(B)\Theta_Q(B^s)Z_t \tag{2.12}$$

donde

$$\Phi_P(B^s) = (1 - \Phi_1 B^s - \ldots - \Phi_P B^{sP})$$

es el operador AR estacional de orden P,

$$\phi_p(B) = (1 - \phi_1 B - \ldots - \Phi_p B^p)$$

es el operador AR regular de orden p,

$$(1 - B)^d$$

representa las diferencias regulares I de orden d,

$$(1 - B^s)^D$$

representa las diferencias estacionales I de orden D,

$$\Theta_Q(B^s) = (1 - \Theta_1 B^s - \ldots - \Theta_Q B^{sQ})$$

es el operador de media móvil estacional de orden Q,

$$\theta_q(B) = (1 - \theta_1 B - \ldots - \theta_q B^q)$$

es el operador de media móvil regular de orden q y Z_t es un proceso de ruido blanco.

Algunos autores utilizan una notación más simplificada, de la siguiente forma:

$$\phi_p(B)\Phi_P(B^s)W_t = \theta_q(B)\Theta_Q(B^s)Z_t \tag{2.13}$$

donde

$$W_t = \nabla^d \nabla_s^D X_t$$

denota la serie diferenciada, d representa el orden de las diferencias regulares, D representa el orden de las diferencias estacionales y s representa la longitud del periodo estacional.

Cuando se está ajustando un modelo estacional a los datos, la primera tarea es determinar los valores de d y D, los cuales eliminan la mayor parte de la tendencia, la estacionalidad y transforman la serie en un proceso estacionario. Los valores de p, P, q y Q necesitan ser determinados por medio de observación de la ac.f. y ac.f. parcial.

2.3. Modelos de regresión dinámica

El objetivo de los modelos de regresión dinámica es relacionar dos o más series temporales elaborando modelos causales de predicción. Se considera la forma de relacionar una serie temporal, denominada *output* en función de una ó varias series temporales, que se denominan *inputs*. También se considera a priori que existe una causalidad unidireccional desde los inputs hacia el output, desechando la posibilidad de retroalimentación (*feedback*).

2.3.1. *Funciones de transferencia con retardos distribuidos lineales*

Funciones de transferencia

Por simplicidad de la explicación se inicia presentando el supuesto de solo un input, aunque es fácilmente aplicable para inputs múltiples. Si Y_t depende de alguna forma de X_t, entonces podemos establecer que:

$$Y_t = f(X_t) \tag{2.14}$$

donde $f(\cdot)$ es una función matemática. La función $f(\cdot)$ se denomina *función de transferencia*, el efecto de un cambio en X_t es transferida a Y_t de la forma especificada por la función $f(\cdot)$. Si nos centramos en el caso donde $f(\cdot)$ es una función lineal de retardos distribuidos, asumimos que Y_t es una combinación lineal de valores actuales (X_t) y pasados $(X_{t-1}, X_{t-1}, \ldots)$ del input. Por lo que una parte fundamental del procedimiento pasa por identificar de la mejor manera posible la forma de $f(\cdot)$ para el conjunto de datos que estemos analizando.

La forma de la ecuación (2.14) es la más básica, de hecho si conociéramos los valores de X_t y la forma de $f(\cdot)$, podríamos predecir sin error. Es obvio que generalmente existen otros factores que provocan variaciones en Y_t además de los especificados en los inputs. Todos esos otros factores se capturan en un proceso estocástico aditivo (N_t), que es el ruido del sistema y que podría estar autocorrelacionado pero que es independiente de la serie X_t. La relación *input-output* puede tener también un término aditivo constante (C) que es un término que captura el efecto de los inputs en el nivel global de Y_t. De esta manera llegamos a modelos de la forma:

$$Y_t = C + f(X_t) + N_t \tag{2.15}$$

Retardos distribuidos lineales

Si se presenta un cambio en X_t en la ecuación (2.15), la respuesta de Y_t muchas veces no solo ocurre durante un periodo de tiempo. Y_t puede reaccionar a un cambio en X_t con un retardo que se distribuye a lo largo de varios periodos de tiempo. Si se asume que esta relación de retardos distribuidos es lineal, podemos escribir la función de transferencia $f(X_t)$ como una combinación lineal de valores actuales y pasados de X_t:

$$\begin{aligned} Y_t &= f(X_t) \\ &= v_0 X_t + v_1 X_{t-1} + v_2 X_{t-2} + \cdots \end{aligned} \tag{2.16}$$

La función de transferencia real subyacente a una muestra de datos podría no ser lineal, sin embargo asumimos linealidad (Pankratz, 1991) porque:

1. Simplifica considerablemente el análisis estadístico
2. A pesar de su relativa simplicidad, los modelos lineales han probado ser de utilidad en una gran variedad de situaciones
3. Un modelo lineal es frecuentemente un primer paso o aproximación

El coeficiente v_0 es un peso que establece la respuesta de Y_t a un cambio en X_t, v_1 establece la respuesta de Y_t a un cambio en X_{t-1}, v_2 establece la respuesta de Y_t a un cambio en X_{t-2} y así sucesivamente. En teoría esta respuesta de retardos distribuida podría ser infinita.

Algunas veces Y_t podría no reaccionar inmediatamente a un cambio en X_t, por lo cual algunos pesos iniciales de v podrían ser cero. Como un ejemplo, el aumento del caudal enviado desde una zona de captación puede demorarse varias horas en verse reflejado en una planta depuradora. El número de pesos de v con valor cero (empezando con v_0) se llama tiempo muerto (*dead time* en la literatura anglosajona) y se simboliza como b. Si utilizamos la notación del operador de retardo, definiendo $v(B)$ como:

$$v(B) = v_0 + v_1 B + v_2 B^2 + v_3 B^3 \tag{2.17}$$

por lo que reorganizando, la ecuación (2.16) podría ser reescrita como

$$Y_t = v(B)X_t \tag{2.18}$$

donde B es el operador de retardo definido de tal forma que $B^k X_t = X_{t-k}$.

La ecuación (2.17) se define como función de transferencia del filtro de Box and Jenkins (1970). Los pesos individuales de v en $v(B)$, $(v_0, v_1, v_2, \ldots)$, se denominan como pesos de respuesta al impulso. El conjunto de los v pesos se conocen como la función de respuesta al impulso. Para que el sistema (2.16) sea estable se debe cumplir que una variación finita en el input produzca también una variación finita en el output. Por lo que se deberá cumplir que:

$$\sum_{j=0}^{\infty} v_j = g \tag{2.19}$$

siendo g finito. El valor de g representa el cambio total en Y_t motivado por un cambio unitario en X_t mantenido indefinidamente en el tiempo.

Agotamiento exponencial de pesos v *- El modelo Koyck*, (Koyck, 1954)

Si suponemos que el tiempo muerto es nulo $(b = 0)$, y suponemos también que conocemos que los pesos v en $v(B)$ están relacionados entre ellos de la siguiente forma:

$$\begin{aligned} v_1 &= \delta_1 v_0 \\ v_2 &= \delta_1 v_1 \\ v_3 &= \delta_1 v_2 \\ &\vdots \\ v_k &= \delta_1 v_{k-1} \qquad k = 1, 2, 3, \ldots \end{aligned} \tag{2.20}$$

Iniciando por la respuesta inicial v_0, cada respuesta subsecuente (desfasada en el tiempo) de Y_t a un cambio en X_t, representada por los pesos v, (v_1,v_2,. . . ,) es una fracción constante (δ_1) de la respuesta del periodo anterior. Los pesos v se agotan exponencialmente como se puede apreciar en el gráfico (2.1).

Cada v_k (excepto v_0) tiene una relación conocida con v_{k-1}. De hecho, todos los pesos v son conocidos si conocemos v_0 y δ_1. Esto es evidente para

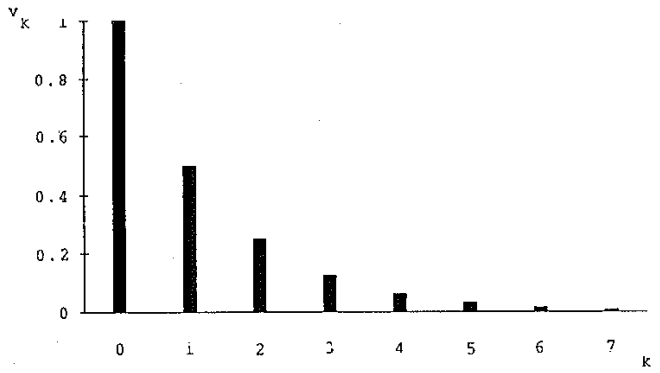


Figura 2.1: Gráfico de pesos como respuesta a un impulso con patrón de agotamiento exponencial simple, $v_0 = 1$, $\delta_1 = 0{,}5$

v_1 en (2.20). La sustitución recursiva resulta en los siguientes pesos v:

$$v_2 = \delta_1 v_1 = \delta_1^2 v_0$$

$$v_3 = \delta_1 v_2 = \delta_1^3 v_0$$

$$v_4 = \delta_1 v_3 = \delta_1^4 v_0$$

$$\vdots$$

el patrón puede ser expresado en una forma mucho más compacta como

$$v_k = \delta_1^k v_0, \qquad k \geq 0 \tag{2.21}$$

por lo que v_0 nos proporciona un valor de inicio para el agotamiento posterior, y δ_1 nos indica la tasa de agotamiento. En la práctica, si podemos estimar δ_1 y v_0, haremos buen uso de una muestra con datos limitados ya que podremos estimar el resto de pesos v usando la ecuación (2.21).

Forma parsimoniosa del modelo Koyck

Si usáramos el método de Koyck para encontrar una forma de la función de transferencia Koyck que tenga solo δ_1 y v_0 como sus parámetros, lo que tendríamos que hacer es sustituir los pesos v en la ecuación (2.18) de acuerdo con el patrón de agotamiento exponencial de la ecuación (2.21) para obtener:

$$Y_t = v_0 X_t + \delta_1 v_0 X_{t-1} + \delta_1^2 v_0 X_{t-2} + \cdots \tag{2.22}$$

para encontrar la respuesta para Y_{t-1}, remplazamos t por $t-1$ en toda la ecuación (2.22) para obtener:

$$Y_{t-1} = v_0 X_{t-1} + \delta_1 v_0 X_{t-2} + \delta_1^2 v_0 X_{t-3} + \cdots \tag{2.23}$$

y multiplicando ambos lados de la ecuación (2.23) por δ_1:

$$\delta_1 Y_{t-1} = \delta_1 v_0 X_{t-1} + \delta_1^2 v_0 X_{t-2} + \delta_1^3 v_0 X_{t-3} + \cdots \tag{2.24}$$

y si sustraemos (2.24) de (2.22), obtendremos:

$$Y_t - \delta_1 Y_{t-1} = v_0 X_t$$

ó

$$Y_t = v_0 X_t + \delta_1 Y_{t-1} \tag{2.25}$$

La ecuación (2.25) es una función de transferencia con solo v_0 y δ_1 como parámetros por estimar. Si los pesos v siguen el patrón establecido en la ecuación (2.21), entonces podemos escribir la función de transferencia (2.18) con unos cuantos parámetros, como en (2.25). Sin embargo esta última ecuación, aunque más compacta, no contiene menos información que la ecuación (2.18) la cual tiene tiene un número infinito de parámetros ($v_0, v_1, v_2, \ldots$).

Retardos distribuidos en forma racional

La función de transferencia de la ecuación (2.25) puede ser escrita utilizando el operador de retardo de la siguiente forma:

$$(1 - \delta_1 B) Y_t = v_0 X_t \tag{2.26}$$

Y para obtener la forma racional, dividimos la ecuación (2.26) por $1-\delta_1 B$ y obtenemos:

$$Y_t = \frac{v_0}{1 - \delta_1 B} X_t \tag{2.27}$$

Si comparamos las ecuaciones (2.18) y (2.27), podemos ver que $v(B) = \frac{v_0}{1-\delta_1 B}$ para el caso especial del modelo Koyck con $b = 0$. La ecuación (2.27),

en el fondo es la misma que (2.25) y (2.26) pero escrita en forma polinomial racional. En (2.27) hemos indicado $v(B)$ como un cociente de polinomios de orden finito en B, y ahora B es tratado como una variable algebraica.

La familia de retardos distribuidos en forma racional

El modelo Koyck de respuesta al impulso es solo un miembro de la familia de modelos de retardos distribuidos en forma racional. Esta familia es un conjunto de funciones de respuesta al impulso $v(B)$ de la forma:

$$v(B) = \frac{\omega(B)B^b}{\delta(B)} \tag{2.28}$$

donde[1]

$$\omega(B) = \omega_0 + \omega_1 B + \omega_2 B^2 + \cdots + \omega_s B^s \tag{2.29}$$

$$\delta(B) = 1 - \delta_1 B - \delta_2 B^2 - \cdots - \delta_r B^r \tag{2.30}$$

y B^b es un parámetro de retardo que representa el tiempo que transcurre antes de que el impulso en la variable input produzca un efecto en la variable output.

Los modelos de retardos distribuidos en forma racional, pueden capturar una gran variedad de patrones de comportamiento de respuesta al impulso con unos cuantos parámetros. Y aunque la familia completa incluye un número infinito de posibles modelos, solamente unos cuantos suelen ocurrir con frecuencia en la práctica, el gráfico (2.2) presenta algunos ejemplos. Es por este motivo que es relativamente fácil identificar un modelo razonable para un determinado conjunto de datos.

El numerador $\omega(B)B^b$ y el denominador $\delta(B)$ de la ecuación (2.28) tienen distintas funciones a la hora de representar patrones de respuesta al impulso.

1. El factor B^b del numerador captura el tiempo muerto

[1]La notación de $\omega(B)$ en (2.29) es un poco diferente a la utilizada en Box et~al. (1976), donde utilizan signos negativos para cada término excepto para ω_0. El uso de signos positivos o negativos no tiene relevancia en las ecuaciones siempre y cuando su utilización sea consistente

2. El factor $\omega(B)$ captura los picos (que no son parte de ningún patrón)

3. El factor $\delta(B)$ representa el patrón de agotamiento

Retardos distribuidos en forma racional con inputs múltiples

La formulación puede fácilmente ser extendida para varios inputs, por ejemplo, con dos inputs ($X_{1,t}$ y $X_{2,t}$) tendríamos que:

$$\begin{aligned} Y_t &= v_1(B)X_{1,t} + v_2(B)X_{2,t} \\ &= \tfrac{\omega_1(B)B^{b_1}}{\delta_1(B)}X_{1,t} + \tfrac{\omega_2(B)B^{b_2}}{\delta_2(B)}X_{2,t} \end{aligned} \tag{2.31}$$

donde:

$v_i(B) = v_{i,0} + v_{i,1}B + v_{1,2}B^2 + \cdots = \frac{\omega_i(B)(B)^{b_i}}{\delta_i(B)}, \qquad i = 1, 2$
$b_i =$tiempo muerto para el input $X_{i,t} \qquad i = 1, 2$
$\omega_i(B) = \omega_{i,0} + \omega_{i,1}B + \cdots + \omega_{i,s_i}B^{s_i}, \qquad i = 1, 2$
$\delta_i(B) = 1 - \delta_{i,1} - \delta_{i,2}B^2 - \cdots - \delta_{i,r_i}B^{r_i}, \qquad i = 1, 2$
$h_i =$ orden de $\omega_i(B), \qquad i = 1, 2$

Y extendiendo esta formulación para M inputs, $i = 1, 2, \ldots, M$, el resultado en forma compacta se escribiría:

$$Y_t = \sum_{i=1}^{M} v_i(B)X_{i,t}$$

$$Y_t = \sum_{i=1}^{M} \frac{\omega_i(B)B^{b_i}}{\delta_i(B)}X_{i,t} \tag{2.32}$$

La forma completa de los modelos de regresión dinámica en forma racional y algunos casos especiales

La ecuación (2.32) representa la forma racional de un modelo con M inputs y M funciones de transferencia. Un modelo de regresión dinámica también puede incluir una constante (C), así como un modelo para el ruido de la serie (N_t):

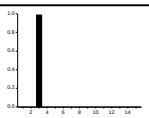
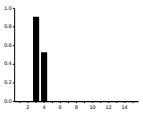
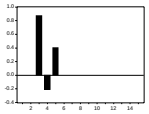
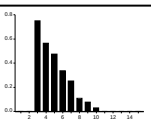
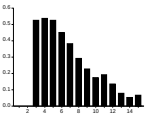
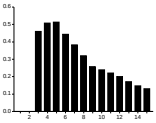
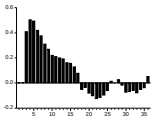
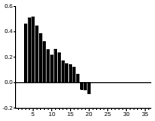
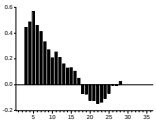
(b, r, s)	Función de Transferencia	Ponderaciones de respuesta al impulso
	$r = 0$	
$(3, 0, 0)$	$v(B)x_t = \omega_0 x_{t-3}$	
$(3, 0, 1)$	$v(B)x_t = (\omega_0 - \omega_1 B)x_{t-3}$	
$(3, 0, 2)$	$v(B)x_t = (\omega_0 - \omega_1 B - \omega_2 B^2)x_{t-3}$	
	$r = 1$	
$(3, 1, 0)$	$v(B)x_t = \frac{\omega_0}{(1-\delta_1 B)}x_{t-3}$	
$(3, 1, 0)$	$v(B)x_t = \frac{(\omega_0-\omega_1 B)}{(1-\delta_1 B)}x_{t-3}$	
$(3, 1, 2)$	$v(B)x_t = \frac{(\omega_0-\omega_1 B-\omega_2 B^2)}{(1-\delta_1 B)}x_{t-3}$	
	$r = 2$	
$(3, 2, 0)$	$v(B)x_t = \frac{\omega_0}{(1-\delta_1 B-\delta_2 B^2)}x_{t-3}$	
$(3, 2, 1)$	$v(B)x_t = \frac{(\omega_0-\omega_1 B)}{(1-\delta_1 B-\delta_2 B^2)}x_{t-3}$	
$(3, 2, 2)$	$v(B)x_t = \frac{(\omega_0-\omega_1 B-\omega_2 B^2)}{(1-\delta_1 B-\delta_2 B^2)}x_{t-3}$	

Figura 2.2: Funciones de respuesta al impulso típicas

$$Y_t = C + \sum_{i=1}^{M} \frac{\omega_i(B)B^{b_i}}{\delta_i(B)} X_{i,t} + N_t \tag{2.33}$$

El término de error no tiene porqué ser necesariamente un ruido blanco. Se puede suponer con carácter general que N_t sigue un proceso ARIMA, aunque sigue siendo independiente de la variable de input X_t. Los procesos ARIMA y su formulación fueron presentados previamente en este documento en la ecuación (2.10) de la página (25) así como en la ecuación (2.12) de la página (26) para el caso en que se considera la estacionalidad. Si integramos estas ecuaciones a la ecuación general del modelo de regresión dinámica se obtiene:

$$Y_t = C + \sum_{i=1}^{M} \frac{\omega_i(B)B^{b_i}}{\delta_i(B)} X_{i,t} + \frac{\theta(B)}{\phi(B)(1-B)^d} Z_t \tag{2.34}$$

y para el caso de tratarse de una serie con estacionalidad:

$$Y_t = C + \sum_{i=1}^{M} \frac{\omega_i(B)B^{b_i}}{\delta_i(B)} X_{i,t} + \frac{\theta_q(B)\Theta_Q(B^s)}{\phi_p(B)\Phi_P(B^s)(1-B)^d(1-B^s)^D} Z_t \tag{2.35}$$

2.3.2. *Análisis de Intervención*

El análisis de intervención es una técnica de modelación estocástica que se utiliza para analizar rigurosamente intervenciones (naturales o inducidas por el hombre) que causan un cambio significativo en el nivel medio de una serie temporal. Box and Tiao (1975) presentaron el análisis de intervención como un caso especial de los modelos de función de transferencia Box-Jenkins (Box et~al., 1976), donde un modelo de respuesta de función de transferencia se usa para relacionar una serie de salida estocástica con una variable de intervención cualitativa. Esta técnica es capaz de manejar variables relacionadas con retardo, variables predictoras autocorrelacionadas y, adicionalmente, ya que un modelo ARMA se construye separadamente para los residuos, la metodología puede fácilmente manejar los residuales autocorrelacionados.

Box and Tiao (1975) denominaron análisis de intervención a la inclusión en un modelo de series temporales de variables ficticias para representar sucesos que producen efectos deterministas. Las variables ficticias más utiliza-

das para representar sucesos cualitativos que afectan a la serie son de dos tipos: *variables impulso* y *variables escalón*. Las variables impulso representan sucesos que ocurren únicamente en un instante, por ejemplo, un accidente, un error de medida o una huelga. Las variables escalón representan acontecimientos que comienzan en un instante conocido y se mantienen a partir de ese instante, por ejemplo, una subida de precios, un cambio legal, la construcción de una presa que modifica el caudal de un río, el establecimiento de programas de detección de fugas en una red hidráulica que se refleja como una disminución de la demanda de un sector hidráulico, etc.

El esquema que se usa para evaluar el efecto de una variable de intervención única es la forma racional del modelo de regresión dinámica que se presentó en el apartado (2.3.1), página (28):

$$Y_t = C + \frac{\omega(B)B^b}{\delta(B)}X_t + N_t \tag{2.36}$$

la definición de las variables y operadores es igual a como se definieron en el apartado anterior. La única diferencia reside en que X_t es una variable determinística binaria en vez de una variable estocástica. Como en los apartados anteriores, N_t puede ser explicado por un proceso ARIMA. Por consiguiente el modelo de intervención es un caso especial del modelo de regresión dinámica (función de transferencia + ruido).

Intervenciones tipo impulso

Una intervención tipo impulso se utiliza cuando ocurre algún suceso en una serie temporal que provoca que un valor de la serie aumente ó disminuya puntualmente y a partir del mismo en adelante, la serie sigue la evolución que presentaba antes del suceso. Supongamos que el evento de intervención tipo impulso ocurre durante el periodo $t = i$. Entonces la variable input X_t se define:

$$X_t = \begin{cases} 0, & t \neq i \\ 1, & t = i \end{cases}$$

X_t es una variable binaria ($0 =$*off*, $1 =$*on*). La naturaleza temporal de una intervención tipo impulso se refleja en que X_t solamente tiene valor *on* durante el periodo i. El efecto de una intervención tipo impulso en la serie Y_t se aprecia en el gráfico (2.3a), donde la serie Y_t tiene una media estacionaria

excepto en el momento del efecto del impulso de intervención $t = i$. Durante este periodo Y_t aumenta (podría también disminuir) alejándose del nivel medio constante anterior. Inmediatamente después de $t = i$, la serie retorna al nivel anterior. El caso de una serie temporal no estacionaria se aprecia en el gráfico (2.3b). La situación de que el impulso dure más de un periodo se aprecia en (2.3c) donde después del momento $i + 1$ la serie retorna a su nivel anterior.

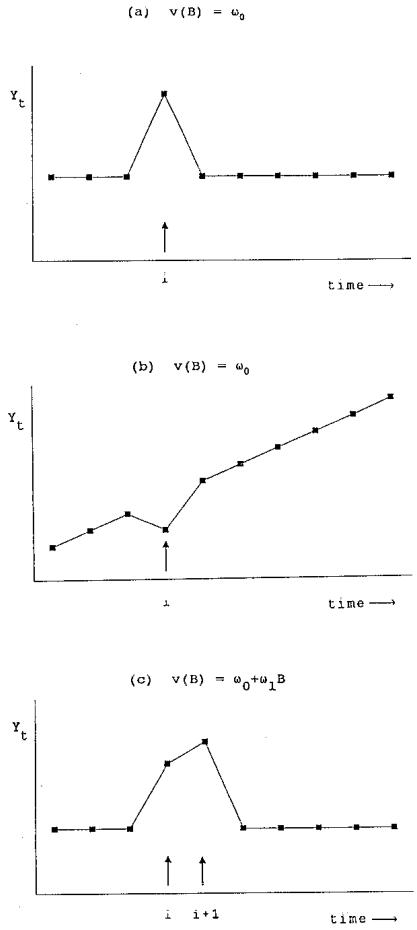


Figura 2.3: Ejemplos de intervenciones tipo impulso para un periodo y multi periodo

El caso (d) del gráfico (2.4) muestra el caso de una serie estacionaria intervenida por un impulso que provoca un aumento brusco en el valor de Y_t, pero que con el paso del tiempo vuelve a su nivel original con un agotamiento del tipo presentado en el modelo de Koyck (ver apartado 2.3.1). El caso (e) es muy similar al caso (b) por el hecho de que ambas series no son estacionarias, con la diferencia de que Y_t reacciona de una forma dinámica al impulso de intervención después del periodo $t = i$, con una disminución de

su impacto en los periodos $i+1, i+2, i+3, \ldots$, de nueva cuenta el fenómeno se reproduce con el modelo de Koyck. Finalmente el caso (f) es un caso especial del impulso tipo Koyck con $\delta = 1$, por lo que la respuesta subsecuente al momento i será 1.0 veces la respuesta anterior, sin agotamiento alguno, lo cual lo convierte en un impulso de respuesta tipo escalón.

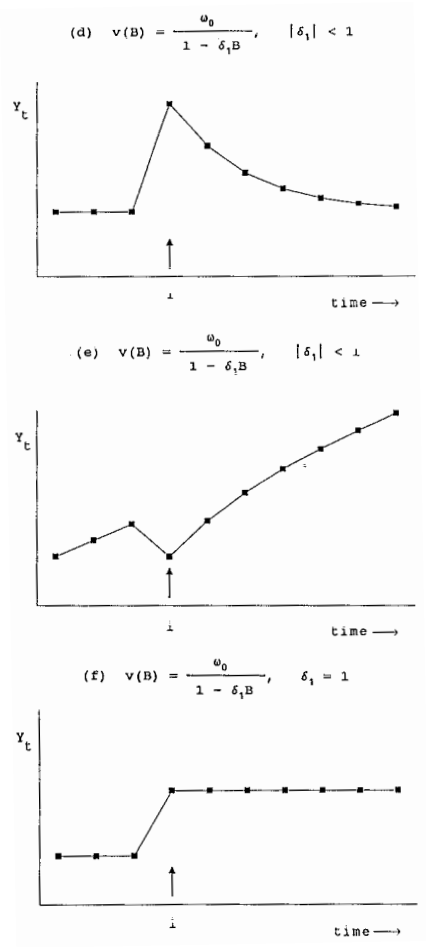


Figura 2.4: Ejemplos de intervenciones tipo impulso multi periodo

Intervenciones tipo escalón

Las intervenciones tipo escalón provocan un cambio permanente en el nivel de Y_t. Para una intervención tipo escalón en el momento $t = i$, tenemos:

$$X_t = \begin{cases} 0, & t < i \\ 1, & t \geq i \end{cases}$$

La naturaleza permanente de una intervención tipo escalón queda reflejado en el hecho que X_t tenga valor "on" continuamente, iniciando en el periodo i. Por lo que, X_t tiene valor 1 cuando $t = i$, conservando ese valor en adelante. El gráfico (2.5a) muestra un proceso de este tipo y el caso (b) es el de un proceso tipo escalón que se mantiene por varios instantes de t para luego volver a su valor original. En cambio el caso (c) presenta una intervención tipo escalón para una serie con media no estacionaria.

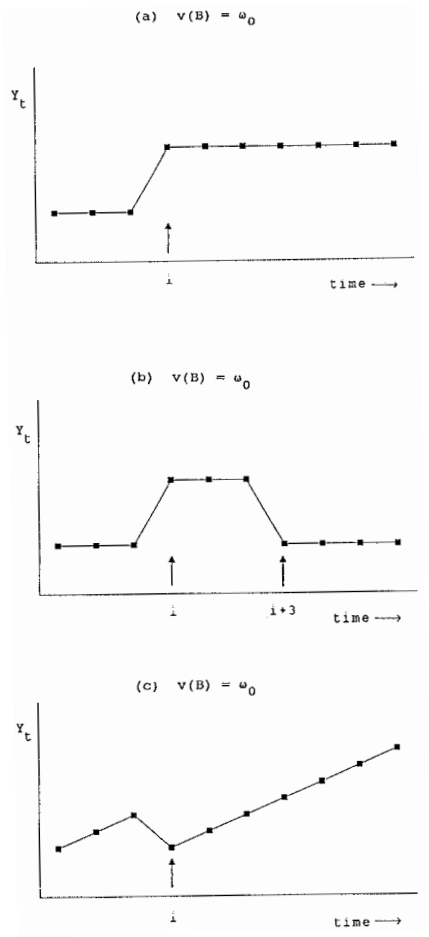


Figura 2.5: Ejemplos de intervenciones tipo escalón

Existen también intervenciones tipo escalón que contienen efectos dinámicos. El gráfico (2.6d) muestra un ejemplo donde Y_t mantiene una media estacionaria excepto por la intervención tipo escalón en el periodo i. La función de transferencia para este caso es $(\omega_0 + \omega_1 B)X_t$ y ω_0 captura el aumento o disminución en Y_t durante el periodo i, mientras que ω_1 captura el aumento o disminución durante el periodo $i+1$. En el caso (d) de ese mismo gráfico se explica con el modelo de Koyck (sección 2.3.1) con $|\delta_1| < 1$. Finalmente el caso d es también explicado con el modelo Koyck para $\delta_1 = 1$, por lo que resulta en una función con apariencia de una rampa en vez de un escalón.

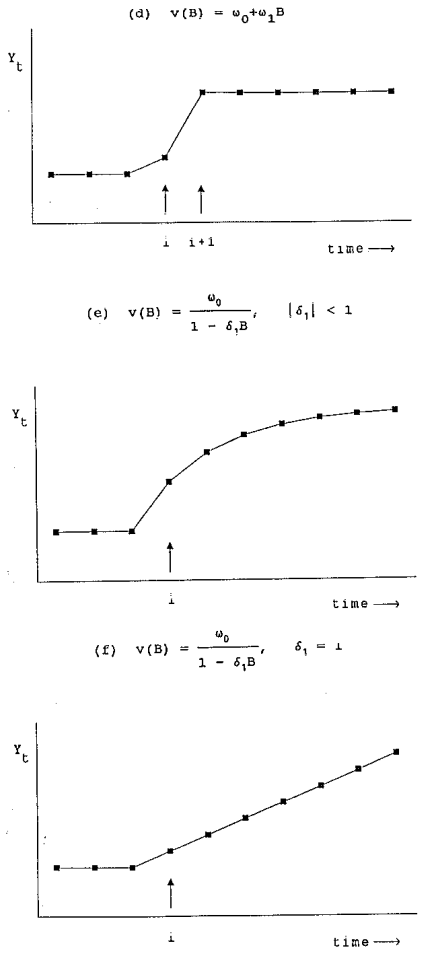


Figura 2.6: Ejemplos de intervenciones tipo escalón muti periodo

2.4. Valores Atípicos

En la sección anterior se presentaron las herramientas para modelar fenómenos o sucesos puntales conocidos por medio de los modelos de regresión dinámica o mediante el análisis de intervención. Sin embargo, en las series temporales reales es común encontrar hechos puntales que desconocemos. Dentro de estos eventos podemos definir los errores de medición, eventos extraordinarios en su magnitud, fallos de los sistemas de registro, errores de transcripción, etc. Las observaciones afectadas por estas intervenciones presentarán una estructura distinta de las demás por lo que aparecerán como valores atípicos u *outliers*. Dependiendo de su naturaleza los valores atípicos pueden tener impactos de moderados a sustanciales. Peña (2005) menciona los principales motivos por los que los valores atípicos deben ser identificados para ser separados o tomados en cuenta en la dinámica habitual de la serie:

1. Si sus efectos son grandes, pueden sesgar las estimación de los parámetros del modelo, lo que producirá malas predicciones futuras.
2. Si el suceso ha ocurrido en la última parte de la serie y alguna observación afectada se utiliza para generar predicciones, estas no serán buenas, incluso si los parámetros han sido bien estimados.
3. Si estos sucesos atípicos pueden volver a aparecer en el futuro y los identificamos y estimamos sus efectos, podemos incorporar esta información en las predicciones y obtener intervalos de predicción más realistas

Liu (2006) añade algunos motivos más:

1. Un mejor entendimiento de la serie bajo estudio. La detección de valores atípicos destacará la ocurrencia de esos eventos extremos que afectan a la serie, su impacto y duración. El desenmascarar esas ocurrencias nos guiará a esclarecer el comportamiento de la serie.
2. Como se mencionó anteriormente los valores atípicos pueden afectar la estimación de los parámetros. Como resultado, si se utiliza un modelo de intervención, se debe estar seguro que los efectos de la intervención no están afectados a su vez por efectos de valores atípicos.

Resumiendo, la identificación de los atípicos consiste en detectar observaciones que parezcan haber sido generadas de forma distinta al resto para después, una vez investigadas las causas que los generaron, evaluar si son eliminados o si son incorporados mediante un modelo explícito.

Existen una gran variedad de trabajos en los cuales se ha evaluado el impacto de los valores atípicos en la estimación de los parámetros de modelos de series temporales así como en sus predicciones. Chang et~al. (1988) propusieron que los valores atípicos pueden ser considerados como generados por modelos de intervención dinámicos en momentos desconocidos y proponen un proceso iterativo para estimar parámetros de series temporales ARIMA en presencia de atípicos innovativos y aditivos. Ledolter (1989) centró su trabajo en evaluar los efectos de los atípicos aditivos en las predicciones con modelos ARIMA. Concluye en su estudio que si un atípico aditivo al final de la serie se ignora, resulta un incremento del error cuadrático medio de las predicciones derivado de un efecto de arrastre del atípico y del sesgo en la estimación de los parámetros del modelo ARIMA. Adicionalmente, concluye que si el atípico ocurre no muy cerca del origen de predicción, las predicciones no se verán muy afectadas. Los outliers inflan la varianza de las innovaciones por lo que los intervalos de predicción son bastante sensitivos a los atípicos aditivos. Peña (1990) presentó una metodología para identificar valores atípicos en series modeladas con modelos ARIMA y propuso estadísticos con los cuales se puede medir la influencia de los atípicos aditivos e innovativos. A su vez el estadístico nos indica la robustez del modelo ajustado. Unos años más tarde Chen and Liu (1991, 1993b) presentaron un metodología para obtener estimaciones conjuntas de parámetros de modelos ARIMA y de valores atípicos del tipo aditivo, innovativo, cambio de nivel y cambio temporal. Con esta metodología se obtienen parámetros de modelos menos contaminados y los efectos de los atípicos, así como su clasificación son estimados conjuntamente. Los parámetros estimados con esta metodología son muy similares a los obtenidos mediante el método de máxima verosimilitud usando un modelo de intervención para incorporar los valores atípicos. En una continuación del trabajo de Ledolter (1989), Chen and Liu (1993a) utilizaron la metodología presentada en Chen and Liu (1991) para evaluar el impacto en las predicciones de valores atípicos ocurriendo cerca o en el origen de la predicción y demuestran que los procedimientos estadísticos existentes no son capaces de determinar efectivamente el tipo de atípico debido a información insuficiente. En castellano encontramos el trabajo de Trivez (1994) en el cual estudia el impacto que tiene la ignorancia de valores atípicos en la exactitud de las predicciones puntuales con modelos ARIMA evaluada en términos del error cuadrático medio de predicción. Concluye que el grado de incidencia depende del tipo de atípico que ocurra, del periodo temporal en el que tengan lugar, del horizonte temporal para el que se realice la predicción, de la magnitud del valor atípico y del proceso subyacente a la serie temporal.

Lo que resta de esta sección la dedicaremos a presentar los distintos tipos de valores atípicos que se han venido mencionando hasta este punto.

Destacaremos sus aspectos matemáticos y las metodologías que se utilizan para detectarlos. Los tipos de atípicos que se han considerado en la literatura (ver Chen and Liu (1993b)) son: el atípico aditivo (AO) producido frecuentemente por errores en las mediciones, el atípico innovacional (IO) propio de situaciones dinámicas, el cambio de nivel (LS) y el cambio temporal (TC). A continuación se presentarán los aspectos más relevantes de cada tipo de valores atípicos, si se desea una visión más completa se recomienda la lectura de (Peña, 2005, Cap.13 Valores atípicos) y también el trabajo de Trivez (1994).

2.4.1. *Aditivos (AO)*

Un atípico aditivo (AO) es un evento que afecta una serie solamente durante un periodo de tiempo. Una situación especial de AO son los errores de registro, por lo que este tipo de errores también se les suele llamar errores burdos. Como ejemplo, si existe un error de medida importante en el instante h, el dato z_h será un atípico aditivo en la serie z_t. El modelo que seguirá una serie observada, z_t, afectada por un AO en $t = h$ será:

$$z_t = \begin{cases} y_t & t \neq h \\ y_t + \omega_A & t = h \end{cases}$$

donde y_t es la serie no afectada por valores atípicos y que sigue un proceso ARIMA:

$$y_t = \psi(B)a_t$$

Por lo que el modelo que reproduce las serie observada, z_t, es

$$z_t = \omega_A I_t^{(h)} + \psi(B)a_t \tag{2.37}$$

donde $I_t^h = 0$, $t \neq h$; $I_h^{(h)} = 1$. Es fácil observar que el modelo es de la misma forma al de intervención con un impulso presentado en la sección anterior. Sin embargo en ese caso la variable $I_h^{(h)}$ se supone conocida y en cambio en este caso el instante h es desconocido.

Efectos de los AO en los residuos

Los (AO) dejan una huella en la serie de residuos debido a las alteraciones que producen en un punto. Si construimos un modelo ARIMA desconociendo su presencia, es posible identificarlos estudiando los residuos del modelo e_t, que serán iguales a las innovaciones verdaderas a_h hasta antes del instante del evento atípico en $t = h$. En los instantes posteriores a h la relación entre ambas variables para un AR(1) es:

$$\begin{aligned} e_h &= \omega_A + a_h && (2.38) \\ e_{h+1} &= -\phi\omega + a_{h+1} \\ e_{h+j} &= a_{h+j}, \quad j \geq 2 && (2.39) \end{aligned}$$

Se deduce pues que un atípico AO en un AR(1) modifica el residuo en el instante $t = h$, que es cuando se produce la mayor afectación. El residuo posterior también se ve afectado, pero en sentido inverso y por una magnitud que es el producto del tamaño del atípico y el parámetro $(|\phi < 1|)$. Si generalizamos para un AR(p), tendríamos un modelo del tipo:

$$\phi(B)z_t = \omega_A\phi(B)I_t^{(h)} + a_t \tag{2.40}$$

La relación entre los residuos calculados con los datos observados, que contienen el efecto de los atípicos y las innovaciones, puede entonces escribirse como:

$$e_t = \omega_A x_t + a_t$$

donde x_t es cero en todo momento salvo para $x_h = 1, x_{h+1} = -\phi, \ldots, x_{h+p} = -\phi_p$. Por lo tanto, los p residuos posteriores a h estarán afectados por la relación

$$e_{h+j} = a_{h+j} - \phi_j\omega_A, \quad j \geq 0$$

con $\phi_0 = -1$. Finalmente podemos deducir que en un AR(p) los p residuos posteriores estarán afectados de manera compleja, que dependerá de los parámetros AR y que la suma de todos los efectos sobre los residuos es $\omega_A(1 - \phi_1 - \ldots - \phi_p) = \omega_A\phi_p(1)$. Tal y como en los procesos no estacionarios $\phi_p(1) = 0$ por tener una raíz unitaria, los efectos de los residuos deben compensarse positivos con negativos. Si generalizamos para procesos ARIMA, la relación entre los residuos calculados en la serie que contiene el efecto de los residuos, e_t, y las innovaciones verdaderas, a_t, es:

$$x_t = \pi(B)I_t^{(h)} = -\sum_{j=0}^{T-h} \pi_i I_t^{(h+j)}$$

con $\pi_0 = -1$. Por lo que tendremos que el número de residuos afectados por un atípico dependerá del orden del polinomio $\pi(B)$. Entonces:

$$e_{h+j} = -\pi_j \omega_A + a_{h+j} \qquad j \geq 0$$

por lo que en el caso que todos los π_j fueran no nulos, todos los residuos posteriores al instante de ocurrencia del atípico estarán afectados. En todo lo anteriormente presentado hemos supuesto que existe un único AO en la serie, por lo que si existieran más de uno sus efectos se podrían superponer y modificarían la estructura de la serie.

Efectos de los AO en la estimación de los parámetros

Los efectos de los AO en modelos ARIMA son complejos aritméticamente. Como una alternativa ilustrativa se presenta a continuación el efecto de un AO en el instante (h) en un AR(1) de media cero. La estimación del parámetro ϕ será:

$$\hat{\phi} = \frac{\sum (z_t - \bar{z}) - z_{t-1}}{\sum (z_t - \bar{z})^2}$$

si $z_h = y_h + \omega_A$, y $z_t = y_t, t \neq h$, entonces

$$\bar{z} = \bar{y} + \frac{1}{T}\omega_A$$

Y si T es grande, podemos suponer que $\bar{z} \simeq \bar{y} \simeq 0$, llegamos a que,

$$\hat{\phi} = \frac{\sum y_t y_{t-1} + \omega_A(y_{h-1} + y_{h+1})}{\sum y_t^2 + \omega_A^2 + 2\omega_A y_h}$$

La magnitud del atípico aparece linealmente en el numerador y al cuadrado en el numerador, de lo que podemos deducir que cuando ω_A sea grande el valor de $\hat{\phi}$ será muy pequeño. Intuitivamente podemos deducir que el efecto de un atípico aditivo depende mucho del tamaño muestral. Para tamaños muestrales muy grandes un AO de tamaño moderado puede tener un efecto pequeño y por contra si la muestra es pequeña, el efecto puede ser importante.

2.4.2. *Innovacionales(IO)*

La principal diferencia entre los atípicos AO y los IO que ahora trataremos, es que el efecto de un atípico AO en una serie observada es independiente del modelo, cosa que no ocurre con los atípicos IO. Esto se puede

apreciar si comparamos la ecuación (2.37) con la del modelo para una serie que sufre un atípico innovativo de magnitud ω_I en el instante h que se presenta a continuación:

$$z_t = \psi(B)(\omega_I I_T^{(h)} + a_t) \tag{2.41}$$

es más evidente si escribimos la ecuación anterior como

$$z_t = y_t + \psi(B)\omega_I I_t^{(h)} \tag{2.42}$$

donde $y_t = \psi(B)a_t$ es la serie sin contaminar que sigue el modelo ARIMA.

Como se puede ver, la estructura del modelo $\psi(B) = \frac{\theta(B)}{(1-B)^d\phi(B)}$ afecta al atípico. Por lo que la relación entre las observaciones contaminadas, z_t, y las originales, y_t, es:

$$z_t = \begin{cases} y_t & t < h \\ y_t + \omega_I\psi_j & t = h + j, \quad j \geq 0 \end{cases} \tag{2.43}$$

Los efectos de un IO en una serie temporal son más intrincados y con estructuras más complejas que los producidos por los otros tipos de atípicos. De acuerdo a la ecuación (2.41) los efectos de un IO depende de los pesos del modelo ARIMA de z_t. Un atípico IO ocurrente en una serie estacionaria tendrá efectos temporales ya que los pesos ψ_j decaen exponencialmente hasta 0. En cambio si el proceso es un ARIMA, entonces el IO se propagará según la estructura del modelo afectando a todos los valores observados después de su ocurrencia. En (Chen and Liu, 1993a, pag.16) se presentan los efectos esperados por un IO en modelos de z_t de bajo orden.

Efectos de los IO en la estimación de los parámetros

El efecto de un IO en la estimación de los parámetros depende del modelo ajustado a la serie. Sin entrar en desarrollos aritméticos (para más información ver (Peña, 2005, pag.376)) podemos concluir que si $\omega_I \to \infty$ entonces $\hat{\phi} \to \phi$, por lo que no se producirán sesgos en las estimaciones de los parámetros. Sin embargo, al igual que en el caso de los otros atípicos, ocasionarán un incremento importante en la amplitud de los intervalos de confianza de las predicciones.

2.4.3. *Cambio de Nivel (LS)*

Un atípico LS es un evento que afecta una serie en un momento dado y cuyos efectos permanecen. Un atípico de cambio de nivel puede reflejar un cambio en el proceso, el cambio en el mecanismo de registro de los datos, etc. La ecuación que caracteriza a una serie que sufre este tipo de atípicos en el instante h es:

$$z_t = \omega_L S_t^{(h)} + \psi(B)a_t \tag{2.44}$$

donde $S_t^{(h)}$ es la variable escalón, $S_t^{(h)} = 1$ si $t \geq h$ y cero en otros casos. Los valores de la serie observada estarán relacionados con la serie sin contaminar por el cambio de nivel mediante:

$$z_t = \begin{cases} y_t & t < h \\ y_t + \omega_L & t \geq h \end{cases}$$

Efectos de los LS en los residuos

Los residuos calculados con los parámetros verdaderos están relacionados con las innovaciones mediante:

$$e_t = \omega_L x_t + a_t$$

donde $x_t = \pi(B)S_t^{(h)}$ es una variable que toma el valor cero antes de la intervención, el valor uno en el instante $t = h$ y los valores $1 - \sum_{i=1}^{j} \pi_i$ posteriormente. Por lo que:

$$e_t = \begin{cases} a_t & t < h \\ a_t + \omega_L(1 - \sum_{i=1}^{j} \pi_i) & t = h + j \end{cases}$$

Se observa que después de un cambio de nivel todos los residuos se verán afectados, pero que el efecto depende: (1) del modelo, siendo más importante para los modelos estacionarios, (2) de la distancia entre el momento de aparición, h, y el final de la serie.

2.4.4. *Cambio Temporal (TC)*

El último de los atípicos que se consideran en la bibliografía es el denominado como cambio temporal ó TC, se define:

$$z_t = \frac{\omega_{TC}}{1 - \delta B} I_t^{(h)} + \psi(B) a_t \tag{2.45}$$

En realidad los casos presentados anteriormente, AO y LS, son casos frontera de los TC en donde $\delta = 0$ y $\delta = 1$ respectivamente. En el caso del TC, el atípico produce un efecto inicial ω_{TC} en el instante t y su efecto se agota gradualmente con el paso del tiempo. El parámetro δ establece la tasa de agotamiento. Habitualmente, en la práctica este tipo de atípico se utiliza estableciendo el valor de δ en 0.7.

2.4.5. Métodos de detección de valores atípicos

Es de esperar que el momento de la ocurrencia, la magnitud y la clasificación (A0, IO, LS, TC) de los valores atípicos presentes en una serie temporal sean desconocidos, por lo que se hace necesaria una metodología para esclarecer estos puntos. En los trabajos de Chang et~al. (1988), así como en Chen and Liu (1993b)[2] ó en Peña (2005) se presentan metodologías para la detección de atípicos.

Las distintas metodologías consiguen dos objetivos simultáneamente: (1) Obtener estimaciones robustas de los parámetros de modelos de series temporales, y (2) Revelar información de la localización y la naturaleza de los efectos de los atípicos. Con pequeñas variaciones siguen los siguientes pasos:

1. Se identifica el modelo de series temporales, sus parámetros (que podrían estar sesgados ya que no toman en cuenta los atípicos presentes) y residuos.

2. Se calculan los efectos de las cuatro clases de atípicos ($\omega_{AO}, \omega_{IO}, \omega_{LS}, \omega_{TC}$) en los residuos para cada instante t.

3. Se calculan estadísticos estandarizados ($\lambda_t^{IO,AO,LS,TC}$ según Peña (2005) o $\hat{\tau}_{IO,AO,LS,TC}(t)$ según Chen and Liu (1993b)) de los efectos de los atípicos mediante σ_a, siendo esta la desviación estándar de los residuos. Se debe tener precaución en eliminar por algún método el sesgo que contiene σ_a por los mismos atípicos.

[2]El software SCA Statistical System tiene implementada la metodología de Chen and Liu (1993b)

4. Si alguno de los estadísticos estandarizados supera el valor crítico C (3-3.5 según la sensibilidad deseada), existe la posibilidad de presencia de un atípico y se corrigen su efectos.

5. Iterativamente se estiman los nuevos parámetros del modelo de series temporales conjuntamente con los atípicos identificados.

La metodología funciona bastante bien para detectar atípicos en series con bajo nivel de contaminación de atípicos y presenta problemas en series con gran cantidad de atípicos. Para un mayor detalle sobre la metodología ver (Liu, 2006, pag. 7.20).

2.5. Las redes neuronales artificiales (ANN)

Las *ANN* quedan englobadas conjuntamente con la lógica difusa y los sistemas expertos dentro de las denominadas técnicas de inteligencia artificial. En los últimos años, la inteligencia artificial está encontrando aplicaciones en numerosas actividades cotidianas de los seres humanos y además ha encontrado un importante campo de aplicación para desarrollarse en áreas del conocimiento científico, tales como la robótica, la visión artificial, técnicas de aprendizaje, gestión del conocimiento, modelación de procesos físicos, etc.

2.5.1. *Conceptos Básicos*

El concepto y la idea original de redes neuronales artificiales (en delante ANN) aparece por primera vez en una publicación pionera presentada por McCulloch and Pitts (1943). Surge con la idea de emular y comprender las funciones que realiza el cerebro humano y su sistema nervioso por medio de una conceptualización del mismo en un modelo matemático de estructuras simples (neuronas) interconectadas. En definitiva, este trabajo fue básico para el posterior desarrollo de la técnica de redes neuronales tal y como ahora lo conocemos, e introduce por primera vez conceptos (con origen en la fisiología humana) como umbral de activación, excitación de neurona, sinapsis, pero ahora aplicados a modelos matemáticos.

A partir de esta primera publicación fueron apareciendo varias más, aunque es hasta la década de los ochenta cuando el tema de las ANN tiene un reimpulso con la publicación de Rumelhart et~al. (1986). En ella se introduce el algoritmo de retropropagación del error que es el más utilizado hasta la actualidad ya que representa un marco teórico matemático riguroso para las redes neuronales. A partir de la aparición de este algoritmo las redes neuronales empezaron a encontrar aplicaciones en áreas tan diversas de la ingeniería como pueden ser la computación, la acústica, la cibernética, el procesamiento de imágenes, solo por mencionar algunas. Un buen resumen de la evolución histórica de las *ANN* se puede encontrar en Haykin (1999) pág. 38, así como en Hilera and Martínez (1995) pág. 3.

Es difícil encontrar una definición unificada para las redes neuronales artificiales (*ANN*). Una que cubre muchos de los aspectos mencionados por la mayoría de los autores podría ser la de (García~Bartual, 2005, pag.24) que las define como un procesador distribuido en paralelo compuesto por unidades de proceso elementales (nodos o neuronas artificiales) masivamente interco-

nectadas entre sí y con organización jerárquica, con capacidad de adquirir conocimiento de tipo experimental con capacidad para simular y predecir procesos no lineales.

Las ANN se basan en las siguientes reglas (ASCE, 2000a):

1. El procesamiento de la información tiene lugar en muchos elementos individuales llamados nodos, también llamados unidades, células o neuronas.
2. Las señales son transmitidas entre los nodos a través de enlaces de conexión.
3. Cada enlace tiene asociado un peso que representa la potencia de su conexión.
4. Cada nodo aplica una transformación no lineal llamada función de activación a su red de entradas para determinar su señal de salida.

Y a su vez se pueden caracterizar por su arquitectura, que representa el patrón de conexión entre nodos, por el método de conexión de los pesos sinápticos y por su función de activación. Una red neuronal típica consta de un número de nodos que están organizados de acuerdo a un arreglo particular. Una forma de clasificar las redes neuronales es por el número de capas: única, bicapa o multicapa. La capa de entrada recibe las variables, una o varias capas ocultas conectadas con la capa anterior y con la siguiente pero que no interconecta sus propios nodos (la excepción son las redes neuronales recurrentes) procesan los datos y finalmente una capa de salida aporta la salida(s) del modelo o predicción del proceso analizado. Las *ANN* pueden también ser categorizadas por la dirección del flujo de la información y del procesamiento. La topología de las *ANN*, el número de capas ocultas y la cantidad de nodos en cada capa, se determinan por medio de prueba y error. La figura 2.7 muestra la configuración de una *ANN* feedforward de propagación hacia adelante.

2.5.2. *Fundamentos matemáticos de las redes neuronales*

Las entradas a un nodo (ver figura 2.8) perteneciente a una capa, serán las variables de entrada a una red o, serán las salidas de los nodos de otra capa dependiendo del sitio que ocupe la capa dentro de la red. Las entradas forman un vector de entradas $X = (x_1, x_2, ..., x_i, ..., x_n)$ y de la misma forma la secuencia de pesos asignada a cada entrada forman un vector de

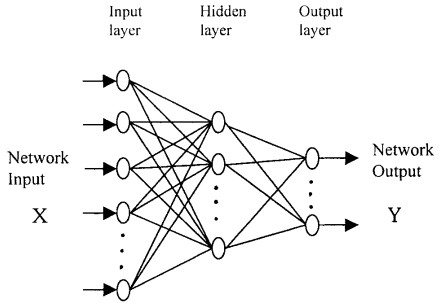


Figura 2.7: Configuración de una red neuronal *feedforwad* de 3 capas, ASCE (2000a)

pesos $W_j = (w_{1j}, ..., w_{ij}, ..., w_{nj})$, donde w_{ij} representa el peso de la conexión proveniente del nodo ith de la capa precedente hacia el nodo actual.

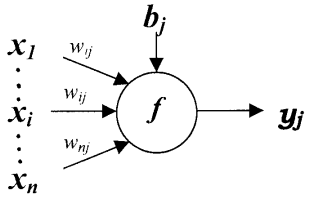


Figura 2.8: Diagrama esquemático de un nodo *j*, ASCE (2000a)

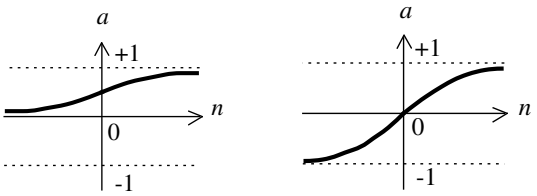
La salida del nodo j, y_j, se obtiene computando el valor de la función f con respecto al producto de los vectores X y W_j menos b_j, siendo b_j el umbral de activación del nodo (*bias* en la nomenclatura anglosajona). Este umbral (b_j) debe ser superado para que el nodo pueda activarse. La ecuación que define esa operación es:

$$y_j = f(X \cdot W_j - b_j) \tag{2.46}$$

La idea central de las *ANN* es que los parámetros (b_j, w_j) pueden ser ajustados para que la red reproduzca algún comportamiento deseado. De este modo, se puede entrenar una red para realizar un trabajo determinando sus pesos w_j y umbrales b_j.

La función f se denomina función de activación y su forma determina la respuesta del nodo al total de la señal de entrada que recibe. Las funciones de activación más comúnmente utilizadas en (2.46) son las del tipo sigmoidal (ver figura 2.9), debido principalmente a que son diferenciables y esto facilita la búsqueda de solución a los algoritmos empleados. Se define como

$$f(t) = \frac{1}{1 + e^{-t}} \tag{2.47}$$



(a) Función de activación logsigmoidal

(b) Función de activación tansigmoidal

Figura 2.9: Funciones de activación tipo sigmoidal, Demuth et~al. (2009) Matlab R2008b

Las funciones sigmoidales son ecuaciones asintóticas con forma de *S* que pueden aceptar entradas de todo el espacio de los números reales ($-\infty$ a $+\infty$) y acota sus salidas entre (0,1) ó (-1,1). También son frecuentemente utilizadas las funciones tipo escalón (definida como hard-lim en Demuth et~al. (2009) Matlab R2008b) figura (2.10) y finalmente las funciones del tipo lineal (ver figura 2.11) que típicamente se ubican en la capa de salida de la red.

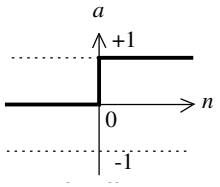


Figura 2.10: Función de activación tipo escalón, Demuth et~al. (2009) Matlab R2008b

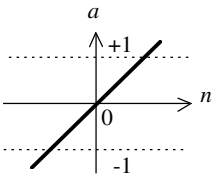


Figura 2.11: Función de activación lineal, Demuth et~al. (2009) Matlab R2008b

2.5.3. *Reglas de Aprendizaje*

Las redes neuronales tiene la propiedad fundamental de ser capaces de aprender de su entorno y mejorar su desempeño mediante el aprendizaje.

En el caso de las ANN el conocimiento se encuentra representado en los pesos sinápticos de las conexiones entre neuronas. Haykin (1999) define el aprendizaje en el contexto de las ANN como:

> *El aprendizaje es un proceso por el cual los parámetros libres de una red neuronal son adaptados mediante un proceso de simulación por el entorno en el cual la red neuronal está embebida. El tipo de aprendizaje se determina por la manera en la cual los cambios en los parámetros tienen lugar.*

Un algoritmo de aprendizaje es un conjunto de reglas bien definidas para la solución de un problema de aprendizaje. No existe un algoritmo de aprendizaje genérico para todas las ANN, en cambio existen un conjunto de algoritmos, cada uno con sus ventajas para determinados problemas y se diferencian entre ellos por la forma en que se formula el ajuste de los pesos sinápticos. Las reglas de aprendizaje se agrupan en dos grandes categorías: el aprendizaje supervisado y no supervisado. En el primer caso, se le proporciona al algoritmo de aprendizaje con un conjunto de ejemplos (el conjunto de entrenamiento) del comportamiento adecuado de la red, es decir se le proporciona una entrada y la respuesta correcta esperada para fines de comparación. Los pesos sinápticos son ajustados iterativamente hasta acercar las salidas de la red a los objetivos. En la segunda categoría, el aprendizaje no supervisado, los pesos son modificados solamente como respuesta a las entradas a la red, no existen objetivos de salida para comparación. Tienen su aplicación principalmente en problemas de clasificación. El algoritmo de aprendizaje supervisado más utilizado es el de retropropagación del error EBLA (error backpropagation learning algorithm). El nombre dado al algoritmo proviene de la forma en que los errores a la salida de la red son propagados hacia atrás atravesando en sentido inverso las distintas capas de la red. El EBLA es una adaptación del algoritmo de Widrow-Hoff. La implementación más sencilla de retropropagación actualiza los pesos de aprendizaje en la dirección en la cual la función de desempeño disminuye más rápidamente, el negativo del gradiente, en base a una tasa de aprendizaje preestablecida y variable. Existen una gran cantidad de variaciones del algoritmo básico. Por ejemplo la que adiciona al algoritmo de retropropagación con un *momento* o inercia con la cual se habilita a la red para responder no solamente al gradiente local, sino también a las tendencias recientes de la superficie de error. Esta incorporación actúa como un filtro que permite a la red ignorar pequeños sucesos en la superficie de error con el fin de prevenir que la red se quede atrapada en un mínimo local.

Los algoritmos de aprendizaje hasta ahora mencionados, si bien es cierto que aportan buenas soluciones, suelen ser lentos y requerir de una gran

cantidad de iteraciones para converger hasta una solución. Existen algoritmos que se basan en otras técnicas estándar de optimización y que suelen ser mucho más rápidos (hasta 100 veces más rápidos), tales como el de gradientes conjugados (éste a su vez tiene sus propias variaciones, e.g. Fletcher-Reeves Update, Polak-Ribiére Update, Powell-Beale Restarts, Scaled Conjugate Gradient), el de Quasi-Newton (BFGS Algorithm y One Step Secant Algorithm) o el de Levenberg-Marquardt y su variación de memoria reducida.

El aprendizaje Hebbiano es la regla de aprendizaje no supervisado más conocida. Se basa en el trabajo del neurocirujano Canadiense Donald Hebb, quien teorizó que el aprendizaje neuronal (i.e. cambio sináptico) es un fenómeno local expresable en términos de la correlación temporal entre los valores de activación de las neuronas. Específicamente, el cambio sináptico depende de las actividades pre-sinápticas así como de las post-sinápticas y establece que el cambio en un peso sináptico es una función de la correlación temporal entre las actividades pre-sinápticas y post-sinápticas. Específicamente, el valor del peso sináptico entre dos neuronas aumenta cuando ambas están en el mismo estado y disminuye cuando ambas están en diferentes estados.

2.5.4. *Ventajas y Limitaciones de las* ANN

Son muchas las ventajas y limitaciones que se podrían argumentar a favor y en contra de las *ANN*. Sin embargo, lo más adecuado sería analizar individualmente cada tipo de red, pero ese no es el objetivo de esta sección ni de este documento. En referencia a las redes con algoritmos tipo EBLA mencionamos las que consideramos más importantes.

Ventajas

1. La ventaja más importante de las *ANN* se encuentra en la resolución de problemas que son demasiado complejos para las tecnologías convencionales, problemas que no tienen una solución algorítmica o problemas en los que la solución algorítmica es demasiado complicada para ser identificada

2. Proporcionan una alternativa analítica a las técnicas convencionales que frecuentemente están limitadas por estrictas hipótesis de normalidad, linealidad, independencia, etc.

3. Por su facilidad para capturar muchas clases de relaciones, permiten al usuario de una forma relativamente fácil modelar fenómenos que con otras metodologías podría resultar muy complicado o incluso imposible.

Limitaciones

1. Las teorías matemáticas usadas para garantizar el desempeño de las *ANN* aplicadas están aún en desarrollo.
2. Inestabilidad para explicar los resultados que obtienen.
3. Problemas de escalabilidad, prueba y verificación.
4. El producto final de la actividad de una *ANN* es una red entrenada que no proporciona ecuaciones o coeficientes que definan una relación (como en los modelos de regresión) que vaya más allá de su propia matemática interna. La red es la ecuación final de la relación.

Parte III

Estado del arte

Capítulo 3

Revisión del estado del arte

El mejor profeta del futuro es el pasado.
Lord Byron
1788 – 1824
Poeta inglés.

Se dice que el presente está preñado del futuro.
François Marie Arouet, Voltaire
1694 – 1778
Escritor y filósofo francés.

A lo largo de este capítulo presentaremos un revisión a modo de resumen de la evolución de las distintas técnicas utilizadas en la modelación y predicción de la demanda de agua potable a las diferentes escalas, ya sea mensual, diaria u horaria.

Es evidente que la introducción de los modelos ARIMA, conjuntamente con los modelos de función de transferencia y el análisis de intervención, popularizados por George Box y Gwilym Jenkins en los primeros años de la década de 1970 con la publicación del libro *Time Series Analysis. Forecasting and Control* (Box and Jenkins, 1970), así como también por el artículo de Box and Tiao (1975), marcaron un hito y dotaron de herramientas estadísticas rigurosas a investigadores del área de la hidráulica y la hidrología para modelar fenómenos que se presentan en la naturaleza. En el campo de los gestores de sistemas de agua potable, esta aportación representó un salto cualitativo a la hora de modelar y predecir las series históricas de demanda de agua potable, ya que introdujo una técnica que permitió hacer un uso mucho más eficiente de los datos. Si bien es cierto que existen publicaciones anteriores, fue David R. Maidment con varios colaboradores (E. Parzen, M. Crawford, S. Franklin, Miaou), el primero en desarrollar y explotar una línea de

investigación con un conjunto de publicaciones de aplicaciones específicas en el campo de la gestión de sistemas de agua potable conteniendo el germen de la metodología de George Box para modelar y predecir la demanda de agua en distintas ciudades.

El capítulo está organizado cronológicamente, por lo que el hilo conductor que organiza y estructura este capítulo será el tiempo. La excepción será la parte correspondiente a las redes neuronales artificiales, ya que estas han tenido una evolución temporal diferenciada de la de los métodos de series temporales. La revisión no está organizada como un agrupamiento de técnicas en base a autores o metodologías, en cambio, se irán describiendo y comentando las publicaciones conforme fueron divulgadas en los distinto medios. Las revistas científicas en las que se identificaron publicaciones relacionadas con la modelación y predicción de la demanda de agua potable y sus temas relacionados se enlistan en el Apéndice A.

Agrupando las evoluciones en el tema de la modelación y predicción de la demanda de agua potable por décadas, tendríamos que:

- Década de 1970
  - Aparición de la metodología Box y Jenkins
  - Modelos ARIMA
  - Modelos de función de transferencia
  - Análisis de intervención
- Década de 1980
  - Modelación de la demanda de agua potable con la metodología Box y Jenkins
  - Principalmente a escala mensual con buenos resultados
  - Modelación a escala diaria como función de series de temperatura y lluvia
  - Modelos de función de transferencia demanda-variables climáticas
- Década de 1990
  - Aparición de la metodología de redes neuronales artificiales (*ANN*)
  - Modelación de la demanda a escala horaria
  - Confrontación de las metodologías de *ANN* contra las de fundamento estadístico - de series temporales
- Década de 2000
  - Metodologías híbridas, *ANN* y de fundamento estadístico
  - Lógica difusa

3.1. Antecedentes

Los primeros trabajos presentados en publicaciones científicas relacionados directamente con la predicción de demandas de agua urbana, fueron los desarrollados por David R. Maidment (Universidad de Texas en Austin) en colaboración con Emanuel Parzen (Texas A & M). Maidment and Parzen (1984b) y Maidment and Parzen (1984a) Fueron publicados casi simultáneamente. Este último es una aplicación práctica de la metodología presentada en el primero. Sin embargo, esto no significa que el tema no haya sido abordado con anterioridad. Las publicaciones presentadas hasta entonces, consistían principalmente en estudios desarrollados por dependencias gubernamentales estadounidenses e integrados como reportes técnicos y en anuarios estadísticos de la AWWA[1] de varias ciudades de los Estados Unidos de Norteamérica (Seidel, 1978). En estos trabajos, el estudio de la demanda de agua urbana se basaba en el uso del concepto del consumo *per cápita* desde donde obtenían los valores de consumos máximos mensuales, diarios y horarios *per cápita* que eran usados con fines de diseño. La tendencia de uso del agua la obtenían multiplicando la dotación *per cápita* por las proyecciones del crecimiento de la población para finalmente obtener el volumen de agua esperado. Otro grupo de trabajos consistían en análisis estadísticos de la evolución de la demanda a escala anual y/ó mensual, así como en construcción de modelos de regresión que involucraban variables socio-económicas (precio del agua, rentas por familia, número de personas por m^2, elasticidad del precio del agua, etc.), es decir, tenían un enfoque econométrico aunque en algunos contados casos se incluyó en estos análisis la precipitación anual como variable secundaria. Estos modelos de regresión estaban pensados para conocer los picos de demanda esperados, para luego utilizarlos en el dimensionamiento de las instalaciones de los sistemas de distribución. Algunos de ellos, como el presentado por Hansen and Narayanan (1978), en el cual construyeron modelos de regresión de la demanda contra variables climáticas, encontraron que la lluvia de verano, la temperatura del aire y las horas de sol son variables explicativas de la demanda mensual en la ciudad de Salt Lake City, UTAH. Anderson et~al. (1980) utilizaron la evapotranspiración, la lluvia efectiva y niveles de embalses para sus modelos de regresión en un verano de sequía en Fort Collins, Colorado. Como se comenta en Maidment and Parzen (1984b) los modelos de regresión estadística mencionados anteriormente, son útiles para depurar los factores que influencian significativamente las diferencias en el uso del agua entre una ciudad y otra, pero no nos aclaran la estructura inherente de los patrones de

[1] American Water Works Association

uso de agua a lo largo del tiempo. En cambio, la metodología propuesta en Maidment and Parzen (1984b) [2] enfoca su trabajo en esta línea.

[2]El primer planteamiento de esta metodología fue propuesto por Salas~LaCruz and Yevjevich (1972) en una publicación interna de la universidad estatal de Colorado, Fort Collins

3.2. El modelo de transformaciones en cascada, Maidment and Parzen (1984b)

La metodología de este trabajo asume que se cuenta con una serie histórica de demandas de agua y una serie de observaciones de variables socio-económicas y climáticas que pueden estar relacionadas. Para su estudio, propone una conjunto de transformaciones consecutivas que las denominaron como en *cascada* (ver el esquemático 3.1). La hipótesis fundamental propuesta es que la serie puede ser descompuesta en componentes de memoria larga y de memoria corta. Las componentes de memoria larga constan de tendencia y estacionalidad, donde la tendencia refleja las variables que varían lentamente a lo largo del tiempo, como pueden ser la población, el precio del agua, las rentas familiares, y la estacionalidad refleja el patrón cíclico de variación en el uso del agua durante el año. Los parámetros de las funciones que describen la tendencia y la estacionalidad, se considera que una vez estimados pueden operar independientemente de los valores de demanda de agua de algún año en particular. En cambio, las componentes de memoria corta no son totalmente predecibles, ya que dependen de las observaciones actuales y pasadas de la demanda. Estas componentes del estudio en cuestión son la autocorrelación y correlación climática. La autocorrelación refleja la perpetuación de las desviaciones de la demanda con respecto a las variaciones de los patrones de largo plazo, y la correlación climática refleja el efecto de eventos climáticos anormales tales como temporadas de sequía, fuertes eventos de lluvia o cambios en la temperatura.

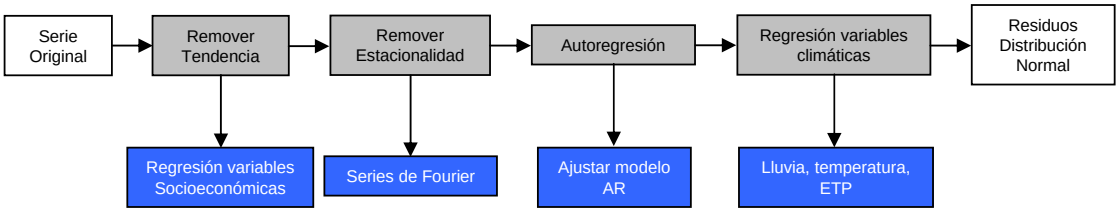


Figura 3.1: Cascada de transformaciones a las series temporales de demanda de agua, (Maidment and Parzen, 1984b)

Cada una de las transformaciones se describe a continuación:

1. **Remover tendencia**: Consiste en hacer una regresión de la demanda media anual frente a variables socioeconómicas como pueden ser la población, el número de conexiones domiciliarias, las rentas por vivienda y/ó el precio del agua. Los residuos de esta regresión se utilizan para

transformar la serie temporal original de demandas mensuales de agua en una serie temporal sin tendencia, en la cual el efecto de los cambios a lo largo del tiempo aportados por las variables socio-económicas han sido removidas.

2. **Desestacionalización**: Se ajusta una serie de Fourier a las medias mensuales de la serie temporal ya sin tendencia. La resultante suavizada ajustada a las medias mensuales es sustraída de la serie temporal transformada para producir una serie temporal desestacionalizada de media cero. Los ciclos estacionales han sido removidos.

3. **Filtrado Autoregresivo**: Consiste en ajustar un modelo autoregresivo (AR) a la serie temporal desestacionalizada para dar cuenta de la dependencia del uso de agua a sus propios valores pasados; los residuos de este modelo están independientemente distribuidos en el tiempo y se denominan datos "*preblanqueados*" (prewhitened) por ser casi un proceso aleatorio puro.

4. **Regresión Múltiple**: Consiste en hace una regresión de los residuos resultantes del modelo autoregresivo frente a las series temporales de precipitación, evapotranspiración y temperatura máxima del aire, las cuales han pasado previamente por el mismo proceso de desestacionalización y filtrado autoregresivo descrito para la demanda de agua. Los residuos de esta regresión, son valores de una variable aleatoria pura de los cuales ha sido removida la correlación con variables climáticas. La distribución de probabilidad de los residuos finales debe ser determinada y verificada su distribución normal.

Matemáticamente, esta metodología se resume de la siguiente manera. La serie original $W_a(t)$, se expresa como la suma de un proceso de memoria larga, $W_l(m, y)$ y un componente de memoria corta, $W_s(t)$:

$$W_a(t) = W_l(m, y) + W_s(t) \tag{3.1}$$

donde $t = 12(y - 1) + m$, t es un indicador de los meses desde el inicio de la serie, $t = 1, 2, ..., T$; m es un indicador mensual dentro del año, $m = 1, 2, ..., 12$; y es un indicador de los años, $y = 1, 2, ..., Y$.
El componente de memoria larga, que consta de tendencia y estacionalidad se expresa como

$$W_l(m, y) = W_a(y)C_s(m) \tag{3.2}$$

donde $W_a(y)$ es el valor medio anual de uso de agua en el año y; y $C_s(m)$ es un coeficiente adimensional que relaciona el uso de agua mensual en el mes m con la media mensual del uso de agua de la serie temporal. Para ejemplificar mostramos la figura (3.2) de descomposición de una serie temporal de demanda en sus componente de memoria corta y larga.

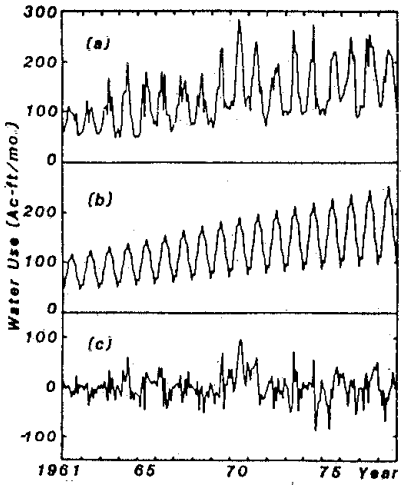


Figura 3.2: Partición de la serie temporal de demandas de agua. (a) Componente de memoria larga (b) Componente de memoria corta (c) $a = b + c$, Datos de Canyon, Texas 1961-1978 (Maidment and Parzen, 1984b)

Una vez que $W_l(m, y)$ es identificada, el componente de memoria corta, $W_s(t)$, se obtiene de

$$W_s = \sum_{i=1}^{p} \alpha_i W_s(t-i) + \sum_{j=1}^{q} \beta_j X_j(t) + \epsilon(t) \tag{3.3}$$

donde el primer sumatorio es una modelo autoregresivo de orden p para $W_s(t)$. El segundo sumatorio es una regresión de valores concurrentes de q variables climáticas preblanqueadas, $X_j(t)$, $j = 1, 2, ..., q$; y $\epsilon(t)$ es un proceso aleatorio puro de los errores residuales de media cero y varianza constante.

3.2.1. *Aportaciones del modelo transformaciones en cascada*

La técnica aplicada en el modelo que nos ocupa, aporta un nuevo concepto distinto al de meramente descubrir los factores que significativamente influencian a la demanda de agua que se obtiene de los modelos de estadísticos de regresión. La técnica de transformaciones en cascada no va

directamente a la búsqueda del resultado que se quiere obtener. En cambio descompone el resultado total —que es el único primeramente deseado— en los componentes parciales de los que surge, en el proceso de su génesis, por tanto, en sus *causas* o fenómenos ingredientes. Por medio de cada una de las transformaciones aplicadas a la serie temporal original, se va eliminando, y por lo tanto conociendo un porcentaje de la varianza atribuible a cada factor: tendencia, estacionalidad, autocorrelación y correlación climática, siendo la varianza explicada de cada factor la diferencia de la varianza de la serie temporal calculada antes y después de aplicar una transformación, presentada como un porcentaje de la varianza de la serie temporal original. En los resultados presentados en Maidment and Parzen (1984b) el método analítico logra explicar en promedio el 80-87 % de la varianza de la demanda mensual, encontrando que esos valores son mejores para las ciudades grandes y con climatología menos extrema. El componente de memoria larga en promedio explica el 70% de la varianza total y el de memoria corta el 15%. Los autores concluyen que la alta proporción del componente de memoria larga en estas ciudades significa que la demanda mensual a un horizonte de varios años es predecible si se cuenta con predicciones precisas de la población.

La mayoría de las publicaciones sobre modelación y predicción de demandas de agua posteriores, se basan en esta metodología, con variaciones en la forma de filtrar las series en cada una de las fases de transformaciones, y también –con mayor o menor éxito– con aplicaciones en distintas escalas temporales, entiéndase horaria, diaria, semanal y mensual.

3.2.2. *Comentarios al modelo de transformaciones en cascada*

La metodología del modelo de transformaciones en cascada asume que con esas transformaciones, la serie llegará a ser estacionaria en primer y segundo orden y también ergódica, requisitos básicos para la modelación estocástica. Sin embargo no debemos olvidar que el tema que nos ocupa, la demanda de agua, es esencialmente un proceso dinámico que depende de la temperatura y es interrumpida por la ocurrencia de lluvia (Sastri and Valdes, 1989) pero también es afectada por fenómenos sociológicos propios de cada comunidad. Estos fenómenos transitorios y de corta duración (lluvia, temperatura, sociológicos) provocan que las series de demandas no sean homogéneas a lo largo de las observaciones por lo que es de esperar que existan distintas correlaciones en distintos momentos. Esta relación de no homogeneidad y no estacionaridad entre las observaciones resulta en que un modelo que es calibrado a partir de un grupo finito de datos de demanda,

solamente es válido para ese conjunto de datos. Si este conjunto de datos contiene alguno de los fenómenos antes citados nos puede guiar a una identificación errónea del modelo.

Es importante recalcar que el modelo requiere el cálculo de 23 parámetros hasta obtener un resultado final, lo que aporta a la predicción un incremento de la incertidumbre con la inclusión de cada parámetro adicional, siendo esta del orden de $1/n$ donde n es el número de observaciones lo que convierte al modelo en poco parsimonioso aunque se pueda contar con un amplio número de grados de libertad. El modelo de transformaciones en cascada, por varios de los motivos antes mencionados, producirá mejores resultados en la modelación de series temporales que de alguna forma enmascaren variaciones bruscas de la demanda por originarse en valores agrupados, como pueden ser las series de demandas anuales, mensuales y en algunos casos semanales que presentan series suavizadas. Se esperarían peores resultados a la hora de utilizar esta técnica en modelación y predicción a escalas diaria y horaria dada la mayor influencia que ejercen sobre ella los fenómenos transitorios antes citados.

3.3. Evolución del Modelo transformaciones en cascada

La primera evolución del modelo de transformaciones en cascada prácticamente se dio conjuntamente con su publicación inicial, ya que los mismos autores publicaron la metodología más en detalle en el artículo Maidment and Parzen (1984a); en el cual la aplican para la ciudad de Canyon, Texas, que a su vez fue parte del grupo de ciudades estudiadas en la anterior publicación. Esta publicación se enfoca a proponer alternativas a las presentadas anteriormente para conseguir los objetivos de cada una de las transformaciones a las que es sometida la serie temporal. Es así que para la primera transformación de la serie, la remoción de la tendencia, ya no solo propone la regresión de la serie frente a las variables que originan la tendencia. En cambio propone como alternativas para estabilizar la tendencia, aplicar logaritmos a los datos, aplicarles la raíz cuadrada o ajustar una función polinomial, etc. Se echa en falta la diferenciación simple entre las propuestas. Ya se verá más adelante que varias de ellas son frecuentemente utilizadas en estudios posteriores. En la segunda transformación de la serie, la desestacionalización, además del ajuste de la serie de Fourier, propone como alternativa construir un modelo autoregresivo cuyos coeficientes reflejen la estacionalidad. Adicionalmente propone hacer una regresión de la serie frente a una variable estacional como puede ser la temperatura del aire. En la transformación correspondiente a la autocorrelación, que tiene como intención eliminar la correlación de la serie con sus propios valores pasados, más que una alternativa, lo que nos propone son pruebas para asegurar que se ha determinado correctamente el orden p del modelo autoregresivo. Así es que para este fin introduce la utilización ya sea del criterio de Parzen (1979) o el de Akaike (1974), que para el momento de la publicación eran relativamente recientes. Por otra parte, y con el fin de identificar por medio de una comparación visual si la serie, una vez filtrada por esta última transformación conserva las propiedades espectrales de la serie original, propone el cálculo de la función de densidad espectral suavizando el periodograma y comparándolo con la de la serie original (ver figura 3.3).

Por último, para revisar la distribución de probabilidad de los residuos finales, después de pasar por las cuatro transformaciones propone un procedimiento no paramétrico llamado *análisis de una muestra* (one simple analysis). Si el proceso de transformaciones ha dado el resultado deseado, deberíamos esperar que los residuos finales se ajusten a una distribución normal de media cero y varianza constante.

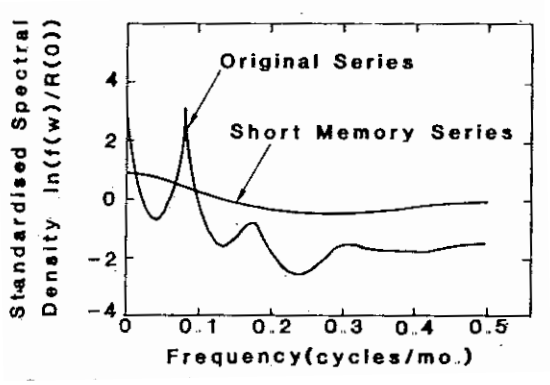


Figura 3.3: Función de densidad espectral de la demanda de agua mensual y su componente de memoria corta, Canyon, Texas 1961-1978. (Maidment and Parzen, 1984a)

Llegados a este punto, la metodología planteada en Maidment and Parzen (1984b) y Maidment and Parzen (1984a) deja abiertas unas líneas de investigación que fueron abordadas en publicaciones posteriores. Estas líneas se mencionan a continuación:

1. Identificación de tests que nos indiquen cuando una transformación adicional en la serie es necesaria para asegurar la estacionareidad de la covarianza

2. La relación de la lluvia y su interacción con el uso del agua. La lluvia se presenta por rachas y requiere de una respuesta dinámica para representar su efecto en la demanda de agua.

3. Los parámetros del modelo son estimados secuencialmente, no simultáneamente.

En Franklin and Maidment (1986) se presenta una aplicación de la metodología para series de demanda urbana a escala semanal y mensual en la ciudad de Deerfield Beach, Florida. El estudio identifica una relación lineal entre la lluvia y la demanda a escala mensual. Sin embargo la inclusión de la lluvia tanto en el modelo mensual como en el semanal, no aportó mejoras significativas en su desempeño. Los máximos errores se obtuvieron cuando se analizó la serie en un periodo de sequía, seguida de lluvias extremas, lo cual evidencia la incapacidad del modelo lineal para reproducir la realidad del efecto de eventos extremos de lluvia en la demanda. La lluvia fue utilizada para plantear distintos escenarios y obtener una variabilidad máxima esperada, es decir para realizar análisis de riesgo.

3.4. Los modelos de función de transferencia de la demanda

La relación de la demanda de agua y su respuesta a la presencia de lluvia fue un tema que quedó patente y sin resolver hasta este punto de nuestra revisión del estado del arte. Las publicaciones existentes trataron el tema casi en su mayoría en intervalos temporales (anual, mensual, semanal) que oscurecen la relación causa-efecto entre la lluvia y la demanda de agua, esta relación es mucho más evidente en la escala diaria (Maidment et~al., 1985). Weeks and McMahon (1973) encontraron que en Australia, el número de días de lluvia por año fue la variable climática más significativa que afecta el uso de agua per-cápita anual. Pero también encontraron que la evapotranspiración semanal y la temperatura máxima media fueron variables explicativas más significativas que la lluvia en un modelo de regresión lineal múltiple que describe las demandas máximas semanales. La razón más probable por la que la lluvia resulta no ser significativa en el análisis semanal, es la de que su relación con la demanda no es lineal ni invariante en el tiempo como asume el análisis de regresión lineal. La evapotranspiración y la temperatura del aire varían en un rango positivo como lo hace el uso del agua y esto podría resultar en coeficientes de regresión que son más significativos que los coeficientes de la lluvia en los modelos de regresión lineal múltiple.

La complejidad del problema y su incompatibilidad con los modelos de regresión múltiple reside en que la lluvia es un proceso puntual que ocurre en pulsos pero es cero la mayoría del tiempo. El problema fue abordado por Maidment et~al. (1985) proponiendo la utilización de funciones de transferencia (Box and Jenkins, 1970) para representar la relación lluvia diaria-demanda de agua, ya que estos modelos tienen la capacidad de representar la respuesta dinámica característica de un evento de lluvia en la variable dependiente, un impulso inmediatamente después del evento de lluvia que se atenúa paulatinamente a lo largo del tiempo. El fenómeno se puede apreciar en el gráfico (3.4) presentado en su publicación.

El planteamiento propuesto en este trabajo se aleja de la metodología presentada en Maidment and Parzen (1984b) y Maidment and Parzen (1984a). Sus principales postulados son:

1. La demanda de agua urbana puede dividirse en uso base y uso estacional, donde el uso base representa la demanda durante los meses de invierno[3]

[3]El supuesto de una demanda base es un concepto de difícil aplicación en patrones de demandas de ciudades españolas, más adelante se abordará el tema

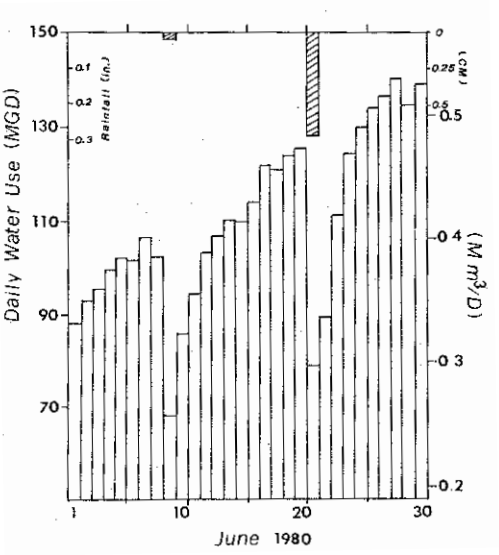


Figura 3.4: Demanda diaria en Austin, Texas, durante Junio de 1980 mostrando el efecto de dos lluvias aisladas. (Maidment et~al., 1985)

2. La variación de la demanda estacional en ausencia de lluvia sigue un patrón que depende de la temperatura del aire

3. La ocurrencia de lluvia provoca una reducción temporal en el uso de agua estacional que se atenúa con el paso del tiempo y eventualmente se vuelve insignificante

3.4.1. *Metodología de modelos de función de transferencia de la demanda*

La metodología supone que la demanda diaria se compone de uso base y uso estacional, ambas exhibiendo una tendencia en el tiempo. El uso estacional tiene dos componentes, uno que varia suavemente durante el año con la temperatura del aire y otro que representa los residuales de memoria corta

$$W(t) = \hat{W}_b(t) + g(t)\left[\hat{W}_p(t) + W_s(t)\right] \tag{3.4}$$

donde

- W es la demanda diaria de agua;

- $\hat{W}_b$ es uso base estimado;
- g coeficiente de tendencia para la demanda pico estacional;
- $\hat{W}_p$ demanda potencial estimada, función de la temperatura;
- W_s demanda de agua de memoria corta;
- t índice del tiempo desde el inicio de la serie

La ecuación del modelo de función de transferencia para la serie de memoria corta es de la siguiente forma:

$$W_s(t) = \bar{W}_s + \sum_{i=1}^{2} \frac{\omega_0^{(T_i)}}{1 - \delta_1^{(T_i)} B} T_i(t) + \sum_{i=1}^{2} \frac{\omega_0^{(R_i)} - \omega_1^{(R_i)} B}{1 - \delta_1^{(R_i)} B} R_i(t) + \\ + \frac{1}{1 - \phi_1 B - \phi_2 B^2 - \phi_7 B^7} a(t) \tag{3.5}$$

donde

- $\bar{W}_s$ es el componente del nivel del modelo de la serie de memoria corta;
- T es el promedio de la temperatura diaria del aire transformada;
- R es la lluvia diaria o alguna variable sustituta para los efectos de la lluvia;
- a es una variable independiente, con distribución normal, con media y varianza cero σ_a^2;
- i índice de la temporada del año o el rango de una variable;
- $\omega_0, \omega_1, \delta_1$ son los coeficientes de la función de transferencia;
- ϕ_1, ϕ_2, ϕ_7 son los coeficientes autoregresivos del modelo de ruido;
- B Operador de retardo

Ecuaciones para la tendencia

La tendencia en los componentes base y estacional se puede observar en los valores de demanda agrupados a escala mensuales. Se debe realizar una regresión para las demandas máxima y mínima mensual contra el tiempo. La demanda base mensual promedio $W_b(m)$ es obtenida del valor de demanda mínimo mensual. Se ajusta un polinomio del tipo

$$\hat{W}_b = a_0 + a_1 m + a_2 m^2 + ... \tag{3.6}$$

El caso presentado en esta publicación corresponde a la ciudad de Austin, Texas. El mes con la demanda mínima es enero, por lo que $m = 1, 13, 25, ...,$ en la ecuación (3.6), aunque típicamente solo a_0 y a_1 son estadísticamente significativos.
La demanda pico mensual se estima de forma similar. La demanda base esperada $\hat{W}_b(m)$ se resta de los datos en los meses que típicamente exhiben los valores de demanda máximos (julio y agosto en Austin, por lo que m=7,8,19,20...). Posteriormente se hace una regresión de la demanda estacional restante $S_p(m)$ con el tiempo, la lluvia mensual $R(m)$ y la temperatura media mensual $T(m)$:

$$\hat{S}_p(m) = b_0 + b_1 m + b_2 m^2 + ... + b_R[R(m) - \bar{R}(m)] + b_T[T(m) - \bar{T}(m)] \tag{3.7}$$

donde $\bar{R}(m)$ y $\bar{T}(m)$ son los valores de las medias aritméticas de $R(m)$ y $T(m)$ respectivamente en ecuación (3.7). Típicamente solo b_0, b_1, b_R y b_T son estadísticamente significativos. Se pueden obtener ecuaciones equivalentes de intervalo diario para la tendencia en el tiempo de la demanda base $\hat{W}_b$. Las ecuaciones (3.6) y (3.7) se utilizan para eliminar la demanda base de los datos y para convertir la componente de demanda estacional en una serie temporal estacionaria. Primero, la demanda estacional $S(t)$ se obtiene restando la demanda base a la demanda total, $W(t)$:

$$S(t) = W(t) - \hat{W}_b(t) \tag{3.8}$$

Segundo, se elige una fecha de referencia, de pivote, t_0 para estandarizar los datos. Se calcula el coeficiente de crecimiento $g(t) = \frac{\hat{S}_p(t)}{\hat{S}_p(t_0)}$ y se utiliza para generar una serie temporal sin tendencia, estacionalmente estacionaria $S_d(t)$ de la demanda de agua estacional:

$$S_d(t) = \frac{S(t)}{g(t)} \tag{3.9}$$

Estas transformaciones se pueden apreciar en el gráfico (3.5).

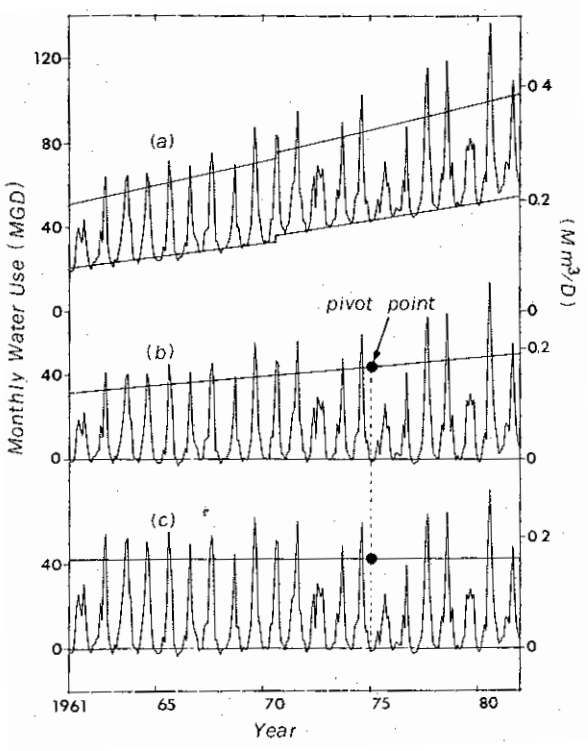


Figura 3.5: Procedimiento para separar y remover tendencia para una serie de demandas mensual en Austin, Texas. (a)Se ajusta una línea de tendencia para la demanda base (b)La demanda base es sustraída y se ajusta una línea de tendencia con los valores máximos mensuales (c)La línea de tendencia de valores máximos se rota respecto a un punto pivote para producir un serie de demanda estacional estacionaria (Maidment et~al., 1985)

Función de calor

Los diferentes valores de demanda que se presentan para cada año, son resultado de la respuesta conjunta de los usuarios del servicio a los efectos de la temperatura y de la lluvia en sus decisiones respecto a regar sus zonas ajardinadas.[4] Para modelar la variación estacional, se asume que existe una relación funcional entre el agua usada y la temperatura del aire, la

[4]Los usuarios del servicio utilizan más o menos agua según las condiciones meteorológicas imperantes. No se debe olvidar que la propuesta fue realizada con datos de la ciudad de

cual es válida en ausencia de lluvia reciente. Esta relación es más evidente durante los periodos secos y menos notoria en periodos húmedos. Por este motivo, es deseable el filtrado de los datos para seleccionar solamente los periodos durante los cuales los efectos de la lluvia son mínimos, con el fin de estudiar la relación entre la demanda de agua y la temperatura del aire. En Austin, se usaron los valores de la demanda media semanal y de la temperatura del aire para estimar la función de calor $H(T)$. Durante el verano (periodo comprendido entre los meses de abril y octubre) se seleccionan los datos de tal forma que no exista evento de lluvia en las dos semanas previas ni en la semana en consideración. Para el resto del año se consideran periodos de 3 días sin lluvia antecedentes a la semana en cuestión para definir un periodo seco. Los datos promedio semanales seleccionados se grafican frente a la temperatura media semanal T. La función de calor también se estudió usando datos diarios y mensuales, pero el intervalo de tiempo semanal llevó a mejores resultados. La función de calor se estima como

$$H(T) = c_0 + c_1 T \tag{3.10}$$

donde c_0 y c_1 son coeficientes que dependen de los tres rangos de T (temperatura del aire) a los que pertenezcan en la gráfica (3.6). Los valores de los coeficientes y los puntos de quiebre son obtenidos por medio de regresión lineal. Los datos observados en el gráfico, indican que existen tres estaciones del año: una de verano (abril a octubre), una de transición (marzo y noviembre), y una de invierno (diciembre a febrero).

La demanda potencial diaria, $W_p(t)$, se estima sustituyendo la temperatura diaria del aire normal que se obtiene de los registros a largo plazo, $T_N(t)$, en la función de calor para cada día del año.

$$W_p(t) = H[T_N(t)] \tag{3.11}$$

Los efectos de memoria corta de la lluvia, la temperatura y los errores aleatorios son obtenidos como:

$$W_s(t) = S_d(t) - W_p(t) \tag{3.12}$$

Para estudiar los efectos de memoria corta de la temperatura del aire usando la ecuación (3.5), la temperatura del aire residual se obtiene en delante:

Austin en el estado de Texas donde una vivienda típica suele tener áreas ajardinadas tanto en el frente como en el patio trasero de la vivienda

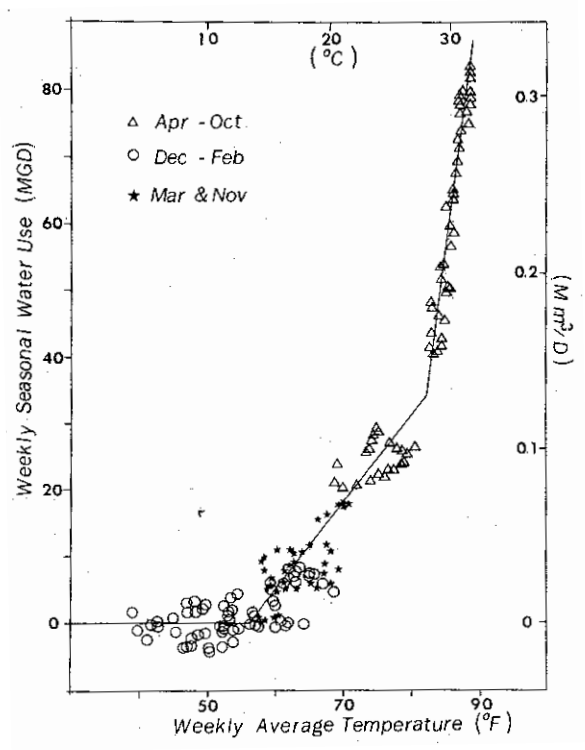


Figura 3.6: Función de calor que relaciona la demanda estacional y la temperatura media del aire. Los datos mostrados corresponden a valores medios semanales durante periodos sin lluvia. (Maidment et~al., 1985)

$$T_i(t) = T(t) - T_N(t) \tag{3.13}$$

El razonamiento para usar esta técnica de desestacionalización en vez de métodos más convencionales como podría ser el de ajustar una serie de Fourier, es que los cambios de la temperatura del aire pueden aumentar o reducir el uso del agua, pero un evento de lluvia solamente puede reducirlo. Un modelo de serie de Fourier de la demanda histórica estacional no representa esos efectos adecuadamente ya que contiene los efectos de la lluvia y la temperatura.

Modelo de función de transferencia

La serie de memoria corta $W_s(t)$ que se obtiene de la ecuación (3.12) una vez removidas tendencia y estacionalidad, se puede observar en la figura 3.7 (b). La serie oscila alrededor de 0, presentando valores positivos cuando la temperatura es mayor que la normal y presenta valores negativos cuando

la temperatura desciende o cuando la lluvia se hace presente. Los efectos de la temperatura y la lluvia son dinámicos, ya que su presencia afecta la demanda en el día en el que ocurre y en varios días subsecuentes.

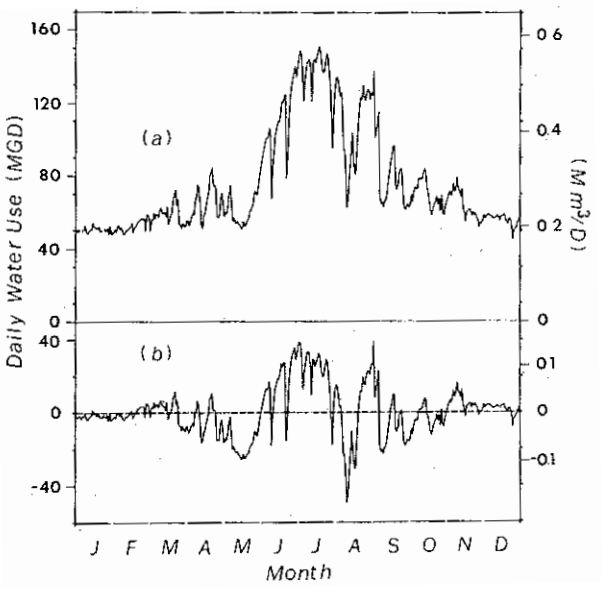


Figura 3.7: (a)Demanda diaria en Austin en 1980 (b) Serie de memoria corta producida por la remoción de tendencia y desestacionalización (Maidment et~al., 1985)

La ecuación del modelo de función de transferencia presentada en (3.5) requiere una adecuada identificación de los parámetros ($W_s(t)$, la temperatura $T_i(t)$ y la lluvia $R_i(t)$) y asume que éstos serán constantes a lo largo del tiempo e independientes del nivel de la serie sin tendencia ni estacionalidad.

La serie de demanda no cumple esos requisitos y sus parámetros son variables porque:

1. La respuesta de la demanda a las variables climáticas varía estacionalmente
2. La relación no es lineal porque la respuesta de la demanda por unidad de cambio en la temperatura o precipitación, depende de la magnitud de las variables climáticas
3. Existe un efecto de interacción entre la lluvia en días consecutivos y la demanda; la respuesta de la demanda no es la misma para el primer día de lluvia que para los días subsecuentes
4. Existe colinealidad entre lluvia y temperatura porque un evento de lluvia provoca una disminución de la temperatura

Estas dificultades obligan a modificar el modelo función de transferencia estándar

1. **Variación Temporal**: Con el fin de permitir la variación temporal de los parámetros se divide el año en las mismas estaciones que se detectaron para la función de calor (Diciembre-Febrero (invierno), Noviembre-Marzo (transición), y Abril-Octubre (verano)).

2. **No linealidad**: La serie se modela como una función lineal de variables independientes transformadas no linealmente.

3. **Interacción**: El efecto de interacción en la respuesta de la demanda a una lluvia en un día determinado, depende de la demanda estacional del día anterior $S_d(t-1)$. Si $S_d(t-1)$ es alta, la respuesta potencial es grande y viceversa. Se emplea $S_d(t-1)$ como una alternativa a la altura de precipitación conjuntamente con unas variables indicadoras cero-uno para distintas magnitudes de $R(t)$. Con esta modificación, la demanda nunca podrá ser menor a la demanda base y en cambio la demanda tiende al valor de base después de varios días continuos con precipitación.

4. **Correlación**: La dependencia de la temperatura del aire con la lluvia hace necesario el preblanqueo de la serie de temperatura previo a la identificación del modelo. Se construye un modelo de función de transferencia para la serie de temperatura diaria y se usa la lluvia como variable explicativa.

3.4.2. *Aportaciones del modelo de función de transferencia para la demanda*

El modelo de función de transferencia aplicado a la demanda, no aporta una nueva forma de descomponer una serie de demandas. Ya se ha explicado anteriormente en la sección 3.2 (página 65) un planteamiento de descomposición de la serie en componentes de memoria corta (autocorrelación y correlación con variables climática) y larga (tendencia y estacionalidad). En este caso se maneja otra terminología y se habla de dividir las series en demanda base y demanda estacional. Sin embargo la demanda base coincide con la definición de los componente de memoria larga y la demanda estacional coincide con el concepto de memoria corta.

La aportación viene por identificar que los modelos de función de transferencia son capaces de modelar la relación existente entre las variables

climáticas y las series de demandas, así como por la forma en que las variables climáticas (lluvia y temperatura) son integradas a la hora de construir un modelo de función transferencia de la demanda. Se solventa así una de las cuestiones mencionadas como líneas de investigación en la sección 3.3 (página 70) en lo que respecta a la lluvia y su efecto dinámico en la demanda de agua. Logran identificar cuantitativamente la disminución en la parte estacional de la demanda derivada de una lluvia por encima y por debajo de un umbral de precipitación.

Por la parte de la relación demanda-temperatura se aporta el concepto de función de calor que resuelve el problema de la no linealidad entre la demanda y la temperatura. La función de calor es una función por segmentos lineales que relaciona demanda y temperatura con distintas pendientes para cada uno de los segmentos. Los extremos de los segmentos (obtenidos arbitrariamente y por observación) representan los umbrales a partir desde los cuales la relación se modifica, ver gráfico (3.6). Es así que cada segmento acotado por sus extremos nos indica la relación que existe entre la temperatura y la demanda diaria. Se utiliza en el proceso de desestacionalización de la serie para obtener una serie de memoria corta $W_s(t)$ sin los efectos de la temperatura ni de la lluvia ya que los datos de días con lluvia no son tomados en cuenta a la hora de obtener la función de calor.

El concepto de dividir la serie temporal en periodos de verano e invierno y en valores mayores y menores que un umbral de precipitación determina la estructura que tendrán los distintos modelos de función de transferencia que se prueban. Es así que se tienen distintos coeficientes para cada parámetro y periodo, dependiendo de la estructura seleccionada. Los coeficientes de la ecuación (3.5) son los siguientes:

1. $W_s(t)$: Componente del nivel, único para cada modelo
2. $\omega_0^{(T_1)}$: Respuesta inmediata en la demanda por variación de la temperatura en el periodo 1
3. $\omega_0^{(T_2)}$: Respuesta inmediata en la demanda por variación de la temperatura en el periodo 2
4. $\delta_1^{(T_1)}$: Efecto de memoria en la demanda por variación de la temperatura en el periodo 1
5. $\delta_1^{(T_2)}$: Efecto de memoria en la demanda por variación de la temperatura en el periodo 2
6. $\omega_0^{(R_1)}$: Respuesta inmediata en la demanda por una lluvia menor que el umbral[5]

[5]En este caso R_1 corresponde a lluvias menores de 0.05 pulgadas/día y R_2 a lluvias mayores a ese valor

7. $\omega_0^{(R_2)}$: Respuesta inmediata en la demanda por una lluvia mayor que el umbral[5]
8. $\delta_1^{(R_1)}$: Efecto de memoria en la demanda por una lluvia menor que el umbral[5]
9. $\delta_1^{(R_2)}$: Efecto de memoria en la demanda por una lluvia mayor que el umbral [5]
10. $\omega_1^{(R_1)}$: Impacto adicional en la demanda por una lluvia menor que el umbral, registrado el día siguiente de la lluvia[5]
11. $\omega_1^{(R_2)}$: Impacto adicional en la demanda por una lluvia mayor que el umbral, registrado el día siguiente de la lluvia[5]

3.4.3. *Comentarios - Modelos función de transferencia de la demanda*

La metodología propuesta, integra tres conceptos principales

1. Un preblanqueado de la serie por medio de las ecuaciones de tendencia (ecuaciones 3.6-3.9)y de la función de calor (ecuaciones 3.10-3.13)
2. Modelos de función de transferencia para captar la relación dinámica existente entre la temperatura (segundo sumando en la ecuación 3.5) y lluvia (tercer sumando en la ecuación 3.5) con la demanda de agua.
3. Una parte autoregresiva que capta la dependencia de la serie de los registros de 2 días anteriores así como a la periodicidad semanal (cuarto sumando en la ecuación (3.5)).

Con esta estructura de escala diaria, el modelo resulta más potente que el modelo de transformaciones en cascada (Maidment and Parzen, 1984b) y logra captar el 97 % de la variabilidad de la serie de demandas de la ciudad de Austin, Texas. Se consigue que solamente un 3 % de la variabilidad de la serie quede sin ser explicada. Sin embargo, la serie temporal estudiada no parece tener variaciones bruscas de la demanda que pudieran aportar complejidad a la serie. Cabría también evaluar la viabilidad del modelo para estudiar series temporales en las cuales el concepto de demanda base ($\hat{W}_b$) es de difícil aplicación. En este modelo la demanda base se obtiene ajustando una función polinomial a la serie para los valores mínimos mensuales (en este caso enero) de cada año. El concepto de demanda base representa el consumo mínimo de agua que el conjunto de usuarios (domésticos, comerciales, industriales y de servicios público-urbanos) de una determinada ciudad requieren para cubrir sus necesidades mínimas independientemente de la climatología imperante, es decir la dotación mínima de una ciudad. Sin embargo, las series de demandas de muchas ciudades españolas y europeas se caracterizan por tener una disminución drástica en el consumo

durante el mes de agosto por motivo de vacaciones de la población. La demanda disminuye entonces a valores muy por debajo de la media esperada y lejos de la demanda global registrada a lo largo de los 11 meses restantes del año. Si se obtuviera ese valor de $\hat{W}_b$ ajustando una función polinomial a todos los meses de agosto de una hipotética serie de demandas con estas características, no sería representativa para el resto del año.

3.5. Evolución - Modelo función de transferencia demanda

En Maidment and Miaou (1986) se aplica la metodología de las funciones de transferencia de la demanda (Maidment et~al., 1985) para 9 ciudades de los estados de Pennsylvania, Texas y Florida en los Estados Unidos de Norteamérica y se comparan los resultados obtenidos. Cada estado aporta al estudio 3 ciudades con características de demanda y climatología similares entre sí, pero diferentes a las de los otros estados. El estudio no estuvo pensado como una evaluación del desempeño de este tipo de modelos en ciudades con distintas características de demanda, sino como una forma de comparar y entender la respuesta de la demanda a la presencia de lluvia así como a cambios en la temperatura. Sin embargo, de los coeficiente de determinación (R^2) obtenidos para las distintas ciudades podemos observar que este tipo de modelos obtiene mejores resultados en ciudades con determinadas características.

Ya se ha mencionado en repetidas ocasiones que la metodología descompone la serie en dos componentes: uso base que se supone insensible a factores climáticos y uso estacional, sensitivo a factores climáticos. En esta primera descomposición se observa que en las ciudades del estado de Pennsylvania, el componente estacional representa el 26 % de la demanda máxima diaria frente al 58 % que registran las ciudades del estado de Texas y el 47 % de las ciudades del estado de Florida. Observando estos datos podemos ver que el componente base (y por lo tanto insensible a factores climáticos) de las ciudades del estado de Pennsylvania es mucho mayor que el de las ciudades de los estados de Texas y Florida. Por decirlo de otra forma, la parte estocástica de la serie de demandas es mucho más pequeña en las ciudades correspondientes al estado de Pennsylvania.

En la parte correspondiente a la función de calor reportan que la demanda de agua aumenta cuando la temperatura del aire supera los 21^oC en todas las ciudades estudiadas, así mismo se observa un cambio en la tendencia que incrementa la demanda entre los 28^o y 32^oC. En las ciudades del estado de Pennsylvania la temperatura no supera los 32^oC por lo que no se registra este último cambio de tendencia.

La parte estacional (y por lo tanto estocástica) de la serie de demandas en ausencia de lluvia para el estado de Pennsylvania, aunque pequeña comparada con las de Texas y Florida, es más sensible a las variaciones de la temperatura del aire. Estas ciudades se muestran más sensibles que las de los otros dos estados, registrando que la demanda diaria varía un 11.2 % respecto

Estado	R^2
Pennsylvania	0.61
Texas	0.73
Florida	0.96

Cuadro 3.1: Coeficiente de determinación (Maidment and Miaou, 1986)

de la demanda potencial pico por cada grado centígrado de modificación de la temperatura del aire (Texas $4{,}7\,\%/^oC$, Florida $5{,}6\,\%/^oC$). No obstante que las temperaturas del verano son más templadas en Pennsylvania, la sensibilidad a cambios en la temperatura del aire es mayor.

Los coeficientes de la función de calor se obtienen con datos de demanda diarios en días sin lluvia y no representan los efectos dinámicos de la relación demanda-temperatura. Para esos fines, es decir para conocer los efectos dinámicos de la temperatura en la demanda a lo largo del tiempo, se obtienen los parámetros de la función de transferencia de la temperatura del aire. Siendo estos parámetros obtenidos con el conjunto completo de datos. Para la obtención de estos parámetros, la metodología propone dividir la serie en verano e invierno. Sin embargo esto no fue posible en las ciudades del estado de Pennsylvania ya que no se observó un comportamiento diferenciado para las dos épocas propuestas.

Algo similar ocurre en lo que se refiere a los parámetros de la función de transferencia para la lluvia. En dos ciudades del estado de Pennsylvania y Texas no fue posible distinguir el umbral de 13 $\frac{mm}{día}$ (0.05 $\frac{pulgadas}{día}$) y por lo tanto no se consiguió estimar los parámetros para precipitaciones mayores y menores que el umbral. Se observa también que para el estado de Pennsylvania, tanto el impacto inmediato de un evento de lluvia como el distribuido en los días posteriores es muy pequeño y representa una variación del 7 % del uso estacional (Florida 42 %, Texas 38 %).

Es bastante claro, independientemente de los valores de R^2 conseguidos por los modelos ajustados para cada estado (ver cuadro 3.1), que la metodología obtiene mejores resultados para ciudades donde la relación de los factores climáticos con la demanda de agua es fuerte. Esta relación podría estar estrechamente vinculada con el hecho de que Texas y Florida son estados que registran temperaturas altas en los meses de verano y la precipitación en esas zonas, por ser poco frecuentes, cuando se presentan afectan fuertemente el consumo conjunto de agua de la ciudad. En cambio, el estado de Pennsylvania presenta un régimen de precipitaciones (lluvia y nieve) muy húmedo a lo largo de todo el año y donde se registran temperaturas medias muy bajas en los meses de invierno así como temperaturas medias que no superan los 32^oC en verano. El modelo funciona mejor para ciudades

grandes en climas áridos con poco uso industrial intensivo de agua (Shaw and Maidment, 1987).

En mayor o menor medida, las características de ciudades del tipo de las de Pennsylvania, provocan que la demanda base sea mucho más importante respecto de la demanda total que la demanda estacional. El modelo de función de transferencia de la demanda basa su fortaleza en la capacidad de modelar y predecir esa componente estacional, por lo que su proporción con respecto a la demanda total afecta su desempeño. En esta metodología la demanda base es considerada como un preblanqueado de la serie original.

3.5.1. *Función de transferencia y análisis de intervención - Demanda*

Hipel et~al. (1975) sugirieron que el análisis de intervención podría ser apropiado para una gran variedad de aplicaciones en el área de los recursos hidráulicos y en la ingeniería ambiental. Propusieron un método de análisis y lo ejemplificaron con un caso práctico para el río Nilo, donde identificaron que la parte dinámica del modelo puede ser caracterizada por una respuesta del tipo escalón de la forma

$$\mathbf{Y}_t = \omega_0 S_t^{(T)}$$

con la cual representaron la variación del caudal medio del río derivado del inicio de operación de la presa de Aswan, obteniendo buenos resultados.

Si bien los modelos de función de transferencia aplicados a la modelación y predicción de la demanda de agua (Maidment et~al. (1985), Maidment and Miaou (1986)) fueron un paso importante en la comprensión de los efectos que las variables climáticas ejercen sobre la demanda, estos no habían sido utilizados para medir y predecir los efectos que las decisiones técnicas tomadas por los gestores de un sistema de agua urbana pueden tener. Estas decisiones pueden ser restricciones al servicio, un encarecimiento del precio del agua, cambios en los códigos técnicos, entre otras.

Shaw and Maidment (1987) presentaron un estudio desarrollado para la ciudad de Austin, Texas, en el periodo entre 1984 y 1985. Durante los cinco años previos, la ciudad presentó un rápido crecimiento de población y empleo, lo que generó que las infraestructuras de saneamiento de la ciudad estuvieran en riesgo de verse superadas por el volumen de agua que

tendrían que tratar. Los gestores del sistema de agua se vieron en la necesidad de legislar una ordenanza que obligaba al cumplimiento de un programa de conservación que limitaba por niveles (según el riesgo de superar los límites preestablecidos para cada nivel) las actividades de uso de agua de los ciudadanos cuando el conjunto de la demanda superara unos límites preestablecidos. Durante los veranos de 1984 y 1985 fue necesaria la aplicación del programa de conservación por parte de los ciudadanos. Durante este periodo los gestores del sistema utilizaron un modelo de predicción de la demanda llamado *WATFORE* (Water use Forecasting) que fue desarrollado por la Universidad de Austin, con el cual generaban predicciones de unos cuantos días basados en las condiciones de temperatura y lluvia presentes y futuras. La metodología del modelo de predicción de la demanda fue la presentada en Maidment et~al. (1985). Con el fin de evaluar la efectividad del programa de conservación aplicado en los años 1984 y 1985 se planteó un análisis que aportara datos que sirviesen para modificar y mejorar el programa para el verano siguiente. Sin embargo los autores del estudio aclaran que es muy complicado separar los efectos del programa de conservación de agua reflejado en las variaciones de la demanda de agua de los imputables a los efectos climáticos.

En este estudio, con el fin de medir el impacto del programa de conservación de agua en la ciudad de Austin utilizan el análisis de intervención. Se utiliza un modelo del tipo presentado en la ecuación (3.5) (Maidment et~al., 1985) e incorporan un componente de intervención del tipo

$$\sum_{j=1}^{k} \mathbf{Y}_{tj} = \omega_j S_t^{(T)}$$

donde k representa el número de intervenciones que afectan a la serie, en este caso el número de veces que se aplicó el programa de conservación en sus distintas niveles, $S = 1$ durante el periodo de restricción y 0 el resto del tiempo.

Con el fin de probar y medir el impacto de un ciclo de riego de jardines que permitiera regar cada 5 días en casas o negocios de forma distribuida evitando picos de demanda no deseados, se incluyó un término de intervención más complicado

$$\sum_{j=1}^{k} \mathbf{Y}_{tj} = (\omega_0 + \omega_1 B + \omega_2 B^2 + \omega_3 B^3 + \omega_4 B^4)(S(t))$$

donde $S = 1$ en los días cuando estaban autorizadas para regar las direcciones con terminación 0 y 1 para el año 1984, ó 9 y 0 para el año 1985, en cualquier otro caso $S = 0$.

Los valores de los parámetros obtenidos aportan una medida del cambio medio en la demanda diaria durante un periodo de restricción. El impacto de la intervención se define como la magnitud de la desviación con respecto a condiciones normales, considerando que las condiciones normales se refieren a la demanda en el periodo durante el cual el cumplimiento de las restricciones no está en efecto. Los parámetros estimados reportados en este estudio, aunque son internamente consistentes, son también en la mayoría de los casos no significativos estadísticamente si se comparan con sus errores estándar. En este caso de estudio obtuvieron información con la cual se propuso un nuevo ciclo de riego de jardines que eliminó los picos de demanda.

Por este motivo los autores comentan las necesidad de investigaciones adicionales sobre métodos cuantitativos para evaluar el impacto de los programas de restricción y conservación en la demanda diaria. Por otra parte agregan que se deben desarrollar métodos estadísticos más rigurosos para determinar la efectividad global de los programas de conservación así como investigar la forma de cuantificar la probable variación en el tiempo de los parámetros del modelo.

En Sastri and Valdes (1989) se presenta una caso de análisis de intervención para aplicaciones "*en línea*". Explican que los procesos subyacentes de las series de demanda diaria del tipo de las analizadas en Maidment and Miaou (1986), no son estacionarios respecto a la media ni a la varianza. Un rasgo común en series de demanda típicas para ciudades del sur de los Estados Unidos de Norteamérica, es el descenso transitorio del nivel de la demanda después de que se presente un evento de lluvia y adicionalmente, las series muestran variaciones influenciadas por cambios estacionales de la temperatura. Con respecto a éstos fenómenos transitorios los autores comentan que:

- Las fluctuaciones transitorias locales de una serie temporal de demandas, no son homogéneas entre las distintas observaciones, por lo que las correlaciones entre las observaciones son diferentes en diferentes momentos
- La interrelación de inhomogeneidad, y no estacionareidad entre las observaciones, significa que un modelo obtenido para un conjunto de datos, solo puede explicar el comportamiento local correspondiente a esas observaciones.

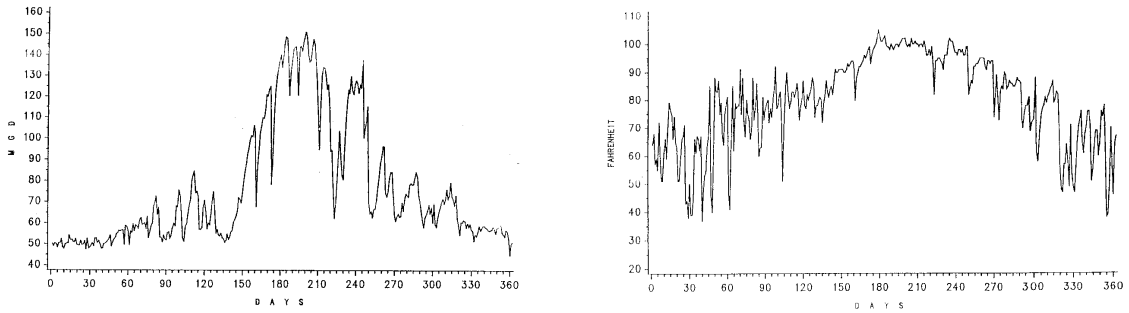
- Los métodos convencionales para remover la no estacionareidad, no son eficientes en estos casos ya que las tendencias locales son temporales.
- Los modelos que asumen homogeneidad y estacionareidad en la covarianza, no tienen una aplicación simple en modelación y predicción de series de demanda de agua que nunca alcanzan el equilibrio estadístico.

A continuación aclaran que, al ser muy frecuentes las predicciones de demanda a lo largo del año, y ya que seguramente imperarán condiciones climáticas diferentes en cada momento en que se realice la predicción, se deberá ajustar un modelo cada vez que se vaya a realizar una predicción para captar la dinámica más reciente del proceso de demanda de agua, así como proporcionar las predicciones climáticas necesarias. Por otra parte, al incurrir en una mala selección del modelo como resultado de los puntos antes mencionados, se obtienen residuos con coeficientes de autocorrelación que no se agotan hasta después de un número grande de retardos. Para solucionar este problema algunos autores (Maidment et~al., 1985) ajustan modelos a grupos parciales de datos, correspondientes a las estaciones del año. Sin embargo esta técnica implica complicar el análisis de los datos y crea varios submodelos con la penalización de tener que estimar parámetros adicionales. De esta forma se vulnera el concepto de parsimonia de los modelos de series temporales planteado por Box et~al. (1976), donde comentan que cualquier procedimiento de identificación de modelo debe tender a identificar modelos parsimoniosos que logren una eficiente estimación de parámetros y predicción de la serie temporal.

Las series de demanda que son esencialmente homogéneas, pero que se ven afectadas por media y varianza variables en momentos puntuales, pueden ser modeladas por medio de funciones determinísticas o pueden ser removidas fácilmente de la serie para transformarlas en estacionarias. La remoción de eventos aislados (lluvias, outliers) que provocan la no homogeneidad de las series es un tema más complicado y las técnicas convencionales como la diferenciación, cuando se aplican a series no homogéneas producen falsos valores de autocovarianzas que tardan varios retardos en agotarse.

La metodología de esta publicación se basa en que los eventos de lluvia, son la principal fuente de inhomegeneidad en las series de demanda de agua, e ignora la que la temperatura del aire pudiera aportar en verano para simplificar el modelo. Las lluvias son eventos puntuales que inician y terminan aleatoriamente, es decir que no ocurren a intervalos regulares a lo largo del tiempo (ver figura 3.9). Es por eso que cualquier evento de lluvia puede ser

representado como un pulso que provoca una caída transitoria del nivel de demanda de agua y puede ser modelado de una manera relativamente fácil con una función de transferencia. La segunda suposición de esta metodología es que las variaciones estacionales del proceso de demanda de agua (figura 3.8(a)) son guiadas por la temperatura máxima del aire (figura 3.8(b)). En los gráficos mencionados se observa que ambas series tienen valores picos máximos contemporáneos durante los meses de verano. Con este análisis los autores proponen para su metodología que la temperatura diaria del aire guía el proceso de demanda, mientras que la lluvia solamente la perturba transitoriamente.



(a) Demanda de Agua potable (Millones de galones por día)

(b) Temperatura máxima diaria (Grados Fahrenheit)

Figura 3.8: Series temporales de la ciudad de Austin, 1980. (Sastri and Valdes, 1989)

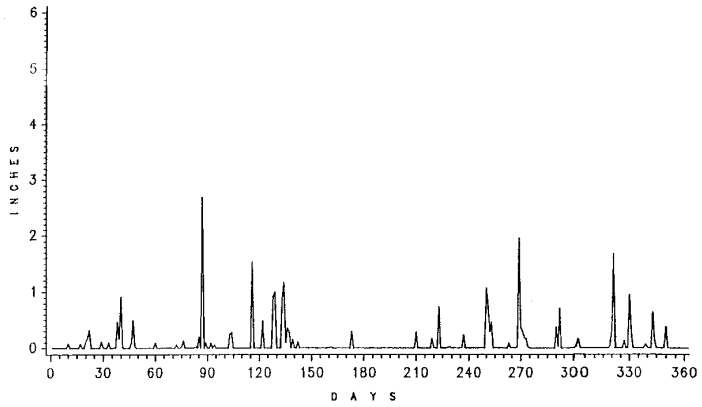


Figura 3.9: Serie temporal de eventos de lluvia diaria de la ciudad de Austin, Texas en el año 1980, (Sastri and Valdes, 1989)

Con el fin de representar el patrón anual de la serie de demandas de la ciudad de Austin, proponen la función sinusoidal sinc (ver figura 3.10):

$$g(t) = \frac{a\ sen\left[\frac{2\pi(t-c)}{b}\right]}{\frac{2\pi(t-c)}{b}} + d;\ t \neq c \tag{3.14}$$

$$g(t) = a + d;\ caso\ contrario \tag{3.15}$$

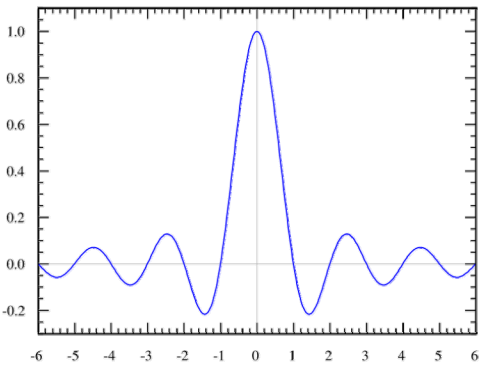


Figura 3.10: Función Sinc, (Sastri and Valdes, 1989)

donde la parte central del gráfico de la función corresponde a la demanda suavizada durante el verano sin existir ninguna intervención por lluvia; las dos colas de la función representan el nivel de demanda esperado durante finales del otoño e inicios de la primavera. Los parámetros de la función sinc significan:

- a=Amplitud
- b=Periodo (equivalente a la mitad del ciclo estacional)
- c=Desplazamiento horizontal del pico de la función
- d=Desplazamiento vertical del nivel base de la serie temporal

Los parámetros c y b tienen unidades de tiempo y pueden ser fácilmente preestablecidos para hacerlos coincidir con el día en que ocurre la demanda pico y el periodo estacional. Por lo tanto, solo se tienen que estimar los parámetros a y d. Esta forma de captar el comportamiento de la serie temporal es una alternativa más simple a la de ajustar una serie de Fourier, propuesta y utilizada en Maidment and Parzen (1984a), Maidment and Parzen (1984b) y Franklin and Maidment (1986).

Las intervenciones por lluvia las clasifican en tres grupos definidos por el concepto de IMS[6], que se explica como el lapso de tiempo que dura la memoria de la intervención en la demanda de agua, provocado por un evento de lluvia:

1. **Aislada:** Provocada por una lluvia aislada, de duración menor a un día, seguido de un periodo (IMS[6]) durante el cual el efecto transitorio en el nivel del agua decrece con el tiempo.
2. **Múltiple:** Provocado por uno o más eventos de lluvia que ocurren durante una periodo de tiempo de menor duración que el IMS[6] del primer evento. Eventos separadas por cuando menos un día sin lluvia, seco.
3. **Continua:** Igual que la Múltiple pero sin días sin lluvia, seco.

Los estudios empíricos reportados en este trabajo sugieren que los efectos de las intervenciones provocadas por la lluvia pueden ser adecuadamente explicados por un modelo lineal que es función del tiempo (t) en que ocurre la intervención, el número de días de lluvia ininterrumpida ($d(t)$), una media móvil de temperaturas máximas diarias ($T(t)$), y una media móvil de los valores más recientes de demanda de agua ($W(t)$). La ecuación del modelo para captar los efectos de la intervención por lluvia es:

$$I(t) = \alpha_0 I_{t-1} + \alpha_1\, x_1(t) + \alpha_2\, x_2(t)\, y_1(t) + \alpha_3\, x_2(t-1)\, y_2(t) + z(t) \tag{3.16}$$

donde

$$x_1(t) = \frac{1}{3}\sum_{i=1}^{3} T(t-i) \qquad x_2(t) = \frac{1}{3}\sum_{i=2}^{4} W(t-i)$$

$$y_1(t) = f[s(t)] \qquad y_2(t) = 0,\ si\ s(t) = d(t)$$

$$y_2(t) = 1,\ si\ s(t) \neq d(t) \qquad s(t) = d(t)\ si\ d(t) < 3$$

$$s(t) = 3,\ si\ d(t) > 3$$

Los parámetros de α_0, α_1, α_2 y α_3 en la ecuación (3.16), se asumen constantes por segmentos y el término de error $z(t)$ es un proceso ARIMA. El comportamiento transitorio de $I(t)$ es aproximado por una función ($f[s(t)]$) de

[6]Intervention Memory Span

agotamiento exponencial discreta en el tiempo, con un agotamiento entre unos cuantos días y dos semanas. Las variables $y_1(t)$, $y_2(t)$ son funciones determinísticas de $s(t)$. La función de $f[s(t)]$ debe ser estimada de los registros diarios de lluvia y de demanda de agua de la ciudad, y una vez que los parámetros de la función han sido determinados, sus valores se consideran constantes. En este caso, se observa que la mayoría de los eventos tienen ($d(t)$) igual o menor que 3 días. El comportamiento transitorio de las intervenciones, puede ser representado por un modelo simple para $y_1(t)$ de solo dos parámetros, del tipo:

$$f[s(t)] = \beta^{s(t)-\gamma} \tag{3.17}$$

Donde β y γ son característicos de la intervención producida por las lluvias en la ciudad.

La metodología propuesta asume que la serie del proceso de demanda se compone de la siguiente forma:

$$W(t) = g(t) + I(t) + N(t) \tag{3.18}$$

donde $N(t)$ es un proceso ARIMA cuya varianza puede ser dependiente del tiempo y $g(t)$, e $I(t)$ como fueron definidas en las ecuaciones 3.15, 3.14 y 3.16. De la estructura de la ecuación (3.18) se observa que los efectos de las intervenciones por lluvia ($I(t)$)se *"sobre-imponen"* al nivel medio del proceso de demanda ($g(t)$). El proceso de cálculo de los parámetros necesarios se realiza de forma iterativa (de ahí que esta metodología se haya definido anteriormente con capacidad de ser utilizado *"en línea"*):

1. *Paso 1*
   Se estiman los parámetros a y d, habiéndose definido antes los valores de b y c para ajustar $g(t)$ (mínimos cuadrados) a la secuencia de observaciones de $W(t)$. La función ajustada se denomina $\hat{g}(t)$ y se establece $\hat{W} = \hat{g}(t),\ t = 1, 2, \ldots, n$, donde n = número total de observaciones.

2. *Paso 2*
   Se calcula $\hat{I}(t) = W(t) - \hat{W}(t),\ t = 1, 2, \ldots, n$ y se estiman los parámetros de la ecuación 3.16, incluyendo los parámetros del modelo ARMA $z(t)$.

3. *Paso 3*
   Se calcula una secuencia mejorada de $[\hat{g}]$, haciendo que $\hat{R} = W(t) - \hat{I}(t),\ t = 1, 2, \ldots, n$ y reajustando $g(t)$ a la secuencia $[\hat{R}(t)]$. La nueva secuencia $[g(t)]$ es una versión mejorada de $[\hat{g}(t)]$ del paso 1. Se repiten los

pasos 1 y 2 hasta que se satisfagan las condiciones de convergencia. En la última iteración se obtienen los estimadores finales $[\hat{g}^*(t)]$ y $[\hat{I}^*(t)]$.

4. *Paso 4*
   Se estiman los parámetros del modelo ARMA para los residuos que resulten de $N(t) = W(t) - \hat{g}^*(t) - \hat{I}^*(t),\ t = 1, 2, \ldots, n$.

Los autores comentan que la ecuación (3.18) no incorpora la influencia directa de las variaciones del día a día de la temperatura del aire en los niveles de demanda diaria. La función sinc, $g(t)$, solo representa un promedio estacional de esa influencia en la demanda de agua. Es por este motivo que $N(t)$ resulta ser dependiente del tiempo. No obstante, una función polinomial simple, puede ser propuesta como función normalizante para describir ese comportamiento variable en el tiempo. La versión final del modelo sería:

$$W(t) = a_0 T(t) + g(t) + I(t) + h(t)\frac{\theta(B)}{\phi(B)}\epsilon(t) \tag{3.19}$$

donde

- a_0 es una constante por segmentos
- $T(t)$ es la temperatura máxima para el día t
- $g(t)$ como ya fue definida en las ecuaciones (3.14) y (3.15)
- $I(t)$ como ya fue definida en la ecuación (3.16)
- $\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \theta_3 B^3$
- $\phi = 1 - \phi_1 B$
- $h(t) = \left[a_1 t + \frac{a_2}{t}\right]^{-1}$ todos los parámetros se consideran constantes por segmentos
- $\epsilon(t)$ es un proceso Gaussiano, ruido blanco de media cero y varianza constante σ^2

Finalmente, β y γ pueden considerarse como parámetros de control que varían según el tipo de lluvias que ocurran y determinarán la contribución que $x_2(t)$ tendrá en $I(t)$.

3.5.2. *Modelos de demanda - Efectos climáticos no lineales*

Es el enfoque tradicional de la modelación de demandas de agua potable, utilizar modelos de series temporales, como los presentados en Maidment et~al. (1985). Considera que la demanda total puede ser dividida en demanda base, que es insensible a factores climáticos y caracterizada como la demanda media en meses invernales, y demanda estacional, que es sensible a factores climáticos, siendo caracterizada por la diferencia entre la demanda total y la demanda base. En este enfoque de modelos de función de transferencia, los efectos de la lluvia sobre la demanda estacional, se entienden como dinámicos y dependientes del estado antecedente. En cuanto al efecto de la temperatura en la demanda estacional, se identifica un punto de quiebre de la temperatura, denominado umbral de temperatura, por debajo del cual la demanda es independiente de la temperatura. La función de calor se utiliza para identificar este umbral por medio de un modelo lineal por segmentos que relaciona las variaciones del uso estacional de la demanda con la temperatura.

Es esperable que la demanda base presente una tendencia debido a factores poblacionales (aumento de la población) o socioeconómicos. La demanda base tradicionalmente se obtiene haciendo una regresión de la demanda frente a los factores antes mencionados, o ajustando un polinomio a los valores mínimos de cada año. Miaou (1990) presenta un trabajo, en el cual se destaca que se pueden dar casos en los que la insensibilidad de la demanda base a factores climáticos podría no cumplirse y propone un esquema más general, el cual permite que tanto la demanda base como la estacional puedan ser dependientes de estos factores. La formulación es como sigue:

$$B_m = \beta_0 + \beta_1 H_\tau(T_m) + \beta_2 G\gamma(R_m) + \sum_i \beta_{i+2} X_{i,m} + \nu_m \tag{3.20}$$

donde $H_\tau()$ y $G_\gamma()$ son las funciones de temperatura y de lluvia efectiva, definidas como:

$$H_\tau(T_m) = T_m - \tau \qquad T_m \geq \tau$$

$$H_\tau(T_m) = 0 \qquad caso \quad contrario$$

$$G\gamma(R_m) = R_m \qquad R_m \leq \gamma$$

$$G\gamma(R_m) = \gamma \qquad caso \quad contrario$$

donde τ y γ son la temperatura de referencia o umbral y lluvia de referencia o umbral respectivamente. τ representa un umbral de temperatura por debajo del cual la demanda es independiente de la temperatura, y γ es un umbral de cantidad de lluvia donde la lluvia en exceso no contribuye más a reducir la demanda de agua. Es evidente, que la identificación de esos umbrales por cualquier técnica estadística, es un punto importante para obtener mejores resultados a la hora de realizar las predicciones.

Histéresis en la demanda

Los efectos que el fenómeno de histéresis puede tener en la modelación de la demanda de agua potable fueron abordados en Miaou (1990). Al modelar series de demanda de agua potable, es bastante frecuente incorporar a las ecuaciones, parámetros obtenidos de series de temperatura, que en muchas ocasiones resultan en una mejora en la relación observado-modelado. Sin embargo, no en todos los casos la temperatura nos aportará mejoras en la modelación. Es importante destacar que para valores similares de temperatura, se pueden registrar distintos valores de demanda potencial, lo cual quiere decir que se presenta el fenómeno de histéresis provocada por marcadas variaciones estacionales de la demanda.

En Miaou (1990) se analiza una serie de demandas de agua potable de la ciudad de San Diego, California. El gráfico (3.11) presenta la demanda mensual (en galones per cápita por día) frente a la temperatura media máxima del mes correspondiente.

El autor comenta que la histéresis mostrada en el gráfico es el resultado de varias posibilidades, por ejemplo, el calendario de actividades de las instituciones de la ciudad, variaciones de la radiación solar caracterizada por la cantidad de horas de sol a lo largo de un día, la persistencia del riego de jardines a lo largo de todo el año. Bajo la misma temperatura los efectos de la histéresis provocan diferentes niveles de demanda estacional, y posiblemente, diferentes intensidades de respuesta durante la primavera y el otoño. En general, la utilización de modelos lineales convencionales para series de demanda de agua con histéresis evidente, como los mostrados en el gráfico (3.11), producirán series de residuos altamente correlacionados. Adoptando los postulados sugeridos en el estudio de demanda diaria de Maidment and Miaou (1986), una serie de demandas se puede descomponer en demanda base y demanda estacional, y a su vez esta última consiste de: (1) demanda potencial estacional (presentado en página 73) que es dependiente de la temperatura en ausencia de lluvia, (2) de una reducción de la demanda provocada por la ocurrencia de lluvia, (3) un componente aleatorio de me-

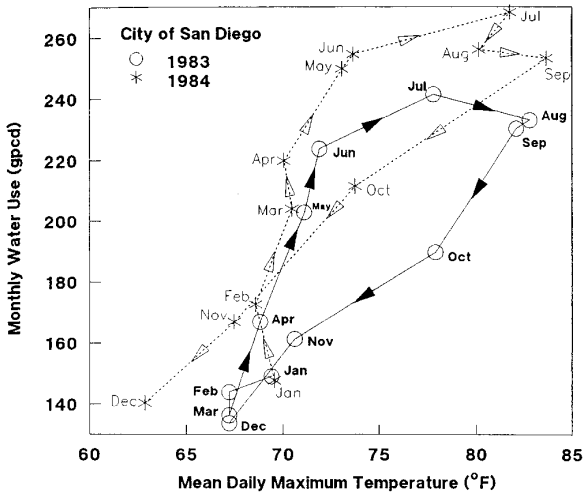


Figura 3.11: Histéresis en la relación temperatura-demanda mensual, en San Diego, California, (Miaou, 1990)

dia cero. La histéresis tiene efecto en la demanda potencial estacional y en el caso de no ser significativa, la demanda potencial W_p puede ser modelada como una función lineal de la función de calor (como fue explicado en página 76):

$$W_m = \beta_0 + \beta_1 H_\tau(T_m)$$

En cambio, cuando los efectos de la histéresis se estiman significativos y se cuenta con un conjunto de datos relativamente grande (más de tres años), las utilización de algún modelo que utilice series de Fourier producirá mejores resultados.

3.5.3. *Predicción de la demanda a escala horaria*

El objetivo del control en línea de un sistema de distribución, es preparar y ejecutar un plan para operar el sistema (Shvartser et~al., 1993). Típicamente los planes de operación se preparan con 24 horas de antelación. Las principales razones para que esto sea así, son primero, porque las demandas reproducen un marcado ciclo diario, segundo porque las tarifas eléctricas dependen de la hora del día y finalmente porque, en ciertas horas del día, algunos estados del sistema (niveles de tanques) pueden ser establecidos con antelación con mucha precisión. Una condición necesaria para prepa-

rar un plan de operación es la predicción de las demandas para el periodo de planificación.

Shvartser et~al. (1993) presentan un modelo para predecir la demanda horaria basado en el análisis de series temporales y el reconocimiento de patrones desarrollado para un sistema de distribución de agua en Israel. Explican que la variación de la demanda a lo largo del día depende de muchos parámetros –temperatura, humedad, tiempo desde la última lluvia, día de la semana, etc—. Sin embargo, son detectables algunos patrones marcadamente similares a lo largo del día, reproduciendo una "firma" característica de cada conjunto de usuarios típicos. En todos los casos se puede detectar un patrón similar: baja demanda por la noche, creciente por la mañana, alta demanda con variabilidad a lo largo del día y finalmente disminución al final del mismo. Generalizando, caracterizan el patrón de la demanda diaria en segmento *creciente*, segmento *oscilante* y segmento de *descenso*. La metodología presentada, se basa en asumir este patrón general, para después identificar los puntos de transición entre segmentos y construir modelos de series temporales para cada segmento. Es decir, que ajustan modelos Box-Jenkins de bajo orden a cada estado del sistema, para después modelar la transición entre un estado al siguiente como un proceso de Markov. La metodología asume condiciones meteorológicas estables. Estos modelos dieron buenos resultados en su aplicación, sin embargo requieren un control constante de la demanda.

Homwongs et~al. (1994) presentaron un modelo que denominaron Adaptive Forecasting System (AFS), para obtener predicciones horarias de la demanda en la ciudad de Arlington, Virginia. La metodología se basa principalmente en el algoritmo de suavizado exponencial de Winters (Winters, 1960), obteniendo resultados aceptables a pesar de no modelar las variabilidades climáticas. El algoritmo Winters es de utilidad en situaciones de actualización recursiva de los datos, así como en la extracción de factores estacionales dependientes del tiempo. Para construir el algoritmo toman en cuenta dos periodos de estacionalidad, el que se produce de un día a otro (24 horas de antelación) y el que se produce los fines de semana, que se repite cada 7 días (168 horas de antelación), o lo que es lo mismo, un algoritmo opera para los días de lunes a viernes y otros para los fines de semana. La estructura completa del algoritmo incluye un complemento de mínimos cuadrados recursivos (RLS) para evitar problemas de inicialización, una detección de *outliers* para reducir la sensibilidad de la metodología de Winters a valores erróneos y finalmente un filtro autoregresivo para minimizar las autocorrelaciones de los residuos, ver gráfico (3.12).

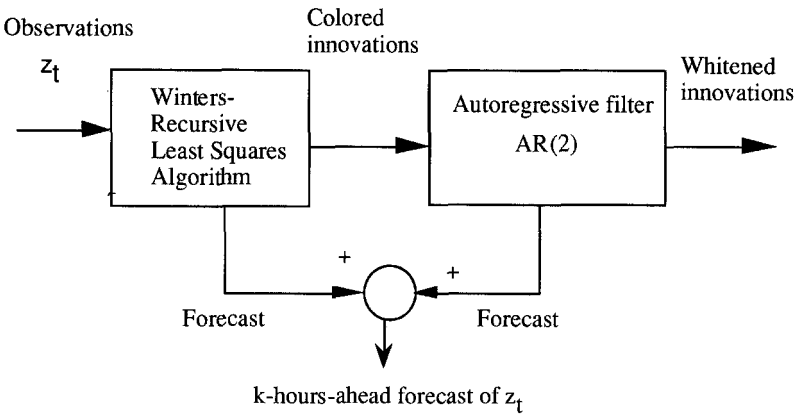


Figura 3.12: Diagrama del sistema de predicción adaptativo (AFS), (Homwongs et~al., 1994)

3.6. Modelación y predicción - Estudios más recientes

La gran mayoría de las publicaciones recientes sobre modelación y predicción de la demanda, subyacen en los avances de la década de 1980 y algunos son meros análisis exploratorios de las series de demanda, su relación con la lluvia, temperatura y alguna otra variable climática, pero particularizados para ciudades o regiones muy concretas con mayor o menor éxito.

Ya en épocas más recientes se han abierto nuevas líneas de investigación que se enfocan a probar el desempeño de las técnicas de inteligencia artificial, entiéndase redes neuronales artificiales y sistemas expertos en la predicción a distintas escalas temporales de series de demandas de agua potable. Al mismo tiempo realizan comparaciones del desempeño de estas técnicas con los modelos estadísticos clásicos. A continuación, se resumen las dos líneas de investigación antes mencionadas.

3.6.1. *Aplicaciones - modelos de series temporales clásicos*

Con el fin de tener un entendimiento más preciso del patrón que compone la serie histórica de demandas diarias de la ciudad de Nueva York, así como su relación con las variables climáticas, Protopapas et~al. (2000) hicieron un estudio del cual obtuvieron como resultado que durante los meses de invierno, la temperatura afecta levemente el comportamiento de la demanda. En cambio, en los meses de verano las precipitaciones causan una disminución de la demanda diaria, especialmente para lluvias tormentosas. Por medio de métodos gráficos, encontraron también que existe un umbral de temperatura ($78°F$ - $26°C$), a partir del cual se presenta una correlación positiva con la demanda. Es decir, que una vez superada esa temperatura, la demanda aumenta linealmente con la temperatura.

Zhou et~al. (2000) se basaron en la metodología de Maidment et~al. (1985) para modelar y predecir la demanda diaria de agua de la ciudad de Melbourne, Australia. Construyeron varios modelos de series temporales en los cual la demanda se considera como la suma de demanda base y demanda estacional (incluyendo componentes para la estacionalidad, para el clima y para la persistencia). Utilizaron también series de temperatura y de precipitación con el fin de mejorar el desempeño del modelo. El mejor de los modelos calibrados, tuvo un $R^2 = 89{,}6\,\%$ y un error estandar absoluto de $\pm 8\,\%$ en fase de validacion. En Zhou et~al. (2002) se repite la metodología ante-

rior pero limitandose al distrito de Chelsea (predominantemente residencial, 350,000 habitantes), en Melbourne, de donde dispusieron de 6 años de datos diarios y horarios. Para mejorar el desempeño de sus modelos introducen a los modelos la ETP como variable climática y utilizan el concepto de API (Antecedent Precipitation Index, mm). Al contar con gran cantidad de datos, obtuvieron un abanico de patrones de demandas semanales, diarios y horarios. Con estos patrones esperados, conjuntamente con el tipo y el día de la semana, hicieron desagregaciones para construir distintos modelos y predecir posteriormente. Las predicciones diarias que realizaron, lograron explicar el 83 $\%$ de la varianza, con un error estandar de 57 $l/p/d$ (el estudio se realizó con dotaciones, no con demandas). En cuanto a la demanda horaria, el mejor de los modelos construidos logra explicar el 66 $\%$ de la varianza, con un error estandar de 162 $l/p/d$.

Aly and Wanakule (2004) proponen la utilización de los algoritmos de alisado exponencial, para predecir la demanda de agua media mensual y diaria de ciudades de los alrededores de Tampa, Florida. Aplican el algoritmo de Winters y utilizan 6 años de datos (1991-1996) para calibrar los coeficientes del modelo reservando dos años para comprobar su capacidad predictiva. Los resultados obtenidos fueron, que en la predicción diaria el modelo explica el 70 % de la variabilidad de la demanda diaria. El 30 % de varianza no explicada lo asumen como aleatoria y aclaran que solamente podrá ser predicha desde un enfoque estocástico. Claramente, un componente aleatorio está presente en los registros de demanda diaria, por lo que no se deberían esperar predicciones muy precisas de un modelo determinístico como este. Sin embargo, ese componente aleatorio es eliminado cuando los datos diarios se agrupan en valores medios mensuales, de esta forma las predicciones esperadas serán mejores.

Gato et~al. (2007b) presentan un modelo en el cual hacen una propuesta para obtener los umbrales de lluvia y temperatura de referencia por debajo de los cuales la demanda es independiente de las variables climáticas y hacen también una comparación entre la utilización de las series de Fourier (Maidment and Parzen (1984a,b); Zhou et~al. (2000)) y la función de calor (Maidment et~al., 1985) para modelar la parte de la demanda correspondiente al uso estacional. Para la identificación del umbral de temperatura, aplicaron a la serie de demandas una transformación del tipo función recíproca, para después ajustar una función polinómica de la temperatura máxima diaria (ver figura 3.13). Posteriormente calcularon la derivada de la función para obtener su mínimo, en este caso ese valor fue de $15{,}27^oC$, siendo este el umbral de temperatura. Por encima de este umbral, la demanda aumenta, pero por debajo del mismo la demanda presenta un comportamiento independiente de la temperatura. Se aplica el mismo método para

obtener el umbral de lluvia, ajustar una función polinómica a la demanda y se calcula la derivada de la función para obtener en este caso su máximo, obteniendo un valor de 4.82 mm como umbral (ver figura 3.14).

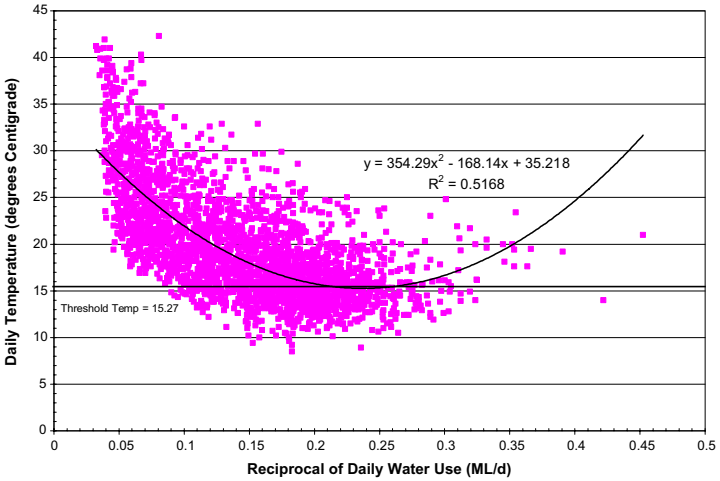


Figura 3.13: Obtención del umbral de temperatura, (Gato et~al., 2007b)

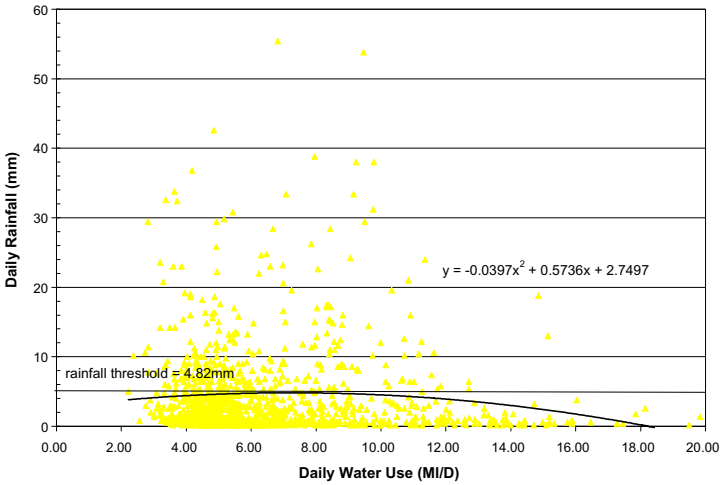


Figura 3.14: Obtención del umbral de lluvia, (Gato et~al., 2007b)

Posteriormente, y utilizando la metodología de series temporales de transformaciones en cascada (Maidment and Parzen, 1984a), construyen varios modelos de entre los cuales el que mejor desempeño ofreció, obtuvo valores de $R^2 = 0{,}81$ en fases de calibración y validación. Este modelo toma en consideración en sus ecuaciones, un indicador para los días de lunes a viernes y los de fin de semana ya que existe un muy evidente ciclo semanal con valores máximos en los fines de semana (ver figura 3.15).

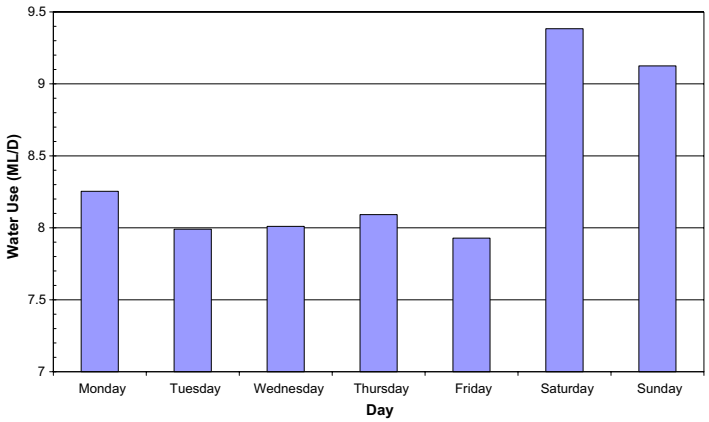


Figura 3.15: Ciclo semanal, (Gato et~al., 2007b)

En (Gato et~al., 2007a, Julio) se presenta un caso de estudio que utiliza la metodología anterior para East Doncaster, ciudad de Melbourne, Australia. El mejor de los modelos estudiados obtuvo un coeficiente de determinación R^2 de 86 % tanto en fase de calibración como en validación.

Alvisi et~al. (2007) propusieron un planteamiento más simplificado para un experimento a pequeña escala en Castelfranco Emilia, provincia de Modena, Italia (23,000 habitantes). Consistió en utilizar los datos obtenidos de un sistema de monitoreo constante de caudales para construir un modelo de predicción de los caudales a corto plazo, a escala diaria en un primer paso y horaria posteriormente. El modelo, que denominaron Modelo de predicción de demanda de agua basado en patrones, consiste de dos módulos, el módulo de demanda diaria (DM) y el módulo de predicción de la demanda horaria (HM). El módulo de demanda diaria, consiste de un componente periódico estacional, $\bar{Q}_m^{d,s}$, un corrector que representa el componente periódico semanal, $\bar{\Delta}_{i,j}^{d,w}$ y finalmente, un componente de desviación para captar la persistencia de la demanda $\bar{\delta}_m^d$, por lo que la ecuación es la siguiente:

$$Q_m^{d,for} = \bar{Q}_m^{d,s} + \bar{\Delta}_{i,j}^{d,w} + \delta_m^d \tag{3.21}$$

El componente periódico estacional, $\bar{Q}_m^{d,s}$, lo modelan por medio de series de Fourier:

$$\bar{Q}_m^{d,s} = a_0 + \sum_{k=1}^{K} \left[a_k \cos\left(\frac{2\pi mk}{365}\right) + b_k \sin\left(\frac{2\pi mk}{365}\right) \right] \tag{3.22}$$

$$m = 1, 2, \ldots, 365$$

donde a_0 es el valor medio del ciclo estacional, a_k y b_k son los coeficientes de Fourier y K es el número de armónicos considerados.

El factor de corrección semanal $\bar{\Delta}_{i,j}^{d,w}$ se define como:

$$\bar{\Delta}_{i,j}^{d,w} = \bar{Q}_{i,j}^{d} - \bar{Q}_{j}^{w} \tag{3.23}$$

donde $\bar{Q}_{i,j}^{d}$ es el valor medio de la demanda observada el día i de la semana $(i = 1, \ldots, 7)$ de lunes a domingo, y j es la temporada $(j = 1, \ldots, 4)$ para invierno, primavera, verano y otoño (ver figura 3.16). $\bar{Q}_{j}^{w}$ es el valor medio de la demanda diaria en la temporada j. Por último δ_{m}^{d}, representa las desviaciones entre la demanda media diaria Q_{m}^{d} y el valor medio estimado en base de los componente periódicos $\bar{Q}_{m}^{d,s}$ y $\bar{\Delta}_{i,j}^{d,w}$, es decir:

$$\delta_{m}^{d,obs} = Q_{m}^{d,obs} - \left(\bar{Q}_{m}^{d,s} + \bar{\Delta}_{i,j}^{d,w}\right) \tag{3.24}$$

El proceso δ_{m}^{d}, es modelado usando un proceso autoregresivo AR(1) (Box et~al., 1976), por lo que para fines de predicción, la ecuación es:

$$\delta_{m}^{d} = \Phi_1 \cdot \delta_{m-1}^{d} \tag{3.25}$$

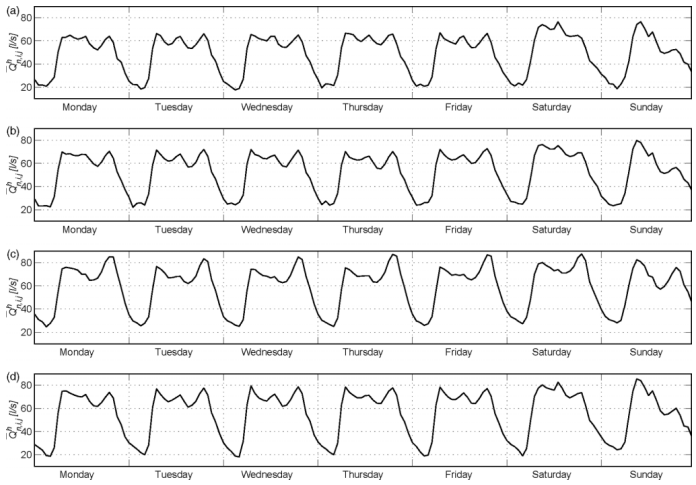


Figura 3.16: Patrones de demanda diarios para los siete días de la semana:(a)en invierno, (b)en primavera, (c)en verano, (d)en otoño. (Alvisi et~al., 2007)

El módulo horario (HM) se basa en el módulo diario (DM) (ver figura 3.17) para hacer sus predicciones. La demanda $Q_{t+k}^{h,for}$ predicha en la $t$ésima hora para k horas de anticipación se obtiene con:

$$Q_{t+k}^{h,for} = Q_{m}^{d,for} + \bar{\Delta}_{n,i,j}^{h} + \epsilon_{t+k} \tag{3.26}$$

donde: $Q_m^{d,for}$ es la demanda media diaria predicha con el módulo diario (DM), hasta la fecha $m-1$.
$\bar{\Delta}_{n,i,j}^h$ es la desviación horaria respecto al patrón diario

$$\bar{\Delta}_{n,i,j}^h = \bar{Q}_{n,i,j}^h - \bar{Q}_{i,j}^d \tag{3.27}$$

$\bar{Q}_{n,i,j}^h$ representa el valor de demanda horaria observado en la hora n ($n = 1, \ldots, 24$, hora del día) del día i en la temporada j.
$\bar{Q}_{i,j}^d$ es el valor de demanda diario observado en el día i en la temporada j
ϵ_{t+k} es el componente de persistencia horaria que se modela usando una regresión de los errores ϵ_{t+k-1} y ϵ_{t+k-24}

$$\epsilon_{t+k} = \Psi_1 \epsilon_{t+k-1} + \Psi_2 4 \epsilon_{t+k-24} \tag{3.28}$$

Los coeficientes son calibrados en base a los errores observados ϵ_t^{obs}, siendo:

$$\epsilon_t^{obs} = Q_t^{h,obs} - (Q_m^{d,obs} + \bar{\Delta}_{n,i,j}^h) \tag{3.29}$$

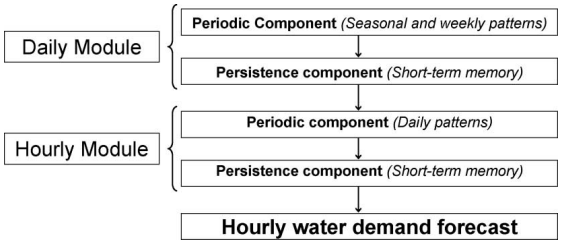


Figura 3.17: Estructura de los dos módulos que conforman el modelo de predicción (Alvisi et~al., 2007)

El desempeño del modelo fue evaluado en términos de varianza explicada (*EV*), error cuadrático medio (*RMSE*) y el error absoluto medio (MAE %). En el cuadro (3.2) se presentan los resultados que obtuvieron para los módulos diario y horario, diferenciando cuando se utiliza el componente de persistencia y sin él, en fase de calibración y validación. De este cuadro podemos concluir que la utilización del componente de persistencia resulta en una mejora en las predicciones en términos de *EV* y *RMSE* en fase de calibración y validación, tanto en el módulo diario como en el horario. Por otra parte se observa también que el desempeño del modelo presenta peores resultados si comparamos los valores de *EV* y *RMSE* obtenidos en fase de validación y

calibración. Hay un evidente deterioro de los resultados en fase de validación.

	Calibración		Validación	
	Diario	Horario	Diario	Horario
EV componente periódico	0.69	0.94	0.36	0.89
EV comp. periódico + persistencia	0.84	0.97	0.76	0.97
RMSE componente periódico	2.83	4.63	13.52	14.64
RMSE comp. periódico + persistencia	2.01	3.17	4.61	4.42

Cuadro 3.2: Varianza explicada y error cuadrático medio de la predicción diaria y horaria (Alvisi et~al., 2007)

Si comparamos el desempeño del modelo por sus predicciones diarias (*EV*=0.76 en el mejor de los resultados) con otros modelos, estas son peores que las obtenidas con técnicas desarrolladas en las décadas de 1980 (Maidment et~al. (1985) *EV*=0.97, Maidment and Miaou (1986) *EV*=0.96, Sastri and Valdes (1989) *EV*=0.982)[6]. Cabe decir sin embargo que este modelo tiene una estructura más simple y un número bajo de parámetros a calibrar desde los datos.

Este tipo de modelos, con una metodología *transparente*, presenta muy buenos resultados en las predicciones horarias, equiparables a las que se obtienen con la utilización de métodos del tipo de caja negra, como son las redes neuronales (Ghiassi et~al. (2008) *EV*=0.97). Es importante destacar la uniformidad de las predicciones, ya que el desempeño del modelo se mantiene estable sin importar la hora del día en la que se realice la predicción. Es adecuado comentar que la metodología ha sido utilizada para realizar predicciones en una zona muy reducida y en consecuencia con caudales muy bajos. En la fase de validación del cuadro presentado, el caudal medio fue de entre $54\frac{l}{s}$ y $68\frac{l}{s}$. Sería importante probar su desempeño en una ciudad con caudales de demanda más importantes y con usuarios de distintas características que aportarían una variabilidad adicional a la demanda.

[6]Todos estos modelos integran variables climáticas para realizar sus predicciones

3.7. Las redes neuronales (ANN)

En los últimos años las redes neuronales han sido usadas cada vez más para predecir el comportamiento de sistemas complejos, tales como fenómenos naturales y físicos (una revisión del estado del arte de la predicción con *ANN* se puede encontrar en Zhang et~al. (1998)). El aumento de su popularidad, se debe en parte a la percepción de que las *ANN* son capaces de sortear muchas de las dificultades que complican la implementación de métodos estadísticos tradicionales. Esto se debe principalmente a que las *ANN* no están sujetas a las reglas restrictivas que gobiernan los modelos estadísticos. El énfasis de la modelación con *ANN* se pone en métodos en los que el resultado y la precisión son el principal objetivo, mientras que los estadísticos apuntan hacia métodos universales óptimos estadísticamente (Maier and Dandy, 2000).

Como ya se ha mencionado en la sección 2.5, las *ANN* empezaron a ser el objeto de investigaciones serias en los años posteriores a 1940, no se popularizaron hasta la década de 1980 y con las mejoras en la arquitectura computacional y de velocidad de cómputo de los años 1990, se logró un progreso significativo en la utilización de las *ANN*.

En la siguiente sección presentaremos diversos casos de aplicación de las *ANN* en los recursos hídricos y en la modelación de la demanda urbana de agua. También analizaremos un grupo de trabajos en los cuales las *ANN* se presentan comparadas con las técnicas clásicas de series temporales y algún caso en el que han sido utilizadas conjuntamente.

3.7.1. *Las ANN en los recursos hídricos*

La técnica de las *ANN* ha ido evolucionando a lo largo de los más de 60 años que han transcurrido desde que se establecieron sus primeras ideas fundamentales y desde entonces, con algunas etapas en las que pareció que la línea de investigación se vería truncada, ha conseguido ser más eficiente —nuevas reglas de aprendizaje, algoritmos, funciones de activación— a la hora de resolver problemas diversos. Al mismo tiempo se fue desarrollando una terminología propia de las *ANN*, es decir su propia jerga. Es así que por dar algunos ejemplos, los residuales de la terminología estadística se denominan errores en la terminología de las *ANN*, las variables independientes son los inputs, las predicciones son los outputs, los parámetros estimados son los pesos, etc.

No es de extrañar que en el campo de los recursos hídricos (en Lingireddy and Brion (2005) se puede encontrar una colección de publicaciones en esta área) y de la hidráulica en general, las *ANN* hayan encontrado un amplio abanico de fenómenos y problemas factibles de ser abordados ya sea como aporte de una solución explicativa de los fenómenos o simplemente como una alternativa susceptible de comparación. Los fenómenos tratados en el ámbito de las ciencias relacionadas con el agua, tienen el común denominador de que responden a leyes físicas o matemáticas que se han ido comprendiendo cada vez más conforme la ciencia nos aporta el conocimiento necesario. Es así que las metodologías de uso consolidado se valen del entendimiento que se tiene de esas leyes para modelar fenómenos —en muchas ocasiones por medio de sistemas de ecuaciones diferenciales en derivadas parciales— integrando tantas variables que afecten el fenómeno como sea posible, asumiendo la no linealidad espacial y temporal que puedan contener. Por ello se pueden encontrar fenómenos que pueden ser modelados con unas cuantas variables y otros en los que la cantidad de variables llega a ser muy grande y con un importante aporte de incertidumbre acumulada integrada al modelo por cada una de ellas.

Si algún estigma o lastre han tenido que soportar las *ANN*, es el escepticismo que generan, ya que su modo de solución es en cierta forma abstracto y se les ha colocado el calificativo de *cajas negras*. Este planteamiento tiene su origen en que las *ANN* no requieren conocer ni contener en su estructura de solución a las formulaciones o leyes que rigen el fenómeno a modelar. En cambio basan su fortaleza en la habilidad para asociar las variables de entrada con la o las salidas identificando las reglas que gobiernan sus relaciones. A su favor se debe decir que es precisamente esa innecesariedad de conocer el proceso lo que las convierte en una herramienta potente y las hace capaces de generalizar una gran variedad de fenómenos. Sin embargo la experiencia ha mostrado que los mejores resultados se obtiene cuando el modelador tiene un buen entendimiento físico del fenómeno a modelar. A pesar de los inconvenientes –ciertos o no– que la metodología de las *ANN* pueden tener, con el paso del tiempo se han ido granjeando un espacio cada vez más importante en las diversas áreas de las ciencias exactas.

Durante la década de 1990 empieza a extenderse su utilización y se dan los primeros pasos para convertirla en una metodología consolidada en el campo de la ingeniería civil. A finales de esta década la *ASCE*[7] organiza un comité de trabajo con enfoque hacia la hidrología y publican dos artículos con los resultados encontrados. La primera de ellas (ASCE, 2000a), presenta las que consideraron pautas básicas para el uso de las *ANN*. Por otra parte también destacan sus fortalezas y limitaciones:

[7] American Society of Civil Engineering

1. Son capaces de reconocer la relación entre entradas y salidas sin contener explícitamente el modelo físico que las relaciona
2. Funcionan bien aún cuando el conjunto de datos utilizados para el entrenamiento contenga ruido y errores de medición
3. Son capaces de adaptarse a soluciones a lo largo del tiempo para compensar circunstancias cambiantes
4. Poseen unas características de procesado de la información y una vez entrenadas son fáciles de usar

Se echa en falta en esa publicación algún comentario acerca de la limitación que presentan las *ANN* cuando una red entrenada con un determinado conjunto de datos –con su rango de mínimos y máximos– es requerida para modelar o predecir un fenómeno que contiene valores fuera del rango del conjunto de datos de entrenamiento. Es decir, es complejo lograr que una *ANN* aprenda y generalice. Sería deseable que una ANN sometida a esta situación entregara valores de interpolaciones suaves para el espacio no entrenado.

La segunda publicación de este par de trabajos (ASCE, 2000b) tiene un enfoque netamente hidrológico y dedica su texto a realizar comparaciones entre distintas metodologías. Maier and Dandy (2000) realizaron una recopilación de 43 artículos en los cuales las *ANN* han sido utilizadas con éxito para modelar y predecir procesos de variables de recursos hídricos. La revisión contempla las publicaciones entre los años 1992 y 1998, las variables modeladas corresponden a procesos muy diversos: concentración de cianobacterias y algas en lagos, caudales en ríos, precipitación, coeficientes de escorrentia en cuencas, salinidad, pH, niveles de agua en embalses y ríos. Las escalas temporales varían desde la minutal hasta la anual y se utilizan tanto datos reales como sintéticos. De esta revisión se puede observar que la cantidad de publicaciones científicas relacionadas con las *ANN* y los recursos hídricos ha ido creciendo desde 1992 cuando se reportan 2 publicaciones hasta las 17 y 10 de los años 1997 y 1998 respectivamente.

3.7.2. *ANN vs. Series Temporales*

Las *ANN* y las series temporales se suelen presentar como técnicas enfrentadas. Es muy frecuente encontrar publicaciones que evalúan sus desempeños realizando comparaciones con los resultados de unas y otras. Limitándonos a la modelación y predicción de demanda de agua podemos citar

a Jain and Ormsbee (2002), Bougadis et~al. (2005), Adamowski (2008), entre otros. En una época en la que las *ANN* eran todavía técnicas relativamente nuevas, Tang et~al. (1991) publican un estudio en el cual realizan una comparación del desempeño de las *ANN* contra modelos del tipo Box-Jenkins. Estudian tres series temporales, una de ellas es la utilizada por Box and Jenkins (1970) para desarrollar su metodología (pasajeros de avión entre 1949 y 1960), otra serie pertenece a la venta de coches en Estados Unidos entre 1966 y 1982, la última pertenece a la venta de coches de fabricación extranjera a los Estados Unidos entre 1966 y 1982. Las series fueron seleccionadas buscando que tuvieran patrones de comportamiento marcadamente distintos para probar y comparar el desempeño de ambas técnicas. Realizaron varias pruebas hasta encontrar una estructura de *ANN* óptima para cada una de las series temporales. Finalmente realizaron comparaciones en términos de la suma del error cuadrático y concluyeron que las *ANN* superaban a las técnicas clásicas tanto en la predicción a corto como a largo plazo.

En cambio, Chatfield (1993) en una primera revisión ya corroboraba empíricamente que las predicciones con *ANN* no eran necesariamente mejores que otras alternativas. Unos años más tarde Faraway and Chatfield (1998) utilizan la serie de Box and Jenkins (1970) para realizar un estudio comparativo entre modelos del tipo Box-Jenkins, Holt-Winters y *ANN*. En su trabajo los autores comentan que contrario a las afirmaciones de grandes éxitos que se venían reportando, su evidencia empírica de predicciones con *ANN* indica un grado de éxito variable y advierten de lo arriesgado que puede resultar –a pesar de que la modelación con *ANN* es en esencia no paramétrico– que procesos completos sean automatizados en un ordenador y que sea utilizado por personas con escaso conocimiento ya sea en *ANN* o en predicción. Con su estudio demuestran que un buen modelo de *ANN* para series de datos temporales debe ser seleccionado combinando habilidades tradicionales de modelación con conocimientos de análisis de series temporales y de los problemas particulares. Los resultados de las predicciones que obtuvieron con las mejores arquitecturas de *ANN* ajustadas para esta serie (en términos de suma de los errores cuadráticos a 1 y 12 pasos, AIC, BIC) no resultan ser mejores que el conocido modelo Box-Jenkins ajustado utilizando un número menor de parámetros. Finalmente, aclaran que es complicado hacer una evaluación del desempeño de esta nueva técnica ya que la vasta mayoría de los trabajos que se habían presentado pueden ser criticados desde un punto de vista estadístico (el artículo de Tang et~al. (1991) es uno de ellos). Los estadísticos tienden a percibir las *ANN* como una técnica menos *ilustrativa* ya que su estructura de *caja negra* es complicada de entender e interpretar, además de que no es posible obtener descriptores del término de error y no hay una forma directa de calcular los intervalos de predicción.

Es bastante obvia la confrontación que existe entre las técnicas clásicas de series temporales –de las cuales los estadísticos de carrera son expertos– y las técnicas de inteligencia artificial como son las *ANN* del domino de los *coneccionistas*[8]. Pero por otra parte es adecuado decir que más que técnicas enfrentadas, se debería ver en ellas a técnicas complementarias. Los modelos de series temporales suelen ser muy potentes modelando series con comportamientos lineales (una vez realizadas las transformaciones que fuesen necesarias) y en cambio, las *ANN* tienen una estructura adecuada para modelar fenómenos no lineales, por lo que resulta una mala elección de recursos utilizar una *ANN* para modelar series con evidentes patrones lineales. De la misma forma es poco lógico modelar fenómenos no lineales con modelos lineales en esencia y esperar buenos resultados.

En esta línea Zhang (2003) publica un artículo en el cual propone una combinación de técnicas con el fin de conseguir mejores predicciones. Esta idea la justifica con los siguientes tres puntos

1. En la práctica es complicado determinar si una serie es generada desde un proceso lineal o no lineal
2. Las series temporales de procesos reales raramente son procesos lineales o no lineales puros.
3. No existe un método de predicción válido para todas las situaciones

Con esta idea, el autor propone la utilización conjunta de modelos del tipo ARIMA y *ANN*, es decir un modelo híbrido. Considera que una serie temporal está compuesta de componentes lineales y no lineales de modo que

$$y_t = L_t + N_t$$

donde L_t representa el componente lineal y N_t el no lineal. El componente lineal es un modelo ARIMA, de esta forma los residuos del modelo lineal contendrán solamente relaciones no lineales

$$e_t = y_t - \hat{L}_t$$

y a su vez estos residuos son modelados usando *ANN*s

$$e_t = f(e_{t-1}, e_{t-2}, ..., e_{t-n}) + \epsilon_t$$

[8]Coneccionismo: Una rama de la ciencia cognitiva que sostiene que los procesos mentales humanos (como el aprendizaje) pueden ser explicados mediante modelación computacional de redes neuronales pensadas para simular las acciones de las neuronas interconectadas en el cerebro. Fuente Merriam-Webster Online Dictionary. 2009

f es una función no lineal determinada por la red neuronal y ϵ_t es el componente de error. Por ello las predicciones se combinan de tal forma que

$$\hat{y}_t = \hat{L}_t + \hat{N}_t$$

Con este planteamiento se explotan las habilidades tanto de los modelos ARIMA como las de las *ANN*s. La metodología propuesta es puesta a prueba con la predicción de tres series de ámbitos marcadamente diferentes y logran mejorar los resultados en las predicciones obtenidas por los modelos ARIMA y de *ANN* utilizadas sin combinarse entre sí en términos de error cuadrático medio.

En esta misma línea de los modelos híbridos del tipo ARIMA-*ANN*, se presentan los trabajos de Rojas et~al. (2008) y de Valenzuela et~al. (2008), aunque estos van más allá y proponen también una metodología para automatizar la determinación de los órdenes de los componentes $p, d,$ y q de los modelos ARIMA, así como de los valores de sus parámetros. En definitiva se trata de un sistema experto que contiene un total de 43 reglas utilizadas para la selección de los modelos ARIMA. Estas reglas se encargan de seleccionar, por medio de un algoritmo genético que optimiza las soluciones, el modelo ARIMA óptimo que cumpla con las condiciones de estacionareidad de la metodología.

3.7.3. *Las ANN en la predicción de demanda*

Las *ANN*s han encontrado también un campo de aplicación en la predicción de la demanda de agua potable. Sin embargo las publicaciones científicas son más bien escasas en comparación con las que se pueden encontrar modelando fenómenos hidrológicos.

Griñó~C. (1991) presentó un trabajo un tanto adelantado para su época, de hecho rompe el orden cronológico de este capítulo. El trabajo tenía como objetivo predecir la demanda diaria con un día de antelación en un sector de la ciudad de Barcelona, España, utilizando la metodología de redes neuronales. Probó varias arquitecturas de redes y métodos de aprendizaje, así como la inclusión de series de intervención (indicadores para pascuas, verano, festivos y fin de semana largo). Los resultados que obtuvo fueron de 4.55 % de error para una red de arquitectura 15-20-1 sin series de intervención (ver figura 3.18), y 4.12 % de error para una red de arquitectura 19-35-1 con series de intervención binarias (4 series binarias (0,1)). La inclusión de las series de intervención eliminaron gran cantidad de errores que aportaban valores pico.

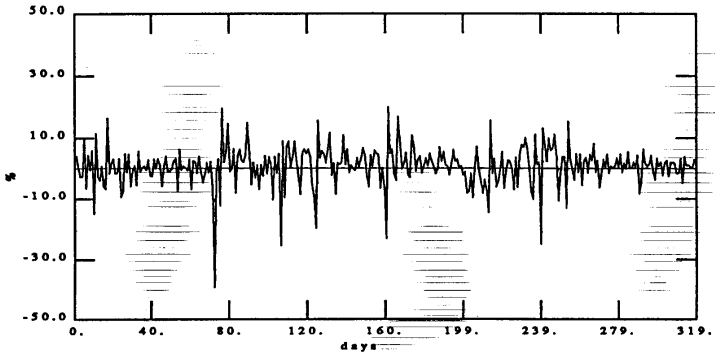


Figura 3.18: Residuos de red con arquitectura 15-20-1 sin series de intervención (Griñó~C., 1991)

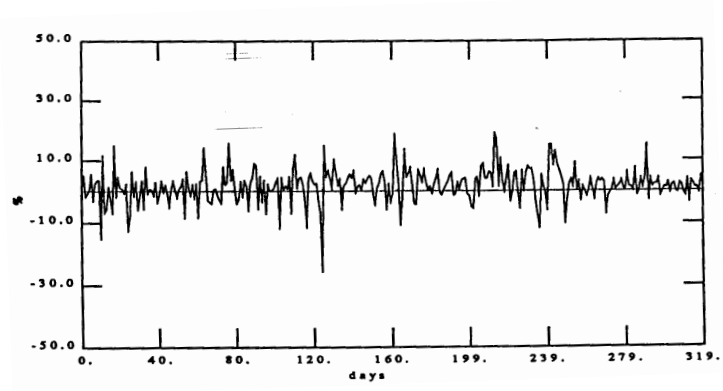


Figura 3.19: Residuos de red con arquitectura 19-35-1 con series de intervención (Griñó~C., 1991)

Jain and Ormsbee (2002) presentaron un trabajo en el cual realizan predicciones a escala diaria para la ciudad de Lexington, Kentucky, que registra una demanda mínima de 102,205 $\frac{m^3}{dia}$ entre los años 1982 y 1992. Construyeron dos modelos de regresión con desagregación semanal con estructura similar entre ellas, dos modelos de series temporales, uno de ellos del tipo AR y el otro es un modelo de regresión con agrupamiento semanal (media móvil) de las variables de temperatura y demanda antecedente (MAR). De entre los que clasificaron como modelos de inteligencia artificial (AI), construyeron dos modelos del tipo de sistema experto (RESM1 y RESM2) y una *ANN* simple con estructura S231NN. Finalmente un híbrido de sistema experto mejorado con una *ANN* (CREN). Los resultados que obtuvieron se presentan en el cuadro (3.3). Se observa en este cuadro que los modelos del tipo AI superan el desempeño de los de técnicas de series temporales y de regresión. Sin embargo es correcto comentar que las estructuras utilizada en las técnicas *clásicas* son de las más simples y no son representativas del mejor de los desempeños posibles para ellas.

Modelo	AARE
Regresión-desagregación 1	12.68-8.33
Regresión-desagregación 2	12.19-8.41
AR	6.25-5.15
MAR	6.11-5.08
S231NN AI	5.91-4.95
RESM1 AI	6.15-4.82
RESM2 AI	6.16-4.87
CREN AI	5.98-4.75

Cuadro 3.3: Desempeño de los modelos construidos para predecir la demanda diaria en Lexington, Ky. (Jain and Ormsbee, 2002)

Joo et~al. (2002) emplearon la técnica de *ANN* para predecir la demanda diaria de la ciudad de Seoul, Corea. Utilizaron como variables explicativas del fenómeno, la temperatura máxima diaria, un indicador de días festivos, la demanda previa y fueron descartadas la velocidad del viento y la humedad por presentar valores bajos de correlación con la demanda diaria. Los valores de la serie de demandas y de las variables explicativas fueron normalizados para limitarlos a valores entre 0 y 1. Construyeron una red de tres neuronas con función de activación logística (ANN331). El mejor de los resultados en fase de calibración fue de R^2=0.87 para el periodo de primavera y un valor global de $R^2 = 0{,}78$ para todo el año. Obtuvieron mejores resultados para la fase de validación con predicción a 1 mes, alcanzando un MAE(%)=1.55 en el mes de octubre.

Bougadis et~al. (2005) construyeron una variedad de modelos de regresión lineal múltiple, modelos ARIMA con diversas estructuras, así como varias *ANN*, para predecir la demanda pico semanal. Los modelos se diferenciaban entre sí por la forma en que integraron en sus ecuaciones las series temporales de demanda pico semanal, temperatura y lluvia acumulada. El estudio fue realizado para la ciudad de Ottawa, Canadá, que en 1992 registraba una demanda diaria de 109,300 m^3. Después de realizar un análisis exploratorio de los datos, construyeron 3 modelos de regresión y regresión múltiple, 7 modelos ARIMA y 3 *ANN*.

Model	AARE	Max ARE	R^2	
			Calibración	Validación
Regresión	18.01	40.78	0.5669	0.4543
ARIMA	13.79	39.09	0.3205	0.2995
ANN	12.19	33.28	0.7241	0.6986

Cuadro 3.4: Desempeño de los modelos construidos para predecir la demanda pico semanal en Ottawa, Canadá. (Bougadis et~al., 2005)

Del cuadro 3.4, se observa que los modelos *ANN* presentan mejores resultados a la hora de predecir la demanda pico semanal. Sin embargo, ninguna de las técnicas consigue valores elevados de R^2, siendo las redes neuronales las que mejores resultados presentan. El proceso que se busca modelar y predecir es la demanda pico semanal y utilizan como variable exógena la temperatura. Es de esperar que la demanda pico semanal no presente una gran variabilidad, como lo presentan la demanda horaria o la diaria, por lo que resulta un tanto extraño que los modelos no hayan conseguido mejores valores.

Zhang et~al. (2006) aplicaron la metodología de las *ANN* para predecir la demanda diaria de la ciudad de Lousville, Kentucky con dos días de antelación. Para ello construyeron *ANN*s para invierno[9] y verano diferenciadas entre ellas por la cantidad de variables utilizadas. La de invierno, con una menor cantidad de variables de entrada a la red, utiliza solamente los valores de demanda de hasta dos periodos antecedentes. Para el verano incluyeron variables climáticas, temperatura media máxima del día en curso, predicciones de la lluvia para los dos días siguientes, lluvia para el día en curso, predicciones de lluvia para los dos días siguientes y finalmente la demanda del día en curso. La arquitectura de red utilizada fue una red con 10 variables de entrada, 13 nodos ocultos y 2 salidas (10,13,2). Con estas *ANN*

[9]No se presenta la arquitectura de red utilizada

consiguieron hacer predicciones a dos días con una precisión del 97.21 % para la temporada de invierno y de 95.89 % para la temporada de verano.

3.8. Conclusiones - Estado del arte

A lo largo de este capítulo hemos presentado los resultados de la revisión del estado del arte de la modelación y predicción de la demanda de agua urbana. Es claro que esta evolución ha ido de la mano de los avances del área de la modelación estadística o bien de la econometría (Winters, 1960; Brown, 1963; Box and Jenkins, 1970). Estas nuevas técnicas fueron poco a poco llevadas al campo de la ingeniería y/ó de la gestión de recursos hídricos (Salas~LaCruz and Yevjevich, 1972; Oh and Yamauchi, 1974; Box and Tiao, 1975; Hipel et~al., 1975; Domokos et~al., 1976; Anderson et~al., 1980) por su adaptabilidad para reproducir el proceso que ocurre en la demanda de agua urbana. En un principio fueron modelos de regresión simples e incluso modelos multivariados a escala mensual que en algunos casos incorporaban variables econométricas como la elasticidad del precio del agua, el incremento de la población, las rentas *percápita*, etc. Sin embargo el grupo de trabajos presentados por David Maidment y sus colaboradores (Maidment and Parzen, 1984a,b; Maidment et~al., 1985; Maidment and Miaou, 1986; Franklin and Maidment, 1986; Shaw and Maidment, 1987) dieron un nuevo enfoque a la forma en que hasta esos años se abordaba la modelación de la demanda de agua urbana y marcaron una pauta a seguir para muchas publicaciones posteriores que han planteado enfoques similares, o que lo han utilizado como referencia y punto de partida para presentar evoluciones (Miaou, 1990; Protopapas et~al., 2000; Zhou et~al., 2000, 2002; Gato et~al., 2007b,a). Es importante destacar que la metodología con base en Maidment(1984a; 1984b; 1985; 1986; 1986; 1987) propone la descomposición de la demanda en demanda base y estacional, donde la primera representa los valores mínimos a lo largo de cada año. Esta hipótesis supone que esos valores se registrarán durante los meses de invierno cuando la demanda será mínima[10]. Este planteamiento –pensado para ciudades del tipo del sureste de los Estados Unidos, Texas– se puede explicar si consideramos que la vivienda típica de esa región es en su mayoría unifamiliar con zonas ajardinadas en la parte delantera y trasera de la finca, que requiere una menor frecuencia de riego en la época invernal, por lo que existe una disminución de la demanda global por el conjunto de las viviendas reduciendo su consumo por una menor evapotranspiración.

[10]En la pág.82 se definió como:

El concepto de demanda base representa el consumo mínimo de agua que el conjunto de usuarios (domésticos, comerciales, industriales y de servicios público-urbanos) de una determinada ciudad requieren para cubrir sus necesidades mínimas independientemente de la climatología imperante, es decir la dotación mínima de una ciudad.

Esta suposición no puede ser extrapolada a la mayoría de las ciudades españolas donde las viviendas se agrupan en conjuntos de edificios compartiendo áreas comunes, que en la mayoría de las ocasiones no cuenta con jardines dentro de sus instalaciones, ergo, no se presenta en las épocas invernales una disminución tan importante como ocurre en las ciudades del área de Texas. En cambio, si que se presenta una disminución que resulta ser muy importante durante los meses de agosto –a pesar de ser uno de los meses más calurosos del año donde cabría esperar demandas máximas– que es generalmente utilizado por los habitantes de las zonas urbanas como Valencia para vacacionar y en consecuencia desplazarse a regiones netamente turísticas, por lo que ocurre también una disminución de las actividades productivas de la ciudad y finalmente una disminución de la demanda de agua. Esta disminución, (ver gráficos 7.26 y 7.27) rompe el vínculo de correlación que se esperaría con respecto a la temperatura. Por esta particularidad de las serie de demandas de agua en ciudades españolas, tampoco puede ser modelada extrapolando la hipótesis que plantean Sastri and Valdes (1989)[11]. Un fenómeno similar se presenta durante las festividades de semana santa, de hecho los valores históricos mínimos de demanda se han registrado durante estos periodos. Los valores mínimos observados en la serie de Valencia no son característicos de los registros de demanda diaria a lo largo del resto del año, es decir, no representan las dotaciones mínimas de la ciudad, en cambio son el resultado de un conjunto de viviendas registrando demanda nula, contraviniendo la hipótesis presentada en Maidment(1984a; 1984b; 1985; 1986; 1986; 1987).

Una de las fortalezas de la metodología que no podemos obviar, es la información que conseguimos extraer de la serie temporal en cuestión. Al descomponer la serie en una secuencia sucesiva de transformaciones (en cascada según Maidment and Parzen (1984a)) obtenemos la varianza explicada como un porcentaje del total de la serie. Es así que por medio de diferencias simples podemos conocer la aportación en términos de varianza explicada de la tendencia, estacionalidad, autocorrelación y correlación climática de la serie temporal estudiada.

Se han encontrado enfoques con un origen diferente en los trabajos de (Sastri and Valdes, 1989; Shvartser et~al., 1993; Homwongs et~al., 1994; Alvisi et~al., 2007) aunque siempre basados esencialmente en modelos estadísticos. Por la parte de las técnicas de inteligencia artificial, concretamente las *ANN* –no sin opiniones divergentes por parte de especialistas de la estadística Chatfield (1993); Faraway and Chatfield (1998),(Chatfield, 2001, pag. 66,166-

[11]Sastri and Valdes (1989) suponen que el proceso de demanda de agua es guiado por la temperatura media del aire y solamente es perturbado transitoriamente por la ocurrencia de lluvia

168)– es indudable que representan una alternativa en la modelación y predicción de demanda de agua urbana. Los resultados obtenidos en Griñó~C. (1991), Jain and Ormsbee (2002), Bougadis et~al. (2005), Adamowski (2008) y Zhang et~al. (2006) son aceptables como para hacer la afirmación anterior. No podemos obviar que la metodología de las *ANN* no intenta modelar el componente error, ni hace ninguna suposición de distribución Gaussiana de los errores y no existe dentro de la metodología una forma directa y de uso extensivo para calcular los intervalos de confianza de las predicciones[12]. Otro aspecto a comentar sobre la metodología de las *ANN* es que el número de parámetros es mucho más grande que el de los modelos tradicionales de series temporales. Para una *ANN* de una capa, el número de parámetros viene dado por $p = (k + 2)H + 1$ donde k es número de variables de entrada y H es número de neuronas ocultas (Chatfield, 2001, pag.69). Por lo que como ejemplo, el trabajo de Griñó~C. (1991) de 19 variables de entrada a la red y 35 neuronas en la capa oculta requiere 736 parámetros para entregar sus predicciones, por lo que el principio de modelos parsimoniosos no se cumple con las *ANN*.

Llegado el momento de seleccionar un determinado modelo de predicción de demanda de agua urbana, debemos tener claro lo siguiente: los objetivos de la predicción, el tipo de datos con que se cuenta, la escala temporal con la que se trabajará, el horizonte de predicción y el nivel de agrupamiento de los datos. Si lo que se busca es un modelo que nos aporte predicciones muy acertadas, tal vez el modelador elegirá un modelo de *ANN*. En cambio si lo que desea es conocer el proceso generador de los datos que se está modelando o prediciendo, es decir el proceso subyacente en la serie temporal y desde luego si se desea conocer la incertidumbre asociada con las predicciones, entonces los modelos basados en las técnicas de series temporales muy probablemente serán los elegidas. Pensando en la implementación de un modelo de predicción de la demanda de agua urbana en tiempo real, tanto los modelos de *ANN* como los de series temporales requerirán que sus parámetros sean recalculados si se presenta un cambio evidente del patrón de demandas.

Es adecuado poner en relieve la cantidad y diversidad de modelos de demanda que se han propuesto hasta la fecha, diferenciados entre ellos por el número y el tipo de variables de entrada que utilizan, por el modelo matemático en el que se basan para describir el comportamiento de la demanda, y por las herramientas de cálculo que utilizan para estimar los parámetros del modelo así como para obtener las modelaciones y/ó previsiones de la demanda de agua urbana. Cada uno de los modelos tiene el inconveniente

[12]El trabajo de Cutore et~al. (2008) presenta resultados que animan al optimismo en lo que respecta a estos últimos puntos

de no ser exportables a otras partes del sistema de agua potable ni tampoco a algún otro sistema de agua potable de otra ciudad, por lo que deberán ser estimados y ajustados a las condiciones particulares. Una vez concluida esta revisión del estado del arte podemos definir algunas lineas de investigación que pueden ser abordadas para su estudio, en el marco de las ciudades españolas y del sur de Europa en general:

1. Modelar los efectos del calendario de festividades y vacaciones en la variabilidad de la demanda y en los registro mínimos.
2. Efectos de los valores atípicos motivados por festividades en la estimación de los parámetros de los modelos.
3. Efectos de los valores atípicos motivados por festividades en la predicción de la demanda urbana.
4. Caracterización de las festividades en base a su impacto y al tipo de día de su ocurrencia.

Una vez concluido esta revisión del estado del arte que ha abarcado la modelación y predicción de la demanda de las últimas décadas, podemos concluir que los modelos desarrollados hasta la fecha con este fin, no son aplicables para ciudades en entornos urbanos densamente poblados como son la mayoría de las principales ciudades de España y muchas de las europeas mediterráneas. Estas ciudades presentan particularidades y un comportamiento marcadamente diferenciado de las ciudades del sur, suroeste de los Estados Unidos de Norteamérica o de Australia, donde han desarrollado los modelos mencionados. Estos modelos contemplan conceptos como demanda base y suponen comportamientos de la demanda diaria que variando lentamente a lo largo del año guiados casi siempre por el ciclo sinusoidal de la temperatura del aire.

En cambio la demanda de las ciudades españolas en entornos densamente poblados, presenta un comportamiento con una variabilidad mucho mayor, con una relación temperatura del aire-demanda no lineal y variable a lo largo del tiempo y del rango de temperaturas registradas (ver sección 7.3.3). Por lo tanto la temperatura no es un predictor eficiente o requiere su incorporación mediante modelos más complejos. Adicionalmente la demanda, presenta súbitas reducciones del consumo provocadas por el calendario de festividades de las ciudades (propio de cada ciudad y región), generando días que registran consumos anómalos, que rompen el patrón regular semanal y que aportan los registros mínimos a lo largo de las series anuales. Por otra parte, la demanda también es afectada en su tendencia anual en el

mes de agosto durante el cual, gran parte de las empresas e industrias (salvo las ligadas al turismo y ocio) suspenden sus actividades y gran parte de la población utiliza este periodo para vacacionar fuera de la ciudades provocando un descenso de la demanda en la época más calurosa del año.

Los modelos que hasta este punto hemos analizado no contemplan ninguna de las peculiaridades mencionadas anteriormente ya que no son propias del proceso de demanda de las ciudades para las que fueron pensados, ya sea por su forma de urbanizar o por sus usos y costumbres. La suspensión de actividades a lo largo de todo un mes y la consecuente disminución de la demanda y en gran medida también la suspensión de actividades por festividades, no son propias de ese tipo de ciudades.

Surge pues, la necesidad de idear una metodología que nos lleve a identificación de un modelo que sea capaz de modelar el proceso generador de la demanda en ciudades densamente pobladas y reproducirlo en predicciones eficientes.

Parte IV

Metodología

Capítulo 4

Antecedentes

The most reliable way to forecast the future
is to try to understand the present.
John Naisbitt
1929 -
Escritor y conferenciante Estadounidense

En el capítulo 3 se ha analizado en detalle los distintos enfoques con lo que la modelación y predicción de la demanda ha sido abordada. De él se evidenció que ninguna de las metodologías analizadas puede ser aplicada directamente a ciudades españolas y europeas mediterráneas en entornos densamente poblados ya que sus planteamientos fueron pensados para otras regiones. En esta sección tenemos como objetivo, la identificación de una metodología que nos lleve a obtener un modelo que se ajuste y sea capaz de reproducir las peculiaridades que ocurren en este tipo de entornos. El modelo deberá ser robusto y estadísticamente riguroso, que logre captar la tendencia, la estacionalidad del proceso, las periodicidades que contiene, así como una descripción del componente error. Además deberá incorporar implícitamente todos le eventos que definiremos como componentes de variabilidad sistemática irregular, siendo estos identificados en la fase de estimación del modelo y caracterizados para su uso en la fase de validación.

La metodología propuesta se valerá de la combinación de 3 técnicas estadísticas:

- Los modelos del tipo Box-Jenkins o modelos ARIMA (presentados en la sección 2.2) para captar la variabilidad sistemática regular de la serie.
- La identificación de valores atípicos (sección 2.4). Para caracterizar los componentes de variabilidad sistemática irregular.

- Los modelos de regresión dinámica en la modalidad de análisis de intervención (sección 2.3.2) para incorporar rigurosamente los componentes que provocan una cambio significativo del nivel medio de la serie.

El primer punto, los modelos Box-Jekins son una herramienta estadística rigurosa con origen en la econometría, pero que han sido llevados a diversas áreas de la ingeniería para la modelación de una gran variedad de fenómenos que se presentan en la naturaleza haciendo un uso eficiente de los datos. En este capítulo no desarrollaremos esta metodología ya que es ampliamente conocida y sus planteamientos fueron presentados ya en la sección 2.2. En el capítulo 7 será aplicada para identificar un modelo ARIMA representativo para la serie que estudiaremos. Es muy importante mencionar que la identificación de este tipo de modelos conlleva una exploración detallada de la serie temporal en estudio, con lo cual se asegura una correcta identificación de la estructura del modelo.

El segundo punto utilizará la técnica de identificación de atípicos con lo cual conseguiremos dos objetivos, el primero será el de depurar la serie temporal y eliminar impactos puntuales en la serie como resultado de observaciones extrañas, imprevisibles, no sistemáticas, relacionadas con acontecimientos extraordinarios o errores en la manipulación de datos que deberán ser removidas de la serie y sustituidas mediante alguna técnica válida, por valores representativos. El segundo objetivo será el de identificar y evaluar la magnitud de las componentes de variabilidad sistemática pero de carácter irregular o de frecuencia anómala. Estas componente corresponderán a los eventos generados por el calendario de actividades y festividades del sitio en estudio. Una vez identificadas, se hará una clasificación de los mismos en base al día de su ocurrencia. Un desarrollo de los impactos que la ignorancia de este tipo de eventos puede tener en la estimación de los parámetros del modelo y en las predicciones se presentará en la siguiente sección.

Y finalmente el tercer punto nos llevará incorporar las componentes de variabilidad sistemática irregular caracterizadas previamente de una manera implícita en un modelo estadístico de predicción. Asumiendo para un escenario de predicción, que un evento generado por el calendario de actividades y festividades ocurriendo en un determinado día de la semana, producirá unos efectos similares a los observados y caracterizados previamente, lo que nos aportará predicciones puntuales más precisas.

Capítulo 5

Componentes de variabilidad sistemática irregular y atípicos en la demanda

El pasado es un prólogo.
William Shakespeare
1564-1616
Dramaturgo, poeta y actor inglés

Cuanto más atrás puedas mirar, más adelante verás.
Winston Churchill
1874-1965
Político británico

- **Componentes de variabilidad sistemática pero de carácter irregular o de frecuencia anómala (al ser sistemáticos son en gran medida previsibles)**
  Dentro de los componentes sistemáticos de carácter puntual podemos clasificar, por ejemplo, a la semana santa que impacta el comportamiento sistemático regular de la serie de demandas. La semana santa es un proceso puntual en el año que no se presenta siempre en la misma semana natural y su efecto no se puede recoger dentro de una estructura ARIMA regular. Dentro de esta misma clasificación podemos englobar a todas las fiestas de distinto carácter, ya sea nacionales, regionales, locales, provinciales,. . . , que modifican su ocurrencia cada año bisiesto.
- **Impactos puntuales en la serie como resultado de observaciones extrañas, imprevisibles, no sistemáticas, relacionados con acontecimientos**

extraordinarios o errores en la manipulación de datos. Dentro de esta clasificación se engloban a todos los errores de registro, fallos de los equipos de registro y manipulación de los datos.

Si nos adherimos a las clasificaciones de atípicos que se presentaron en las sección 2.4, los que se consideran como rasgos peculiares de la serie de demandas se ajustan principalmente a la definición de atípicos aditivos ó AO[1], ya que las festividades son principalmente eventos de carácter puntual y casi siempre periódica. La excepción se presentará cuando el evento consista en un encadenamiento de días festivos (que considerados individualmente serían atípicos aditivos, AO) como es el caso de la semana santa, en donde la definición de atípico transitorio o de cambio temporal (TC)[2] podría resulta más adecuada. Como se presentó en secciones anteriores, existen otros tipos de atípicos, sin embargo la mayoría de la bibliografía sugiere que los del tipo aditivo son los más frecuentes.

5.1. Efectos de los valores atípicos en la serie de demandas

Al obtener la estructura de un modelo ARIMA y estimar sus parámetros desde los datos, hemos asumido que cada uno de los datos que componen las serie de la que disponemos son representativos del fenómeno que estamos intentando modelar. Sin embargo, las series de demandas de agua urbana en entornos dénsamente poblados suelen contener rasgos *peculiares* que no son representativos del conjunto de los datos y que pueden ser considerados y tratados como atípicos. Esos son propios de cada región, provincia, país. Los estadísticos de una serie temporal que contenga este tipo de peculiaridades estarán sesgados.

La ignorancia de este tipo de rasgos pueden llegar a ser importante tanto en la estimación de los parámetros de los modelos ARIMA, como en los resultados de las predicciones de demanda de agua que obtenemos con las ecuaciones de los modelos de este tipo. De una manera muy somera podemos iniciar por decir que su grado de incidencia dependerá de:

- De la clasificación del rasgo peculiar (como fue explicado en la sección 2.4 de valores atípicos)

[1] Es un evento que afecta una serie solamente durante un periodo de tiempo

[2] Es un evento que produce un efecto inicial en un instante y su efecto se agota gradualmente con el paso del tiempo

- Del instante de su ocurrencia, es decir, si está más menos cerca del origen de la predicción
- Del horizonte de predicción
- De la magnitud de la peculiaridad
- Del proceso subyacente a la serie de demandas

5.1.1. *Efectos de los atípicos en las predicciones puntuales de la demanda*

Veamos los efectos de valores atípicos considerando que se conocen los coeficientes del modelo ARIMA y que un atípico (AO) ha sido ignorado. Entonces, la predicción l periodos hacia adelante de mínimo error cuadrático medio Y_{t_0+k+l}, desde el origen t_0+k, será, según Box et~al. (1976):

$$\hat{Y}_{t_0+k}(l) = \pi_1^{(l)} Y_{t_0+k} + \pi_2^{(l)} Y_{t_0+k-1} + \pi_3^{(l)} Y_{t_0+k-2} + \ldots = \sum_{j\geq 0} \pi_{j+1} Y_{t_0+k-j} \tag{5.1}$$

si los parámetros ya han sido estimados entonces $\pi_j^l = \pi_j$, y si

$$\pi(B) = 1 - \pi_1 B - \pi_2 B^2 - \ldots = \frac{\phi_p(B)\Phi_P(B^s)(1-B)^d(1-B^s)^D}{\theta_q(B)\Theta_Q(B^s)} \tag{5.2}$$

calculando π_j desde 5.2, entonces:

$$1 - \sum_{j\geq 1} \pi_j B^j = \frac{\phi_p(B)\Phi_P(B^s)(1-B)^d(1-B^s)^D}{\theta_q(B)\Theta_Q(B^s)} \tag{5.3}$$

siendo

$$\pi_j^{(l)} = \pi_{j+l-1} + \sum_{h=1}^{(l-1)} \pi_h \pi_j^{l-h},\ j = 1, 2 \ldots \tag{5.4}$$

Los pesos de predicción $\pi_j^{(l)}$ determinan hasta que punto un valor atípico de magnitud ω afecta la predicción. Si un atípico aditivo ha ocurrido en el periodo t_0 ($K \geq 0$ periodos antes del origen de la predicción), el error de predicción l-periodos adelante sería:

$$Y_{t_0+k+l} - Y_{t_0+k}(l) = e_{t+k}(l) - \omega \pi_{k+1}^{(l)} \tag{5.5}$$

donde

$$e_{t_0+k}(l) = a_{t_0+k+l} + \Psi_1 a_{t_0+k+l-1} + \ldots + \Psi_1 a_{t_0+k+l+1} \tag{5.6}$$

y Ψ_j, $j = 0, 1, 2, \ldots$, son los coeficientes de B^j en:

$$\Psi(B) = 1 + \Psi_1 B + \Psi_2 B^2 + \ldots = \sum_{j\geq 0} \Psi_j B^j = \frac{\theta_q(B)\Theta_Q(B^s)}{\phi_p(B)\Phi_P(B^s)(1-B)^d(1-B^s)^D} \tag{5.7}$$

Entonces como demostró Ledolter (1989), el incremento relativo en el error cuadrático medio (IECM) de la predicción l periodos hacia adelante viene dado por:

$$IECM(l; k, \omega) = \left(\frac{\omega}{\sigma_a}\right)^2 \frac{\left(\pi_{k+1}^{(l)}\right)^2}{\sum_{j=0}^{l-1} \Psi_j^2} \tag{5.8}$$

De esta ecuación podemos observar que la afectación a las predicciones dependerá de la estructura del modelo ARIMA que se esté utilizando, por ejemplo si el proceso es un autoregresivo, el valor atípico afectará a las predicciones cuando ocurran en las p observaciones más recientes, o en las $p+d$ si el proceso es un ARI(p,d), o en las $P + S$ si el proceso es un ARIMA(p,d,q X P,D,Q)S. Para el caso de los procesos integrados de media móvil, el efecto de un atípico sobre el ECM, no se anula cuando ocurre un número determinado de periodos anteriores al origen de la predicción. Sin embargo a medida que el número de periodos (k) aumenta, el efecto del atípico sobre el grado de exactitud de las predicciones en el ECM, disminuye. Finalmente podemos decir que existe una relación directa positiva entre la magnitud del atípico ($\frac{\omega}{\sigma_a}$) y el efecto cuantitativo del mismo en IECM. Las perturbaciones de los atípicos en las predicciones comentados anteriormente se verán aumentadas por el motivo de que en la práctica, los verdaderos valores de los parámetros de los modelos ARIMA no se conocerán y deberán ser estimados desde el conjunto de datos. Las series temporales que registran la demanda de agua de las ciudades son muy susceptibles a presentar algún tipo de atípico, ya sea generados por eventos de periodicidad irregular o por errores de medición, almacenamiento, etc.

5.1.2. *Efectos de los valores atípicos en la estimación de los parámetros*

Ledolter (1989) y Trivez (1994), han demostrado analíticamente y mediante ejercicios de simulación para modelos del tipo AR(1) y IMA(1,1) los siguientes efectos:

1. Los sesgos en las estimaciones de los coeficientes autoregresivos o de medias móviles que producen una especificación errónea del modelo, es decir, que se derivan del desconocimiento de la presencia de los eventos atípicos, pueden llegar a ser muy elevados.

2. Para cualquier tipo de atípico, e independientemente del instante de ocurrencia, los sesgos en la estimación del parámetro σ_a^2 cuando el modelo ignora la inclusión de los atípicos, son muy elevados y positivos.

Se debe destacar que existe una relación directa entre la varianza del ruido blanco (σ_a^2) y las varianzas de los errores de predicción lo que resulta en un incremento importante de la amplitud de los intervalos de confianza de las predicciones.

5.1.3. *Efectos de los valores atípicos en los intervalos de confianza de las predicciones*

Es deseable a la hora de realizar predicciones, el proporcionar no solo los valores puntuales sino acompañarlos de argumentos de incertidumbre, esto se hace normalmente en la forma de intervalos de predicción. Estos intervalos nos indicarán indirectamente, que tan bueno puede ser un modelo. No se debe esperar que las predicciones sean perfectas y los intervalos de predicción enfatizan este punto. Uno de los principales postulados de la metodología Box-Jenkins es el de que los residuos obtenidos después de ajustar un determinado modelo seguirán una distribución normal, que si se presenta estandarizada será *N*(0,1), de $\mu = 0$ y $\sigma = 1$. Una justificación de la frecuente aparición de la distribución normal es el teorema central del límite, que establece que cuando los resultados de un experimento son debidos a un conjunto muy grande de causas independientes, que actúan sumando sus efectos, siendo cada efecto individual de poca importancia respecto al conjunto, es esperable que los resultados sigan una distribución normal (Peña, 2008).

Si una serie a la que se le ha ajustado un determinado modelo contiene datos atípicos, entonces la estimación de σ^2, es decir la varianza (ecuación 5.9) de los residuos estará afectada. Existe una relación muy directa entre el valor de σ^2 y los intervalos de predicción, resultando en una mayor amplitud de estos.

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} \left(X_i - \bar{X}\right)^2 \tag{5.9}$$

Si una gran cantidad de los X_i valores son atípicos, entonces se obtendrá una σ^2 inflada. Los intervalos de predicción para la metodología de Box-Jenkins se basan en la desviación estándar σ, la raíz cuadrada de σ^2, ya que ésta representa las desviaciones de los residuos de las predicciones. Como se mencionó antes, para obtener los intervalos de predicción se asume que estos residuos siguen una distribución normal con media cero. Bajo este postulado, el intervalo de predicción para la siguiente observación será:

$$Y_{t+1} \pm z\sqrt{\sigma^2} \tag{5.10}$$

El valor de z determina el ancho y probabilidad del intervalo de predicción. Los valores de z están tabulados, como ejemplo si deseamos que el valor predicho tenga un 95 % de probabilidades de estar en ese intervalo entonces deberemos utilizar un valor de $z = 1{,}96$, o $z = 2{,}5$ si el intervalo deseado es el de 99 %. Es teoría estadística básica conocida y se puede comprobar que aproximádamente dos tercios de las observaciones estarán contenidas dentro de $\mu \pm \sigma$, el 95.5 % de la distribución se encuentra contenida en $\mu \pm 2\sigma$ y $\mu \pm 3\sigma$ contiene al 99.7 % de la distribución, ver figura 5.1.

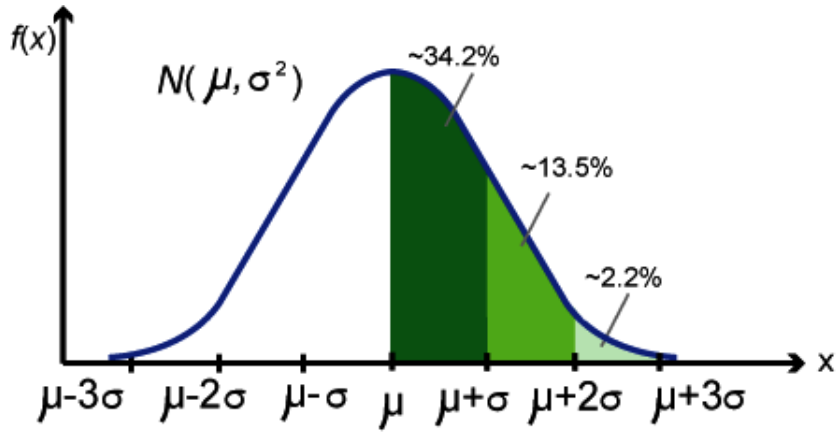


Figura 5.1: Distribución normal

Resulta bastante evidente de las ecuaciones 5.9 y 5.10 que el ignorar o no tener en cuenta la presencia de datos atípicos afectará directamente los intervalos de predicción que se producirán, provocando que el desempeño de un modelo pueda parecer peor de lo que realmente es. Ledolter (1989) encontró mediante simulaciones que para procesos de $n = 100$ representados mediante modelos ARIMA(0,1,1) con atípicos aditivos de magnitud 5σ, que para un $\theta = 0{,}5$ el valor de σ^2 se aumenta en un 32 % y el ancho del intervalo de predicción en un 15 %. No podemos ignorar que cuanto mayor sea la longitud de la serie los impactos en los diferentes estadísticos producidos por valores atípicos serán menores.

5.1.4. *Identificación y caracterización de los componentes de variabilidad sistemática irregular de la demanda*

Los eventos de variabilidad sistemática irregular pueden ser identificados fácilmente. Una aproximación gruesa podría ser la de calcular la relación entre el valor registrado y la desviación estándar de la serie en estudio, por ejemplo cualquier valor superior a 2 ó 3 desviaciones estándar se consideraría atípico. Sin embargo de esta forma solo conseguiríamos identificar la ubicación del atípico y seguiríamos desconociendo su magnitud y sus estadísticos. El germen de las metodologías de Chen and Liu (1993b), Chang et~al. (1988) o Peña (1990) está implementado en algunos de los programas de análisis estadísticos más comerciales (SPSS, SCA, SAS, etc.), con las cuales es posible identificar simultáneamente estimaciones robustas de los parámetros de los modelos de series temporales y revelar tanto la localización como la magnituds de los eventos atípicos. De una manera muy somera las metodologías mencionadas logran identificarlos ajustando un modelo de series temporales y analizando los residuos uno a uno calculando un estadístico que relaciona la magnitud del residuo y la desviación estandar de la serie de residuos. Si uno o varios datos obtenidos superan un valor crítico predefinido, el valor más alto es eliminado o corregido su efecto y se repite iterativamente el proceso hasta que todos los datos obtenidos están por debajo del valor crítico establecido.

La ocurrencia de eventos de variabilidad sistemática irregular en una serie de demandas producirán siempre una disminución significativa de la demanda ya que son el resultado de la suspensión de actividades de un gran número de personas, escuelas, comercios, etc, por lo que reproducirán los efectos de un valor atípico anómalo aún y cuando no lo son. Son más bien rasgos peculiares del proceso de demanda que más que identificarlos para su eliminación, nos interesa conocer sus magnitudes para poder prever sus efectos para el caso que se repitan en un escenario futuro.

Tenemos pues, que si contamos con una serie temporal de registros de demanda diaria suficientemente larga, podremos obtener un conjunto representativo de eventos de variabilidad sistemática irregular. Es así que podremos caracterizar los efectos de un día no laborable aconteciendo en Lunes, Martes,..., Domingo. La variabilidad irregular se evidencia en este punto, ya que no ocurrirá por ejemplo que todos los 1 de enero sean siempre un sábado como ocurrió el año 2001, o el 6 de enero en jueves para el mismo año, etc., y así para el resto de los eventos identificados. Por este motivo un evento de este tipo variará sus efectos según el día de la semana de su ocurrencia independientemente de que el evento sea generado por el mismo día no laborable, 1 de enero, 6 de enero, etc. La excepción para esta última

afirmación serán los días de semana santa, que son siempre jueves, viernes, sábado, domingo y en algunos sitios inclusive el lunes. En este caso los días de la semana en que ocurren serán siempre los mismos pero variarán la fecha en que acontecen, ya que están ligados al calendario lunar, por lo que de nueva cuenta para este tipo de eventos, su variabilidad es irregular en el tiempo.

Una vez identificados todos los eventos de variabilidad sistemática irregular, podremos caracterizarlos en base al día de la semana de su ocurrencia, obteniendo los impactos significativos típicos para un evento de este tipo para cada día de la semana. Hemos de asumir que esta caracterización será más representativa de los impactos para cada día de la semana cuanto más larga sea la serie temporal de que disponemos. Es poco riguroso esperar que los eventos atípicos de las series de demanda produzcan siempre los mismos impactos, ya que siempre existirán variables no consideradas en un modelo estadístico que aportarán su parte de incertidumbre (el tiempo $\rightarrow$ día con lluvia o sin lluvia, la temperatura imperante, las condiciones de la economía regional, etc.). Sin embargo debemos considerar que es una buena aproximación para mejorar la estimación de predicciones puntuales de predicción de la demanda.

Capítulo 6

Los modelos de intervención para la demanda

6.1. Antecedentes

Don't never prophesy: If you prophesies right,
ain't nobody going to remember and if you prophesies wrong
ain't nobody going to let you forget
Samuel Langhorne Clemens, alias Mark Twain
1835 - 1910
Humorista y escritor estadounidense

Los modelos de función de transferencia y de intervención han sido utilizados en la modelación de la demanda de agua urbana para diferentes escalas. En la revisión del estado del arte se han presentado varios trabajos localizados principalmente en Texas, secciones 3.4 y 3.5. Hipel et~al. (1975) sugirieron que esta metodología podía ser empleada para modelar fenómenos del área de la hidrología y presentaron un ejemplo de aplicación para el río Nilo. Maidment et~al. (1985) construyeron un modelo de función de transferencia que incorporaba la temperatura y la lluvia como variables independientes obteniendo una varianza explicada del 97 % para predicciones a un día para la ciudad de Austin. Shaw and Maidment (1987) desarrollaron un modelo con intervenciones para medir los impactos de los programas de racionamiento de agua potable en la ciudad de Austin. Sastri and Valdes (1989) construyeron un modelo preparado para operar en línea recalibrando sus parámetros y que incorpora la temperatura y la lluvia mediante intervenciones para reproducir los descensos de demanda cuando se presenta un evento de lluvia para la ciudad de Austin.

6.2. Incorporación de intervenciones para la demanda

Es claro que los modelos de intervención han ofrecido buenos resultados para modelar fenómenos que afectan a la demanda de agua urbana. Con este enfoque se propone la incorporación de los eventos de variabilidad sistemática irregular a un modelo estadístico de predicción del tipo ARIMA para predecir la demanda de agua en ciudades densamente pobladas españolas, ya que estos eventos aportan una gran variabilidad y sus valores casi siempre están lejos del patrón de demandas que se suele repetir cada 7 días. Sus efectos en un escenario de predicción no se limitan al día de su ocurrencia, sino que también afectan a las predicciones posteriores porque los residuos están también afectados.

Con el conjunto de eventos caracterizados como se ha definido en el punto anterior podremos anticipar e incorporar la magnitud de las desviaciones con respecto a las condiciones normales de predicción entregadas por el modelo ARIMA definido en la fase de estimación.

La predicción de la demanda se realizará de la forma tradicional de los modelos ARIMA, pero tendrá acopladas unas intervenciones que estarán inactivas siempre que el día que se vaya a predecir sea un día de actividad normal, y se activarán cuando el día a predecir corresponda con un día atípico. En esa estimación puntual la predicción del modelo ARIMA será afectada por la magnitud caracterizada para ese día de la semana.

Es así que como ejemplo el día 12 de octubre de 2004, día de la hispanidad, le corresponderá el descenso caracterizado para los días martes. Al ocurrir un evento de variabilidad sistemática irregular en un entorno de predicción con modelos ARIMA, estarán afectados los residuos posteriores generados por el modelo en un orden que dependerá de la estructura del modelo ARIMA que se haya identificado, por lo que se deberán corregir sus efectos afectándolos por los parámetros autoregresivos y de media móvil regular y estacional que forman el modelo para corregir las predicciones posteriores. Los efectos serán menores cuanto más alejado esté el evento del punto del origen de la predicción. Por lo que para predicciones a corto plazo, por ejemplo a 1 día, y el evento habiendo ocurrido uno o dos periodos antes, las correcciones en los ordenes de los parámetros de orden regular influenciarán fuertemente las predicciones. En cambio para predicciones a más largo plazo, por ejemplo 7 días, las correcciones en los ordenes estacionales serán más relevantes porque estos estarán guiando el proceso de predicción.

6.2.1. *Modelo de predicción de la demanda con intervenciones*

Las series temporales de demanda de agua potable suelen ser modeladas mediante modelos ARIMA estacionales, cuya ecuación característica es:

$$\phi_p(B)\Phi_P(B^s)(1-B)^d(1-B^s)^D X_t = \theta_q(B)\Theta_Q(B^s)Z_t \tag{6.1}$$

una explicación completa de esta ecuación se puede encontrar en 2.2.3, página 26 de este documento. Algunas formas alternativas de escribir la anterior ecuación de una forma simplificada es

$$Y_t = \psi(B)Z_t \tag{6.2}$$

donde

$$\psi(B) = 1 + \psi_1 B + \psi_2 B^2 + \ldots = \frac{\theta_q(B)\Theta_Q(B^s)}{\phi_p(B)\Phi_P(B^s)(1-B)^d(1-B^s)^D} Z_t \tag{6.3}$$

Y definiendo que los eventos de variabilidad sistemática irregular que nos interesa incorporar reproducirán los efectos de un atípico aditivo (AO) como fue definido en 2.4.1 de la página 44, entonces tendremos que

$$z_t = \omega_j I_t^h + \psi(B)Z_t \tag{6.4}$$

$$I_t^h = \begin{cases} 0 & t \neq h \\ 1 & t = h \end{cases}$$

donde t=h cuando se anticipe la ocurrencia de un evento de variabilidad sistemática irregular para el instante de la predicción.

Finalmente

$$\omega_j = \begin{cases} 1 \rightarrow & Lunes \\ 2 \rightarrow & Martes \\ 3 \rightarrow & Miércoles \\ 4 \rightarrow & Jueves \\ 5 \rightarrow & Viernes \\ 6 \rightarrow & Sábado \\ 7 \rightarrow & Domingo \end{cases}$$

donde ω_j tomará los valores para cada día de la semana de la caracterización que se habrá realizado previamente.

6.2.2. *El modelo de intervenciones paso a paso*

Para llegar a obtener un modelo de predicción de la demanda de agua potable se deberán seguir los siguientes pasos.

1. Realizar una análisis preliminar del cual se obtengan los estadísticos descriptivos de la serie temporal en estudio. Se buscará identificar los ciclos a nivel anual y semanal

2. Se deberán identificar el conjunto de eventos de variabilidad sistemática irregular que pudieran afectar al patrón regular de la demanda

3. Se aplicará la metodología de Box-Jenkins para la identificación de un modelo ARIMA que reproduzca la periodicidad sistemática regular, los ciclos y la tendencia que presente la serie de demandas. Los parámetros del modelo deberán ser calculados conjuntamente con una identificación de valores atípicos para conseguir que estos sean robustos y no sean sesgados por los valores atípicos. Habremos identificado entonces el modelo ARIMA de la ecuación 6.1. Se deberá verificar que los estadísticos descriptivos de los residuos nos indiquen que los residuos siguen una distribución cercana a la gaussiana.

4. Se verificará si los valores atípicos identificados coinciden en el tiempo con el conjunto de eventos de variabilidad sistemática irregular que se habrán identificado previamente. De existir coincidencia temporal serán caracterizados en base al día de la semana de su ocurrencia. Para el caso de que no exista coincidencia temporal, esos valores serán eliminados y sustituidos por valores característicos de la serie temporal.

5. Se calcularán los valores medios de las magnitudes de las desviaciones con respecto a las condiciones normales producidos por eventos de variabilidad sistemática irregular. Habremos obtenido entonces el catálogo de valores de ω_j.

6. Una vez identificado el modelo ARIMA y el catálogo de valores de ω_j podremos utilizarle para realizar las predicciones empleando la formulación 6.4. Se deberá tener precaución de no solo afectar las predicciones en los momentos de ocurrencia del evento, sino también en los $P+S$ predicciones posteriores a la ocurrencia del mismo.

El resto de los pasos será el análisis que se aplica a cualquier modelo de series temporales, un análisis exhaustivo de los residuos para observar si las periodicidades, las dependencias de la serie de sus propios valores pasados es decir las autocorrelaciones han sido captadas por el modelo de predicción. Con esta metodología deberemos conseguir obtener valores de correlación y de determinación superiores a los que obtendría un modelo ARIMA estacional operando por si solo, primero porque los parámetros estimados serán más robustos y segundo porque las predicciones serán más precisas a lo largo del proceso de predicción pero sobre todo en el momento de ocurrencia de las intervenciones, donde los residuos serán mucho menores. Tendremos entonces a la vez, unos intervalos de confianza de las predicciones más estrechos, ya que los valores de σ de los residuos se habrán reducido. En el capítulo 7 se ha desarrollado una aplicación detallada de la metodología aquí descrita y a la vez se aplicará la metodología de las redes neuronales para emplearlo como una base comparativa.

Es evidente que la aplicación de esta metodología nos llevará a identificar un modelo y un conjunto de eventos que serán propios únicamente para la serie temporal en estudio. Los valores obtenidos solo serán representativos para esta serie y por ningún motivo deberán ser exportados a otros puntos de un sistema de distribución de agua ni a otro sistema de distribución de otra área o ciudad. En cambio, el proceso mencionado líneas arriba se deberá repetir hasta obtener un modelo que se ajuste a las condiciones de este otro sistema.

Parte V

Caso de estudio

Capítulo 7

Caso de estudio

**Mientras cada individuo puede ser un enigma insoluble,
un conjunto de ellos se comporta con exactitud matemática.**
Sherlock Holmes
Personaje ficticio creado en 1887 por Sir Arthur Conan Doyle
Destaca por su inteligencia y hábil uso
de la observación y el razonamiento deductivo.

**Trying to predict the future is like trying to drive down a country road
at night with no lights while looking out the back window.**
Peter Drucker
1909 – 2005
Abogado y tratadista austríaco

7.1. Planteamiento del Problema

Un problema importante en la gestión de sistemas de suministro y distribución de agua potable, es la predicción de la demanda diaria con el fin de programar en las fuentes de captación los caudales que serán demandados, preparar los bombeos necesarios buscando minimizar los costes energéticos y evitar sobrepresiones en la red de distribución. El caso de la ciudad de Valencia no es la excepción. El sistema suministra agua potable a 807,396 habitantes (Valencia-Ayuntamiento, 2006b) en su zona urbana y contaba al 2005 con 417,868 abonados, de los cuales 372,580 eran domésticos y 45,288 industriales (Valencia-Ayuntamiento, 2006a), el área metropolitana que incluye a l'horta nord, l'horta sud, l'horta oest y l'horta centro también son abastecidas por este sistema.

ETAP	Capacidad de Tratamiento ($\frac{m^3}{seg}$)	Capacidad de Tratamiento ($\frac{m^3}{día}$)	Capacidad de Almacenamiento(m^3)
La Presa	3.2	276,000	90,000
El Realón	3.0	259,000	100,000
TOTAL	6.2	535,000	190,000

Cuadro 7.1: Resumen de estaciones de tratamiento de agua potable que abastecen a Valencia

El sistema cuenta con dos estaciones de tratamiento de agua potable (en delante ETAP) que suministran agua a la red de distribución. La ETAP de nombre *La Presa-Manises* tiene a su vez dos fuentes de abastecimiento, el sistema hidrológico del río Turia y el canal Júcar-Turia, que como su nombre lo indica suministra agua desde el sistema hidrológico del Júcar a la ETAP *La Presa-Manises*. Los volúmenes que no pueden ser tratados en esta estación son vertidos al río Turia, aunque esto sucede en muy pocas ocasiones. El cuadro 7.1 presenta un resumen de las capacidades de tratamiento y almacenamiento de las ETAPs existentes.

La ETAP *Picassent-El Realón* se abastece exclusivamente del sistema hidrológico del río Júcar que es regulado en el embalse de *Tous*. Los caudales aportados por las dos ETAP son casi iguales. En el año 2005 la ETAP *La Presa-Manises* aportó a la red 62,607 miles de m^3 (1.98 m^3/seg) mientras que la ETAP *Picassent-El Realón* aportó 60,140 miles de m^3 (1.91 m^3/seg). Además de las dos ETAP, existen pozos de emergencia que suministran agua directamente a la red que no se operan constantemente a lo largo del año y que aportaron en el año 2005, 1,091 miles de m^3 (0.03 m^3/seg). El sistema de abastecimiento de agua de Valencia no cuenta con depósitos de regularización importantes en los cuales se pudiera almacenar un volumen de agua suficiente en el supuesto que se presentara un evento de fallo en las fuentes principales, una disminución o un aumento importante de la demanda. Las ETAP cuentan con depósitos anexos, en el caso de *La Presa-Manises* los depósitos Montemayor de 70,000 m^3 y Collado de 20,000 m^3. *Picassent-El Realón* tiene un depósito con dos módulos del mismo nombre y 100,000 m^3 de capacidad. También podríamos considerar como volumen de regularización el de las propias ETAP. Sin embargo estas instalaciones son dinámicas y no pueden contener el agua mucho más tiempo de lo que dura su proceso de tratamiento. Considerando el volumen de los depósitos mencionados y la demanda media de alrededor de 321,000 m^3 diarios, en caso de un evento de fallo se podría continuar abasteciendo el sistema solamente por 13 horas a partir del momento del fallo. Otro punto importante a mencionar es que el tiempo de viaje del agua suministrada desde la obra de toma en la presa de Tous por el canal Júcar-

Turia es muy grande, de alrededor de las 24 horas, por lo que si se varía el caudal suministrado desde ese punto, las ETAP lo notarán 24 horas después. No es difícil darse cuenta que sin una bien ejecutada operación, el sistema está sujeto a un riesgo significativo de fallo, por ejemplo una situación de sobrepresiones a la red de distribución que resulte en pérdidas por fugas o un vertido del canal Júcar-Turia al río Turia por no existir una demanda en la ETAP.

7.2. Objetivos del análisis

En este ejercicio práctico de investigación, se plantea utilizar las herramientas que el análisis estadístico nos proporciona para construir varios modelos estadísticos sensibles, así como un conjunto de redes neuronales, para predecir las demandas de agua de la ciudad de Valencia, siempre pensando en la utilidad práctica del modelo como una herramienta de decisión y por lo mismo tratando de mantener la simplicidad para una sencilla implementación. El horizonte de predicción será definido en base a las características del sistema y del modelo resultante, buscando que sea siempre el máximo en el cual el valor predicho pueda ser de utilidad en la toma de decisiones.
Se realizará como un primer paso, un análisis de la serie basado en la estadística descriptiva para llegar a familiarizarnos con las series con que se cuenta para luego utilizar metodologías enfocadas al análisis de series temporales y de redes neuronales para identificar los modelos adecuados.

7.3. Análisis Preliminar

En esta sección se realizará un análisis cualitativo y cuantitativo de la serie de demandas, considerando la serie completa así como series individualizadas por año, utilizando herramientas de la estadística clásica y métodos gráficos, con el fin de identificar características relevantes que nos pudieran explicar rasgos del comportamiento de la serie. Los datos que serán analizados consisten en una serie de 4 años de datos diarios de demandas que fueron proporcionados por el departamento técnico de Aguas de Valencia. Los datos corresponden a los volúmenes diarios tratados en cada una de las ETAP *La Presa-Manises* y *Picassent-El Realón* en el periodo comprendido entre el 1 de Enero del 2001 al 31 de Diciembre del 2004, siendo un total de 1,461 datos diarios. Se cuenta también con la serie de temperaturas medias, máximas y mínimas para el periodo antes mencionado y de la misma forma

serán analizadas en busca de patrones de correlación climática con la serie de demandas.

7.3.1. *Estadísticos Básicos y patrones predominantes de la demanda*

Tendencia

Con el fin de observar los patrones predominantes en la serie de demandas de la ciudad de Valencia se ha graficado la serie completa en la figura 7.1. Al agregar una línea de tendencia se observa que esta tiene una pendiente de 28.11 m^3/día o 10,260 m^3/año y nos da una magnitud de la tendencia a la alza de la serie a largo plazo, aunque no se descarta que si se contara con una serie más larga, en un futuro la tendencia pudiera ser invertida por ejemplo mediante programas de recuperación de perdidas por fugas en las redes, racionamiento o por una sensibilización de la población en el uso del agua. En el mismo gráfico se puede observar que existe un patrón estacional similar en todos los años con valores mínimos en los tercios de cada uno de los años y valores máximos a la mitad del año, aunque cada uno de estos máximos y mínimos con distinta magnitud.
En la observación individual de la serie del año 2001 (fig. 7.2), a la cual se le ha agregado una línea de media móvil de 7 días se puede ver que el valor medio de la serie oscila en los 300,000 m^3 con valores mínimos en los meses de abril y agosto y valor máximo a finales del mes de junio e inicios de julio. El cuadro 7.2 resume los estadísticos básicos de la serie completa y también los de cada uno de los años. Se incluyen también los gráficos de los años 2002 al 2004 (figuras 7.3, 7.4, 7.5) para que los patrones puedan ser comparados.

Estadístico	2001	2002	2003	2004	2001-2004
Número de Datos	365	365	365	366	1461
Valor Medio (m^3/día)	307,995.92	311,240.03	331,019.34	335,998.80	321,573.40
Mediana (m^3/día)	309,212.45	314,793.21	333,687.36	337,154.905	324,001.70
Desv. Est. (m^3/día)	24,346.60	23,983.69	22,758.83	22,794.36	26,409.83
Varianza ((m^3/día)^2)	5.9E+008	5.8E+008	5.2E+008	5.2E+008	7.0E+008
Mínimo (m^3/día)	218,212.26	230,605.47	238,398.11	247,273.77	218,212.26
Máximo (m^3/día)	367,439.62	367,319.70	386,866.98	386,276.42	386,866.98
Rango (m^3/día)	149,227.36	136,714.23	148,468.87	139,002.65	168,654.72
Asimetría	-0.49	-0.61	-0.54	-0.525	-0.44
Curtosis	0.31	0.49	1.23	0.0492	0.358

Cuadro 7.2: Resumen de estadísticos básicos de la serie de demandas

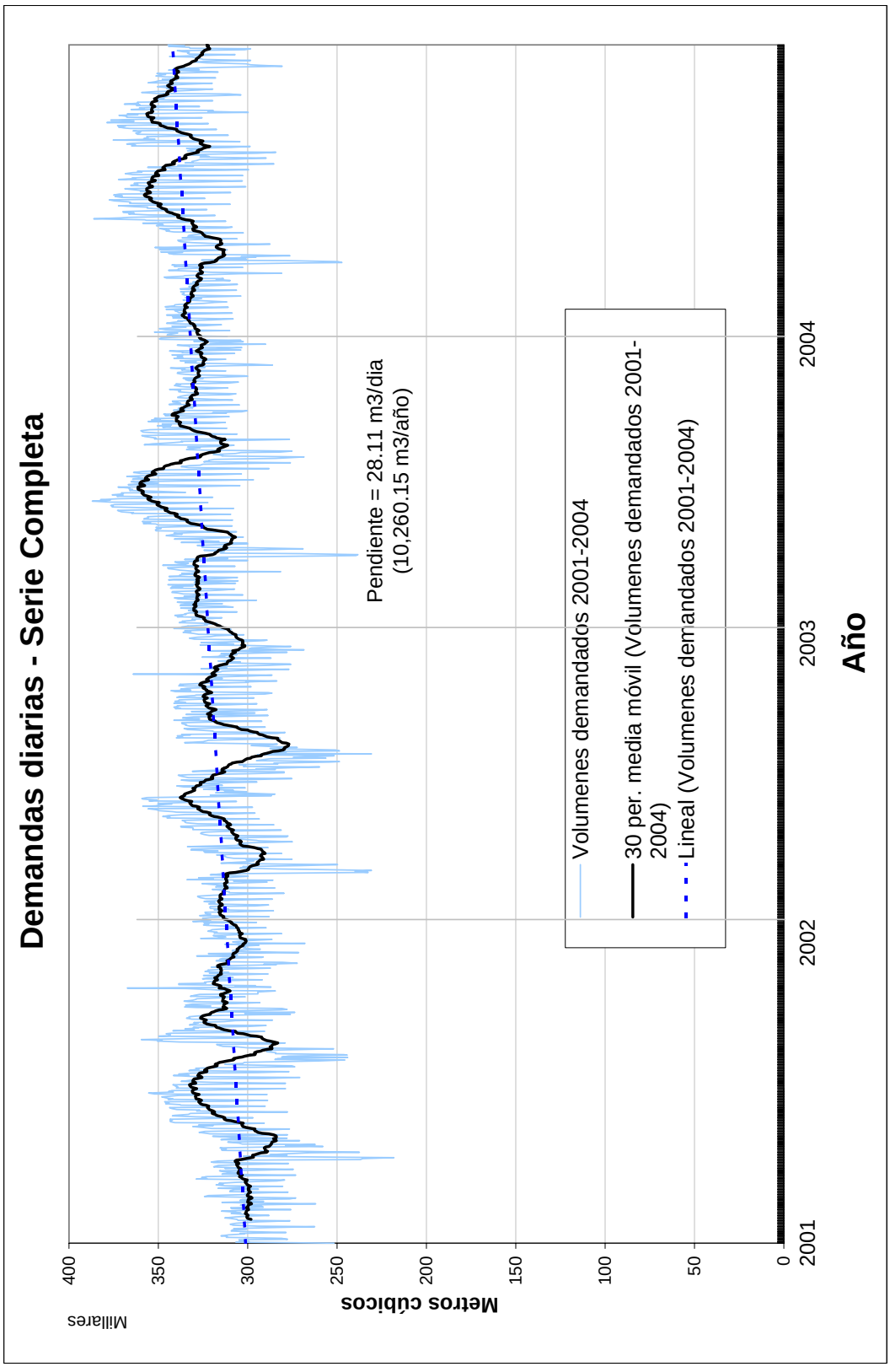


Figura 7.1: Serie de demandas de la ciudad de Valencia de Enero del 2001 a Diciembre del 2004

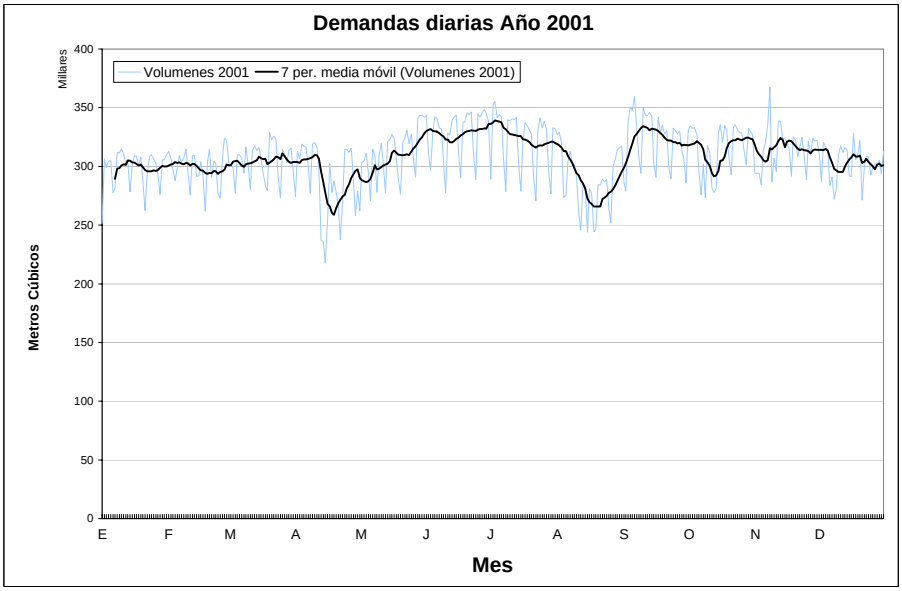


Figura 7.2: Serie de demandas de la ciudad de Valencia, año 2001

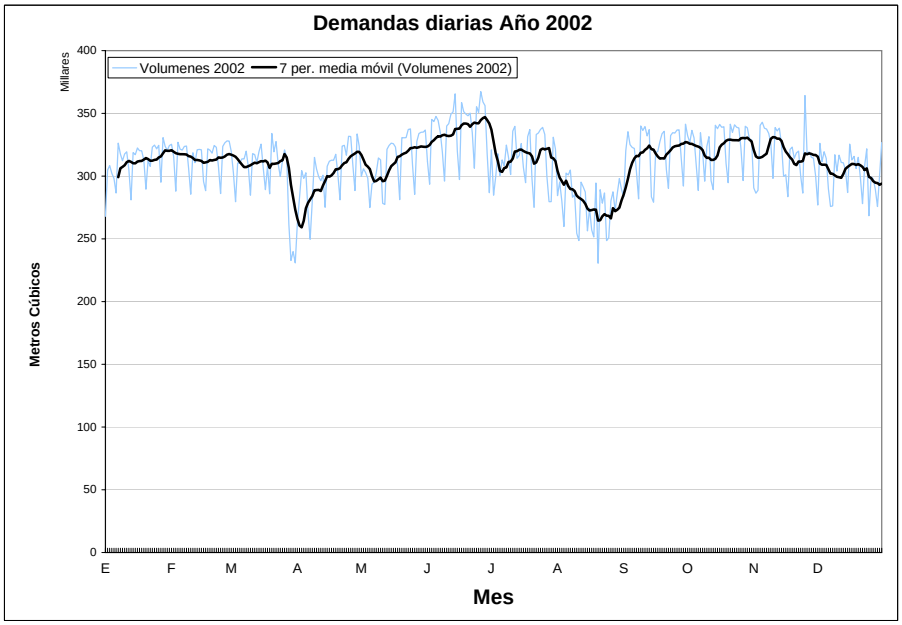


Figura 7.3: Serie de demandas de la ciudad de Valencia, año 2002

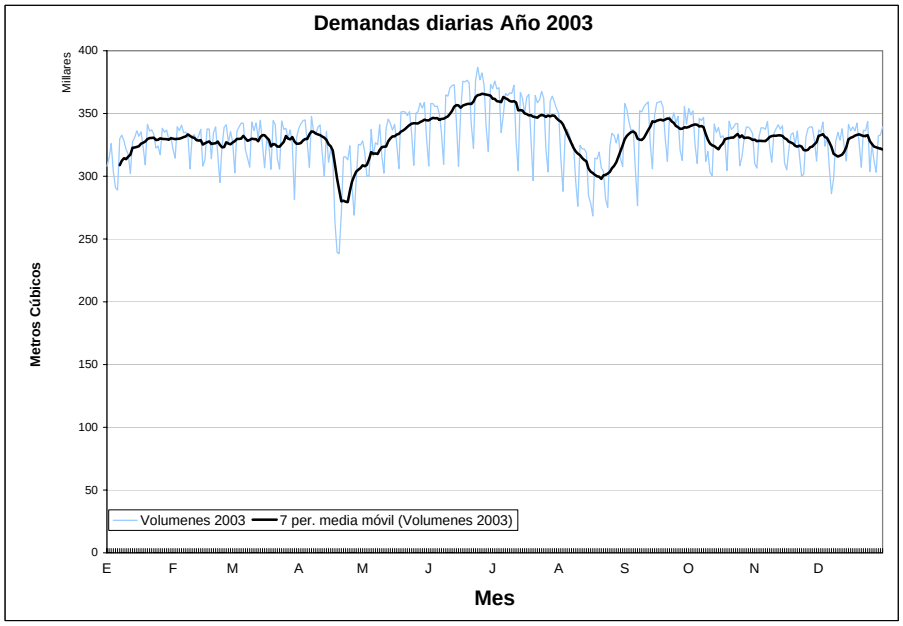


Figura 7.4: Serie de demandas de la ciudad de Valencia, año 2003

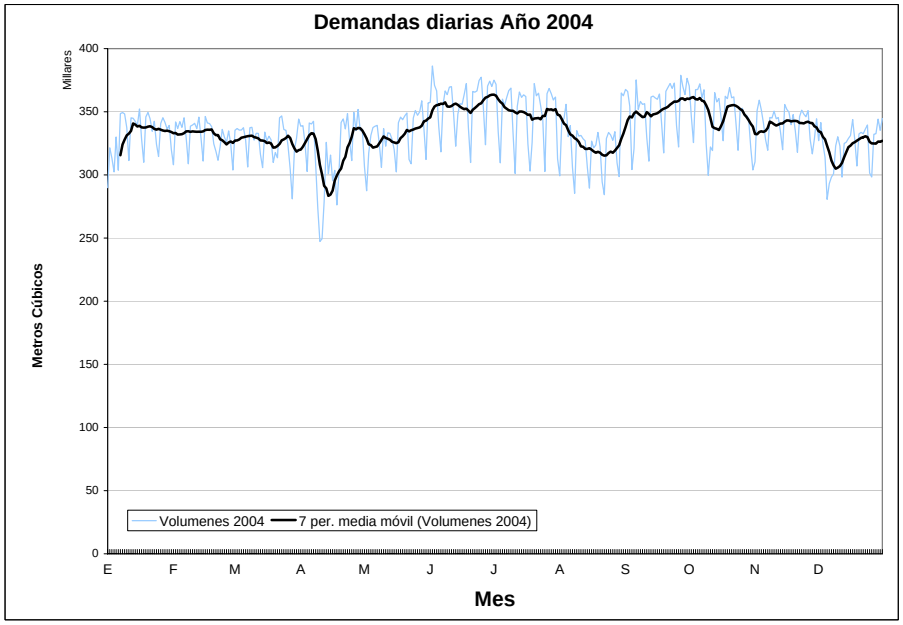


Figura 7.5: Serie de demandas de la ciudad de Valencia, año 2004

Estacionareidad de la serie de demandas

Una definición matemática de un proceso estacionario fue explicada en la sección 2.2.1. No obstante, un concepto intuitivo de una serie estacionaria se cumpliría si no existiese en la serie un cambio sistemático en la media (tendencia) y si no existiese una variación sistemática de la varianza. En otras palabras las propiedades de una sección de datos son muy similares a cualquier otra sección de la serie.
Las series no estacionarias pueden ser detectadas calculando el coeficiente de la función de autocorrelación (ACF) definido por (Box et~al., 1976)

$$r_k = \frac{\left(\sum (x_{(t)} \quad x_{(t+k)})\right) - \left(\frac{1}{n-k} \sum x_{(t)} \quad \sum x_{(t+k)}\right)}{\left[\sum x_{(t)}^2 - \frac{1}{n-k}(\sum x_{(t)})^2\right]^{1/2} \left[\sum x_{(t+k)}^2 - \frac{1}{n-k}(\sum x_{(t+k)})^2\right]^{1/2}} \tag{7.1}$$

donde x es la variable, t es el tiempo y las sumatorias se realizan desde $t = 1$ a $t = n - k$. La ecuación (7.1) determina el grado de correlación entre observaciones que están separados k unidades de tiempo.

El gráfico 7.6 representa el correlograma o (ACF) de la serie que estamos analizando. Se puede observar que existe una dependencia fuerte con los valores anteriores o lo que es lo mismo, los coeficientes de r_k no se reducen rápidamente hacia cero, que es el comportamiento del correlograma de una serie estacionaria. Por lo tanto podemos considerar que nuestra serie es no estacionaria. La interpretación del correlograma tiene más usos que el que ahora le estamos dando; más adelante nos será de utilidad a la hora de identificar un modelo adecuado para nuestra serie temporal.

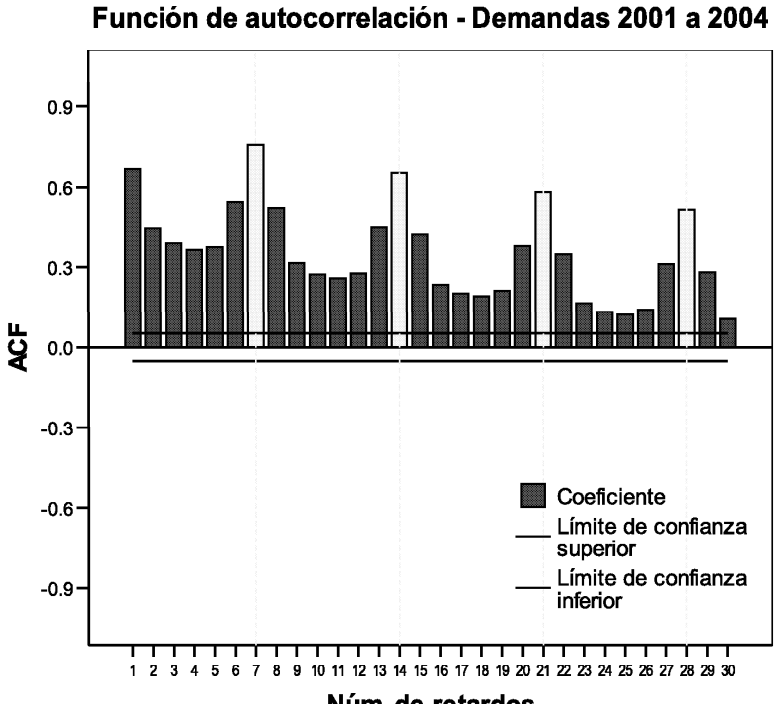


Figura 7.6: Gráfico de la función de autocorrelación de la serie de demandas del 2001 al 2004

Estacionalidad, Demandas Máximas y Mínimas

De la inspección visual de los gráficos y del cuadro de estadísticos básicos podemos observar que la serie de demandas de agua de la ciudad de Valencia presenta una muy marcada variación estacional influenciada por factores climáticos y sociales como son los periodos vacacionales y festivos. Posteriormente se analizará más en detalle la influencia que tienen estos factores en la demanda. Es esperable que se presenten demandas mínimas en los meses de Abril y Agosto ya que tradicionalmente estos meses son utilizados por los usuarios para vacacionar, abandonar la ciudad para desplazarse a los pueblos de los alrededores, por lo que esa disminución de la demanda puede verse reflejada en otros pueblos como un incremento. En el caso de la disminución de la demanda en los meses de Agosto, éstas no son coincidentes con las temperaturas calurosas de esos meses. Se podría considerar que el efecto del factor social supera al climático en este caso. Los valores de demanda observados en esas épocas son de alrededor de los 233,622.40 m^3/día (obtenido como el valor medio de las demandas mínimas de los 4 años de serie), siendo la demanda media diaria de alrededor de los 321,573.40 m^3/día, es decir una disminución del orden del 27 % de la demanda. El valor obtenido anteriormente tiene como fin solamente el de tener un número grueso de la magnitud de las reducciones que se producen.
En cuanto a las demandas máximas se observa que estas se presentan en los meses de Junio y Julio con los valores máximos a inicio de este último mes. Es bastante obvio que el factor climático es determinante en estos picos de demanda ya que coinciden con los meses que registran la evapotranspiración máxima y lluvias casi nulas en la región. Los valores de demandas de esos meses oscilan alrededor de los 376,975.68 m^3/día, resultando en un incremento del orden del 17 % con respecto a la demanda media diaria. De este análisis de la estacionalidad, se observa que las variaciones en la demanda media diaria resultantes de disminuciones por periodos vacacionales, son mayores en magnitud pero no en duración que las resultantes en aumento de la demanda por factores climáticos. Es entendible que esos picos máximos sostenidos someten a un mayor estrés como un primer punto, a las fuentes de captación y como segundo a la red de distribución del sistema. En cuanto a las demandas mínimas, se entiende que más que un problema para el organismo operador, representan un ahorro de agua y se puede conocer con antelación sin la necesidad de un modelo, el inicio, fin y por lo tanto la duración que tendrán. Sin embargo la magnitud de la reducción se desconocerá y será más acusada. Estos eventos igualmente implican decisiones de operación del sistema, en este caso de disminución de los caudales entregados a la red con el fin de evitar sobrepresiones innecesarias que pongan en riesgo de fallo a la red de distribución.

Descomposición de la serie temporal

La descomposición de series temporales, aunque ha sido usada por si sola en algunas ocasiones como método de predicción (se pueden ver análisis similares en Makridakis et~al. (1997)), la usaremos en nuestro caso solo como una herramienta gráfica para adentrarnos en la serie temporal y conocerla mejor.
En este caso hemos utilizado una descomposición aditiva, de la forma

$$X_t = S_t + T_t + Et$$

donde

- X_t es el valor de la serie temporal en el periodo t
- S_t es el componente estacional o índice en el periodo t
- T_t es el componente de tendencia-ciclo en el periodo t, y
- E_t es el componente residuo al periodo t

Hemos obtenido el gráfico 7.7 que corresponde a la descomposición de la serie de demandas del año 2001. Se ha graficado solamente la parte correspondiente al año 2001 por motivos de escala, los patrones son repetitivos para el resto de la serie. El propósito de la descomposición es separar la serie de demandas en componentes de tendencia-ciclo, estacional y residuos.

Los valores del gráfico superior, se obtuvieron ajustando a la serie una media móvil de orden 7 (una semana). El segundo gráfico, que representa la estacionalidad, se obtuvo restando la serie de media móvil a los datos originales. Los índices estacionales para cada día de la semana se computan promediando los valores, por ejemplo de todos los lunes existentes en la serie, y así para todos los días de la semana. Los valores de los índices estacionales fueron escalados para que un día promedio tenga valor 0. Los índices varían desde un mínimo de -29,377.30 m^3de los domingos a un máximo de 9,788.02 m^3de los miércoles. Esto indica que hay una oscilación estacional semanal de -29,377.3 m^3 por debajo a 9,788.02 m^3 por arriba del día promedio a lo largo de un ciclo completo de una semana. Finalmente el gráfico inferior representa el componente de los residuos o la parte irregular de la serie, donde los valores se han escalado para que el promedio de los residuos sea 0.

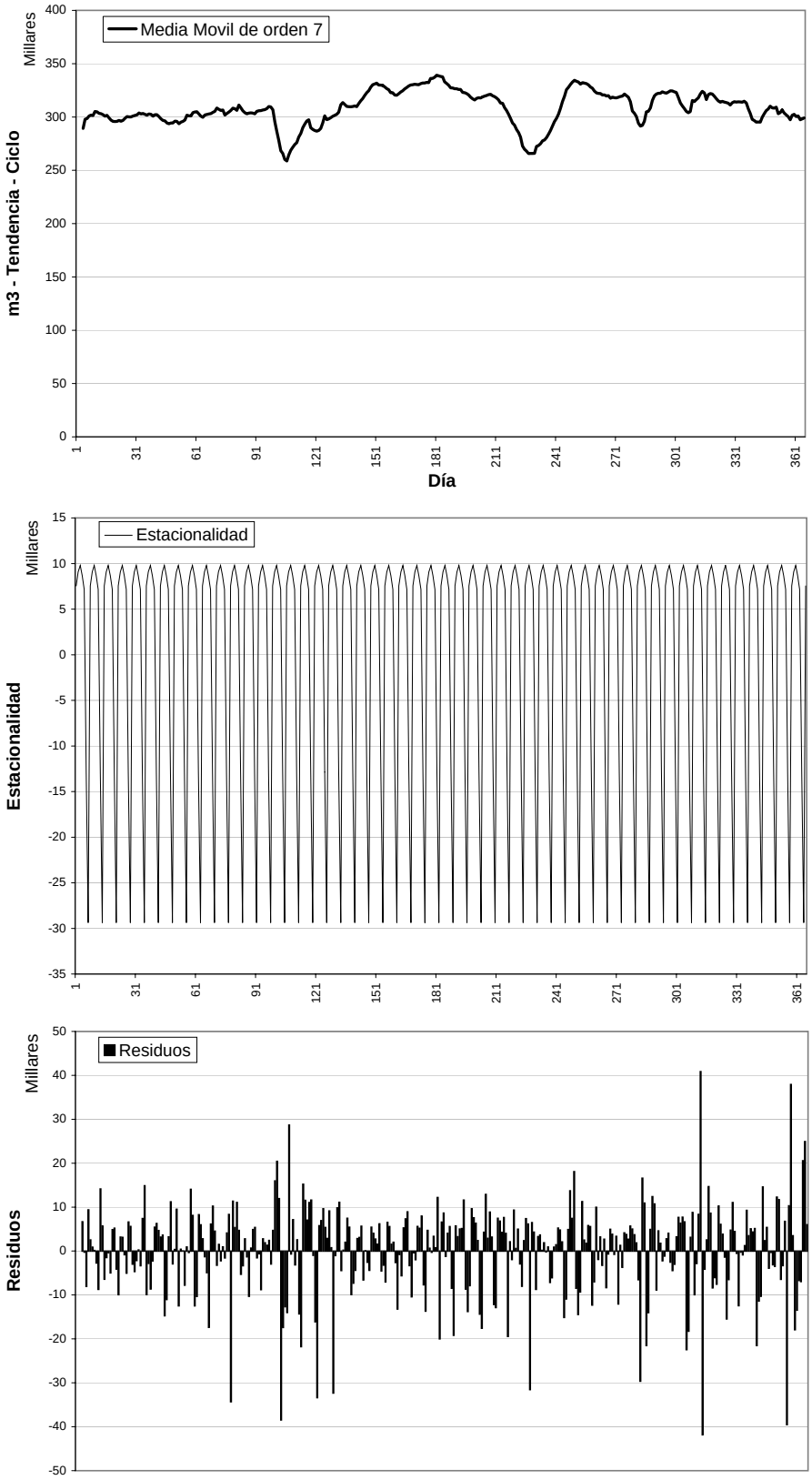


Figura 7.7: Descomposición Aditiva de la Serie de demandas de la ciudad de Valencia, año 2001

Componentes estacionales de la demanda

En el estudio de series temporales se utilizan principalmente dos técnicas para identificar componentes periódicos, es decir, componentes que se repitan varias veces en un intervalo de tiempo definido. Un análisis basado principalmente en la estimación de la función autocorrelación (o la autocovarianza) se denomina análisis en el dominio del tiempo. Un análisis basado principalmente en el espectro es llamado un análisis en el dominio de las frecuencias o análisis espectral (Chatfield, 2001).
Los componentes estacionales o periódicos pueden ser detectados usando el análisis de Fourier, como le define (Bougadis et~al., 2005)

$$x(t) = a_0 + \sum_{m=1}^{N/2} a_m \cos(\frac{2\pi mt}{N}) + b_m \sin(\frac{2\pi mt}{N}) \tag{7.2}$$

donde a_0 es la media del conjunto de datos, m es el número de armónicos, N es el valor de la longitud de la serie, y $t = 1, 2, ..., N$ es el tiempo. Los coeficientes de la serie de Fourier (a_m y b_m) cuando $m \neq \frac{N}{2}$ se definen como

$$a_m = \frac{2}{N} \sum_{m=1}^{N} x_t \cos\left(\frac{2\pi mt}{N}\right) \qquad b_m = \frac{2}{N} \sum_{m=1}^{N} x_t \sin\left(\frac{2\pi mt}{N}\right) \tag{7.3}$$

donde para $m = \frac{N}{2}$

$$a_m = \frac{1}{N} \sum_{m=1}^{N} x_t(-1)^t \qquad b_m = 0 \tag{7.4}$$

Si s^2 es la varianza de la serie temporal $x(t)$, entonces la parte de la varianza explicada (C_m^2) por el m-ésimo armónico se define por

$$C_m^2 = \frac{a_m^2 + b_m^2}{2s^2} \tag{7.5}$$

Como se ha venido haciendo, se ha fraccionado la serie por años y se ha hecho un análisis individual utilizando la formulación del análisis de Fourier, los resultados se presentan en los cuadros 7.3, 7.4, 7.5, 7.6. Análisis similares se pueden encontrar en Bougadis et~al. (2005).

De estos cuadros podemos observar que la componente periódica más predominante en todas las series es el patrón semanal, es decir la de duración 7 días, armónico 52. Esta componente para los años 2001 al 2004 nos produce una varianza explicada de 26, 17, 27 y 29% respectivamente. Es de esperar que si el armónico 52 es significativo, también lo serán probablemente sus múltiplos (104, 156,...), y así ocurre en los cuatro años analizados.

En la sección (7.3.1) se obtuvo el gráfico (7.6) que representa el correlograma de la serie del cual se puede observar que existe un patrón repetitivo cada 7 retardos, lo cual confirma los resultados obtenidos por el análisis de Fourier. El componente de duración 121.7 días no puede ser observado en el correlograma por motivos de escala, tendríamos que generar uno con 242 o 363 retardos para que este patrón pudiera ser observado, resulta por demás innecesario si realizamos un análisis de Fourier.

De este análisis podemos concluir que existe una marcada estacionalidad en las series y que en su momento deberá ser considerado a la hora plantear las ecuaciones del modelo o a la hora de elegir las variables de entrada de las redes neuronales.

Año 2001			
Armónico	**Frecuencia**	**Periodo (días)**	**Varianza Explicada**
52	0.1425	7.0	26%
3	0.0082	121.7	11%
104	0.2849	3.5	10%
1	0.0027	365.0	6%
6	0.0164	60.8	6%
4	0.0110	91.3	5%
2	0.0055	182.5	4%
7	0.0192	52.1	2%
9	0.0247	40.6	2%
156	0.4274	2.3	1%

Cuadro 7.3: Resultados encontrados del análisis de Fourier, Año 2001

Año 2002			
Armónico	**Frecuencia**	**Periodo (días)**	**Varianza Explicada**
3	0.0082	121.7	24%
52	0.1425	7.0	17%
104	0.2849	3.5	9%
2	0.0055	182.5	6%
8	0.0219	45.6	5%
156	0.4274	2.3	2%
5	0.0137	73.0	2%
7	0.0192	52.1	1%
11	0.0301	33.2	1%
19	0.0521	19.2	1%

Cuadro 7.4: Resultados encontrados del análisis de Fourier, Año 2002

Año 2003			
Armónico	**Frecuencia**	**Periodo (días)**	**Varianza Explicada**
52	0.1425	7.0	27%
3	0.0082	121.7	14%
104	0.2849	3.5	11%
1	0.0027	365.0	7%
4	0.0110	91.3	6%
2	0.0055	182.5	5%
6	0.0164	60.8	3%
156	0.4274	2.3	3%
7	0.0192	52.1	2%
10	0.0274	36.5	2%

Cuadro 7.5: Resultados encontrados del análisis de Fourier, Año 2003

Año 2004			
Armónico	**Frecuencia**	**Periodo (días)**	**Varianza Explicada**
52	0.1425	7.0	29 %
3	0.0082	121.7	16 %
1	0.0027	365.0	11 %
104	0.2849	3.5	10 %
157	0.4274	2.3	3 %
6	0.0164	60.8	2 %
12	0.0329	30.4	1 %
14	0.0386	26.1	1 %
4	0.0603	91.3	1 %
22	0.0627	16.6	1 %

Cuadro 7.6: Resultados encontrados del análisis de Fourier, Año 2004

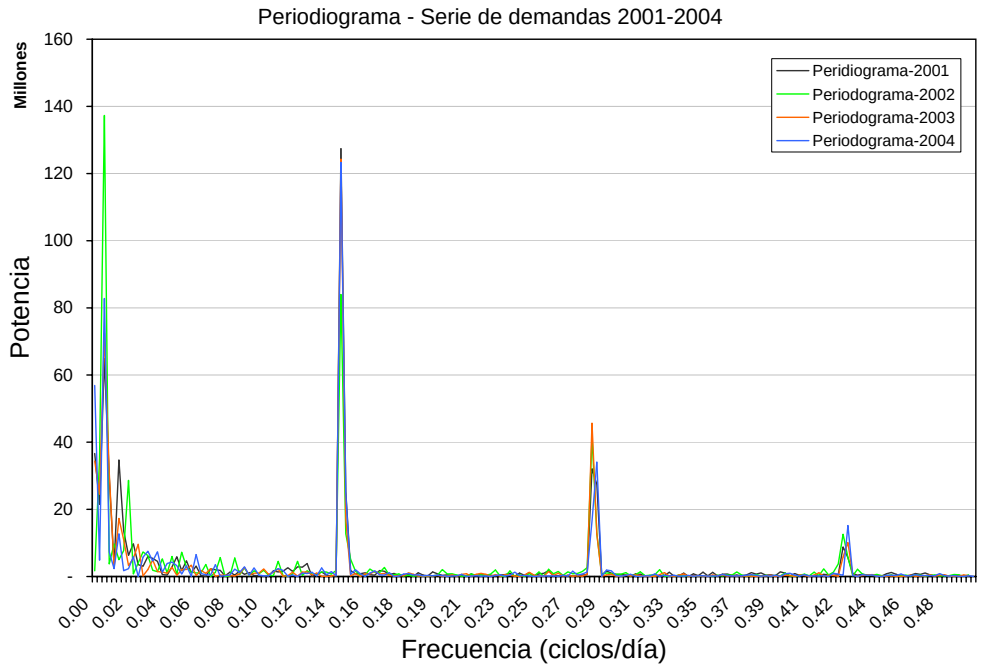


Figura 7.8: Periodograma de frecuencias de la serie de demandas de la ciudad de Valencia, año 2001 al 2004

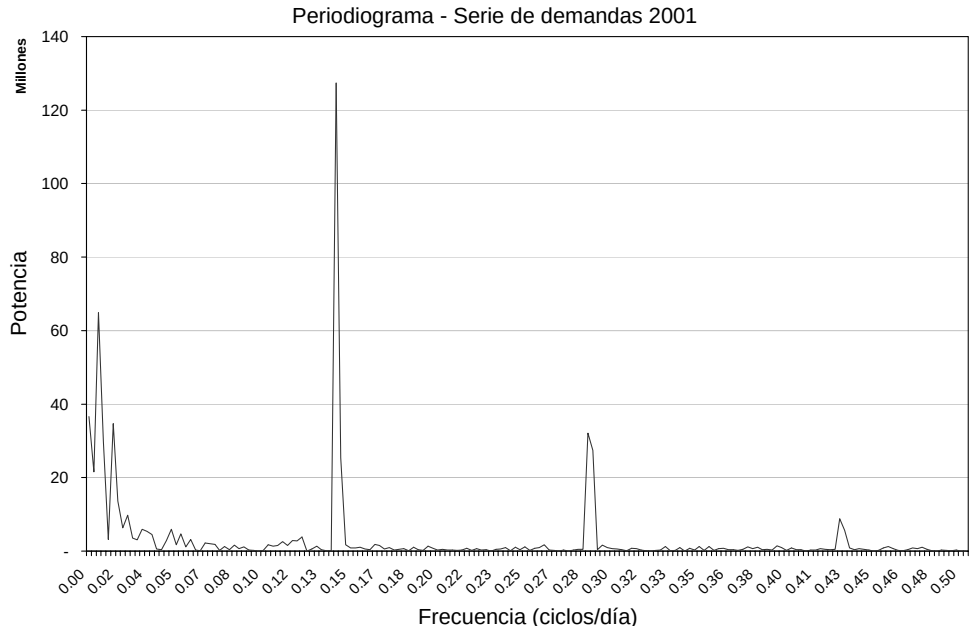


Figura 7.9: Periodograma de frecuencias de la serie de demandas de la ciudad de Valencia, año 2001

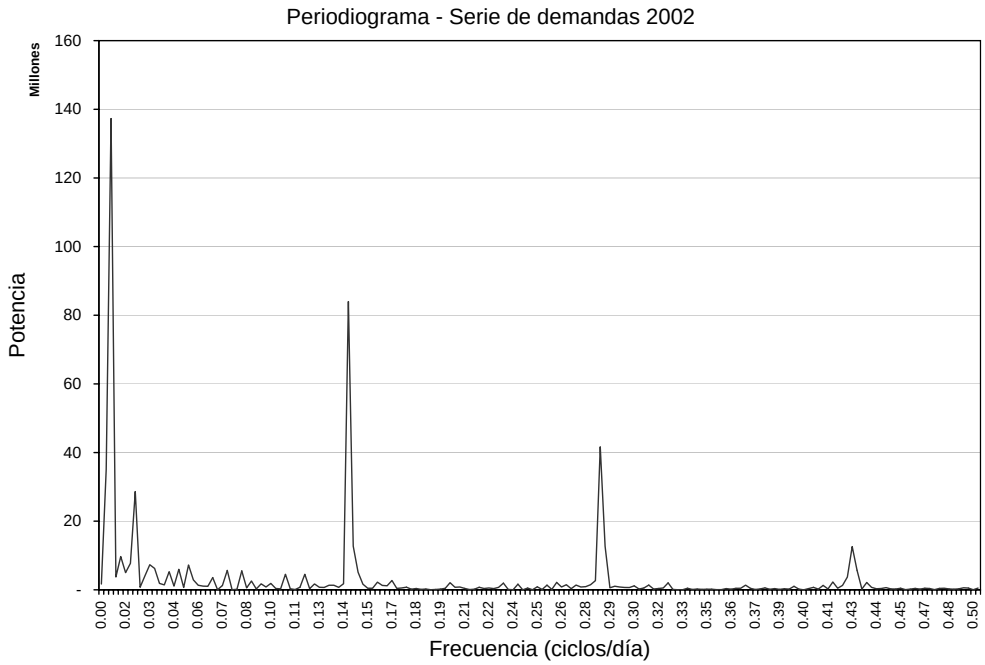


Figura 7.10: Periodograma de frecuencias de la serie de demandas de la ciudad de Valencia, año 2002

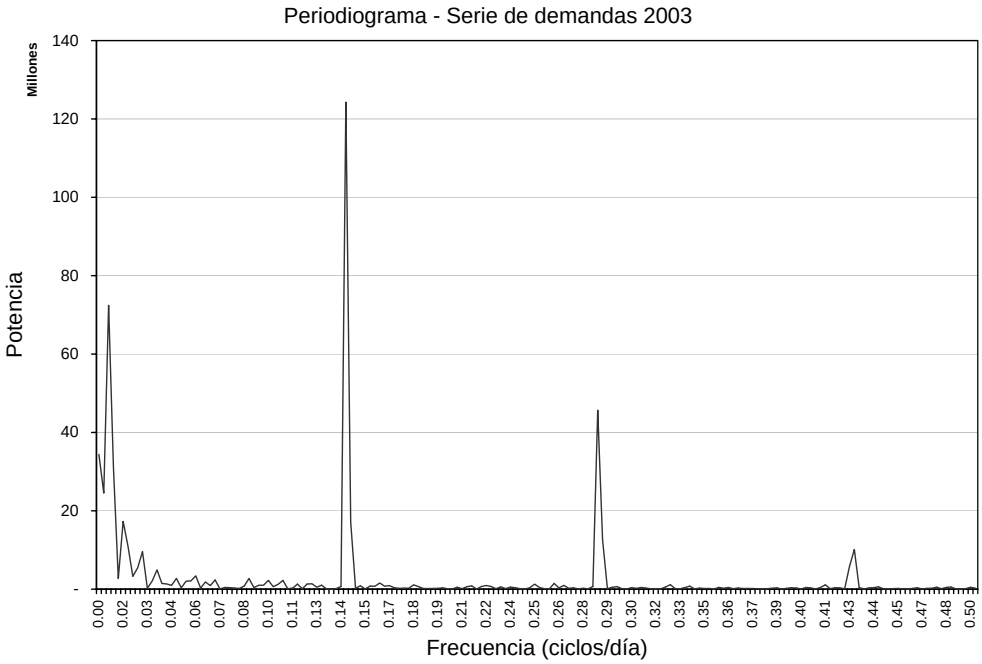


Figura 7.11: Periodograma de frecuencias de la serie de demandas de la ciudad de Valencia, año 2003

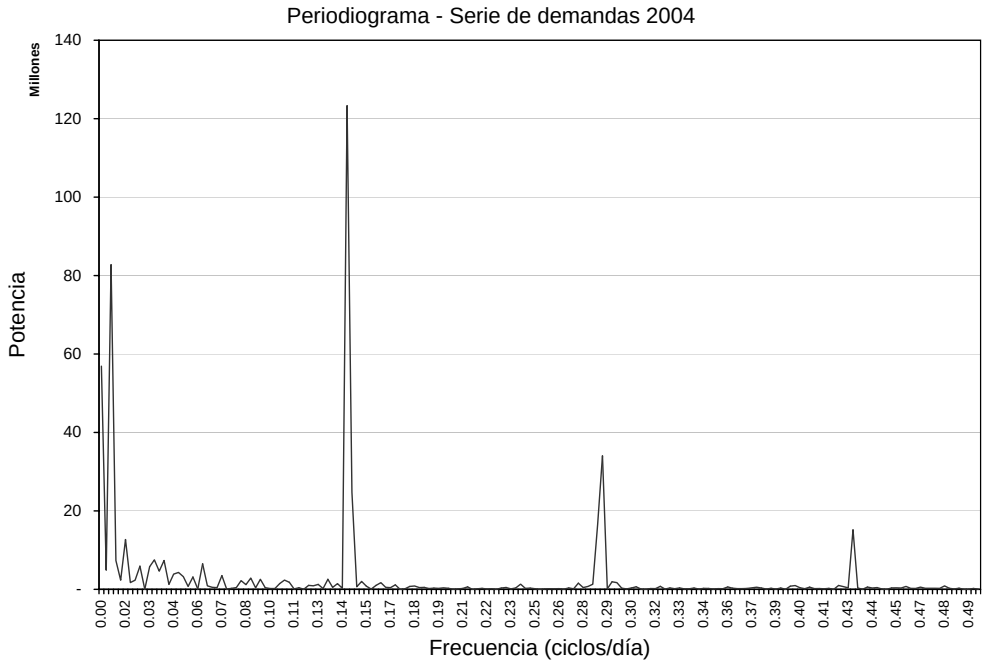


Figura 7.12: Periodograma de frecuencias de la serie de demandas de la ciudad de Valencia, año 2004

7.3.2. *Estadísticos básicos y patrones predominantes de la Temperatura*

Las series de temperaturas para el periodo analizado fueron obtenidas en el Instituto Nacional de Meteorología (INM) y los datos corresponden a la estación meteorológica que se localiza en las instalaciones del INM, en el parque viveros municipales de la ciudad de Valencia.
El cuadro 7.7 resume los estadísticos básicos de la serie de temperaturas completa (temperatura media diaria) y también los de cada uno de los años. Se incluyen los gráficos de la serie de temperatura completa y de los años 2001 al 2004 individualmente (figuras 7.13, 7.14, 7.15, 7.16 y 7.17) para que los patrones puedan ser comparados y su evolución a lo largo de los meses observada.

Estadístico	2001	2002	2003	2004	2001-2004
Número de Datos	365	365	365	366	1461
Valor Medio ($^{\circ}C$)	19.22	18.85	19.11	17.99	18.79
Mediana ($^{\circ}C$)	20.00	18.80	18.00	17.40	18.50
Desv. Est. ($^{\circ}C$)	5.67	4.94	6.27	6.22	5.81
Varianza ($^{\circ}C^2$)	32.25	24.44	39.33	38.69	33.85
Mínimo ($^{\circ}C$)	4.00	9.80	5.40	5.20	4.00
Máximo ($^{\circ}C$)	30.50	28.00	31.90	30.20	31.90
Rango ($^{\circ}C$)	26.50	18.20	26.50	25.00	27.90

Cuadro 7.7: Resumen de estadísticos básicos de la serie de temperaturas medias diarias

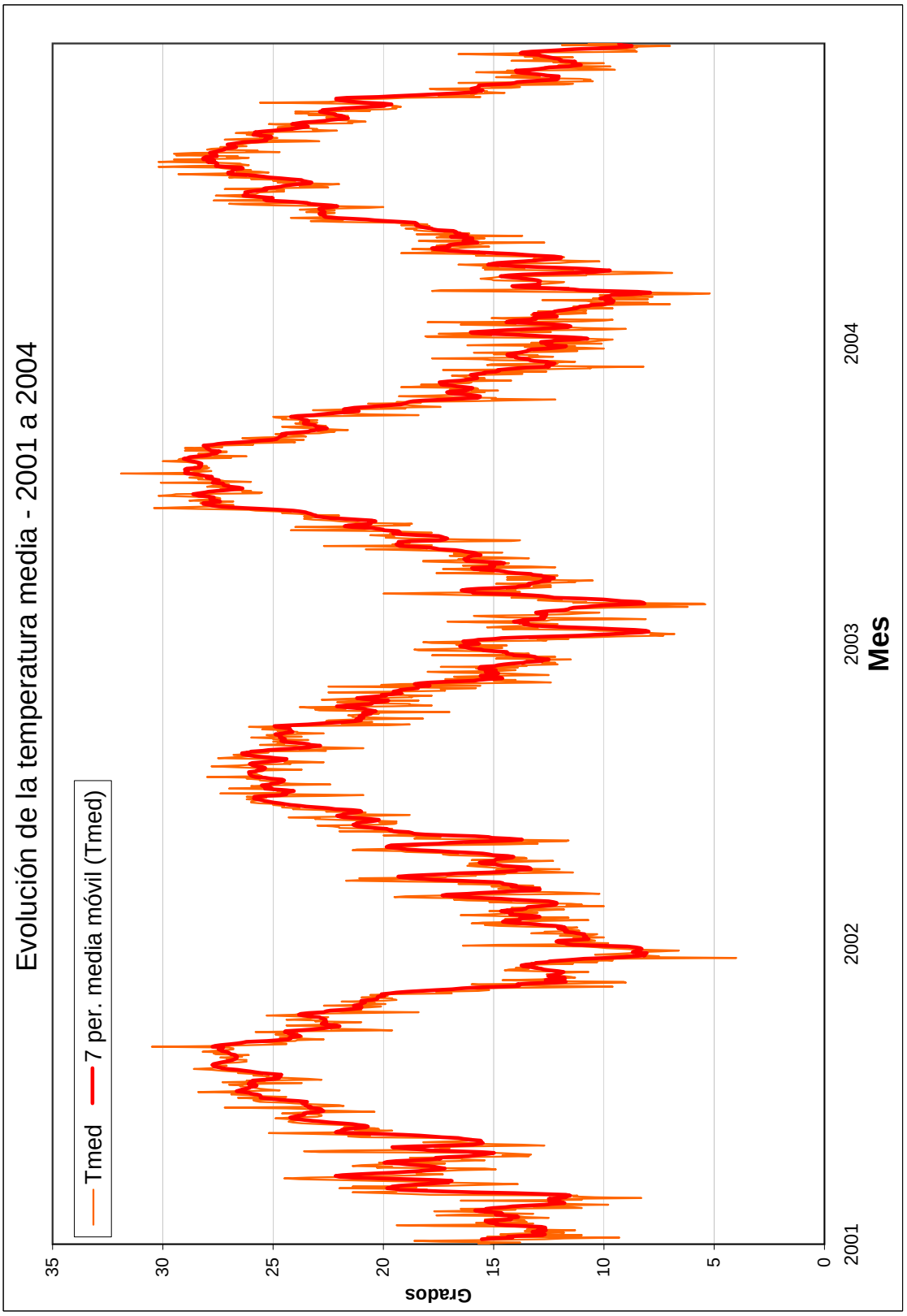


Figura 7.13: Evolución de la temperatura media, en la ciudad de Valencia del 1 de Enero de 2001 al 31 de Diciembre de 2004. Grados Centígrados

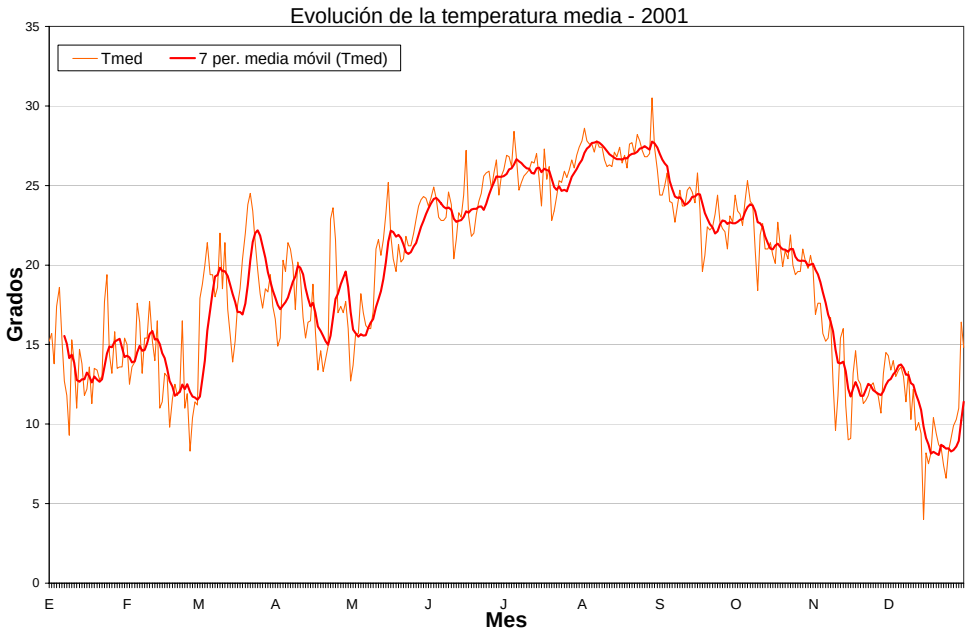


Figura 7.14: Evolución de la temperatura media, en la ciudad de Valencia, año 2001. Grados Centígrados

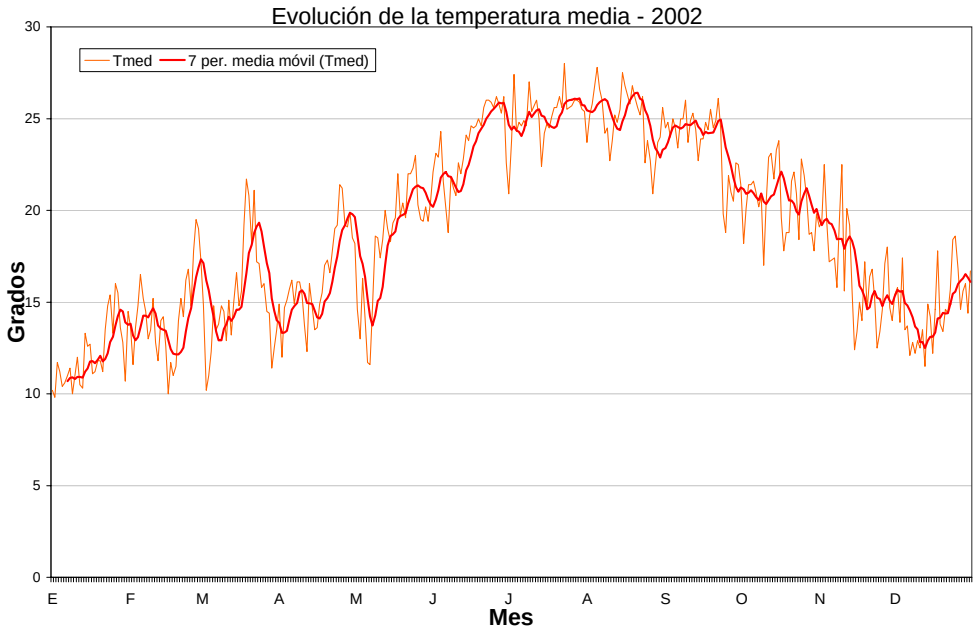


Figura 7.15: Evolución de la temperatura media, en la ciudad de Valencia, año 2002. Grados Centígrados

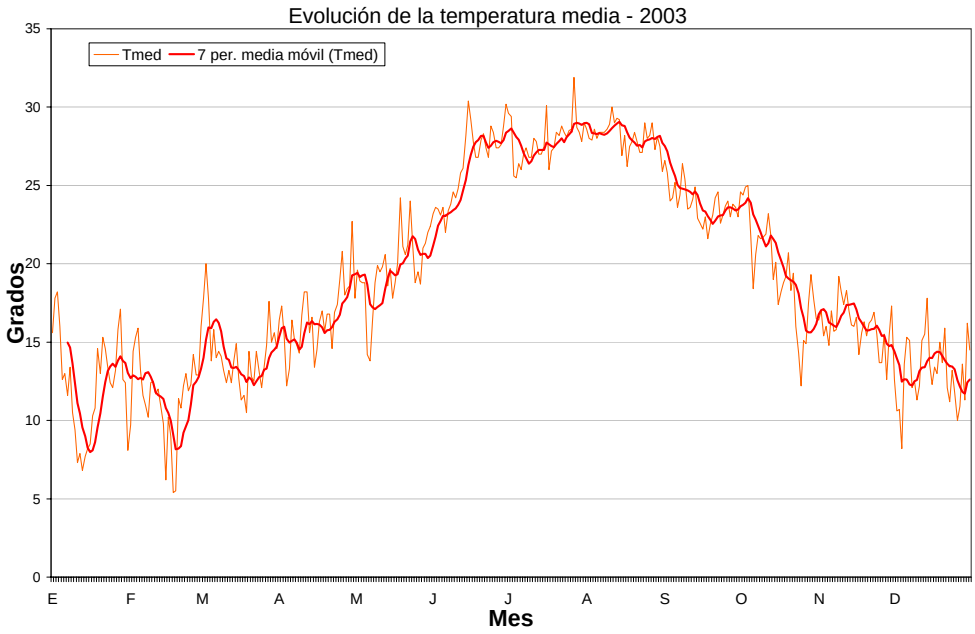


Figura 7.16: Evolución de la temperatura media, en la ciudad de Valencia, año 2003. Grados Centígrados

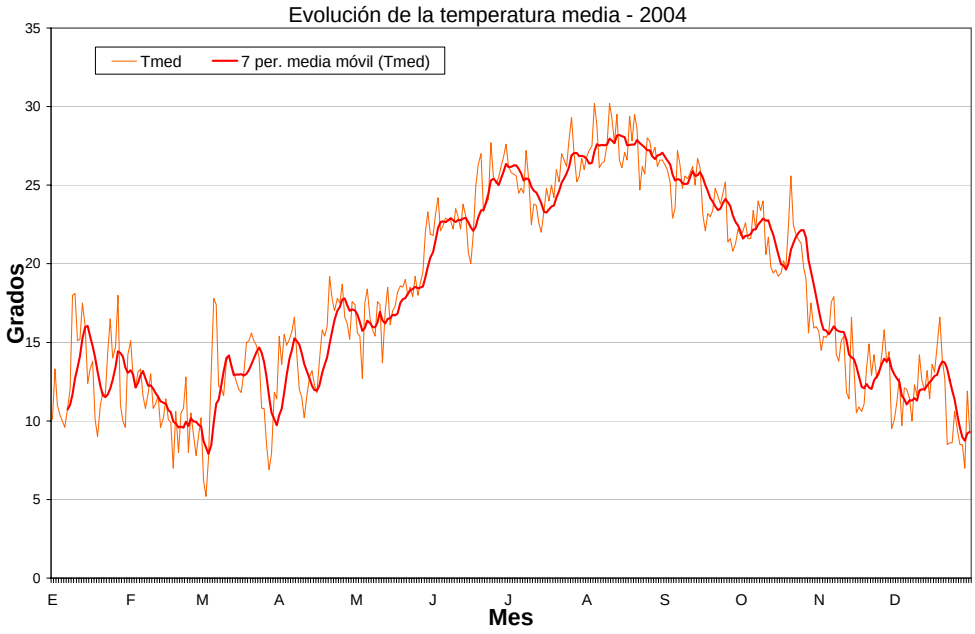


Figura 7.17: Evolución de la temperatura media, en la ciudad de Valencia, año 2004. Grados Centígrados

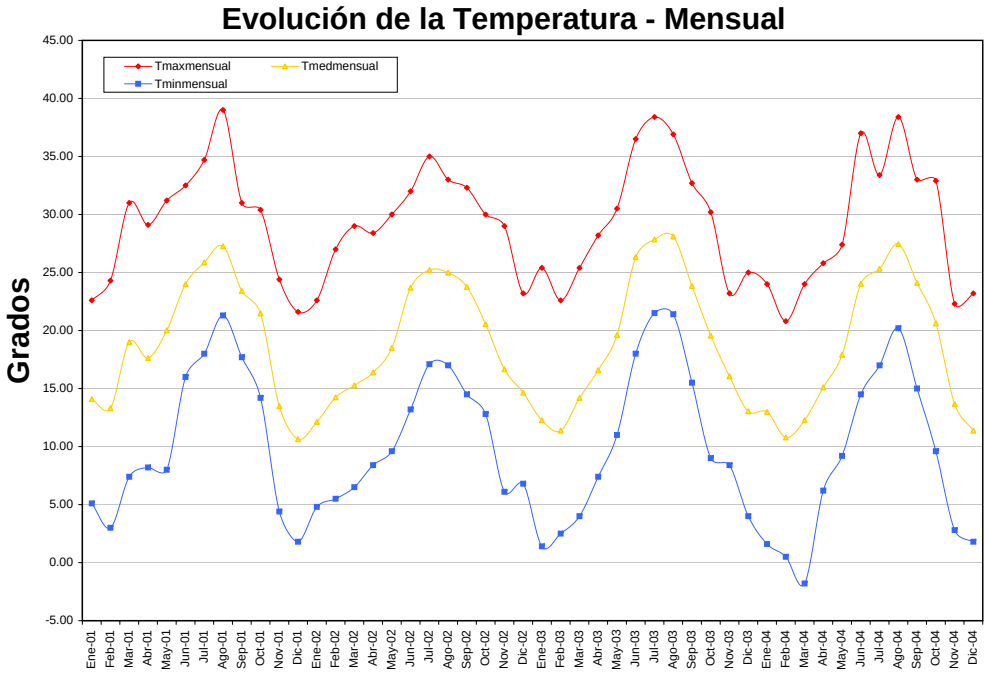


Figura 7.18: Evolución de temperaturas máximas, mínimas y medias mensuales, 2001-2004. Grados Centígrados

Se puede observar del gráfico 7.18 y del cuadro 7.8 que las temperaturas más altas ocurren en los meses de julio y agosto; en cambio las temperaturas mínimas ocurren en los meses de diciembre, enero y febrero, coincidiendo –como es esperable– con las estaciones del verano e invierno respectivamente. La temperatura más alta observada en la serie estudiada ocurrió en el mes de Agosto de 2001, siendo de 39°C y la mínima fue de -1.80°C. Es evidente que el rango de variación de las temperaturas es muy grande. Del histograma de frecuencias de temperaturas medias (figura 7.19), se observa que existen dos grupos de clases de temperaturas que se presentan con más frecuencia. El primer grupo con temperaturas entre los 12°C y 16.5°C y el segundo entre los 22.5°C y los 27°C, abarcando tres clases cada uno de los grupos. Las temperaturas extremas tanto máximas como mínimas ocurren con poca frecuencia y los valores medios, entre los 16.5°C y los 22.5°C tienen una ocurrencia media.

Mes	Temperatura $^{\circ}C$	Año			
		2001	2002	2003	2004
Enero	Máxima	22.60	22.60	25.40	24.00
	Mínima	5.10	4.80	1.40	1.60
	Media	14.90	12.13	12.26	12.98
Febrero	Máxima	24.30	27.00	22.60	20.80
	Mínima	3.00	5.50	2.50	0.50
	Media	13.30	14.24	11.38	10.70
Marzo	Máxima	31.00	29.00	25.40	24.00
	Mínima	7.40	6.50	4.00	-1.80
	Media	18.99	15.27	14.19	12.27
Abril	Máxima	29.10	28.40	28.20	25.80
	Mínima	8.20	8.40	7.40	6.20
	Media	17.64	16.38	16.58	15.10
Mayo	Máxima	31.20	30.00	30.50	27.40
	Mínima	8.00	9.60	11.00	9.20
	Media	20.02	18.50	19.61	17.91
Junio	Máxima	32.50	32.00	36.50	37.00
	Mínima	16.00	13.20	18.00	14.50
	Media	24.00	23.69	26.33	24.02
Julio	Máxima	34.70	35.00	38.40	33.40
	Mínima	18.00	17.10	21.50	17.00
	Media	25.86	25.22	27.85	25.32
Agosto	Máxima	39.00	33.00	36.90	38.40
	Mínima	21.30	17.00	21.40	20.20
	Media	27.26	24.99	28.11	27.45
Septiembre	Máxima	31.00	32.30	32.70	33.00
	Mínima	17.70	14.50	15.50	15.00
	Media	23.40	23.76	23.84	24.11
Octubre	Máxima	30.40	30.00	30.20	32.90
	Mínima	14.20	12.80	9.00	9.60
	Media	21.47	20.54	19.56	20.62
Noviembre	Máxima	24.40	29.00	23.20	22.30
	Mínima	4.40	6.10	8.40	2.80
	Media	13.48	16.66	16.05	13.67
Diciembre	Máxima	21.60	23.20	25.00	23.20
	Mínima	1.80	6.80	4.00	1.80
	Media	10.64	14.64	13.04	11.37

Cuadro 7.8: Resumen de temperaturas máximas, mínimas y medias mensuales, 2001-2004

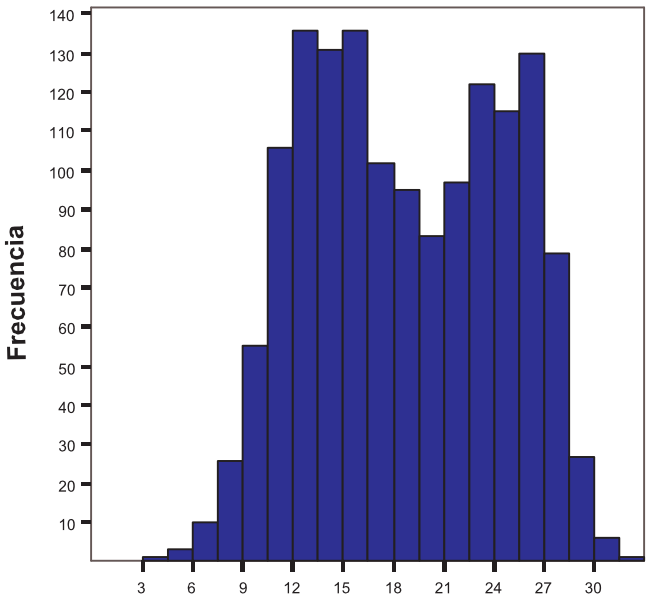


Figura 7.19: Histograma de Frecuencias de la temperatura media, en la ciudad de Valencia, 2001-2004

7.3.3. *Análisis conjunto Demandas-Temperatura*

En secciones anteriores se ha hecho un análisis por separado del comportamiento de la demanda diaria (sección 7.3.1) y de la temperatura media, máxima y mínima registrada en Valencia (sección 7.3.2). Con el fin de conocer si existe una relación del comportamiento de la demanda de agua potable que pueda ser explicado con la temperatura se ha hecho un análisis conjunto de las series. Los resultados se presentan en las secciones siguientes. Maidment et~al. (1985) utilizaron la temperatura y la precipitación como variables independientes para construir un modelo para predecir la demanda de la ciudad de Austin, Texas.

Variaciones de la demanda diaria respecto a la media anual

Como una herramienta más de comparación se han calculado las desviaciones diarias de la demanda con respecto a la media anual y han sido graficadas contra las temperaturas medias diarias registradas. La intención de este tipo de análisis es observar el comportamiento de las desviaciones de la demanda diaria conforme varía la temperatura media diaria. Análisis similares se pueden encontrar en Protopapas et~al. (2000).
Los resultados se pueden observar en los gráficos 7.20, 7.21, 7.22, 7.23 para los años 2001, 2002, 2003 y 2004 respectivamente. Cada gráfico consiste en tres secciones, la parte A) corresponde al graficado de las desviaciones, la B) corresponde al graficado de las desviaciones pero se han clasificado con colores los días festivos, "*días puente*", semana santa y vacaciones de agosto y finalmente, en el C) se han removido los días que han sido clasificados y que aportan mayor variación.

En la parte C de los gráficos se han removido los días festivos y vacaciones, eliminando gráficamente la variabilidad que aportan al comportamiento de la demanda. Estas variaciones son explicados por patrones sociológicos del conjunto de los habitantes de la ciudad y no por el patrón típico de demandas. Una vez removidos esos valores es más fácil apreciar si existe una relación o una tendencia en el comportamiento de la temperatura-demanda. Como punto a resaltar de los resultados de la observación de los gráficos, se tiene que: a lo largo de los 4 años de serie con que contamos, se presenta una tendencia de la demanda diaria a permanecer en torno a la demanda media anual de cada uno de los años cuando la temperatura es inferior a 18^{o}C. En cambio, cuando este umbral de temperatura es superado, la demanda diaria tiende a superar el valor medio anual. Esto nos indica la existencia de dos zonas en la relación demandas-temperaturas, una donde

la demanda es prácticamente insensible a la variación de la temperatura y una segunda zona, a partir de los 18^{o}C, donde la demanda es sensible al aumento de la temperatura. Más adelante se evaluará la magnitud de esta relación, ya que por ahora solo la hemos identificado gráficamente. Conocer estos datos, nos será de utilidad para decidir sobre la inclusión o no de la temperatura en algún modelo, así como de la forma en que su inclusión nos podría aportar los mejores resultados a la hora de predecir.

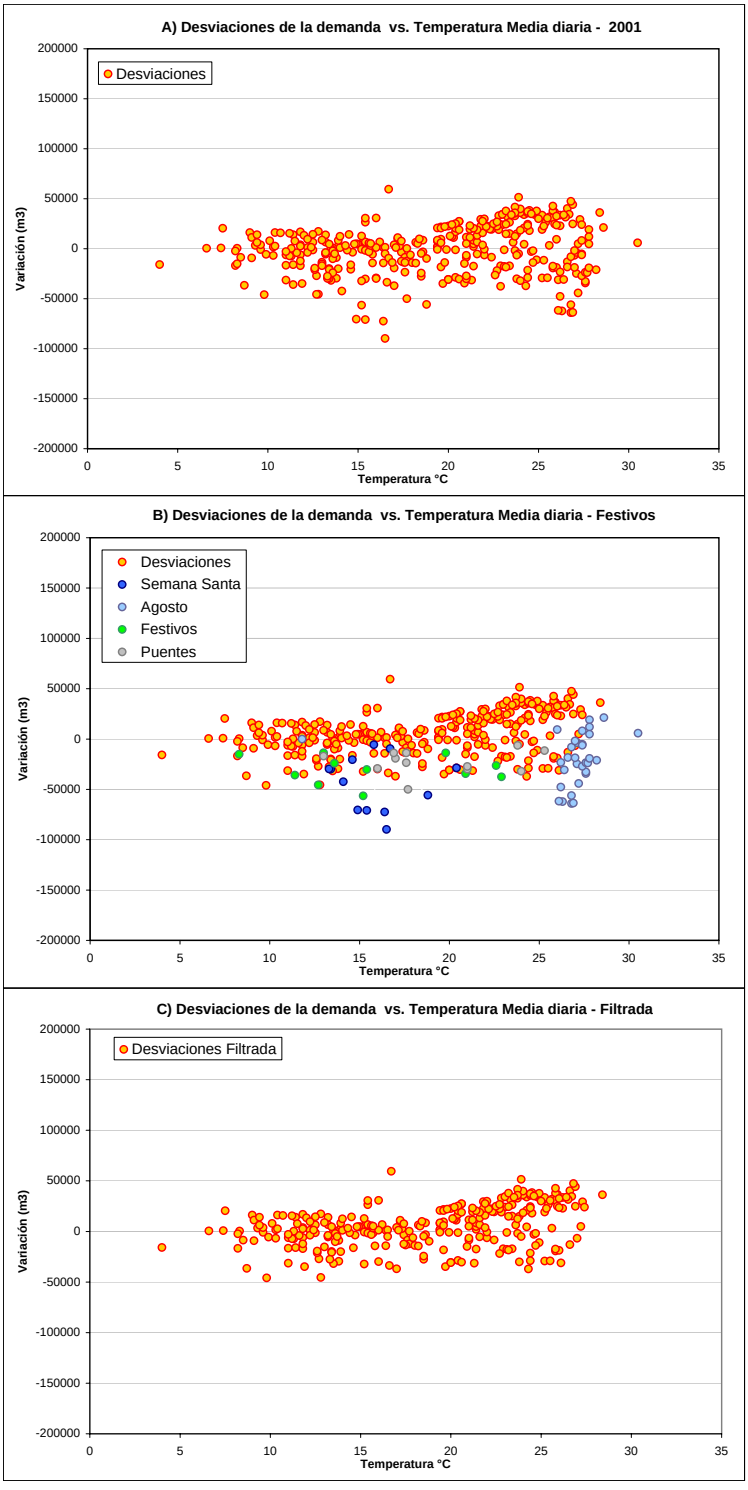


Figura 7.20: A) Gráfico de desviaciones de la demanda con respecto a la demanda media anual vs. Temperatura media diaria, B) Festivos clasificados, C) Serie filtrada, Año 2001

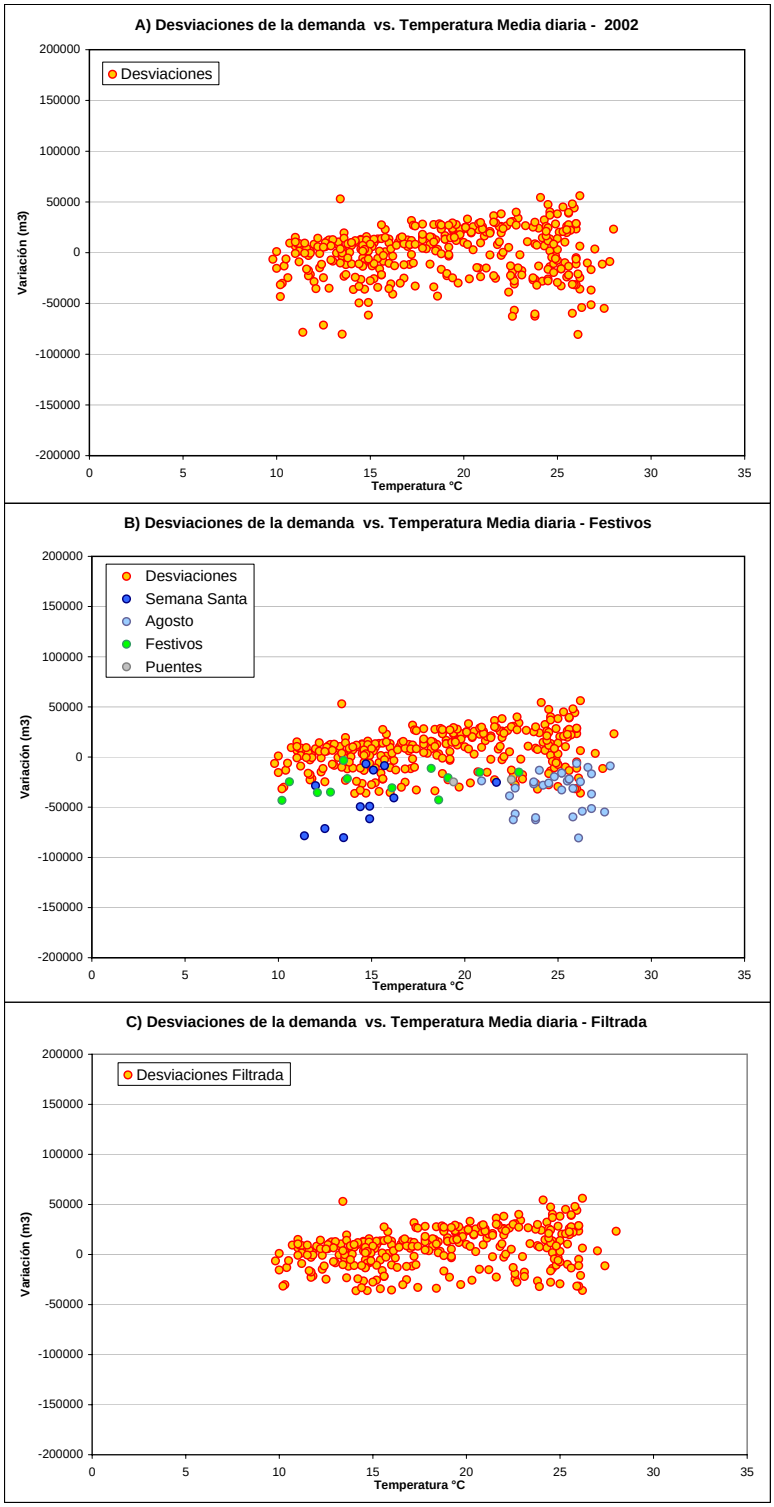


Figura 7.21: A) Gráfico de desviaciones de la demanda con respecto a la demanda media anual vs. Temperatura media diaria, B) Festivos clasificados, C) Serie filtrada, Año 2002

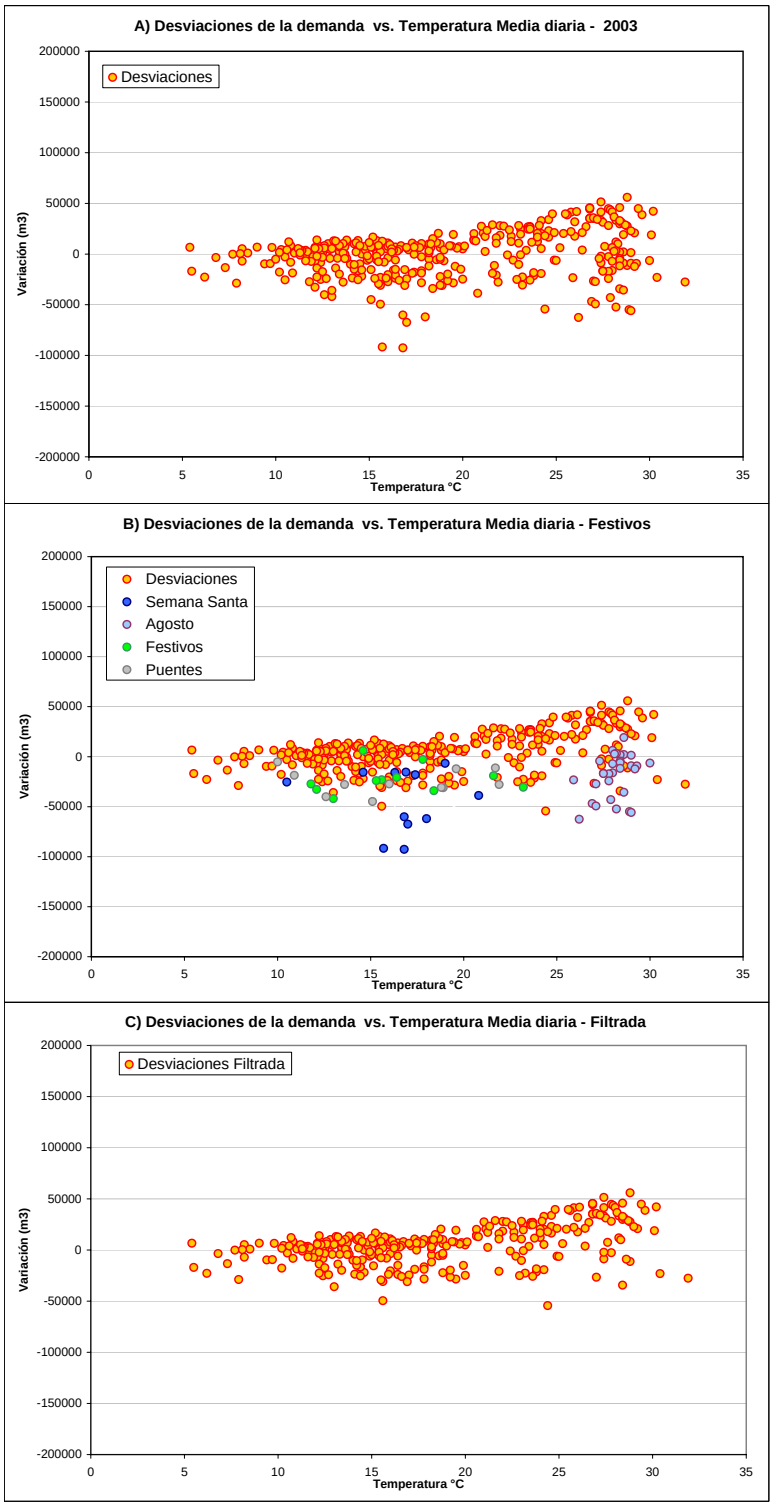


Figura 7.22: A) Gráfico de desviaciones de la demanda con respecto a la demanda media anual vs. Temperatura media diaria, B) Festivos clasificados, C) Serie filtrada, Año 2003

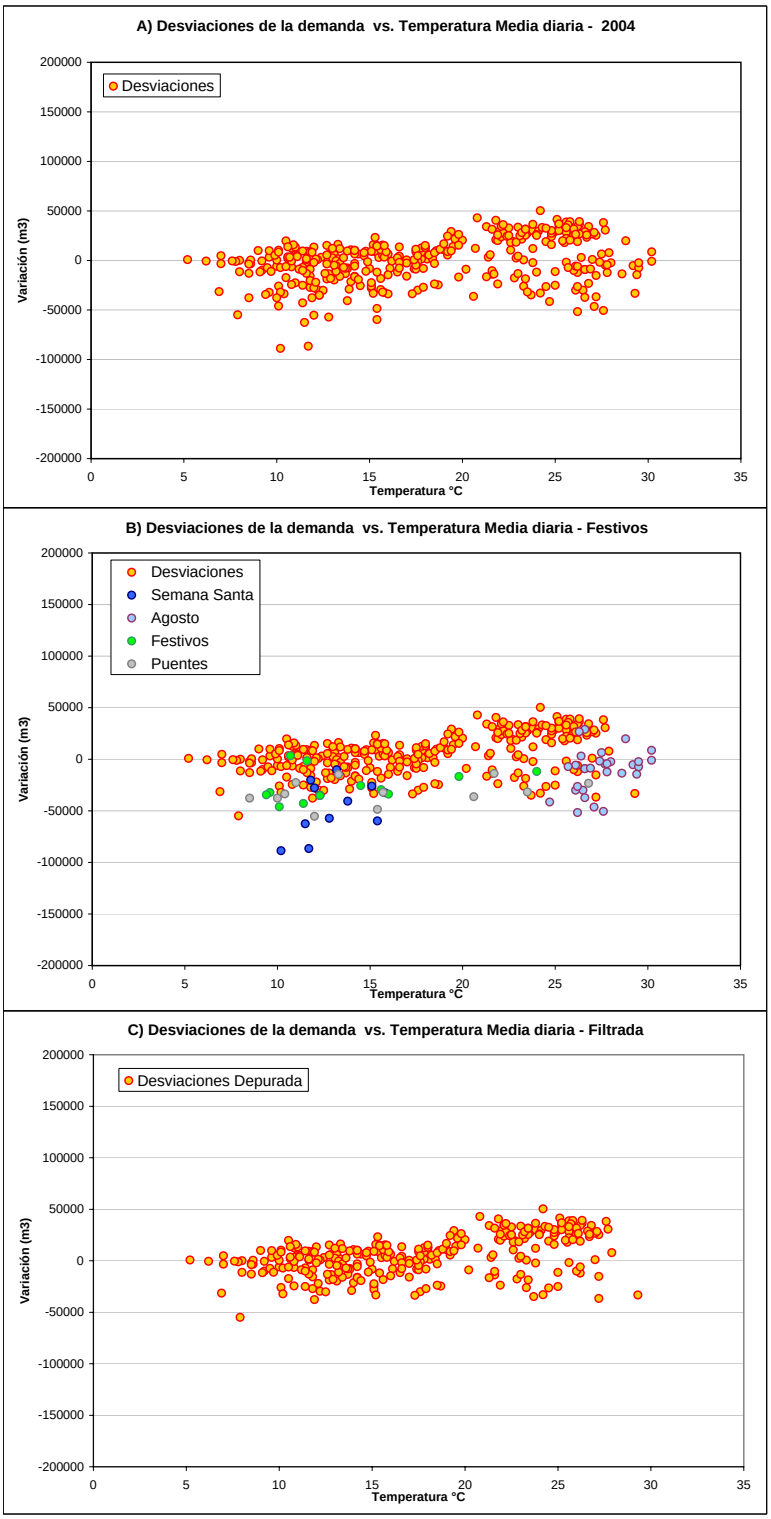


Figura 7.23: A) Gráfico de desviaciones de la demanda con respecto a la demanda media anual vs. Temperatura media diaria, B) Festivos clasificados, C) Serie filtrada, Año 2004

Distribución conjunta de la Temperatura y la demanda diaria

En esta sección se ha realizado análogamente un análisis de la distribución conjunta de la temperatura y de la demanda diaria. Se ha clasificado la demanda de agua en 18 intervalos de 10,000 m^3 y la temperatura en 20 intervalos de 1.5° C. Se hizo un recuento del número de ocurrencias en las series para cada par de valores de demanda (D_i) y temperatura (T_j). Los resultados se presentan en el cuadro 7.9. La moda en la demanda diaria es el intervalo de los 333,000 a 343,000 m^3 (259 ocurrencias) y en la temperatura el intervalo de los 12.01 a 13.50°C (138 ocurrencias). Sin embargo en los intervalos de temperaturas más cálidas el rango de los 25.51 a 27.00°C también ocurrió con mucha frecuencia (130 ocurrencias). Los estadísticos condicionados estimados para los diferentes intervalos de temperatura y para la demanda diaria se presentan en el cuadro 7.10 donde se han calculado los valores máximos, mínimos y medios para cada intervalo de temperatura de 1.5° C, así como también la desviación estándar y el coeficiente de variación. Los resultados han sido reflejados también en el gráfico 7.24. De este se comprueba que al igual que en la sección anterior, en donde las series fueron analizadas año por año, la demanda media diaria para cada intervalo de temperatura tiende a ser superior a la media a partir de los 18°C, aunque en este caso la media considerada es la global (temperatura media de los 4 años). La demanda máxima en cambio, tiene un crecimiento prácticamente lineal con el crecimiento de la temperatura. La demanda mínima presenta un comportamiento irregular. Se presentan valores bajos de la demanda mínima tanto a temperaturas bajas como en temperaturas correspondientes a los meses más calurosos, por lo que la temperatura no parece influenciar su comportamiento. Por otra parte en este gráfico se aprecia la magnitud de las desviaciones que se han reportado en el cuadro 7.10. Las desviaciones son mayores en el intervalo de temperaturas medias de 25.51 a 27° C.

Temp. Media (C°)	Demanda diaria (miles de m^3)																	
	213	223	233	243	253	263	273	283	293	303	313	323	333	343	353	363	373	383	$\sum$
3.01-4.50	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
4.51-6.00	0	0	0	0	0	0	0	0	0	0	1	0	2	0	0	0	0	0	3
6.01-7.50	0	0	0	0	0	0	0	0	0	4	1	3	2	0	0	0	0	0	10
7.51-9.00	0	0	0	0	0	1	1	2	3	2	1	8	7	1	0	0	0	0	26
9.01-10.50	0	0	0	1	1	1	2	1	8	12	8	11	6	3	1	0	0	0	55
10.51-12.00	0	1	0	1	0	1	5	6	10	17	16	17	21	11	0	0	0	0	106
12.01-13.50	0	1	3	0	2	0	8	9	10	21	30	19	26	8	0	1	0	0	138
13.51-15.00	0	0	1	1	2	1	4	9	11	27	27	24	16	8	0	0	0	0	131
15.01-16.50	1	0	3	1	0	1	9	5	14	28	18	15	27	13	1	0	0	0	136
16.51-18.00	0	0	1	0	1	4	3	3	15	12	14	14	25	9	0	1	0	0	102
18.01-19.50	0	0	0	1	0	1	3	3	8	14	11	18	23	11	1	1	0	0	95
19.51-21.00	0	0	0	0	0	0	9	4	7	3	15	14	19	7	4	0	1	0	83
21.01-22.50	0	0	0	0	0	1	1	4	5	11	11	20	14	6	12	11	1	0	97
22.51-24.00	0	0	0	3	1	1	5	12	7	10	11	9	23	16	17	7	0	0	122
24.01-25.50	0	0	0	0	0	1	6	6	10	12	6	17	14	17	8	16	1	1	115
25.51-27.00	0	1	0	6	3	1	7	12	8	10	3	12	20	9	15	17	6	0	130
27.01-28.50	0	0	0	0	1	1	5	8	6	5	9	12	11	2	5	9	5	0	79
28.51-30.00	0	0	0	0	0	0	2	0	2	0	7	5	2	2	4	1	1	1	27
30.01-31.50	0	0	0	0	0	0	0	0	0	1	1	0	1	2	0	0	1	0	6
31.51- +	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
$\sum$	**1**	**3**	**8**	**14**	**11**	**15**	**70**	**85**	**124**	**190**	**190**	**218**	**259**	**125**	**68**	**64**	**16**	**2**	

Cuadro 7.9: Número de ocurrencias para la demanda diaria y la temperatura media

Temperatura °C	Número Ocurrencias	Valor Mínimo	Valor Máximo	Valor Medio	Desviación Estándar	Coeficiente de Variación
3.01-4.50	1	292,117.70	292,117.70	292,117.70	0.00	0.00
4.51-6.00	3	314,026.40	337,628.90	329,471.27	13,382.73	0.04
6.01-7.50	10	304,567.90	340,676.00	321,223.65	13,253.29	0.04
7.51-9.00	26	271,403.80	345,914.90	319,282.51	20,169.04	0.06
9.01-10.50	55	247,273.80	355,736.00	314,330.23	21,388.60	0.07
10.51-12.00	106	232,734.90	351,809.30	317,742.36	22,081.79	0.07
12.01-13.50	138	230,834.20	364,203.20	316,100.90	22,451.83	0.07
13.51-15.00	131	237,490.60	346,733.60	314,822.24	20,922.33	0.07
15.01-16.50	136	218,212.30	359,167.70	314,683.19	25,328.74	0.08
16.51-18.00	102	238,398.10	367,439.60	317,701.22	23,630.30	0.07
18.01-19.50	95	252,231.30	365,166.20	322,486.77	20,939.46	0.06
19.51-21.00	83	273,149.40	378,892.60	321,250.72	23,542.16	0.07
21.01-22.50	97	272,404.00	376,508.30	332,642.01	23,813.64	0.07
22.51-24.00	122	248,570.00	372,346.00	325,816.02	29,254.71	0.09
24.01-25.50	115	270,832.80	386,276.40	330,067.54	27,913.73	0.08
25.51-27.00	130	230,605.50	376,517.00	325,559.05	37,010.15	0.11
27.01-28.50	79	256,357.20	382,364.20	325,643.11	31,162.81	0.10
28.51-30.00	27	275,174.00	386,867.00	332,285.33	27,076.89	0.08
30.01-31.50	6	307,974.50	373,142.50	337,438.38	24,137.09	0.07
31.51- +	1	303,438.90	303,438.90	303,438.90	0.00	0.00

Cuadro 7.10: Estadísticos condicionados de la demanda con temperatura

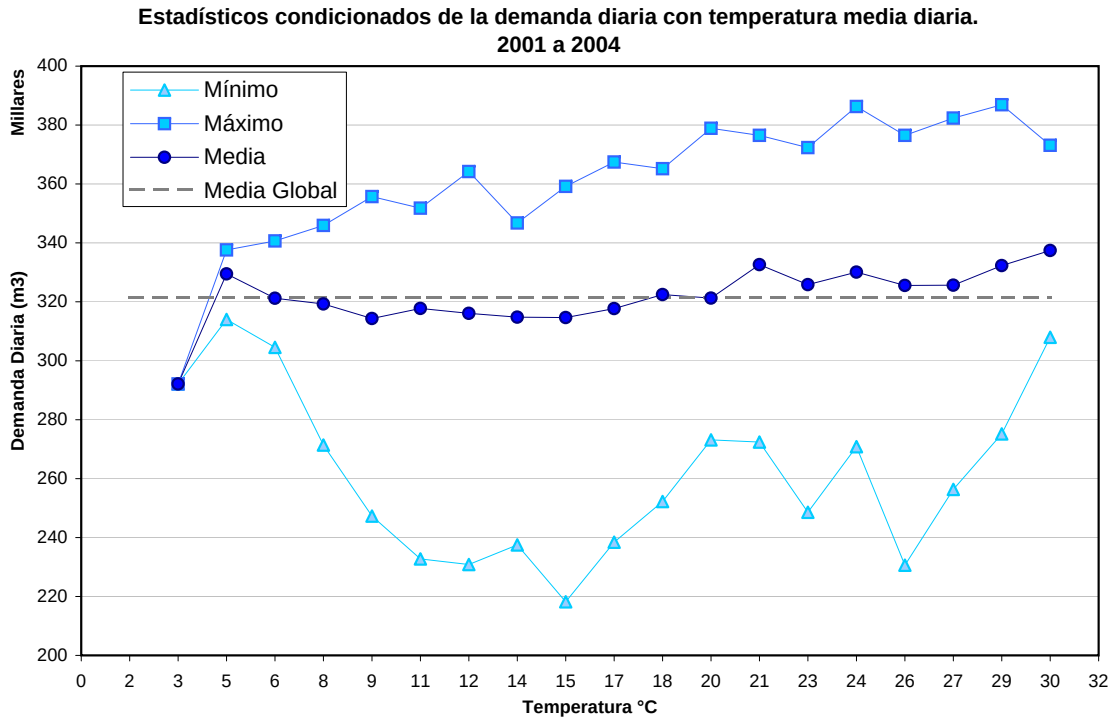


Figura 7.24: Gráfico de estadísticos condicionados de demanda diaria y temperatura

Correlación lineal Demanda diaria - Temperatura

Con el fin de determinar si existe una relación lineal entre el comportamiento de la serie de demandas diaria y la serie de temperaturas, se calculó el coeficiente de correlación lineal de Pearson entre ambas series. Se pueden ver análisis similares en Bougadis et~al. (2005). Como un primer paso se condicionó la demanda diaria con la temperatura en el tiempo *(t)*, después para la temperatura del día anterior *(t-1)* para investigar si impacta en la demanda del día siguiente, la temperatura máxima eliminando el mes de agosto para el tiempo *(t)* (por seguir evidentemente un comportamiento atípico en esta relación), la temperatura media para el tiempo *(t)* y finalmente se calculó la correlación (autocorrelación) de la demanda *(t)* con la demanda *(t-1)* para observar si la demanda del día anterior impacta la del día siguiente. Los resultados se observan en el cuadro 7.11.

Basándose en los resultados obtenidos, se puede ver que la demanda diaria está fuertemente correlacionada con la demanda del día anterior con una coeficiente de 0.6686. En cambio la temperatura media diaria del día actual como la del día anterior obtuvieron coeficientes de correlación muy bajos, 0.1834 y 0.1722, indicando que existe una correlación lineal muy débil entre la temperatura y la demanda diaria en la ciudad de Valencia. La temperatura máxima así como la media, obtienen valores de correlación de 0.3489 y 0.3325 respectivamente cuando no se toman en consideración los valores de los meses de agosto. Es de esperar obtener correlaciones más altas, si se calcula la correlación demanda-temperatura sin considerar los valores cuando el patrón de demandas semanal es atípico.

La débil correlación existente entre temperatura y demanda puede ser explicada si se observa el patrón semanal de demandas diarias (gráfico 7.25), donde se tiene un comportamiento cíclico con valores aproximadamente iguales de lunes a viernes y con valores de demandas muy bajos para el sábado y sobre todo para el domingo. Obviamente la temperatura no sigue ese patrón. La temperatura oscila en una escala más larga (estaciones, meses) como se puede observar en los gráficos 7.26 y 7.27 que representa la evolución de la demanda y la temperatura a lo largo de los 4 años de la serie y del año 2001 respectivamente. Se observa que aunque existen intervalos de tiempo donde se presenta una evolución similar de la temperatura y la demanda, en otros la relación es prácticamente inversa, por ejemplo el mes de agosto, donde a pesar de ser uno de los meses más calurosos y que se esperaría que la demanda fuera máxima, ésta presenta valores muy por debajo de la media. La influencia de los patrones sociológicos (vacaciones y festivos) de la población de Valencia impactan fuertemente en la demanda. En ciudades típicas estadounidenses y australianas, la temperatura y la demanda presentan una correlación más importante ya que los patrones sociológicos

Variable	Correlación con la Demanda (t)
Temperatura med (t)	0.1834
Temperatura med (t-1)	0.1722
Temperatura max (t) sin agostos	0.3489
Temperatura med (t) sin agostos	0.3325
Demanda (t-1)	0.6686

Cuadro 7.11: Coeficiente de correlación entre la demanda diaria y la temperatura, Serie 2001 a 2004

no son tan marcados como en la ciudad de Valencia y en España en general. Algunos de los artículos que plantean sus modelos de predicción basados en la relación directa de la temperatura con la demanda, y que consideran una demanda mínima de base, como vimos son entre otros Maidment and Parzen (1984b), Maidment and Miaou (1986), Franklin and Maidment (1986), Zhou et~al. (2000), Zhou et~al. (2002).

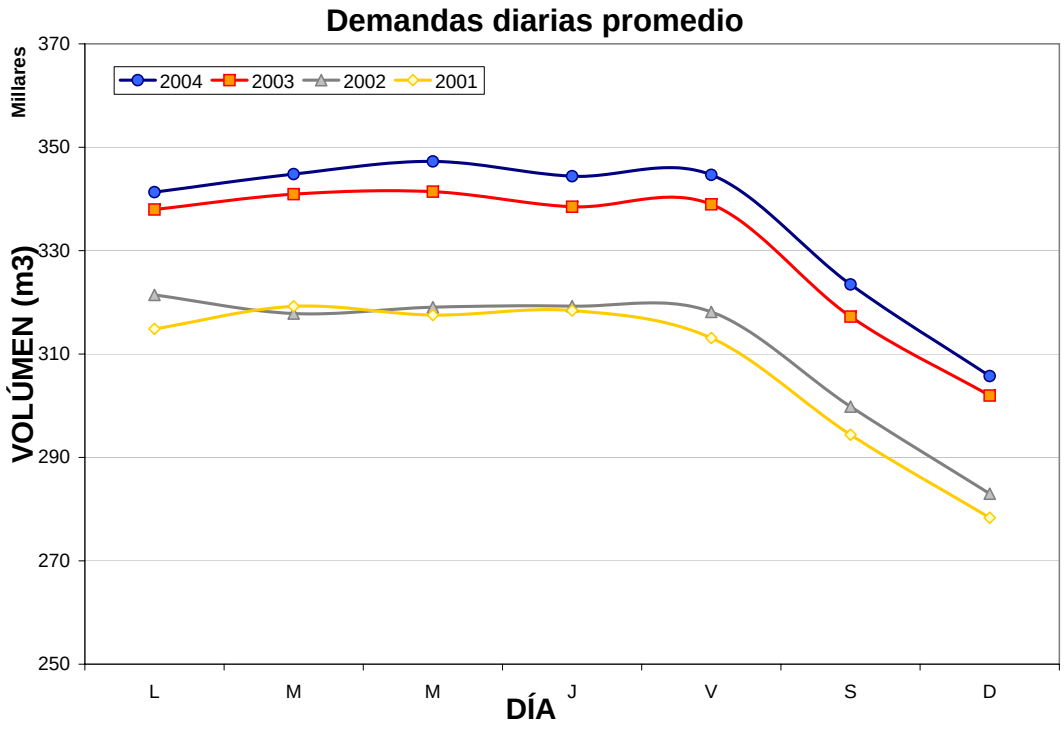


Figura 7.25: Patrón de demandas semanal, Años 2001 al 2004

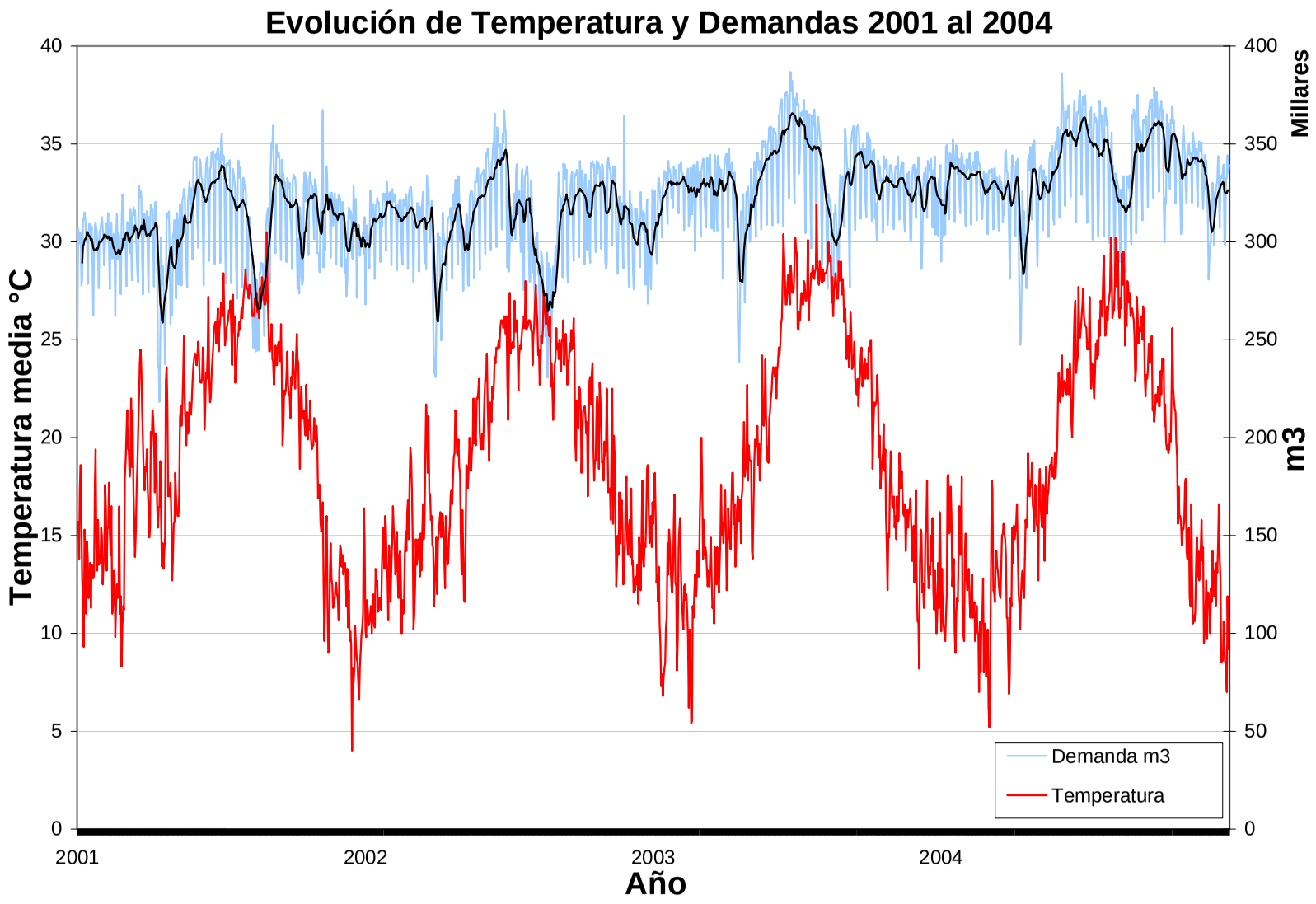


Figura 7.26: Evolución de la temperatura media y de la demanda diaria, Años 2001 al 2004

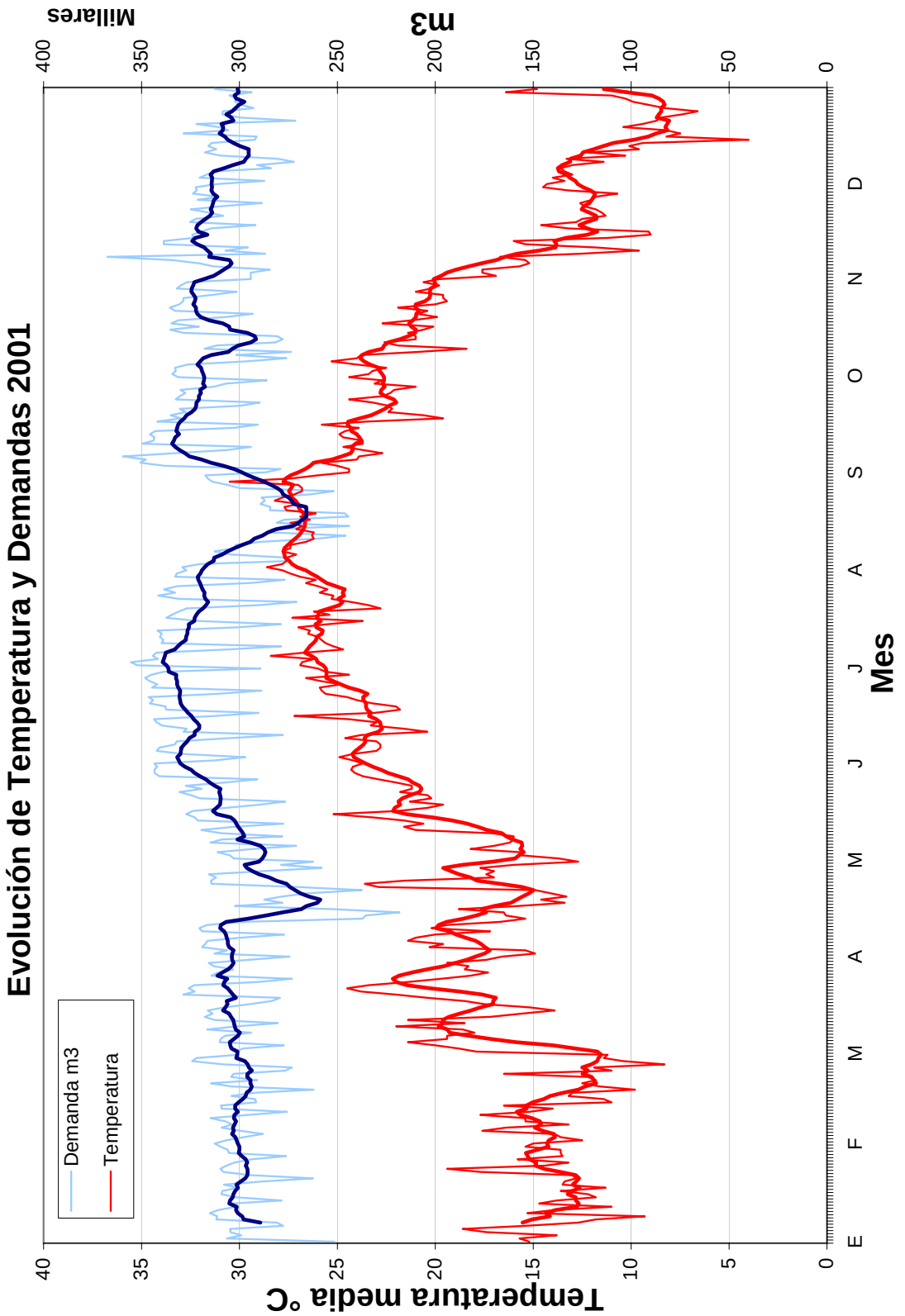


Figura 7.27: Evolución de la temperatura media y de la demanda diaria, Año 2001

7.4. Identificación del modelo de predicción

Esta sección se dedicará a la determinación del modelo del tipo Box-Jenkins más adecuado para describir y predecir la demanda de agua potable en la ciudad de Valencia y área metropolitana. Las herramientas básicas para este fin son la observación del correlograma o ACF, y el correlograma parcial o PACF. De ellos podemos obtener información del número de términos (*p, d, q, P, D, Q*) necesarios para construir el modelo. En la sección 7.3.3 se hizo un análisis de la evolución conjunta de la temperatura y demanda diaria. En base a los resultados gráficos obtenidos se demuestra que la temperatura incide en la variación de la demanda. Sin embargo la correlación lineal obtenida es baja. Ante esta indefinición se deberá evaluar la conveniencia de incluir la temperatura (máxima, mínima o media) en el modelo como variable *exógena* y evaluar su aportación real en la mejora de la capacidad predictiva del modelo.

7.4.1. *Selección del modelo de series temporales*

Determinación del orden de diferenciación

Para identificar el modelo ARIMA apropiado para una serie temporal, se debe iniciar por identificar el orden de diferenciación necesario para hacer la serie estacionaria, si no lo fuera, (Box et~al., 1976) e incluso una transformación de la serie para estabilizar la varianza si fuera necesario. En el caso que estamos analizando, ya hemos observado en la sección 7.3.1 que nuestra serie no es estacionaria en la media por lo que habremos de aplicar algún grado de diferenciación no estacional y/ó estacional hasta lograr que lo sea. El patrón que se observa en el correlograma es el típico de puente suspendido (suspension bridge) característico de las series no estacionarias y con estacionalidad fuerte, por lo cual algún grado de diferenciación estacional también será necesario. Una forma de determinar el grado óptimo de diferenciación, además de observar el correlograma y la apariencia de la serie diferenciada, es comúnmente el grado de diferenciación con el cual la desviación estándar de la serie es la mínima. En nuestro caso calcularemos el RMSE (Root Mean Square Error) de la serie para cada grado de diferenciación.

El *software Statgraphics Plus* nos ofrece la posibilidad de ajustar modelos ARMA o ARIMA de modo que podemos hacer varias pruebas antes de seleccionar el grado de diferenciación óptimo. En este caso iremos ajustando

Modelo	RMSE
ARMA (0,0,0)cc	26414.8
ARMA (0,1,0)cc	21488.2
ARIMA (0,0,0)x(0,1,0)cc	18241.2
ARIMA (0,1,0)x(0,1,0)	16827.6

Cuadro 7.12: Valor del RMSE para cada grado de diferencias regulares o estacionales

modelos solamente con orden *d* y/ó *D* y observaremos los residuos y el correlograma de los residuos sin darle importancia al resto de los parámetros. El cuadro 7.12 presenta los resultados del RMSE de 4 modelos. El modelo sin grado de diferencia de ningún tipo presenta el RMSE más grande. Al incorporar una diferencia no estacional el RMSE se reduce, y si realizamos solo una diferencia estacional se reduce aún más. Aplicando una diferencia estacional y una no estacional, el valor del RMSE es el menor de los cuatro. Normalmente no se recomienda hacer más de dos diferencias en total (estacionales o no estacionales) por lo que nos limitaremos a estos modelos. Tampoco se ha probado con los modelos ARMA (0,2,0) ya que este no eliminaría la fuerte estacionalidad de la serie, ni tampoco el modelo ARIMA(0,0,0)x(0,2,0) ya que este no nos produciría una serie estacionaria.

A continuación se presentan los resultados gráficos de los modelos ARMA (0,1,0)cc y ARIMA (0,1,0)x(0,1,0). Para el primer caso (gráfico 7.28), la serie tiene una apariencia más o menos estacionaria, pero en el autocorrelograma de los residuos (gráfico 7.29) aún hay una fuerte autocorrelación en el periodo estacional, la constante representa el valor medio de la serie diferenciada, por lo que los residuos son las desviaciones con respecto a la media. El periodo estacional es $s = 7$ días (obtenido en la sección 7.3.1). Se observan fuertes puntas de autocorrelación en los periodos $s, 2s, 3s, 4s,$ etc. La misma situación se observa en el autocorrelograma parcial (gráfico 7.30). Por este motivo, es recomendable aplicar también un orden de diferenciación estacional.

Los resultados de aplicar una diferencia no estacional y una estacional se observan en los gráficos de autocorrelación (7.31) y autocorrelación parcial de los residuos (7.32). De la observación de los gráficos de ACF, PACF y del cuadro de RMSE de los modelos analizados hemos concluido que el modelo ARIMA (0,1,0)x(0,1,0)7, es decir el modelo con una diferencia estacional y una no estacional, es el que minimiza el RMSE. Sin embargo los gráficos nos indican también que es necesario la incorporación de términos AR y/ó MA porque los valores de autocorrelación se han vuelto negativos y no han desaparecido del todo.

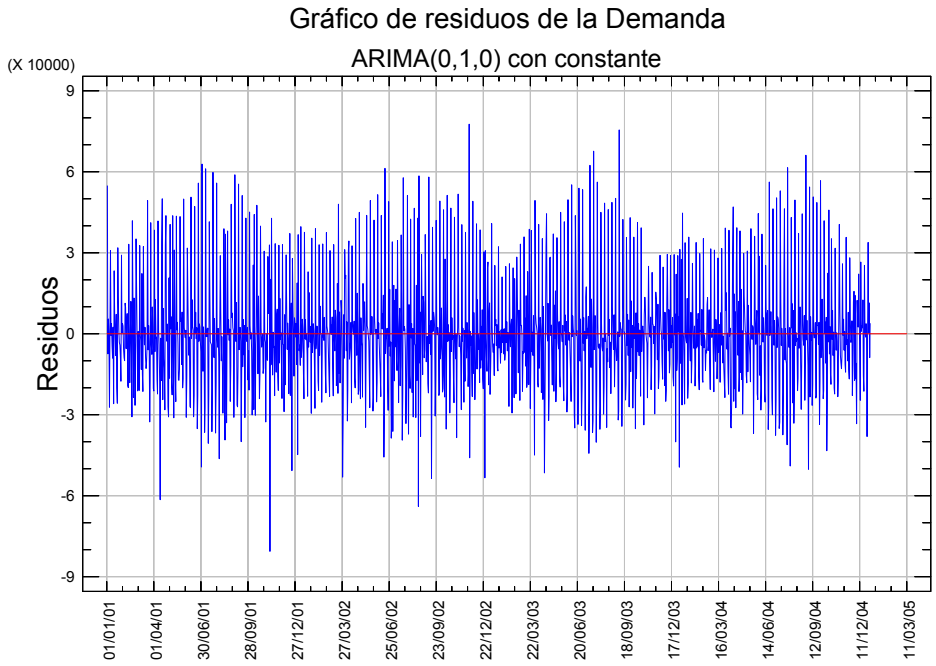


Figura 7.28: Residuos del modelo ARIMA(0,1,0) con constante

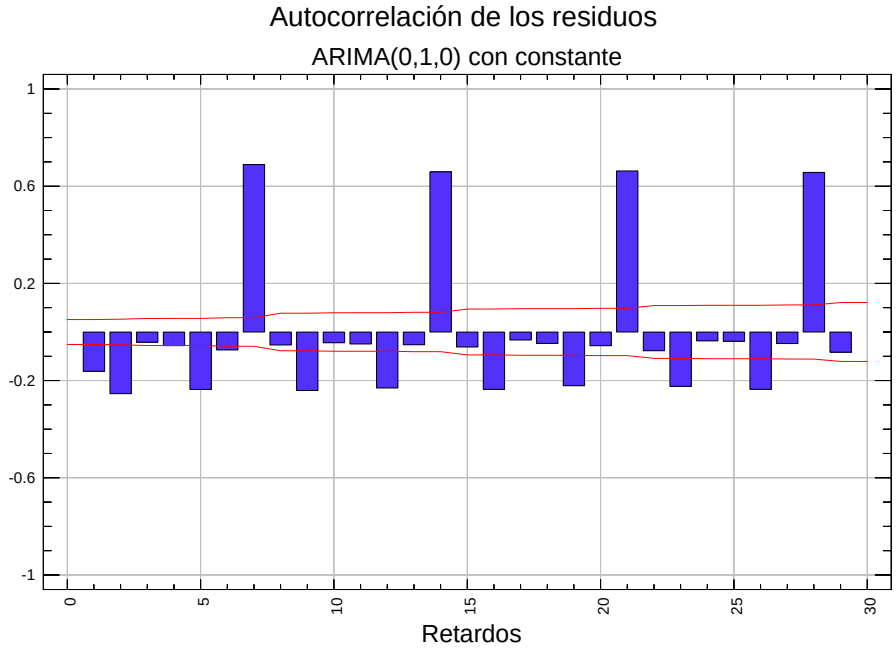


Figura 7.29: ACF de los residuos del modelo ARIMA(0,1,0) con constante

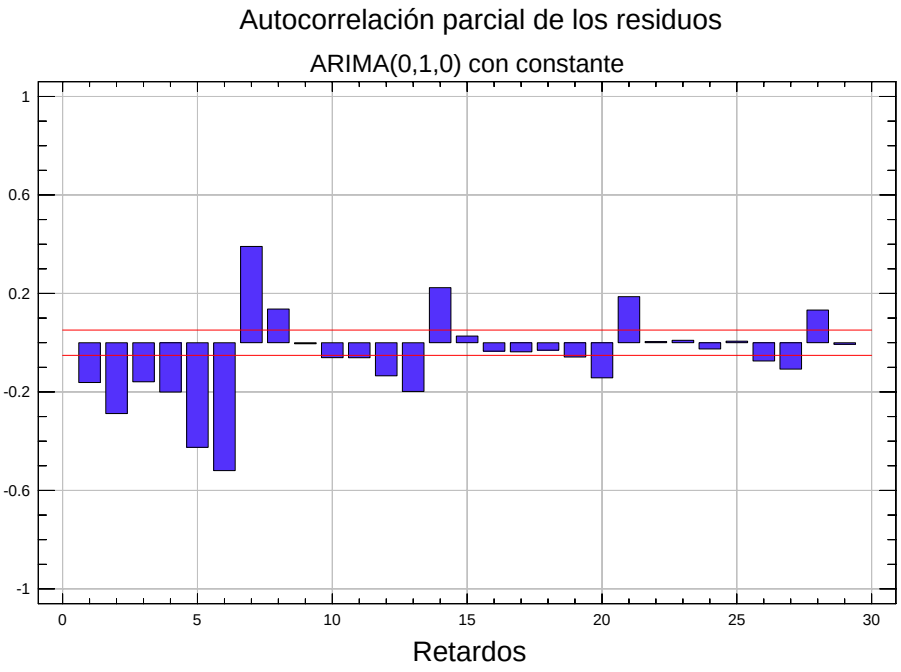


Figura 7.30: PACF de los residuos del modelo ARIMA (0,1,0) con constante

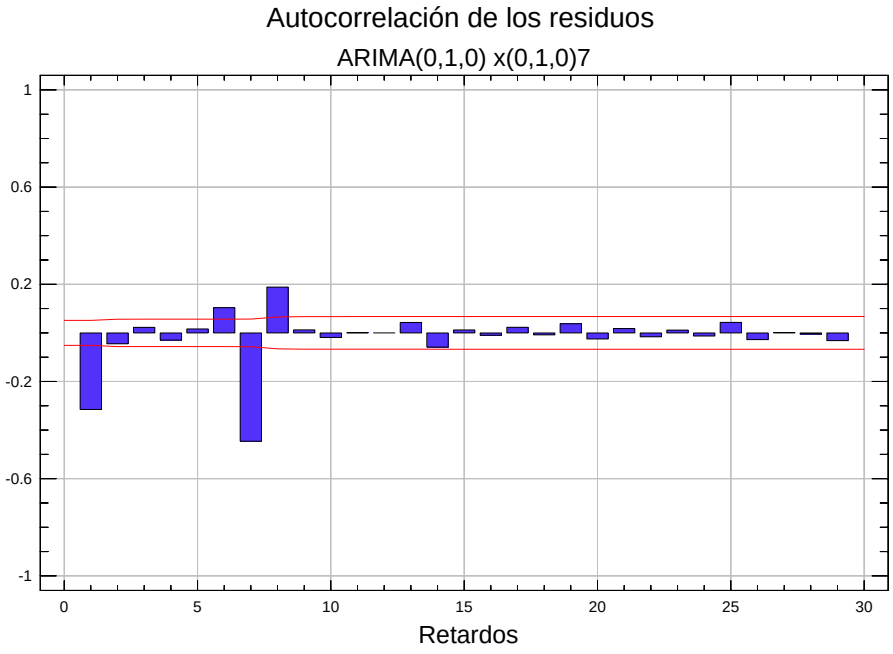


Figura 7.31: ACF de los residuos del modelo ARIMA(0,1,0)x(0,1,0)7

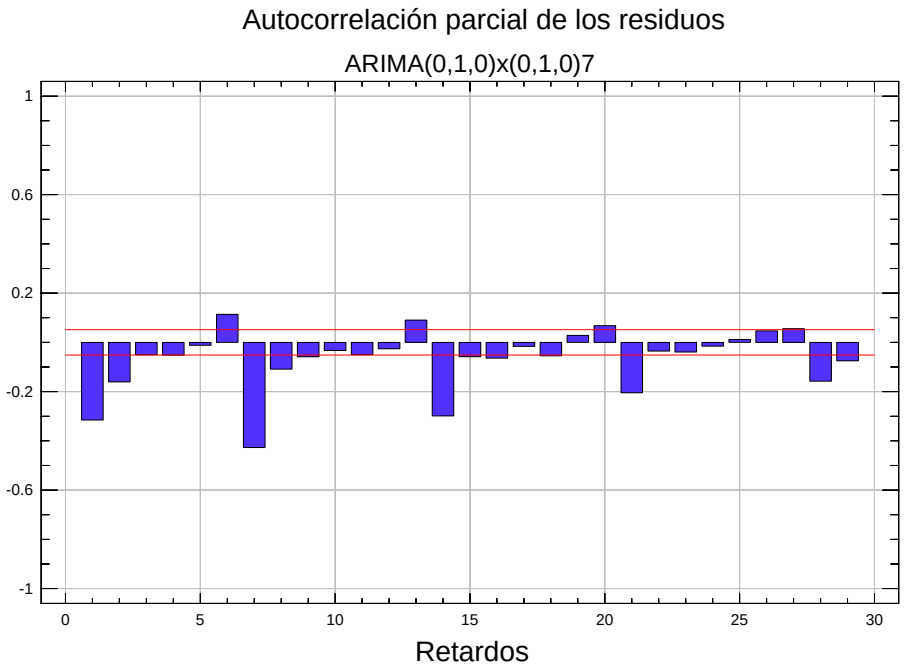


Figura 7.32: PACF de los residuos del modelo ARIMA(0,1,0)x(0,1,0)7

Identificación del número de términos AR y/ó MA

Una vez que hemos conseguido que la serie sea estacionaria por medio de diferenciación, el siguiente paso para ajustar un modelo ARIMA es determinar si se requieren términos AR ó MA para captar la autocorrelación que todavía pudiera existir en la serie.

Programas como *Statgraphics Plus* nos permiten hacer pruebas de distintas combinaciones de términos hasta identificar cual entrega mejores resultados. Sin embargo observado el gráfico de ACF (gráfico 7.31) y el de PACF (gráfico 7.32) se puede identificar el número de términos AR y/ó MA necesarios.

En el caso que estamos analizando los gráfico de ACF y PACF nos indican una leve sobre diferenciación ya que los valores de correlación se han vuelto negativos. Estos presentan el patrón característico que nos indica la necesidad de uno o varios componentes de media móvil (MA) no estacionales (*MA signature*): un corte rápido del ACF y una disminución lenta del PACF (es decir que cuenta con varios valores significativos) y de valor negativo. El número de valores de PACF fuera de los límites críticos nos señalan que 2 componentes MA no estacionales deben ser incluidos en el modelo. Es evidente también el patrón característico que nos indica la necesidad de incluir en el modelo un componente de media móvil estacional (SMA), ya que la autocorrelación en el periodo estacional es negativa. En base a lo anterior, el modelo que incluía inicialmente solo los órdenes de diferencias se ha transformado en un ARIMA (0,1,2)x(0,1,1)7.

Calibración y Validación del modelo

La serie de demandas con la que contamos corresponde a 1461 datos diarios de demandas, 4 años, desde Enero de 2001 a Diciembre de 2004. Hemos separado el conjunto de los datos en dos grupos, 3 años (2001, 2002 y 2003) serán utilizados para calibrar los modelos ARIMA y 1 año (2004) se utilizará para validar el modelo y analizar sus desempeños.

Independientemente de los datos que nos aporta la observación del correlograma, hemos analizado 4 modelos ARIMA distintos que han sido los que mejores resultados han aportado en un análisis preliminar. Los modelos se presentan en el cuadro 7.13. La diferencia entre ellos reside en la inclusión o no inclusión de un segundo componente de media móvil no estacional y de la temperatura como variable exógena. Del análisis preliminar también se determinó que la temperatura máxima (Tmax) es la variable que más información nos aporta al modelo, por lo que la temperatura media y mínima no serán consideradas en adelante para la selección del modelo.

Modelo	ARIMA(p,d,q)x(P,D,Q)s	No. de Parámetros
A	ARIMA (0,1,1)x(0,1,1)7	2
B	ARIMA (0,1,1)x(0,1,1)7 + Tmax	3
C	ARIMA (0,1,2)x(0,1,1)7	3
D	ARIMA (0,1,2)x(0,1,1)7 +1 Tmax	4

Cuadro 7.13: Modelos ARIMA seleccionados y sus parámetros

Modelo (A) ARIMA (0,1,1)X(0,1,1)7 El modelo (A) es el más simple y es por mucho el modelo que se utiliza con más frecuencia para describir series temporales no estacionarias. Fue el utilizado por Box et~al. (1976) para desarrollar su metodología. La diferencia con nuestro caso es que el componente periódico es $s = 7$. El modelo incluye un orden de diferencia simple ($d = 1$) en la serie, un componente de media móvil de orden $q = 1$, una diferencia estacional de orden $D = 1$ y un componente de media móvil estacional de orden $Q = 1$. Es decir que el modelo para su predicción, utiliza los errores de predicción del día anterior *t-1* y los errores de predicción de una semana antes *t-7*. Los resultados del modelo A, tanto para la estimación como para la validación se presentan en el cuadro 7.14.
Los primeros tres estadísticos (RMSE, MAE, MAPE) miden la magnitud de los errores, por lo que nos interesa que sus valores sean mínimos. Los últimos dos estadísticos miden el sesgo. Entonces un buen modelo será el que nos entregue valores cercanos a 0. El presente modelo y sus parámetros (MA(1) y SMA(1)) fue estimado con los primeros 1095 datos. Los últimos 366 datos fueron reservados para validar el modelo. Los valores de los errores obtenidos son aceptables en la fase de estimación e incluso mejoran (se reducen) en la fase de validación. Sin embargo los gráficos de ACF y PACF (7.33 y 7.34) nos indican que al menos 6 de los 30 coeficientes de correlación estimados son significativos para el intervalo de confianza del 95% por lo que los residuos no son completamente un ruido blanco como se esperaría si el modelo estimado captara completamente la estructura de la serie. Esta afirmación se confirma si se observa también el gráfico del periodograma de los residuos (7.35) donde se aprecia que todavía existen ciclos o determinadas frecuencias importantes que destacan sobre el resto. De igual forma el gráfico (7.36) donde se presenta el periodograma acumulado, nos indica que las frecuencias acumuladas de los residuos superan los límites críticos del 95 y 99% que contendrían a las frecuencias si los residuos fueran un ruido blanco. Por otra parte el modelo falla en el test de aleatoriedad de los residuos por rachas (long term cycles) y de autocorrelación (Box-Pierce) de los residuos.
El cuadro 7.15 presenta los resultados de los valores obtenidos para los parámetros MA y SMA. Los dos parámetros son altamente significativos (significativamente diferentes de 0) porque obtienen valores de P muy pequeños por lo que se justifica su inclusión en el modelo.

Estadístico	Periodo de Estimación	Periodo de Validación
RMSE	12,465.10	11,143.56
MAE	8,841.01	7,932.46
MAPE	2.876	2.40995
ME	-118.286	-17.6564
MPE	-0.145004	-0.086111

Cuadro 7.14: Desempeño del modelo (A), estimación y validación

Parámetro	Estimado	Error Stnd.	t	P-value
MA(1)	0.461215	0.0268051	17.2062	0.000000
SMA(1)	0.883966	0.0149446	59.1517	0.000000

Cuadro 7.15: Resumen del modelo (A)

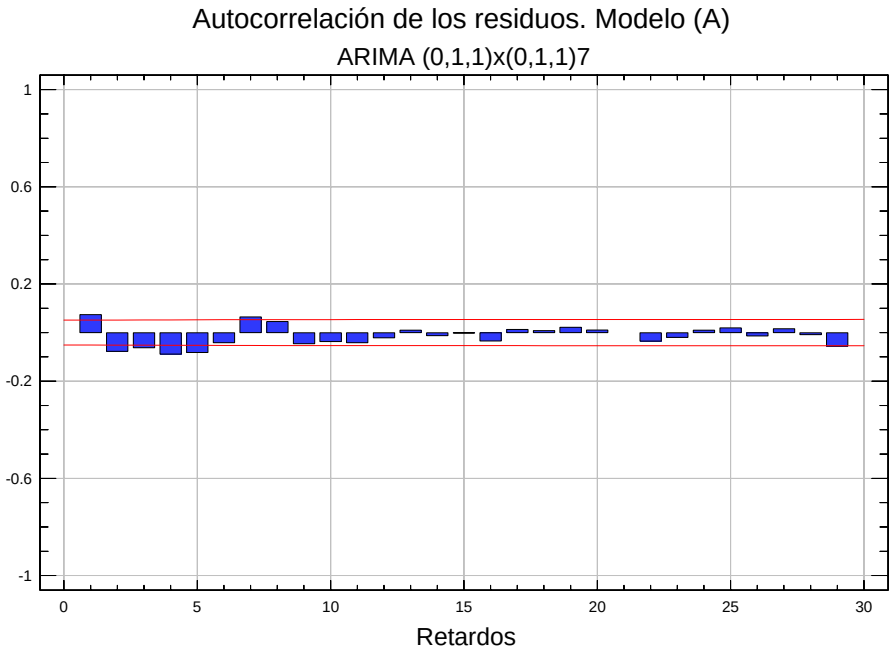


Figura 7.33: ACF de los residuos del modelo (A) ARIMA (0,1,1)x(0,1,1)7

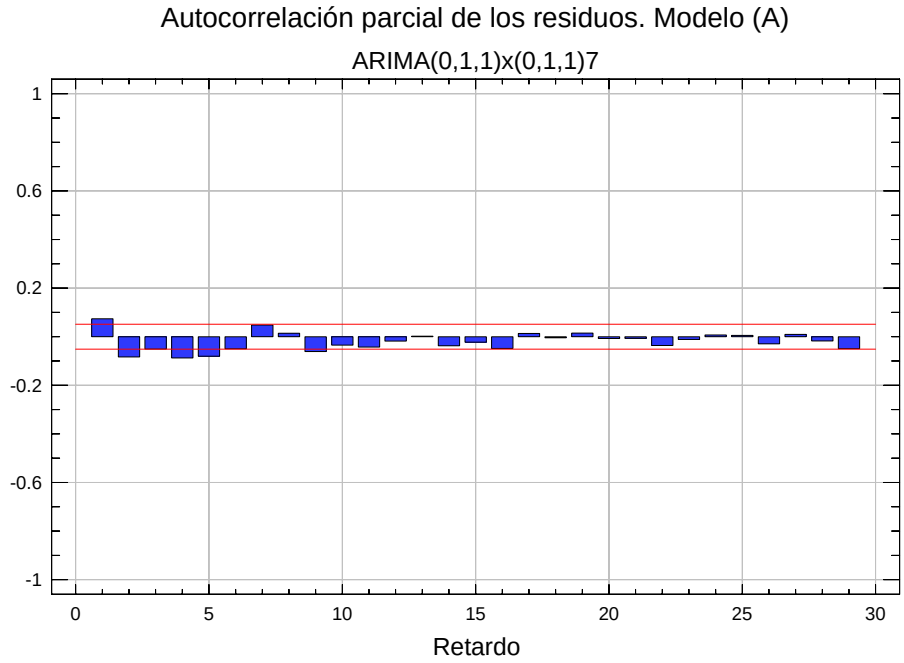


Figura 7.34: PACF de los residuos del modelo (A) ARIMA (0,1,1)x(0,1,1)7

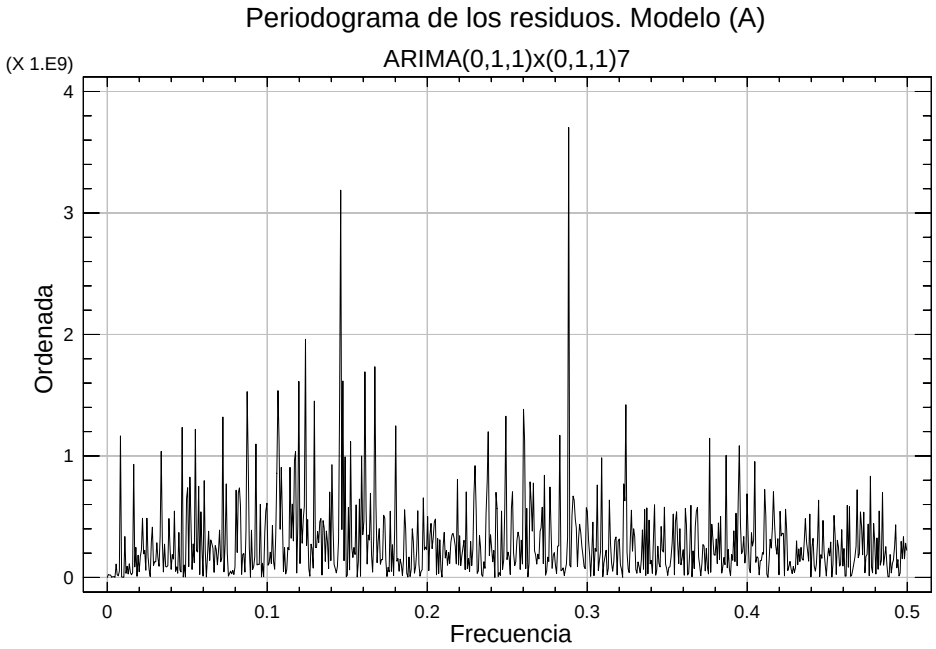


Figura 7.35: Periodograma de los residuos del modelo (A) ARIMA (0,1,1)x(0,1,1)7

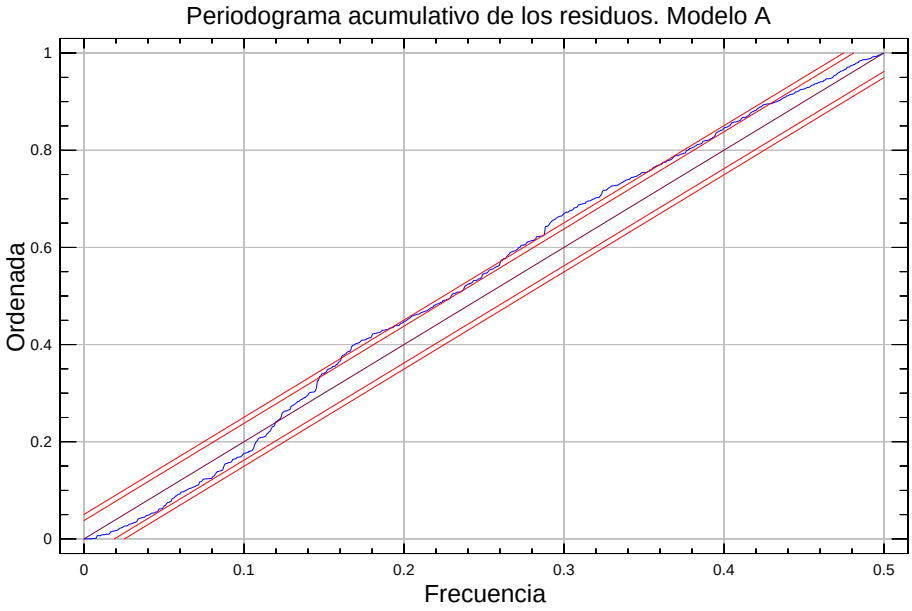


Figura 7.36: Periodograma acumulativo de los residuos del modelo (A) ARIMA (0,1,1)x(0,1,1)7

Modelo (B) ARIMA (0,1,1)X(0,1,1)7 + Tmax El modelo (B) tiene la misma estructura del modelo (A) pero le ha sido incluida la temperatura máxima (Tmax) como variable exógena con el fin de mejorar el modelo con la información que nos aporta la relación temperatura-demanda. El cuadro 7.16 de desempeño del modelo nos indica que en la fase de estimación el modelo presenta peores resultados con respecto a los obtenidos por el modelo (A). Los estadísticos que miden la magnitud de los errores (RMSE, MAE, MAPE) e incluso los estadísticos que miden el sesgo (ME, MPE) presentan peores resultados. En cambio, en la fase de validación la mejoría es significativa con respecto a la fase de estimación y también si comparamos los resultados con los obtenidos con el modelo (A). En cuanto a los gráficos de ACF Y PACF (7.37 y 7.38) presenta los mismos fallos que el modelo (A), lo cual es esperable ya que como se dijo antes, la estructura del modelo es la misma. La inclusión de la temperatura no aporta mejoras en la autocorrelación de los residuos en los retardos periódicos. El periodograma (7.39) y el periodograma acumulativo de los residuos (7.40) evidencian mejoras en el desempeño del modelo con respecto a los del modelo (A). Los valores de la ordenada se han reducido. En cambio el modelo falla en el tests de aleatoriedad de los residuos por rachas(de tendencias) y de autocorrelación (Box-Pierce).
La inclusión de la temperatura máxima (Tmax) como variable exógena en el modelo se justifica ya que como se observa en el cuadro 7.17, es altamente significativa estadísticamente.

Estadístico	Periodo de Estimación	Periodo de Validación
RMSE	12,631.40	10,640.84
MAE	8,912.25	7,629.09
MAPE	2.89865	2.31316
ME	-465.374	2.322
MPE	-0.245292	-0.06569

Cuadro 7.16: Desempeño del modelo (B), estimación y validación

Parámetro	Estimado	Error Stnd.	t	P-value
MA(1)	0.40622	0.0275149	14.7636	0.000000
SMA(1)	0.79270	0.0187152	42.3562	0.000000
Tmax	939.853	133.484	7.04095	0.000000

Cuadro 7.17: Resumen del modelo (B)

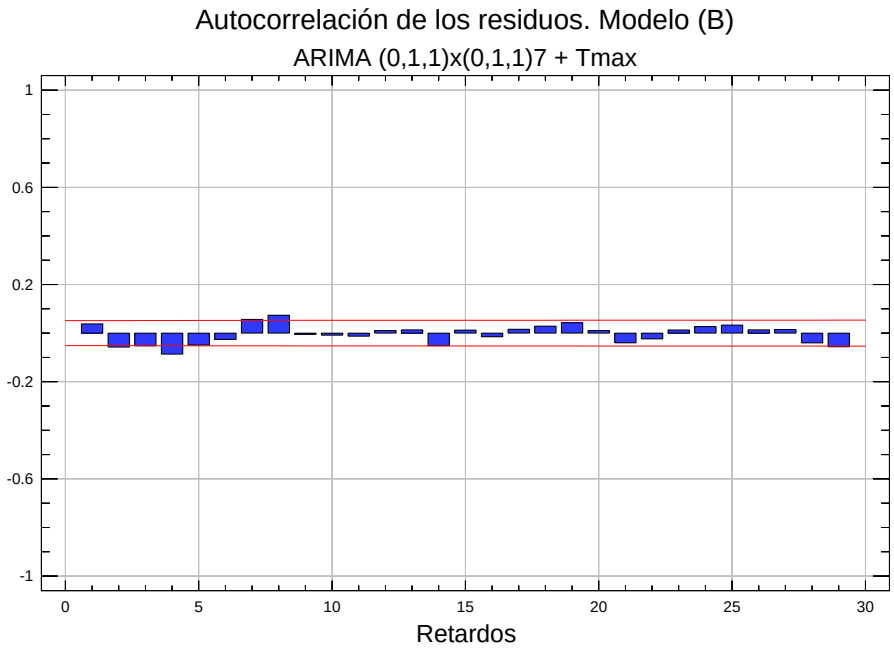


Figura 7.37: ACF de los residuos del modelo (B) ARIMA (0,1,1)x(0,1,1)7 + Tmax

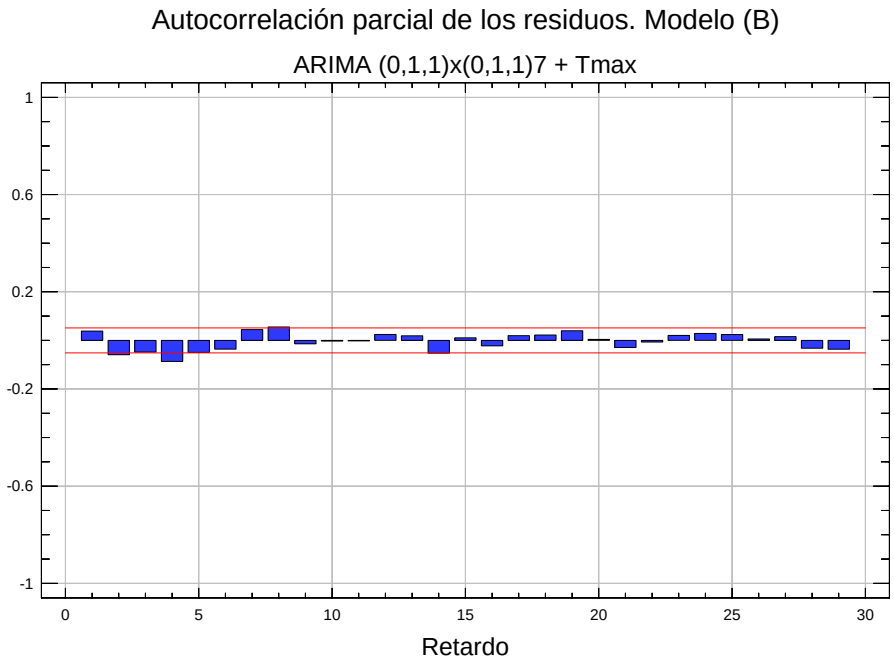


Figura 7.38: PACF de los residuos del modelo (B) ARIMA (0,1,1)x(0,1,1)7 + Tmax

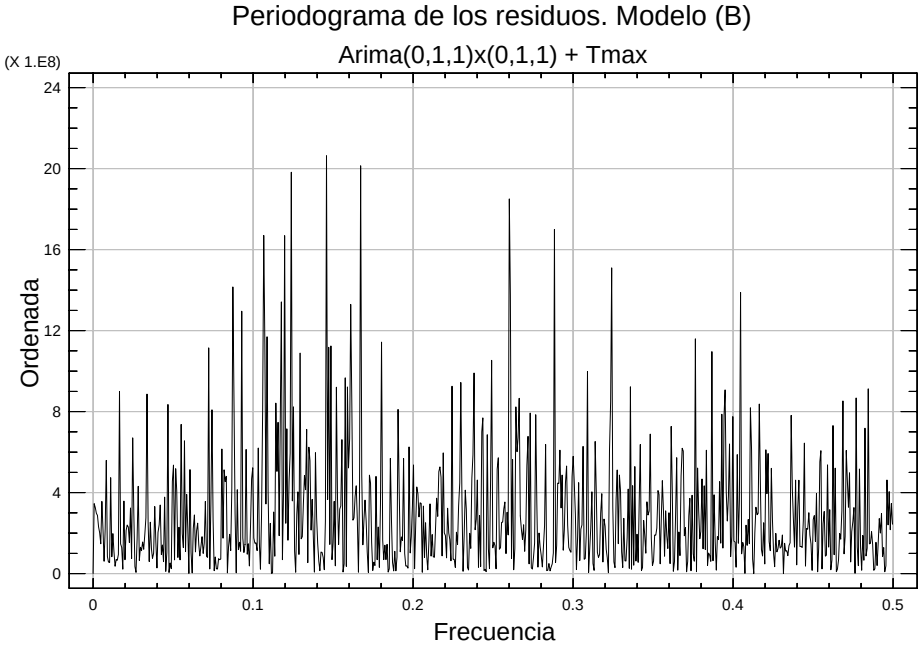


Figura 7.39: Periodograma de los residuos del modelo (B) ARIMA (0,1,1)x(0,1,1)7 + Tmax

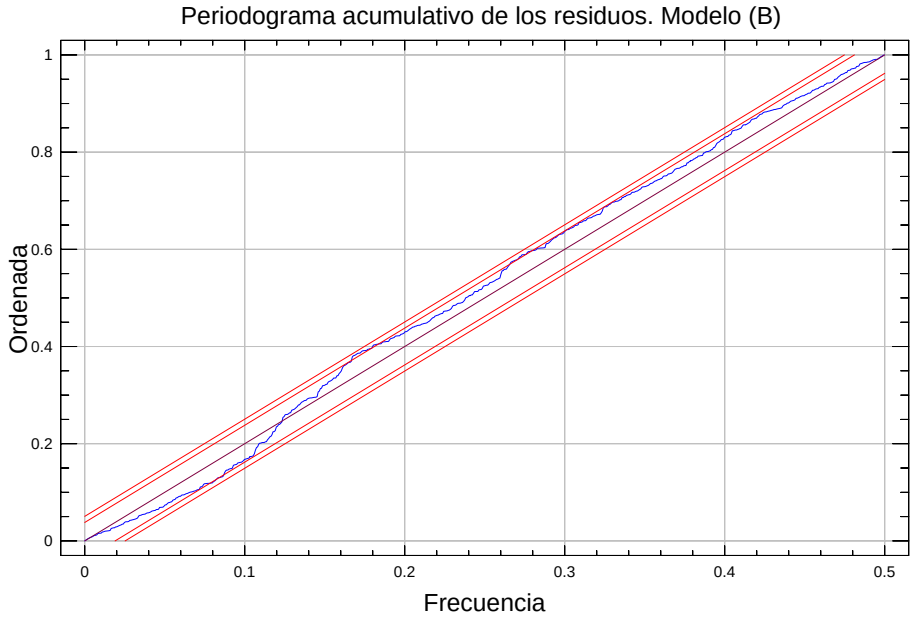


Figura 7.40: Periodograma acumulativo de los residuos del modelo (B) ARIMA (0,1,1)x(0,1,1)7 + Tmax

Modelo (C) ARIMA (0,1,2)X(0,1,1)7 Siguiendo las conclusiones a las que llegamos en la sección 7.4.1, donde el PACF de la serie con una diferencia regular (d) y una estacional (D) nos indica que es necesaria la inclusión en el modelo de dos componentes de media móvil regular MA(2), se ha construido el modelo que hemos denominado (C). Se puede decir que este modelo es diferente en su estructura respecto a los modelos (A) y (B) ya que utiliza dos componentes de media móvil en vez de una para intentar captar la estructura de la serie de demandas de la mejor manera. Es decir que utiliza los errores en la predicción para el momento *t-1* y *t-2* y para la parte estacional del modelo utiliza los errores de predicción de una semana atrás, es decir *t-7*. Con esta estructura el modelo supera las 5 pruebas realizadas a sus residuos de aleatoriedad (long term cycles, tendencia), autocorrelación (Box-Pierce) así como también el análisis de las medias y varianzas de la primer y segunda mitad de la serie de residuos para investigar si estas diferencias son significativas, que en este caso no lo fueron.

Si comparamos el presente modelo con los modelos (A) y (B) notamos inicialmente que supera las pruebas de aleatoriedad y homocedasticidad de los residuos mientras que los otros han fallado en algunas de ellas, por lo que podríamos decir que en efecto logra captar la estructura de la serie de mejor manera que los modelos anteriores. En cuanto al desempeño del modelo (ver cuadros 7.18 y 7.19), en la fase de estimación supera al modelo (A), sin embargo los resultados en validación son muy similares. Ahora si los resultados son comparados con el modelo (B), entonces el modelo (C) es superado en su desempeño, no en la fase de estimación sino en la de validación. Es importante recordar que el modelo (B) cuenta con la temperatura máxima (*Tmax*) en su estructura y el modelo (C) no la tiene. De cualquier forma ambos modelos cuentan con la misma cantidad de términos en su estructura, tres.

De la observación de los gráficos de ACF Y PACF (7.41 y 7.42) se observa que son 5 retardos de 30 los que superan los límites del 95 % y por lo tanto existe correlación estadísticamente significativa en ellos. El periodograma (7.43) y el periodograma acumulativo de los residuos (7.44) evidencian mejora en el desempeño del modelo respecto a los del modelo (A) y (B), permaneciendo siempre los valores dentro de los límites críticos del 95 y 99 %.

Estadístico	Periodo de Estimación	Periodo de Validación
RMSE	12,297.30	11,101.84
MAE	8,684.44	7,926.92
MAPE	2.82554	2.41304
ME	-123.628	-24.5575
MPE	-0.154342	-0.095516

Cuadro 7.18: Desempeño del modelo (C), estimación y validación

Parámetro	Estimado	Error Stnd.	t	P-value
MA(1)	0.407641	0.029783	13.687	0.000000
MA(2)	0.170632	0.029837	5.719 2	0.000000
SMA(1)	0.907968	0.013348	68.0207	0.000000

Cuadro 7.19: Resumen del modelo (C)

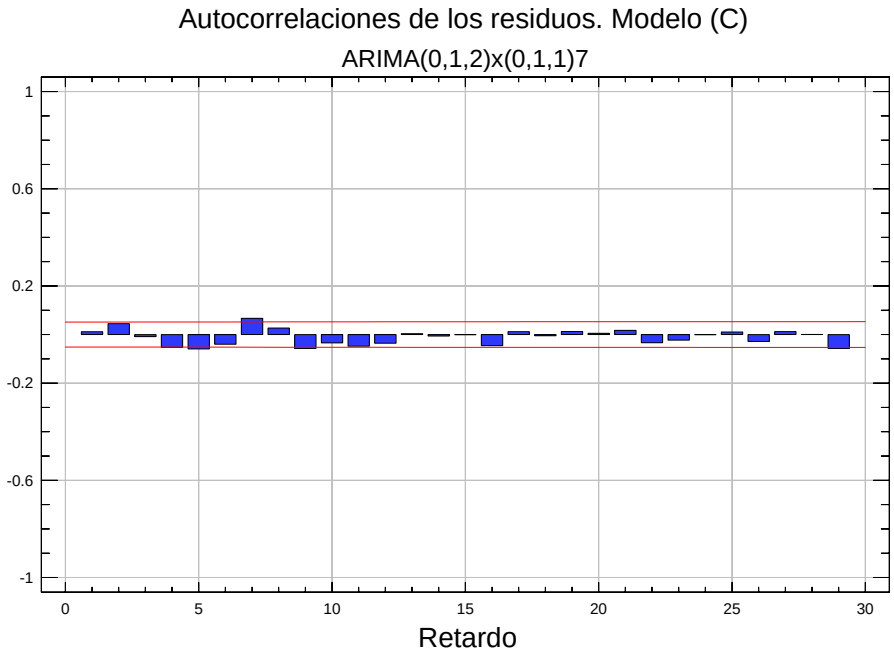


Figura 7.41: ACF de los residuos del modelo (C) ARIMA (0,1,2)x(0,1,1)7

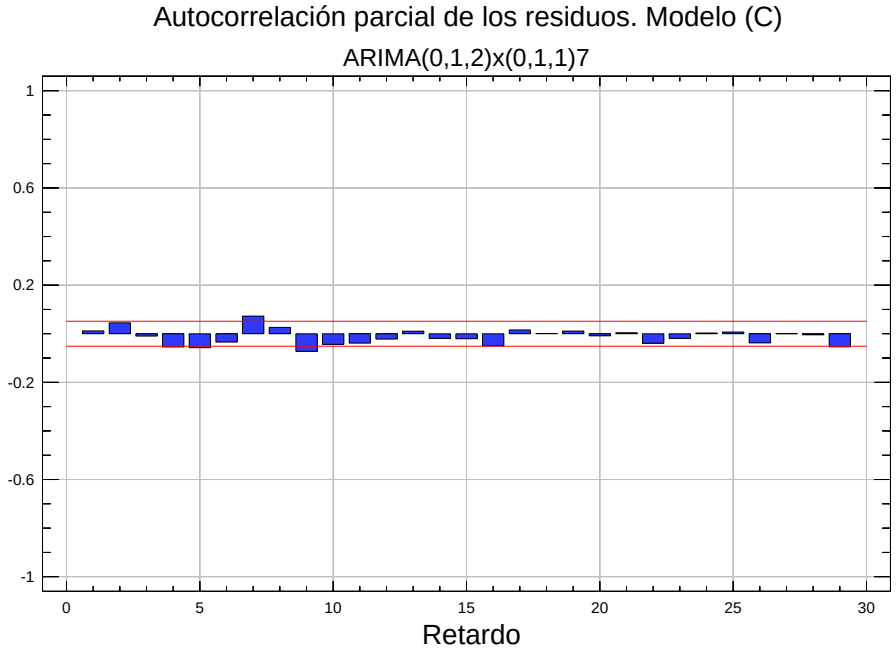


Figura 7.42: PACF de los residuos del modelo (C) ARIMA (0,1,2)x(0,1,1)7

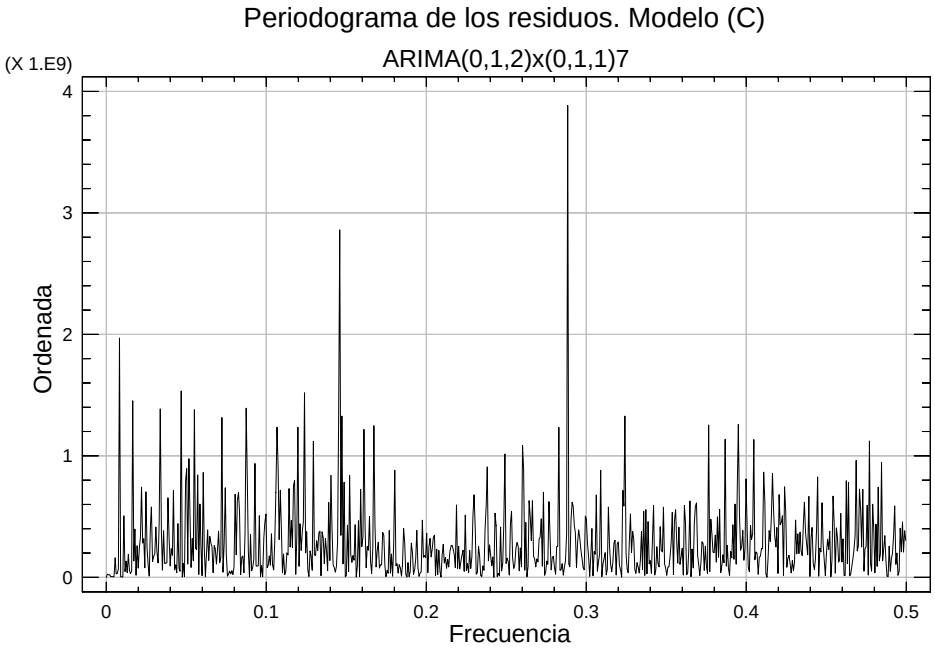


Figura 7.43: Periodograma de los residuos del modelo (C) ARIMA (0,1,2)x(0,1,1)7

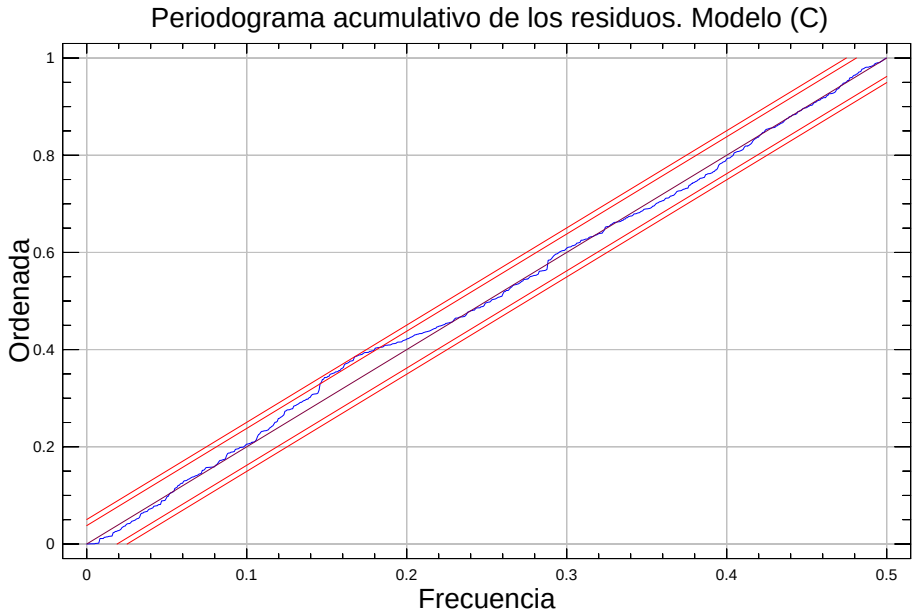


Figura 7.44: Periodograma acumulativo de los residuos del modelo (C) ARIMA (0,1,2)x(0,1,1)7

Modelo (D) ARIMA (0,1,2)X(0,1,1)7 + Tmax Este modelo tiene la misma estructura que el anterior (C), es decir dos componentes de media móvil regular (MA2) que ya vimos que aportaron una mejora en la serie de residuos obtenidos. Además hemos incluido también la temperatura máxima (*Tmax*) como variable exógena en busca de introducir en el modelo la relación temperatura-demanda que ya vimos que fue de utilidad en el desempeño del modelo (B). Por lo tanto este modelo (D) cuenta con 4 parámetros que deberán ser estimados.
Al igual que en los anteriores modelos se ha analizado el desempeño (ver cuadros 7.20 y 7.21) del modelo en las fases de validación y estimación. El modelo con la estructura actual supera en la fase de estimación a todos los modelos anteriores y en la fase de validación presenta los menores valores en los estadísticos que miden la magnitud de los errores (RMSE, MAE y MAPE) y es superado por el modelo (B) en los estadísticos que miden el sesgo (ME y MPE). El modelo obtuvo un valor de -9.9 m^3 para el error medio (ME) lo que indica que tiende a subestimar las predicciones en esa cantidad de metros cúbicos, lo cual es un valor mínimo si tomamos en cuenta que el valor medio de la demanda diaria es de 321,573 m^3 (ver cuadro 7.2). De cualquier forma en la siguiente sección veremos que este criterio no es suficiente a la hora de tomar una decisión sobre el modelo a elegir.

Estadístico	Periodo de Estimación	Periodo de Validación
RMSE	12,082.80	10,603.67
MAE	8,517.06	7,553.24
MAPE	2.77203	2.29683
ME	-53.1731	-9.90066
MPE	-0.127269	-0.081778

Cuadro 7.20: Desempeño del modelo (D), estimación y validación

Parámetro	Estimado	Error Stnd.	t	P-value
MA(1)	0.409113	0.0300746	13.6033	0.000000
MA(2)	0.167618	0.0301579	5.55802	0.000000
SMA(1)	0.874529	0.0157205	55.6299	0.000000
Tmax	993.559	134.449	7.38983	0.000000

Cuadro 7.21: Resumen del modelo (D)

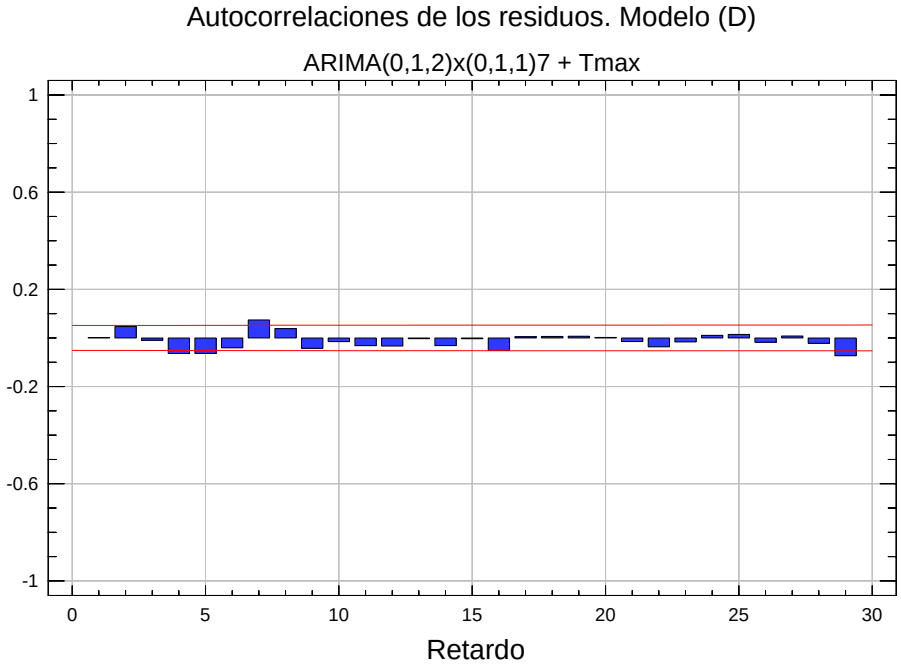


Figura 7.45: ACF de los residuos del modelo (D) ARIMA(0,1,2)x(0,1,1)7 + Tmax

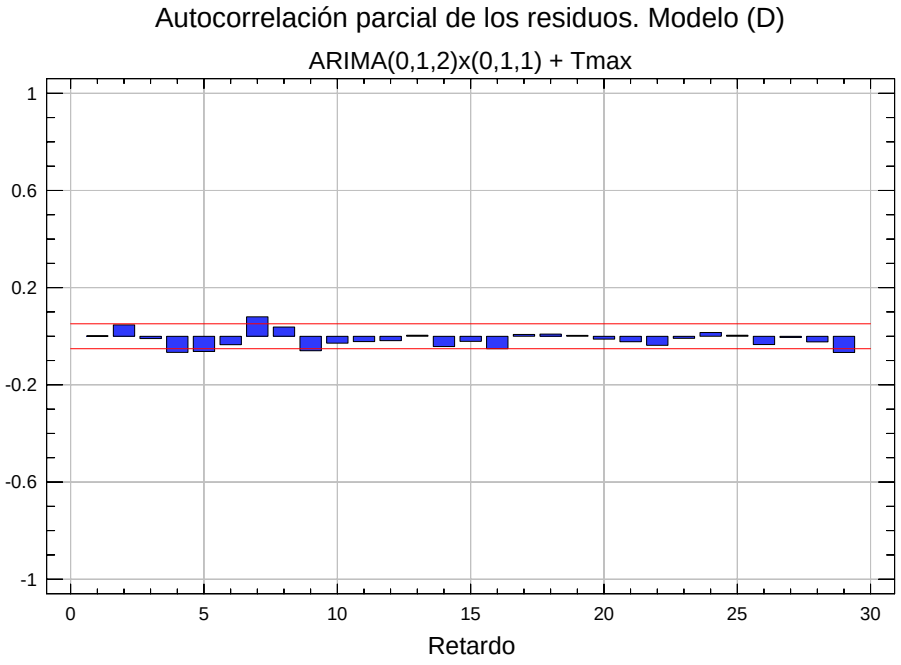


Figura 7.46: PACF de los residuos del modelo (D) ARIMA(0,1,2)x(0,1,1)7 + Tmax

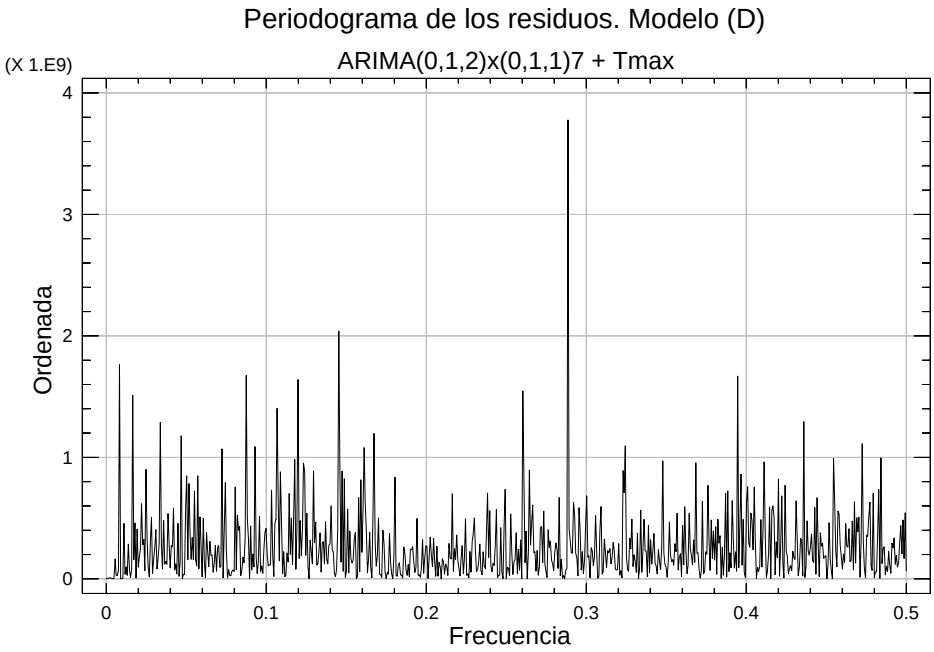


Figura 7.47: Periodograma de los residuos del modelo (D) ARIMA(0,1,2)x(0,1,1)7 + Tmax

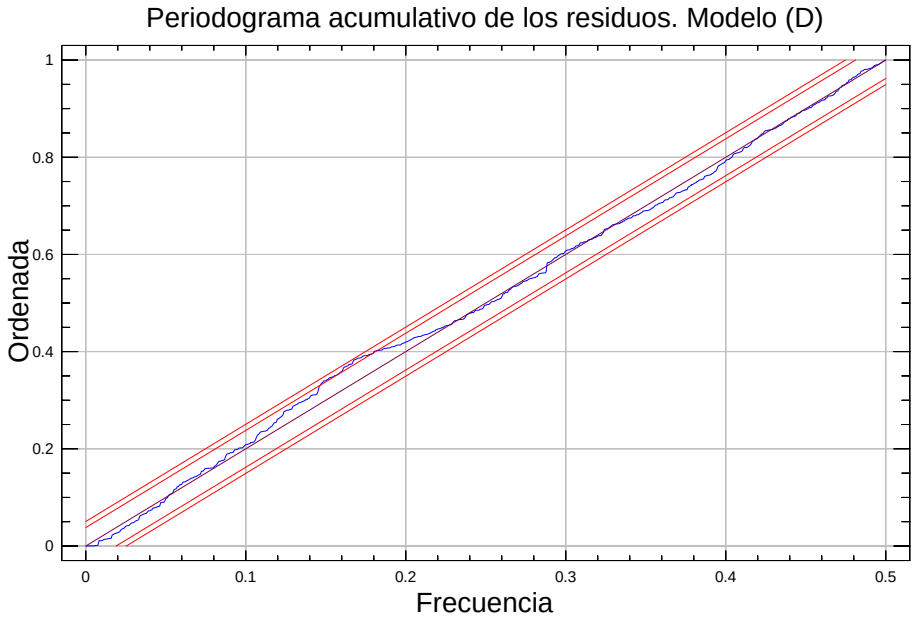


Figura 7.48: Periodograma acumulativo de los residuos del modelo (D) ARIMA(0,1,2)x(0,1,1)7 + Tmax

Comparación de modelos

El cuadro 7.22 presenta el resumen del análisis de los residuos para los cuatro modelos para distintos tests. Si un modelo ha sido bien ajustado a una serie y se ha logrado captar su estructura, entonces esperaríamos que los residuos formen una secuencia de números aleatorios o ruido blanco.

Modelo	RMSE	RUNS	RUNM	AUTO	MEAN	VAR
A	11,143.56	**	OK	***	OK	OK
B	10,622.31	OK	**	**	OK	OK
C	11,101.84	OK	OK	OK	OK	OK
D	10,603.67	OK	*	OK	OK	OK

RUNS=Test de rachas subidas y bajadas
RUNM=Test de rachas mayores y menores que la mediana
AUTO=Test de Box-Pierce para excesiva correlación
MEAN=Test de homogeneidad de medias de 1^{er} y 2^{da} mitad
VAR=Test de homogeneidad de varianzas de 1^{er} y 2^{da} mitad
OK=No significativo (p$\geq$0.05)
*=Marginalmente significativo $(0{,}01 < p \leq 0{,}05)$
**=Significativo $(0{,}001 < p \leq 0{,}01)$
***=Muy significativo $(p \leq 0{,}001)$

Cuadro 7.22: Resumen de comparación de test de los residuos de modelos (A), (B), (C), (D) fase de validación

Como se puede observar, solamente los modelos que incluyen en su estructura un segundo componente de media móvil (MA), es decir los modelos C y D obtienen resultados satisfactorios en el test de autocorrelación de los residuos. Los modelos con estructura más simple, es decir los modelos A y B fallan el test de autocorrelación además de fallos significativos en los test de rachas. Todos los modelos superan los test de homogeneidad de media y varianza (homocedasticidad).

Estadísticos del desempeño de los modelos

Estadísticos más comunes Hasta este punto hemos utilizado mediciones estadísticas *estándar* que nos indican la bondad de ajuste para evaluar un mejor o peor desempeño de los modelos analizados. Sin embargo tomar una decisión basándonos solamente en estos estadísticos nos podría llevar a cometer un error a la hora de seleccionar el mejor modelo.
Por ejemplo, el error medio ME de los residuos por si solo (Makridakis et~al., 1997)

$$ME = \frac{1}{n}\sum_{t=1}^{n} e_t \tag{7.6}$$

tiende a ser un valor pequeño ya que los errores positivos y negativos tienden a anularse mutuamente (ecuación 7.6). De hecho, el ME solamente indica si existe una sub o sobre predicción, pero no nos aporta mucha información a cerca del tamaño de los errores típicos.

El MAE (Error absoluto medio) y el MSE (Error cuadrático medio)

$$MAE = \frac{1}{n}\sum_{t=1}^{n} |e_t| \tag{7.7}$$

$$MSE = \frac{1}{n}\sum_{t=1}^{n} e_t^2 \tag{7.8}$$

evitan que los errores se anulen entre ellos haciéndolos positivos, ya sea tomando su valor absoluto, como lo hace el MAE (7.7), o haciéndolos cuadráticos como el MSE (7.8). El MAE tiene la ventaja de ser más interpretable y fácil de explicar para los no especialistas en el tema. El MSE tiene la ventaja de ser más manejable matemáticamente y por eso es utilizado frecuentemente en la optimización estadística. Los estadísticos anteriores tratan la precisión del modelo y su magnitud depende de la escala de los datos, por lo tanto no proporcionan un punto de comparación entre diferentes series temporales o diferentes intervalos. Es necesario entonces utilizar medidas relativos o porcentuales de error.

$$PE_t = \left(\frac{Y_t - F_t}{Y_t}\right) 100 \tag{7.9}$$

El MPE (Error medio porcentual) y MAPE (Error absoluto porcentual medio) son medidas relativas de error muy usadas

$$MPE = \frac{1}{n}\sum_{t=1}^{n} PE_t \tag{7.10}$$

$$MAPE = \frac{1}{n}\sum_{t=1}^{n} |PE_t| \tag{7.11}$$

Sin embargo tanto el MPE (7.10), como el ME, tienden a ser pequeños porque los PEs (errores porcentuales) positivos y negativos tienden a anularse. En cambio el MAPE (7.11) que utiliza los valores de los PEs absolutos nos aporta información más concreta. Es de más utilidad, por ejemplo saber que el MAPE de un modelo es 5% que simplemente saber que un determinado modelo tiene un MSE de 185.

Criterios alternativos Por todo lo comentado en el punto anterior, a la hora de seleccionar el mejor modelo surge la necesidad de utilizar otros métodos. El modelo perfecto no existe, puesto que todos constituyen simplificaciones de la realidad y siempre son preferibles modelos con menos variables (parsimonioso), puesto que además de ser más sencillos, son más estables y menos sometidos a sesgo. Es por eso que se han propuesto otras medidas de contraste entre modelos. Un criterio plausible para seleccionar un modelo ARIMA podría ser seleccionar el que tenga el valor más pequeño de la sumatoria de los errores cuadráticos, o el valor más grande de verosimilitud. Desafortunadamente esto no siempre funciona, ya que el MSE puede hacerse más pequeño y la verosimilitud más grande simplemente incrementando el número de términos en el modelo (siempre podríamos desarrollar polinomios o modelos AR ó MA suficientemente grandes para ajustarlo de mejor manera a una serie temporal, ver sección 2.2.2) aunque en este caso estaríamos sobreajustando el modelo a los datos.
Es por eso que es muy utilizado el criterio de Akaike (1974) o AIC. En este método la verosimilitud es penalizada por cada término adicional en el modelo. Si el término adicionado no mejora la verosimilitud en un monto mayor que el monto de penalización, no es recomendable su inclusión. La formulación es la siguiente:

$$AIC = -2\,log\,L + 2m \tag{7.12}$$

donde L denota la verosimilitud, aunque una aproximación a la formulación original del AIC se obtiene con (Makridakis et~al., 1997)

$$-2\,log\,L \approx n(1 + log(2\pi)) + n\,log\,\sigma^2$$

donde σ^2 es la varianza de los residuos y n es el número de observaciones en la serie. La obtención de σ^2 es sencilla y casi obvia en este tipo de modelos así que el AIC se puede obtener aproximadamente usando la fórmula:

$$AIC \approx n(1 + log(2\pi)) + n\,log\,\sigma^2 + 2m \tag{7.13}$$

Existen modificaciones del AIC que también se usan, como serían el BIC de Schwarz (Bayesian Information Criterion) ó el FPE (Final prediction error). Una diferencia en los valores del AIC de 2 ó menos no se considera sustancial y deberíamos considerar la opción de seleccionar un modelo más simple. El cuadro 7.23 es un resumen de los resultados obtenidos para cada modelo, incluyendo además una columna donde se ha calculado el AIC para cada uno de ellos. En base a estos, si no tomáramos en cuenta los que hemos denominado como criterios alternativos y observáramos solamente los estadísticos, sería muy complicado elegir el mejor modelo entre (B) y (D), ya que el primero obtiene valores mínimos para los estadísticos ME y MPE, y el segundo valores mínimos para RMSE, MAE y MAPE. En cambio si observamos los valores de Log-verosimilitud, AIC y BIC, podemos ver que el modelo D obtiene los resultados más satisfactorios aun y considerando que este modelo incluye un término adicional por el que fue penalizado.

Modelo	RMSE	MAE	MAPE	ME	MPE	Log-verosim	AIC	BIC
A	11,143.56	7,932.46	2.409	-17.65	-0.0861	-15,738.40	31,480.87	31,491.43
B	10,662.31	7,629.09	2.323	2.32	-0.0656	-15,696.90	31,399.77	31,415.61
C	11,101.84	7,926.92	2.413	-24.56	-0.0955	-15,723.20	31,452.44	31,468.29
D	10,603.67	7,553.24	2.297	-9.90	-0.0817	-15,680.70	31,369.31	31,390.43

Cuadro 7.23: Comparación de modelos en fase de validación

Modelo seleccionado

En las secciones anteriores hemos desarrollado modelos ARIMA con estacionalidad con diferentes estructuras y hemos comparado su desempeño en base a varios criterios. Si nos referimos a la información gráfica, el ACF, el PACF, el periodograma de los residuos y el periodograma acumulativo de los residuos han sido de utilidad. El ACF y PACF nos han proporcionado información con la cual hemos afinado los modelos en cuanto al número de elementos a ser incluidos, además de indicarnos si las periodicidades de la serie fueron captadas por los modelos y como su nombre lo indica nos aportaron valores de las autocorrelaciones de los residuos obtenidos para cada modelo. La información de los periodogramas nos ha servido también para verificar las periodicidades de los residuos con una técnica distinta al correlograma. Además el periodograma acumulativo nos ha dado una evidencia gráfica de la normalidad de los residuos de cada modelo. Si tuviéramos que seleccionar el modelo solamente en base a la información gráfica obtenida, la elección final estaría entre los modelos C y D, ya que sus ACF y PACF presentan menos retardos que sobrepasen los límites de confianza del 95% y aunque aún se observa una leve autocorrelación en los retardos periódicos esta se puede considerar aceptable, no se aprecian ciclos en los valores de las correlaciones. El periodograma de éstos modelos tiene una apariencia más uniforme para las distintas frecuencias, salvo por las correspondientes a 7 y 3.5 días (confirmando la información aportada por ACF y PACF). Finalmente el periodograma acumulativo nos muestra que los residuos siguen una evolución que se mantiene siempre dentro de los intervalos de confianza establecidos, indicando que los residuos siguen una distribución muy cercana a la normal, que es lo esperable para estos modelos.
Si seleccionáramos el mejor modelo en base a los estadísticos estándar (RMSE, ME, MAE, MPE, MAPE), ya hemos dicho en la sección anterior que son los modelos B y D (los que incluyen la temperatura como variable exógena) los que han obtenido los mejores valores. Y si observamos los criterios alternativos que utilizan la verosimilitud (L) y el número de términos para identificar el modelo (AIC, BIC), es el modelo D el que obtiene los mejores resultados.
Concluyendo lo anteriormente comentado, hemos seleccionado el modelo D como la mejor opción y más precisa para predecir las demandas de agua de la ciudad de Valencia. Es importante señalar que el modelo seleccionado es el más complejo en su estructura, ya que es un ARIMA (0,1,2)x(0,1,1)7 + Tmax, por lo que existen 3 parámetros a estimar pertenecientes al modelo ARIMA más el coeficiente de regresión de la temperatura máxima (*Tmax*). Una cuestión importante, es que el modelo requiere de la temperatura para hacer la predicción, sin embargo, para predicciones de más de un día ya no contaremos con datos de temperatura. Por lo que tendríamos que conseguir

mediante algún método, obtener predicciones de la temperatura esperada para la cantidad de días futuros a predecir, lo cual implica una incertidumbre adicional que tendría que ser considerada. De esta forma el modelo podría hacer predicciones de hasta 7 días como lo hace el modelo C. Los gráficos 7.49 y 7.50 presentan los resultados de la validación para las demandas observadas en el año 2004. En general los resultados obtenidos con el modelo D logran reproducir los valores observados a lo largo del año, siguiendo sin problemas la tendencia al alza que presenta la serie, los ciclos ligados a la variación de la temperatura y los ciclos semanales. El coeficiente de correlación (r) entre la serie observada y los valores obtenidos en la validación es de r=0.89.

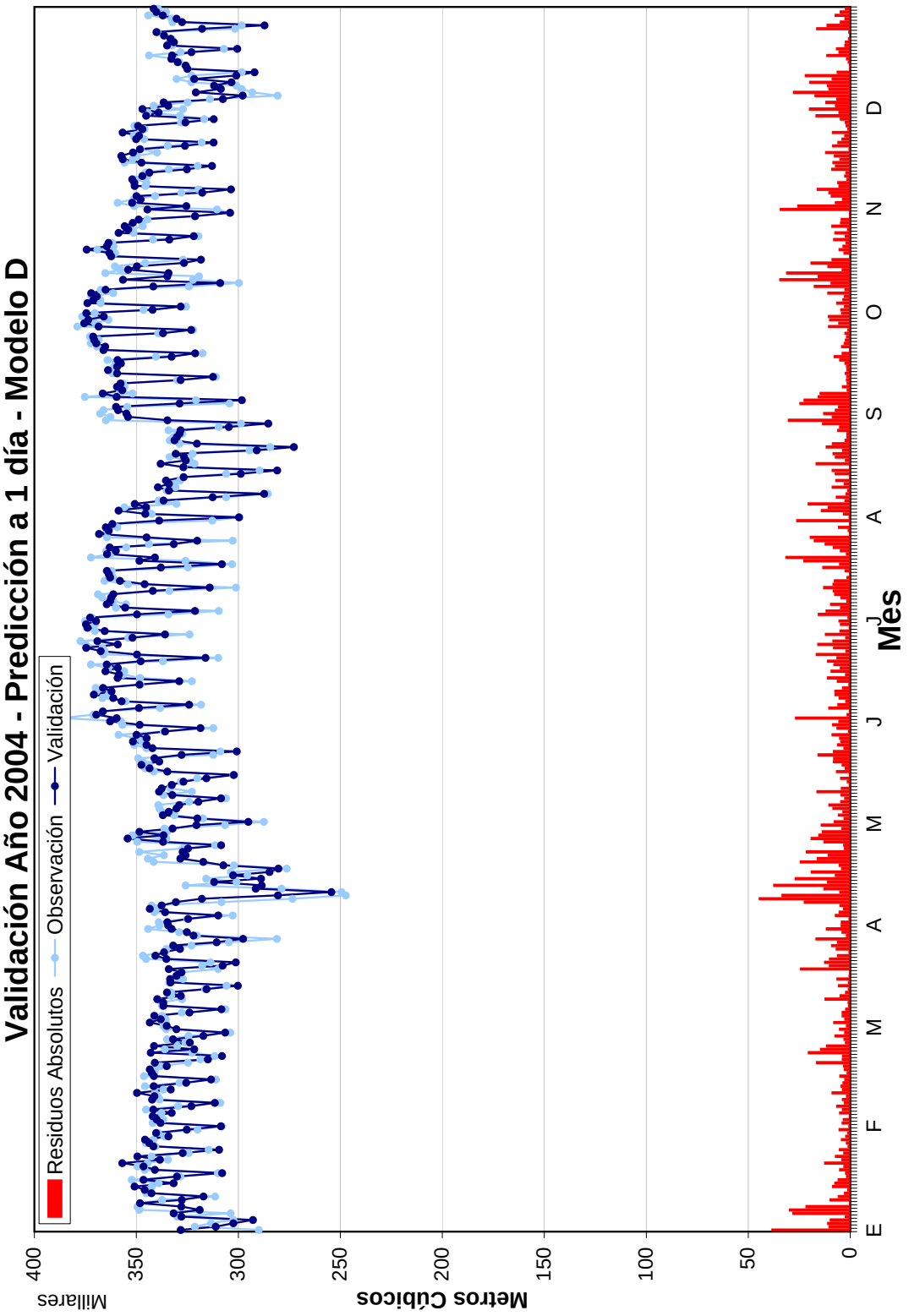


Figura 7.49: Gráfico de Validación vs. Observación del Año 2004

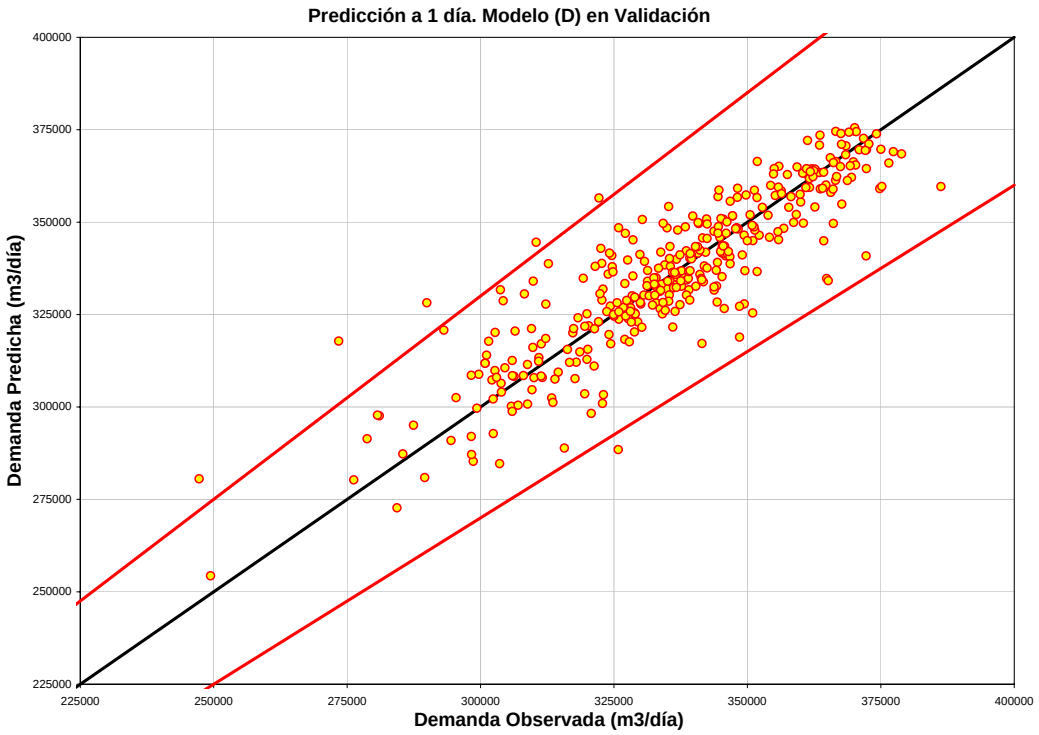


Figura 7.50: Gráfico de Demanda observada vs. Demanda Predicha, Validación, Modelo D. Las líneas rojas representan los límites de errores $\pm$ 10% de la demanda observada

Si lo que estamos buscando es un modelo con un horizonte de predicción más amplio, el modelo C es el más indicado, ya que tiene la misma estructura del modelo D pero no necesita de la temperatura para realizar su predicción. Con este modelo se pueden hacer predicciones de hasta 7 días. Los resultados de la fase de validación para los horizontes de predicción de uno y siete días se presentan en los gráficos 7.51 y 7.53 respectivamente. Se aclara que la predicción a siete días inicia el 1 de enero del 2004 y se obtienen las predicciones a 1, 2, 3, 4, 5, y 6 días previamente, por lo cual la verdadera predicción a siete días será la del 1 de enero, 8 de enero, 15 de enero y así sucesivamente. Los gráficos 7.52 y 7.54 presentan la evolución en el tiempo de la demanda observada y la demanda predicha para los horizontes de un día y siete dias. Es evidente que en la predicción a siete días el modelo tiene problemas para reproducir los cambios bruscos en la demanda diaria. En general y al igual que con el modelo D, los resultados obtenidos logran reproducir los valores observados a lo largo del año, siguiendo sin problemas la tendencia al alza que presenta la serie, los ciclos ligados a la variación de la temperatura y los ciclos semanales. Los coeficientes de correlación del modelo C contra la serie observada para predicción a un día es de r=0.88 y de r=0.76 para siete días.

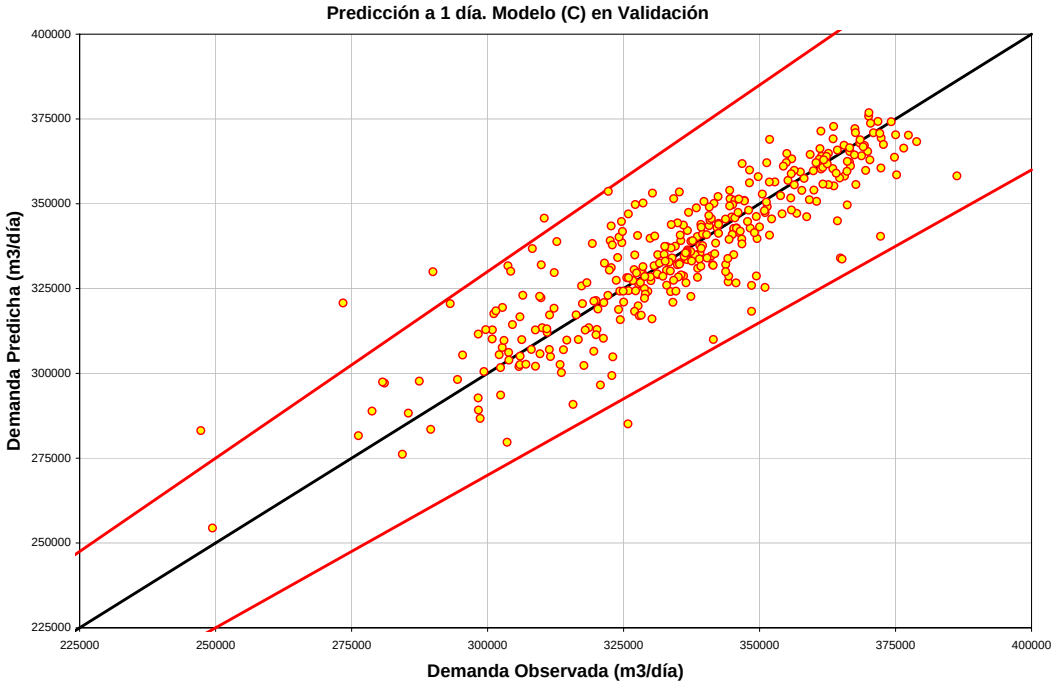


Figura 7.51: Predicción a 1 día. Gráfico de Demanda observada vs. Demanda Predicha, Validación, Modelo C. Las líneas rojas representan los límites de errores $\pm$ 10% de la demanda observada

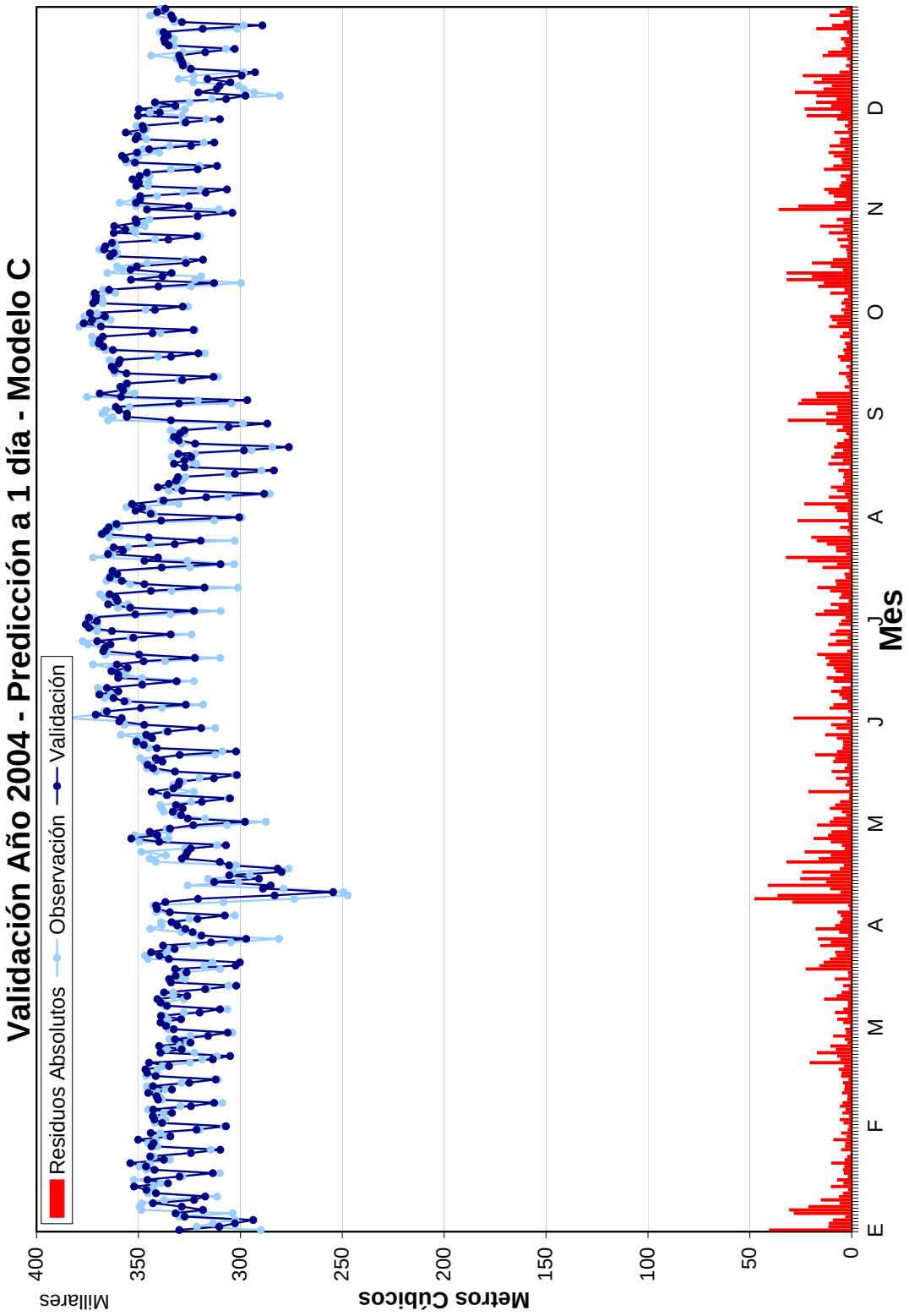


Figura 7.52: Predicción a un día. Gráfico de Validación vs. Observación del, Modelo C. Año 2004

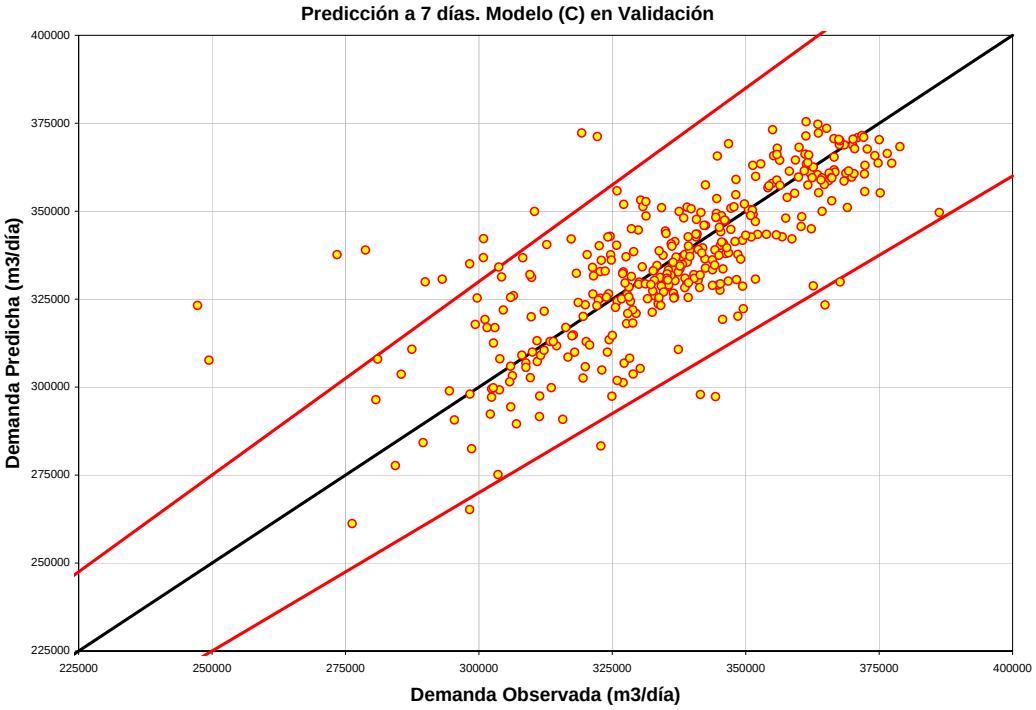


Figura 7.53: Predicción a 7 días. Gráfico de Demanda observada vs. Demanda Predicha, Validación, Modelo C. Las líneas rojas representan los límites de errores $\pm$ 10% de la demanda observada

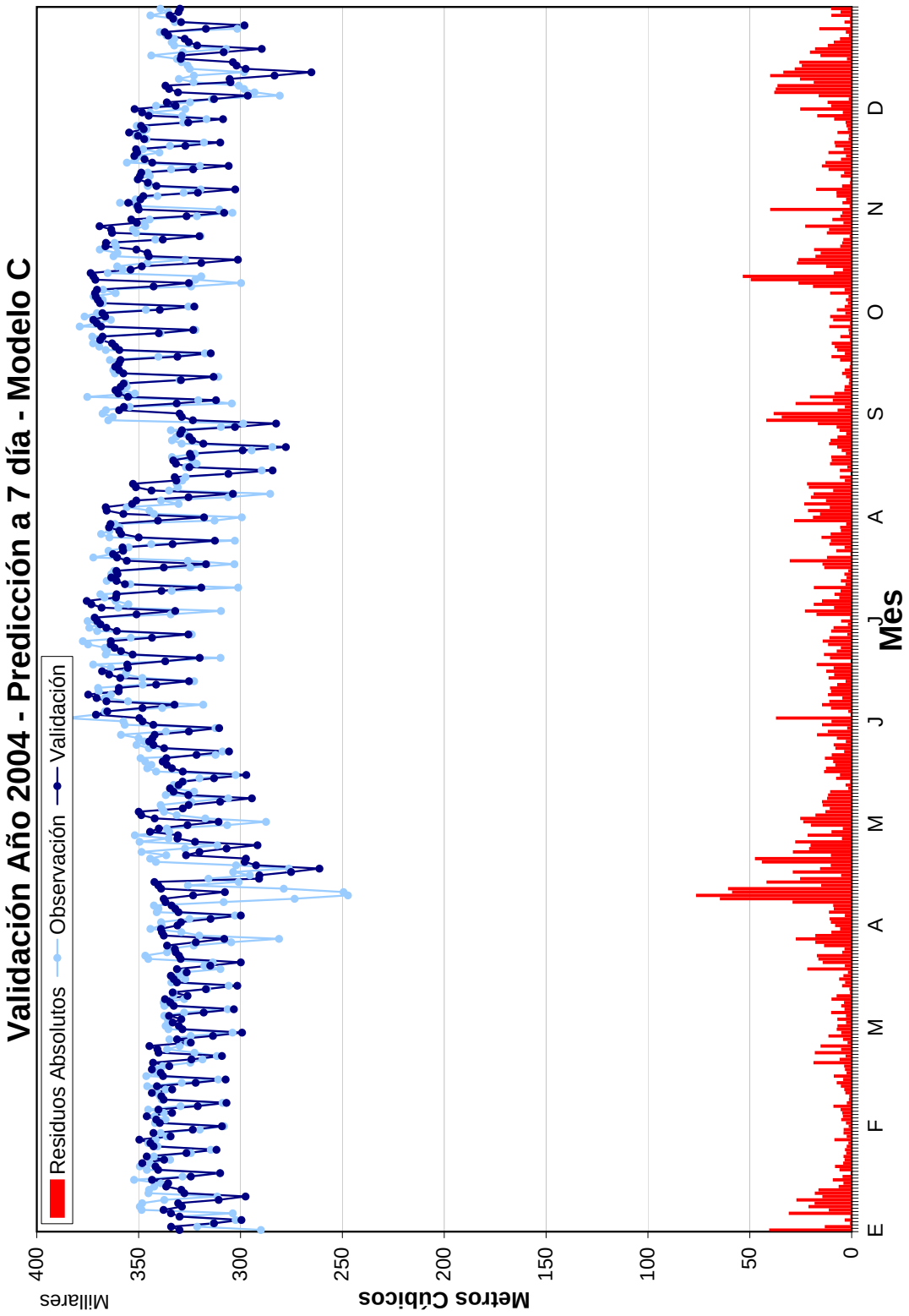


Figura 7.54: Predicción a siete días. Gráfico de Validación vs. Observación del Modelo C. Año 2004

Limitaciones del modelo seleccionado

Antes de comentar las limitaciones que presenta el modelo, se debe analizar que es lo que el modelo está haciendo para realizar su predicción. La ecuación (7.14) representa el modelo utilizando el operador de retardo:

$$(1-B)(1-B^7)Y_t = (1-\theta_1 B-\theta_2 B^2)(1-\Theta_1 B^7)e_t \tag{7.14}$$

Si desarrollamos esta ecuación y con el fin de usarla para predecir aumentamos por uno el orden de los subíndices obtenemos:

$$Y_{t+1} = Y_t + Y_{t-6} - Y_{t-7} + e_{t+1} - \theta_1 e_t - \Theta_1 e_{t-6} + \theta_1\Theta_1 e_{t-7} - \theta_2 e_{t-1} + \theta_2\Theta_1 e_{t-8} \tag{7.15}$$

donde

- $\theta_1 = 0{,}409113$ correspondiente al MA(1)
- $\theta_2 = 0{,}167618$ correspondiente al MA(2)
- $\Theta_1 = 0{,}874529$ correspondiente al SMA(1)

El modelo utiliza los valores de demanda en el instante Y_t (la última demanda con que se cuenta) y obtiene la variación de la demanda que existió entre los instantes Y_{t-6} y Y_{t-7}. Después aparece en la ecuación e_{t+1} que representarían los errores futuros de predicción y ya que es imposible conocerlos se suponen que seguirán una distribución normal con media 0. El resto de la ecuación utiliza los errores en el instante e_t (los errores de la última predicción), los errores en el instante e_{t-6} y e_{t-7} y los suaviza al multiplicarlos por los valores de los parámetros obtenidos para el modelo, además de combinarlos con el coeficiente de regresión correspondiente a la temperatura máxima (T_{max}) en el caso que el modelo la incluya en su ecuación. Es evidente que el modelo está siempre extrayendo valores del pasado (*viendo hacia atrás*), de donde obtiene la información para poder hacer las predicciones. Debido a esto, el modelo será incapaz de realizar una buena predicción cuando se presente un cambio brusco en el patrón de demandas semanal (ver gráfico 7.25), es decir cuando el patrón de demandas de la semana anterior sea muy distinto al de la semana que estamos prediciendo. Este tipo de cambios bruscos se presentan cuando la demanda diaria de uno o varios días

aumenta o disminuye debido a por ejemplo, días no laborables propios de cada comunidad, festividades locales, aumentos o disminuciones de temperatura, aunque estos últimos suelen ser menos bruscos. Estos (por llamarlos de alguna manera) patrones sociológicos, como ya se dijo, son propios de cada ciudad y es información que se conoce con antelación por corresponder a fechas fijas en la mayoría de los casos y algunas otras que no ocurren siempre en la misma fecha como la pascua, se puede conocer cuando ocurrirán con antelación. Esta información podría ser incorporada a un modelo para obtener una mejor precisión a la hora de hacer una predicción. El cuadro del apéndice B (página 317), presenta un listado de los días no laborables en la ciudad de Valencia. Los días no laborables cobran importancia cuando ocurren en un día de la semana que facilite la prolongación de un fin de semana, por ejemplo un día no laborable en jueves o viernes inevitablemente afectarán el patrón de demandas semanal, e igualmente afectará un día no laborable ocurriendo en lunes o martes. La figura 7.55 nos presenta un ejemplo de lo anteriormente comentado. El día 19 de marzo del 2004 (día de San José), festivo en Valencia y por tanto no laborable, tuvo lugar en un viernes. La demanda de ese día sufrió una disminución importante, a tal grado que fue incluso menor que el sábado y domingo posteriores cuando normalmente no ocurre así. El modelo, que está obteniendo su información del día anterior y de la semana anterior, comete un fallo en la predicción mayor que los registrados en los días anteriores. Algo similar ocurre en semana santa, que en el año 2004 tuvo lugar durante el mes de abril. Los días santos del 8 al 11 de abril y posteriormente el 12 de abril (día de san Vicente Ferrer, también festivo en Valencia) presentan valores de demanda muy distintos a los que se venían registrando, alejándose del patrón de demandas y en consecuencia provocando que el modelo cometa errores de gran magnitud. De hecho el viernes santo, 9 de abril, es el día en que el modelo comete el error más grande de la fase de validación.

Finalmente, el modelo C tiene un horizonte de predicción máximo de 7 días. Esto se explica de nuevo con la ecuación 7.15. Con una predicción mayor de 7 días, el modelo no cuenta con información de errores cometidos para calcular la predicción. En cambio, el modelo D que resulta ser el más preciso en sus predicciones, está limitado a predecir a un día por la disponibilidad del dato de temperatura y hasta 7 días si esta también es predicha.

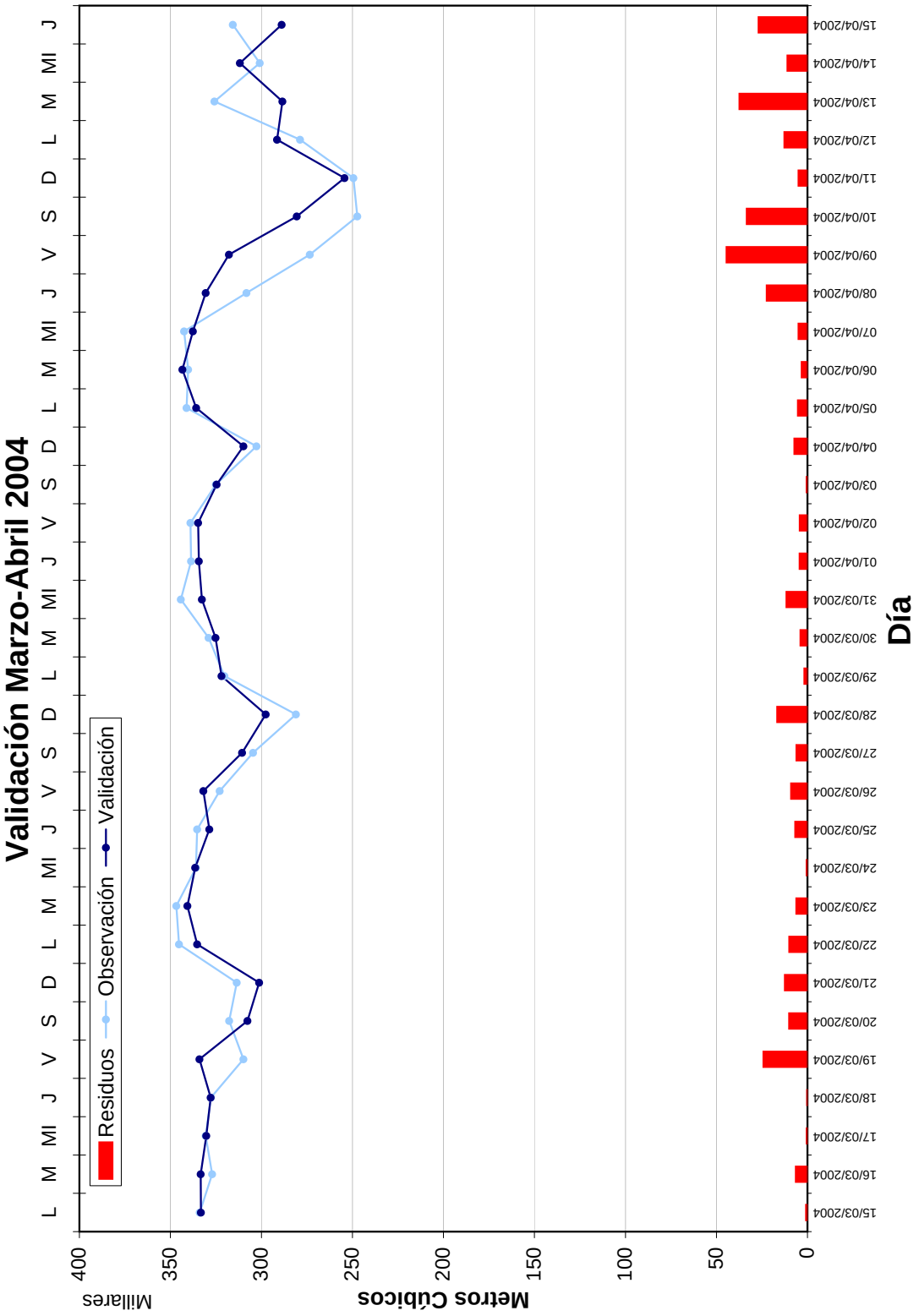


Figura 7.55: Gráfico de Validación vs. Observación del Año 2004

Parte VI

Mejoras Metodológicas

Capítulo 8

Modelos basados en Redes Neuronales

8.1. Introducción

Las redes neuronales artificiales han venido siendo utilizada en los últimos años para la resolución de problemas en el área de los recursos hídricos y más recientemente en la modelación y predicción de la demanda de agua potable (Griñó~C., 1991; Jain and Ormsbee, 2002; Joo et~al., 2002; Bougadis et~al., 2005; Zhang et~al., 2006). A lo largo de esta sección utilizaremos el mismo conjunto de datos con los que se construyeron los modelos ARIMA de la sección anterior para identificar un grupo de *ANN* que logren captar el proceso que sigue la demanda diaria de la ciudad de Valencia, España, en el periodo de 2001 a 2004. El objetivo del desarrollo de esta sección es el de realizar una comparación efectiva e imparcial en igualdad de datos disponibles de los modelos *ANN* con las metodolodías de series temporales y no el de innovar en los planteamientos de análisis de la demanda de agua con estas técnicas. El trabajo numérico de las *ANN* se realizó con el software *MATLAB R2008b* y su *toolbox* de redes neuronales.

La identificación de la arquitectura óptima de una *ANN* requiere del mismo análisis exploratorio del conjunto de datos que se ha desarrollado en la sección 7.3 como apoyo para la identificación de los modelos de series temporales ARIMA. Faraway and Chatfield (1998) recomiendan que un buen modelo de *ANN* para series de datos temporales debe ser seleccionado combinando habilidades tradicionales de modelación con conocimientos de análisis de series temporales y de los problemas particulares. Al haber realizado en las secciones 7.3 y 7.4 un análisis preliminar (estacionareidad, estacionali-

dad, análisis de Fourier, valores máximos y mínimos, etc.), y haber identificado previamente varios modelos ARIMA estacionales para predecir la demanda, contamos ya con un conocimiento de la serie y argumentos desde los cuales partir para la identificación de una o varias redes neuronales.

Un modelo de demanda basado en redes neuronales considera a la demanda de agua como una serie temporal $Y_t(t = 1, 2, \ldots, N)$, y se intenta estimar los valores futuros, Y_{N+1}, a partir de la información que conllevan los valores del pasado. Se obtiene la función f (ecuación 2.46) que relaciona la demanda en el instante $N + 1$ con la demanda de los instantes $1, 2, \ldots, N$.. Es posible además tomar en cuenta otras variables que aporten información del proceso de demanda de agua, como pueden ser la temperatura atmosférica y/ó variables categóricas que expliquen variaciones y ciclos del proceso de demanda.

Ya se ha mencionado en otras secciones que el buen desempeño de las *ANN* radica en su capacidad de aprender las relaciones entre las entradas (datos históricos de la demanda de agua, temperaturas, datos categóricos, etc.) y sus salidas (demanda observada), para después generalizar y predecir la demanda dentro del conjunto de datos observados. Es sumamente importante la correcta elección de las variables o vectores de entrada a la red, ya que es conocido que impacta directamente en la eficiencia de las predicciones de los modelos de redes neuronales. No menos importante resultan ser la topología de la red, el algoritmo de entrenamiento y en una menor medida lo es también el número de iteraciones.

8.2. Variables de entrada a las *ANN*

Las variables de entrada que se utilizaron para entrenar, probar y validar las *ANN* serán en un principio las mismas que se utilizaron en los modelos ARIMA con estacionalidad. Se tiene claro que los valores de demanda en el instante Y_{t+1} están fuertemente correlacionados con el valor en el instante Y_{t-1}, así como al instante Y_{t-7} (ver figura 7.6, pag. 152). Con estos datos tendremos un vector de entrada de longitud 2, formado por el valor de la demanda del día antecedente y la de 7 días antes. Adicionalmente se integrarán a este vector, para *ANN* con estructuras más complejas, variables categóricas binarias (0,1) mediante las cuales se incorpora información de días festivos, laborables y no laborables (ciclo semanal) y un indicador del mes de agosto, esta última por la ya comentada brusca reducción de la demanda durante todo el mes. El cuadro 8.1 resume las variables que se utilizarán.

Variable	Tipo
Demanda antecedente (Y_{t-1})	Numérica
Demanda antecedente (Y_{t-7})	Numérica
Demanda predicha (Y_{t+1})	Numérica
Temperatura máxima $(Tmax_{t-1})$	Numérica
Festivos	Binaria
Laborable/No laborable	Binaria
Indicador Agosto	Binaria

Cuadro 8.1: Variables utilizadas en los vectores de entrada a las redes neuronales

8.3. Elección de *ANNs*

Para la elección de la topología y arquitectura de las *ANNs* que mejor desempeño presenten, iniciaremos por estructuras simples e iremos aumentando su complejidad en cuanto a la dimensión del vector de entrada, la cantidad de neuronas por capa y el horizonte de predicción. Se ha dividido el conjunto de datos de la misma forma en que se hizo para los modelos ARIMA, y se han utilizado 3 años para el entrenamiento de la red y un año se ha reservado oculto para validar el desempeño con la red entrenada. Las registros de demanda diaria han sido preprocesados y escalados para que presenten valores entre (-1,1) y estén acorde con la función tansigmoidal (ver figura, 2.9, pag. 54) que se utiliza en la capa oculta de las redes neuronales propuestas. El algoritmo de entrenamiento utilizado en todos los casos es el de Levenberg-Marquardt por ser el método más rápido para entrenar redes neuronales *feedforward* de tamaño moderado, (Demuth et~al., 2009, pag. 5-31)

8.3.1. *Red 231*

Como una primer intento hemos construido una *ANN* 231, o lo que es lo mismo, un vector de dos entradas, 3 neuronas en la capa oculta y una neurona en la capa de salida, así como un vector que contiene el conjunto de datos objetivo correspondiente a cada día que se va a predecir. El cuadro 8.2 presenta las principales características de la red y la figura 8.1 presenta su topología. Se ha probado su desempeño para predecir la demanda diaria a un día.

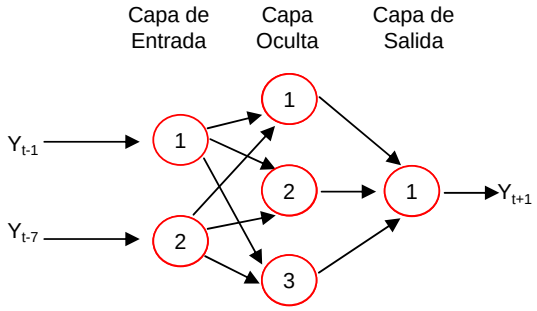


Figura 8.1: Estructura de la red neuronal 231

Dato	Característica
Entradas	2
Salidas	1
Capas	3
Neuronas capa oculta	3
Neuronas capa salida	1
Función de inicializado de red	Nguyen-Widrow
Función de transferencia	Tansigmoidal/lineal

Cuadro 8.2: Variables utilizadas en los vectores de entradas a la red neuronal *ANN*231

Red 231. Predicción a 1 día

Los resultados obtenidos se presentan en el cuadro 8.3 para las fases de entrenamiento y validación. En ellos destaca que se presenta un empeoramiento de los estadísticos en la fase de validación con respecto a la fase de entrenamiento, –hecho esperable–, y lo mismo ocurre con los valores de los coeficientes de correlación r y de determinación R^2.

La figura 8.2 presenta el desempeño de la red entrenada para cuando las entradas corresponden al año 2004 –conjunto de datos reservados para validación–. Es evidente que la red consigue captar la tendencia a largo plazo, el ciclo anual, así como también la periodicidad semanal de la serie de demandas. Los errores de la parte baja de la figura se mantienen en valores bajos salvo en momentos puntuales. Sin embargo, al realizar un análisis más profundo de estos errores, encontramos que la apariencia del periodograma acumulativo y la ACF (figuras 8.3 y 8.4) nos indica que estos no tienen la apariencia de un ruido blanco y aún conservan algún rastro de autocorrelación en los retardos 1 y 2.

Estadístico	Periodo de Entrenamiento	Periodo de Validación
RMSE	14,207.18	14,395.23
MAE	10,393.00	10,379.10
MAPE	3.38	3.15
ME	0.00	1,103.90
MPE	-0.22	0.146
Error Máx. Abs.	72,628.6	67,676.40
r	0.833	0.776
R^2	0.694	0.603

Cuadro 8.3: Desempeño de la *ANN*231, estimación y validación a un día

Las figuras 8.5 y 8.6 presentan el desempeño de la red en las fases de entrenamiento y validación respectivamente. Las líneas rojas indican los límites de errores superiores e inferiores al 10% de la demanda observada. Los valores se mantienen en su mayoría dentro de la banda mencionada y se agrupan alrededor de la línea negra que representa una predicción cien por ciento correcta.

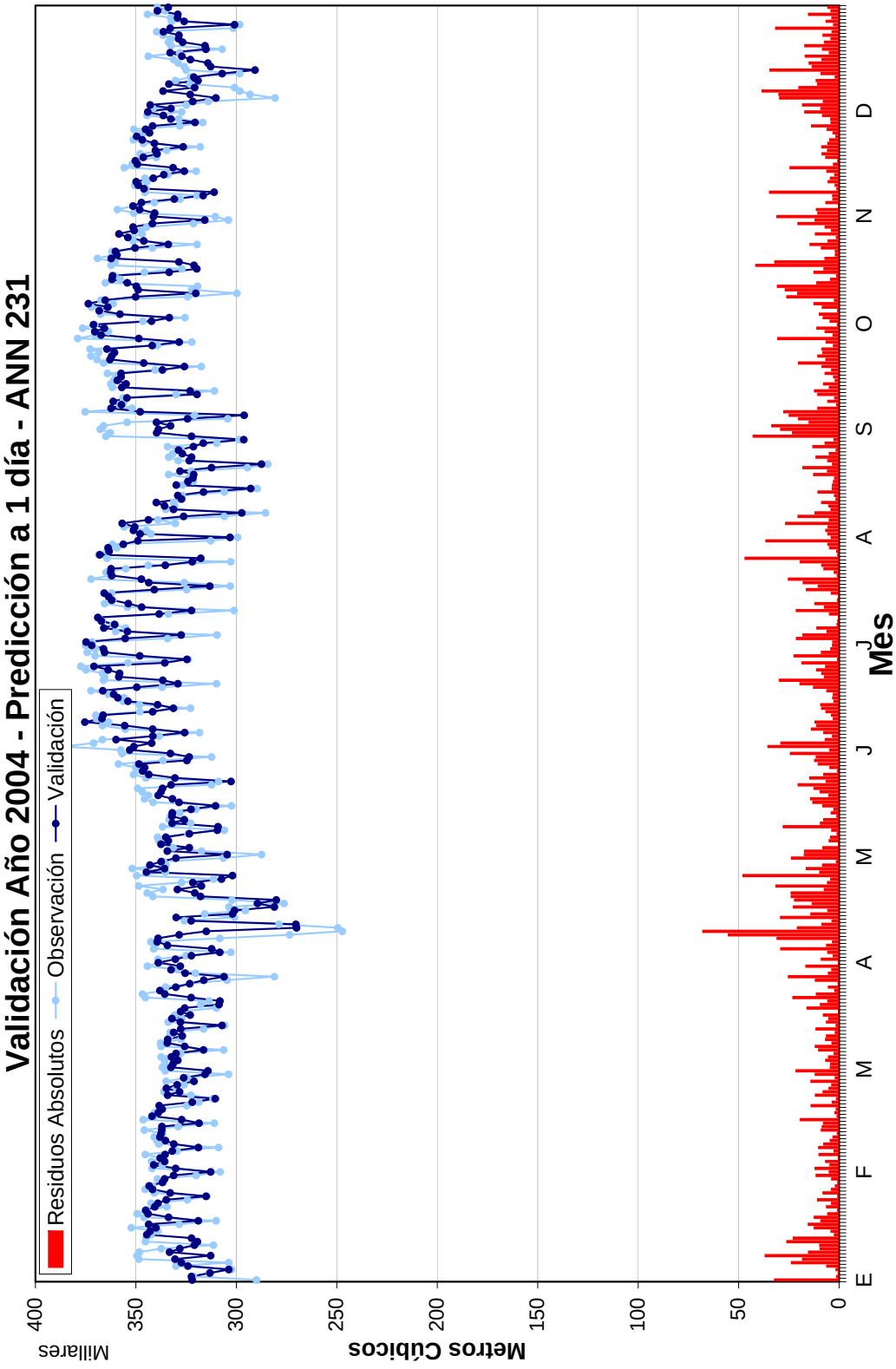


Figura 8.2: Gráfico de Validación vs. Observación, Año 2004. Predicción a 1 día Red 231

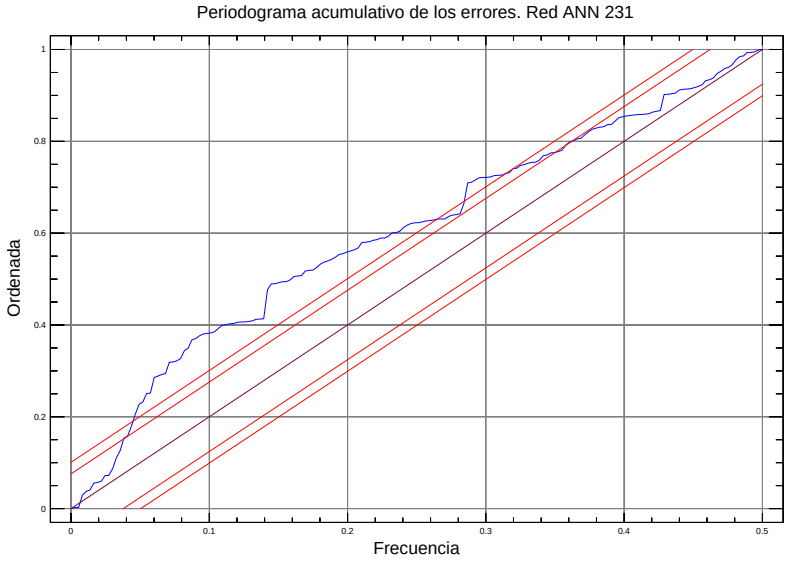


Figura 8.3: Periodograma acumulativo de los residuos, *ANN*231, predicción a 1 día

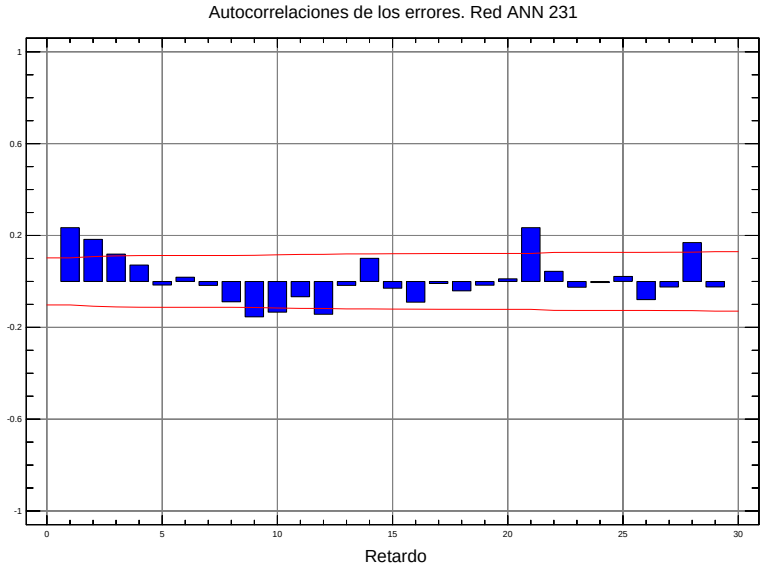


Figura 8.4: ACF de los errores, *ANN*231, predicción a 1 día

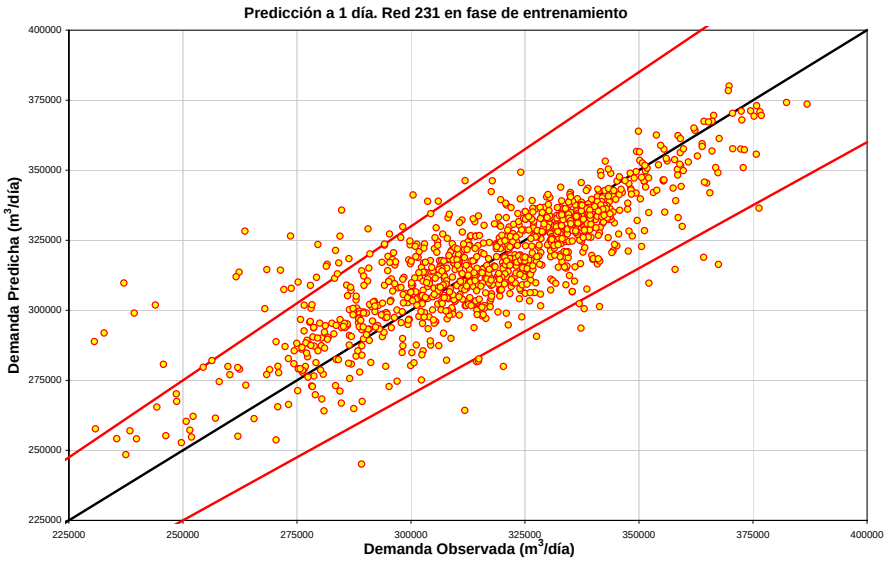


Figura 8.5: Gráfico de Demanda observada vs. Demanda Predicha en fase de entrenamiento, Red 231 y predicción a 1 día. Las líneas rojas representan los límites de errores $\pm$ 10% de la demanda observada

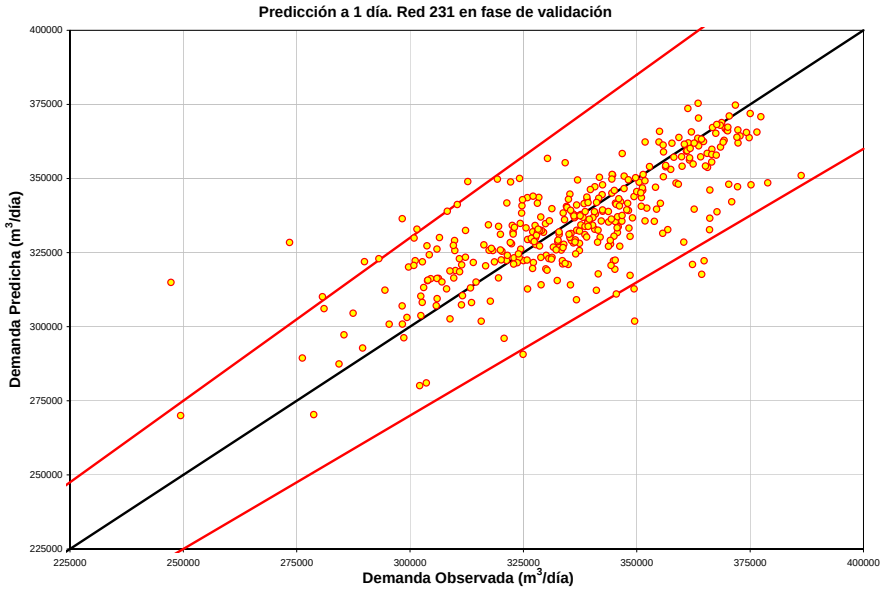


Figura 8.6: Gráfico de Demanda observada vs. Demanda Predicha en fase de Validación, Red 231 y predicción a 1 día. Las líneas rojas representan los límites de errores $\pm$ 10% de la demanda observada

8.3.2. *Red 441*

Una vez vistos los resultados de la red anterior y comprobado que los errores aún conservan rastros de autocorrelación según el ACF presentado. Se ha construido una red más compleja, *ANN* 441, con un vector de 4 entradas, 4 neuronas en la capa oculta y una neurona en la capa de salida, así como un vector que contiene el conjunto de datos objetivo correspondiente a cada día que se va a predecir. Con esta estructura de red, no solo se le proporciona información de los valores de la demanda registrada en 1 y 7 días antecedentes. Adicionalmente se le proporciona información cualitativa del ciclo semanal, con valores asignados de valor 1 a los días laborables (lunes a viernes) y valor 0 para los días del fin de semana (sábado y domingo). Finalmente se le proporciona un indicador de días no laborables regido por el calendario de festividades de la comunidad y que alteran el patrón semanal que presenta la demanda. De esta forma, los días como por ejemplo, San José (19 marzo), semana santa (fechas variables), etc. son considerados con valor 0, mientras que el resto de los días se les asigna el valor 1 (ver Apéndice B, página 317 del calendario de festividades de la ciudad de Valencia).

El cuadro 8.4 presenta las principales características de la red y la figura 8.7 presenta su topología. Se ha probado su desempeño para predecir la demanda diaria a un día.

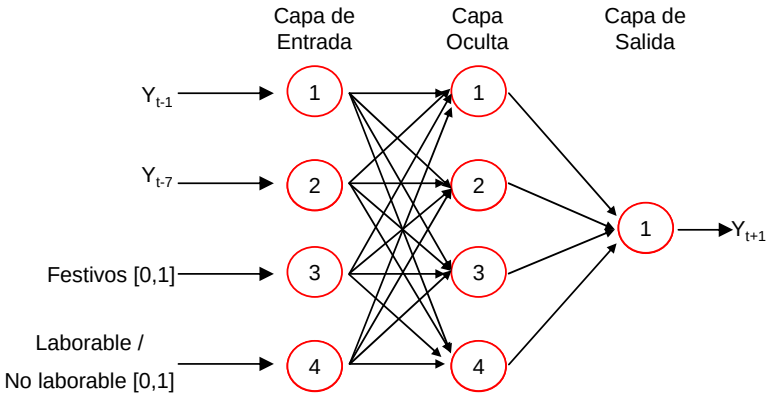


Figura 8.7: Estructura de la red neuronal 441

Dato	Característica
Entradas	4
Salidas	1
Capas	3
Neuronas capa oculta	4
Neuronas capa salida	1
Función de inicializado de red	Nguyen-Widrow
Función de transferencia	Tansigmoidal/lineal

Cuadro 8.4: Variables utilizadas en los vectores de entradas a la red neuronal *ANN*441

Red 441. Predicción a 1 día

Los resultados obtenidos se presentan en el cuadro 8.5 para las fases de entrenamiento y validación. En ellos destaca que se presenta una empeoramiento de los estadísticos en la fase de validación con respecto a la fase de entrenamiento, –hecho esperable–, lo mismo ocurre con los valores de los coeficientes de correlación r y de determinación R^2.

Estadístico	Periodo de Entrenamiento	Periodo de Validación
RMSE	10,922.64	11,359.23
MAE	7,775.30	8,332.94
MAPE	2.50	2.49
ME	0.00	3,859.20
MPE	-0.12	1.09
Error Máx. Abs.	67,682.83	42,232.96
r	0.90	0.88
R^2	0.81	0.78

Cuadro 8.5: Desempeño de la *ANN*441, estimación y validación a un día

La figura 8.8 presenta el desempeño de la red entrenada para las entradas correspondientes al año 2004 –conjunto de datos reservados para validación–. Es evidente que la red consigue captar la tendencia a largo plazo, el ciclo anual, así como también la periodicidad semanal de la serie de demandas. Los errores absolutos de la parte baja de la figura se mantienen en valores inferiores a 50,000 m^3 y la gran mayoría de los picos que se observaban con la red *ANN*231 han desaparecido. En esta ocasión, del análisis más a detalle de los errores, encontramos que la apariencia del periodograma acumulativo y el ACF de los errores (figuras 8.9 y 8.10) nos indica que los errores tienen la apariencia de un ruido blanco. Solamente se aprecia una leve autocorrelación en los retardos periódicos de 7 días.

Las figuras 8.11 y 8.12 presentan el desempeño de la red en las fases de entrenamiento y validación respectivamente. Las líneas rojas indican los límites de errores superiores e inferiores al 10% de la demanda observada. Los valores se mantienen en su mayoría dentro de la banda mencionada y se agrupan alrededor de la línea negra que representa una predicción cien por ciento correcta.

Con este análisis gráfico y estadístico podemos concluir someramente, que las variables categóricas incluidas en el vector de entradas han sido de utilidad y producen una mejora tanto en la fase de validación como en la de

entrenamiento en términos de los estadísticos del cuadro 8.5 y finalmente en la apariencia de los diferentes gráficos presentados.

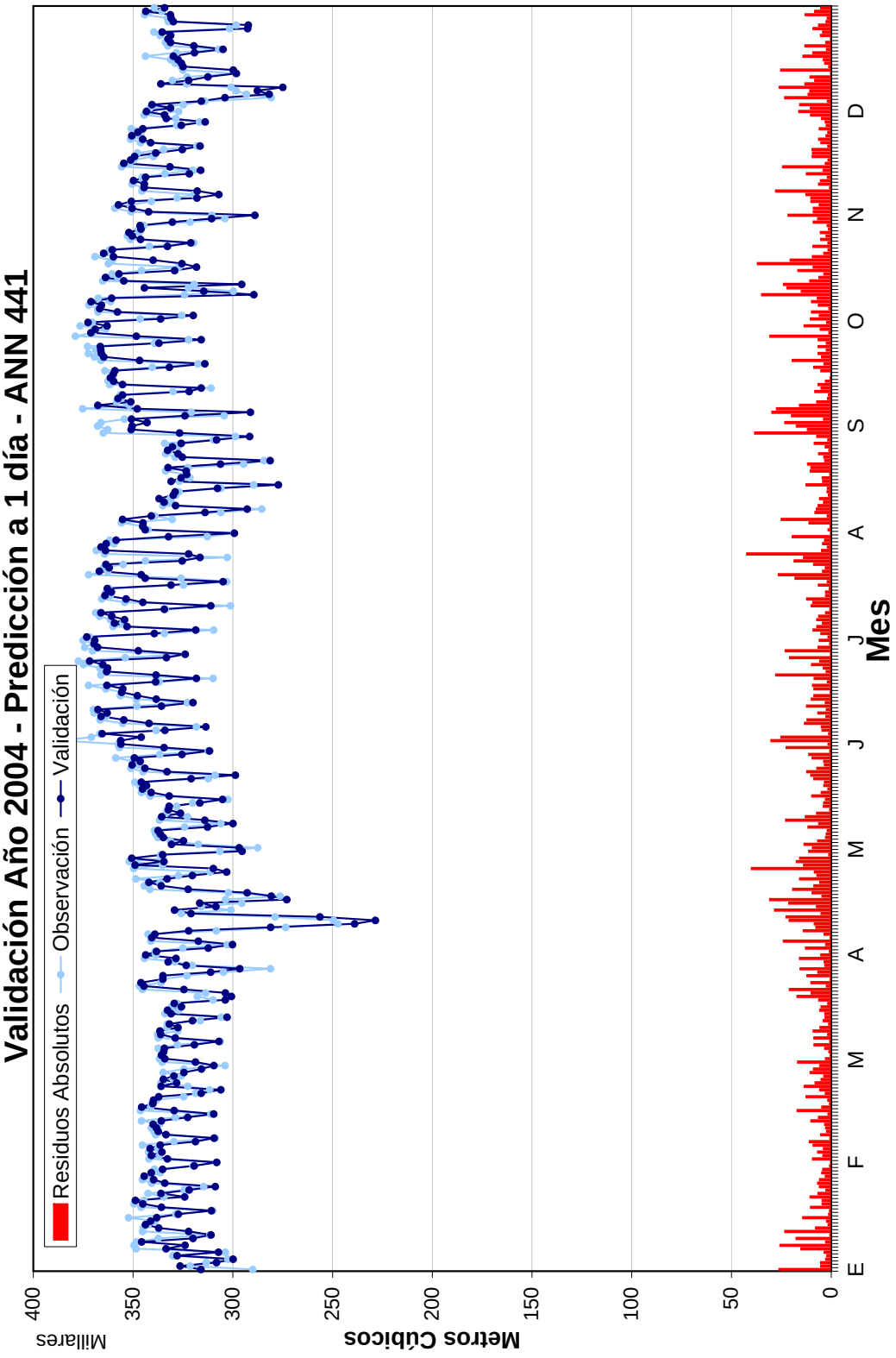


Figura 8.8: Gráfico de Validación vs. Observación, Año 2004. Predicción a 1 día Red 441

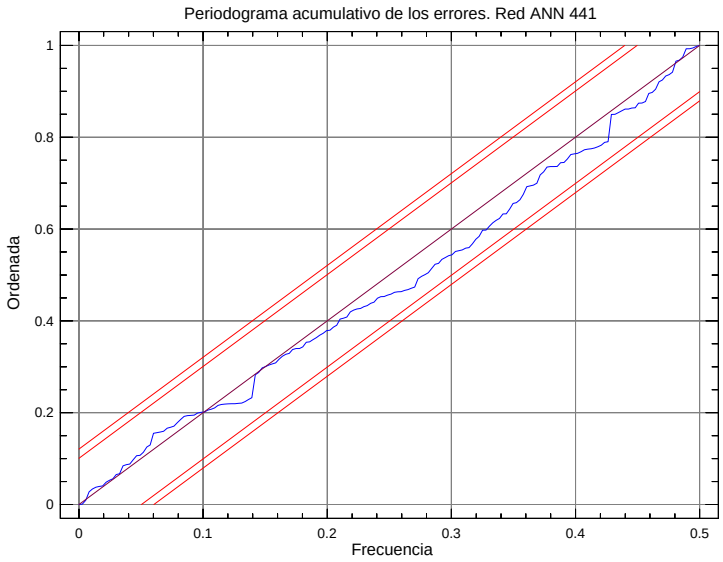


Figura 8.9: Periodograma acumulativo de los residuos, *ANN*441, predicción a 1 día

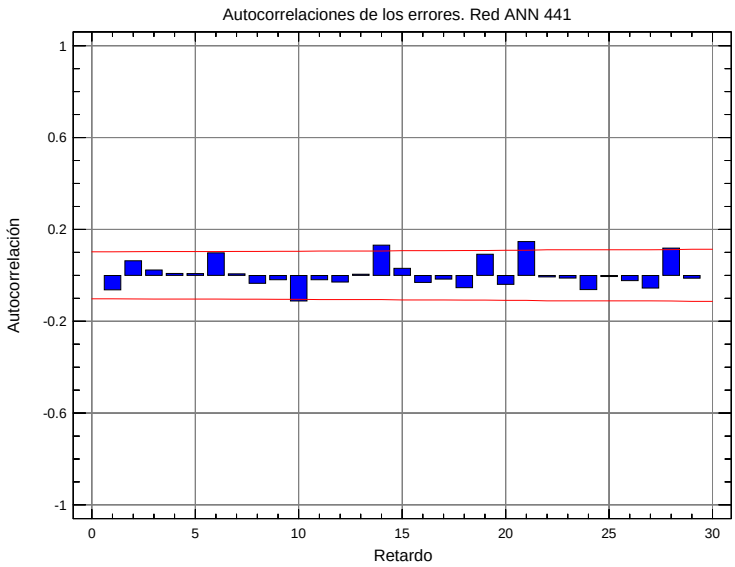


Figura 8.10: ACF de los errores, *ANN*441, predicción a 1 día

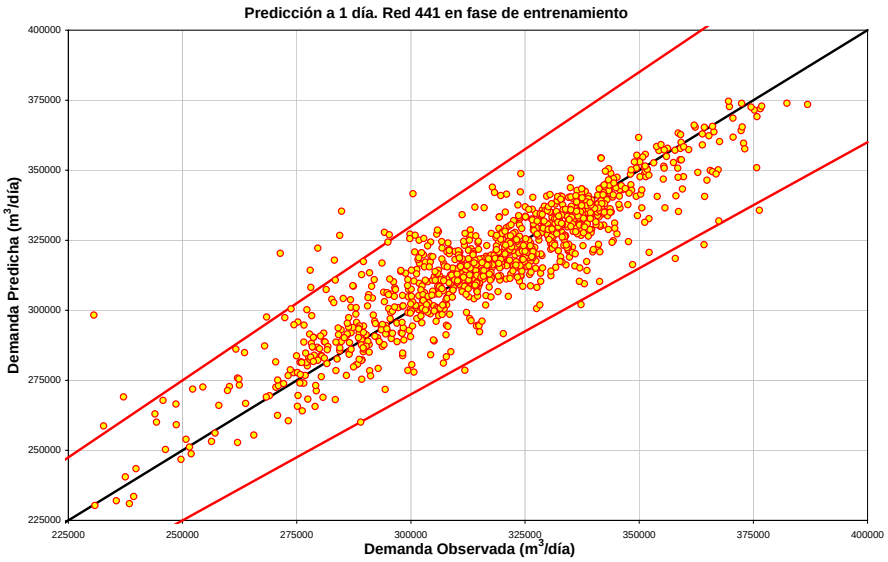


Figura 8.11: Gráfico de Demanda observada vs. Demanda Predicha en fase de entrenamiento, Red 441 y predicción a 1 día. Las líneas rojas representan los límites de errores $\pm$ 10% de la demanda observada

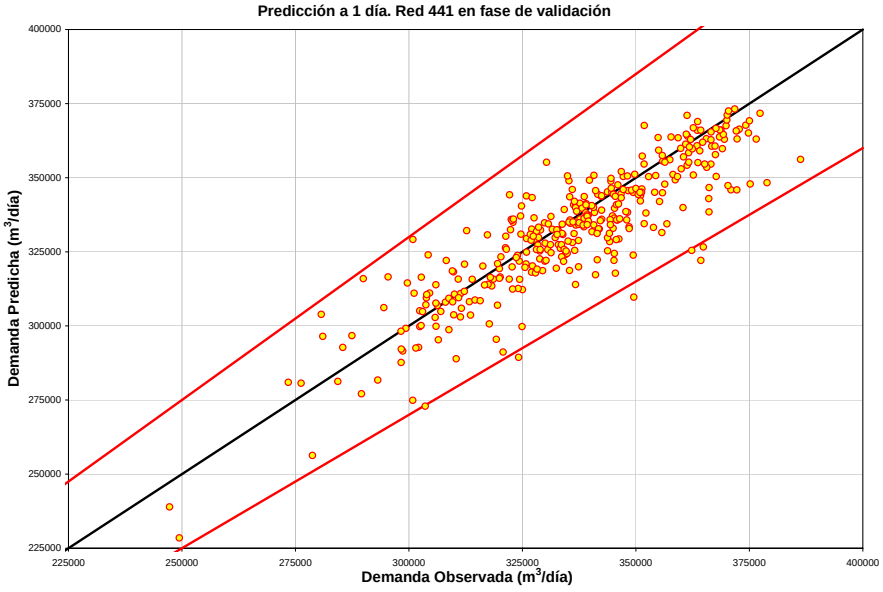


Figura 8.12: Gráfico de Demanda observada vs. Demanda Predicha en fase de Validación, Red 441 y predicción a 1 día. Las líneas rojas representan los límites de errores $\pm$ 10% de la demanda observada

8.3.3. *Red 451*

Si bien los resultados de la red anterior (*ANN* 441) han mejorado sensiblemente respecto a la inicial (*ANN* 231), se ha buscado una red un grado más compleja y se le ha añadido una neurona adicional en la capa oculta. La intención es evaluar si la adición de una neurona resulta en una mejora en las predicciones realizadas.

Una vez vistos los resultados de la red anterior y comprobado que los errores aún conservan rastros de autocorrelación en los retardos periódicos según el ACF presentado. Se ha construido una red más compleja, *ANN* 451, con un vector de 4 entradas, 5 neuronas en la capa oculta y una neurona en la capa de salida, así como un vector que contiene el conjunto de datos objetivo correspondiente a cada día que se va a predecir. El vector de entradas conserva la estructura anterior.

El cuadro 8.6 presenta las principales características de la red y la figura 8.13 presenta su topología. Se ha probado su desempeño para predecir la demanda diaria a un día.

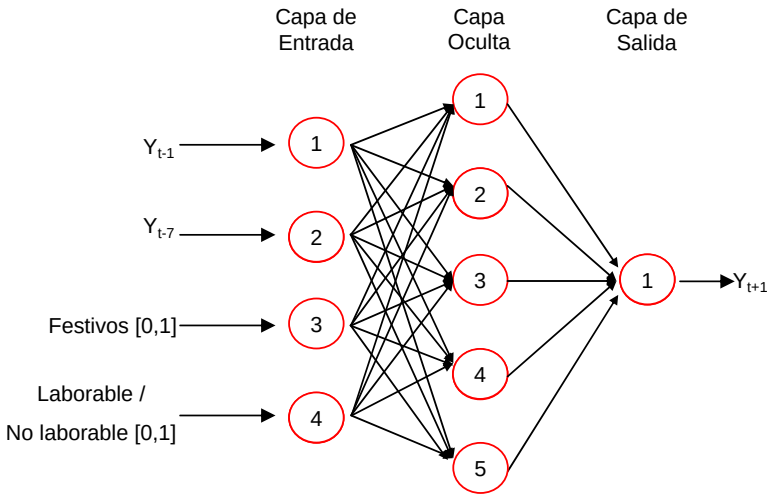


Figura 8.13: Estructura de la red neuronal 451

Dato	Característica
Entradas	4
Salidas	1
Capas	3
Neuronas capa oculta	5
Neuronas capa salida	1
Función de inicializado de red	Nguyen-Widrow
Función de transferencia	Tansigmoidal/lineal

Cuadro 8.6: Variables utilizadas en los vectores de entradas a la red neuronal *ANN*451

Red 451. Predicción a 1 día

Los resultados obtenidos se presentan en el cuadro 8.7 para las fases de entrenamiento y validación. Si se comparan los estadísticos de los errores con los de la red *ANN*441, se puede apreciar una mejoría en términos de RMSE, MAE, MAPE, ME y MPE. Sin embargo los coeficientes de correlación y de determinación no han mejorado. Con estos resultados no se podría argumentar que el aumento del número de neuronas más allá de 4 en la capa oculta, produzca mejores resultados en este caso en particular.

Estadístico	Periodo de Entrenamiento	Periodo de Validación
RMSE	10,841.71	11,459.44
MAE	7,686.65	8,234.82
MAPE	2.47	2.46
ME	0.00	3,565.81
MPE	-0.12	1.00
Error Máx. Abs.	68,081.78	45,260.81
r	0.90	0.88
R^2	0.82	0.78

Cuadro 8.7: Desempeño de la *ANN*451, estimación y validación a un día

La figura 8.14 presenta el desempeño de la red entrenada para las entradas correspondientes al año 2004. Al igual que en los anteriormente presentados, es evidente que la red capta correctamente los patrones y ciclos presentes en la serie de demandas diarias de la ciudad de Valencia. Los residuos mantienen una buena apariencia, siempre en valores mínimos. En el análisis más a detalle de los errores, encontramos que la apariencia del periodograma acumulativo y la ACF de los errores (figuras 8.15 y 8.16) nos indica que los errores tienen la apariencia de un ruido blanco. La autocorrelación en los retardos periódicos de 7 días aumenta con esta nueva estructura de red.

Las figuras 8.17 y 8.18 presentan el desempeño de la red en las fases de entrenamiento y validación respectivamente. Las líneas rojas indican los límites de errores superiores e inferiores al 10% de la demanda observada. Los valores se mantienen en su mayoría dentro de la banda mencionada y se agrupan alrededor de la línea negra que representa una predicción cien por ciento correcta.

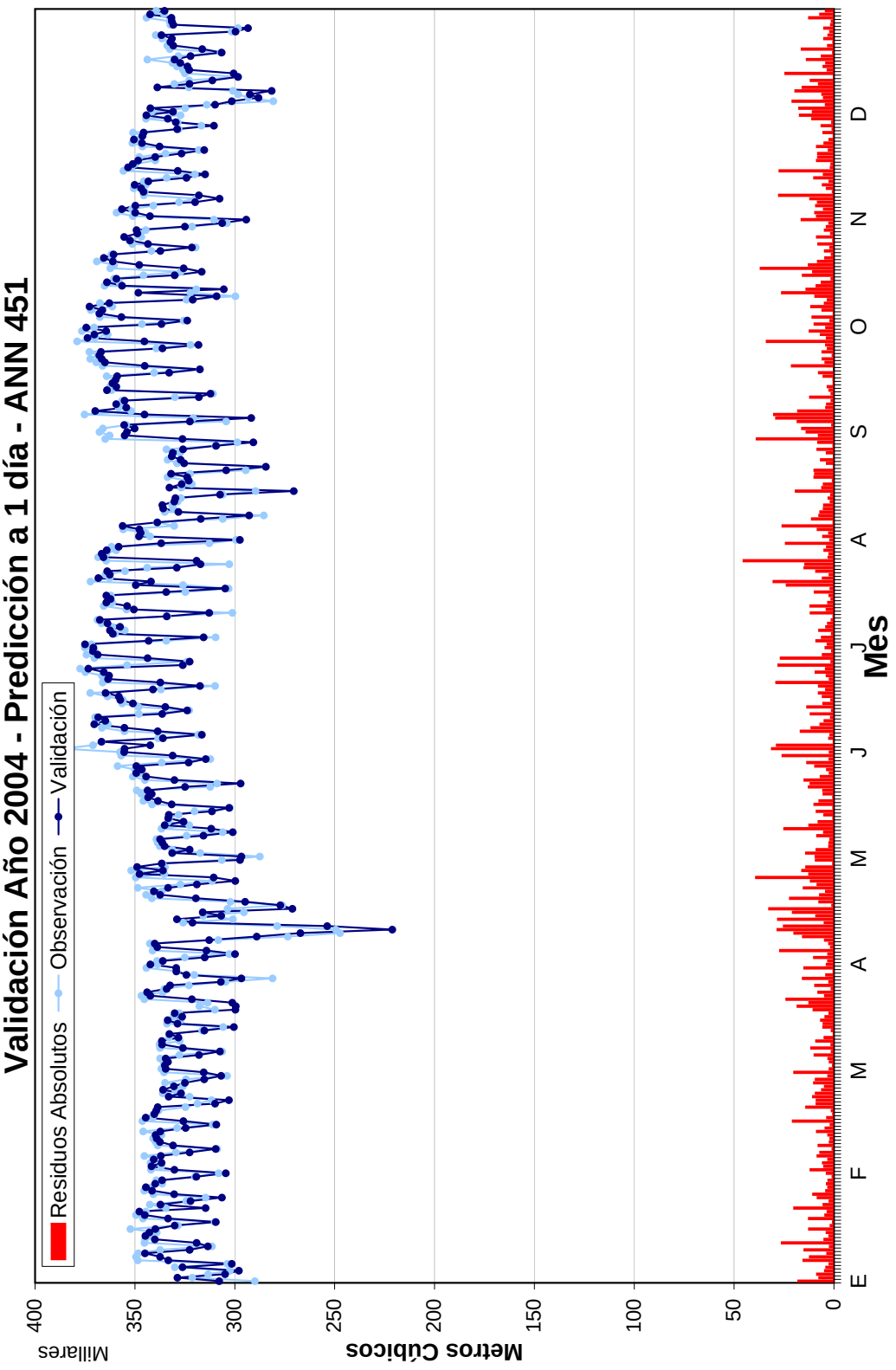


Figura 8.14: Gráfico de Validación vs. Observación, Año 2004. Predicción a 1 día Red 451

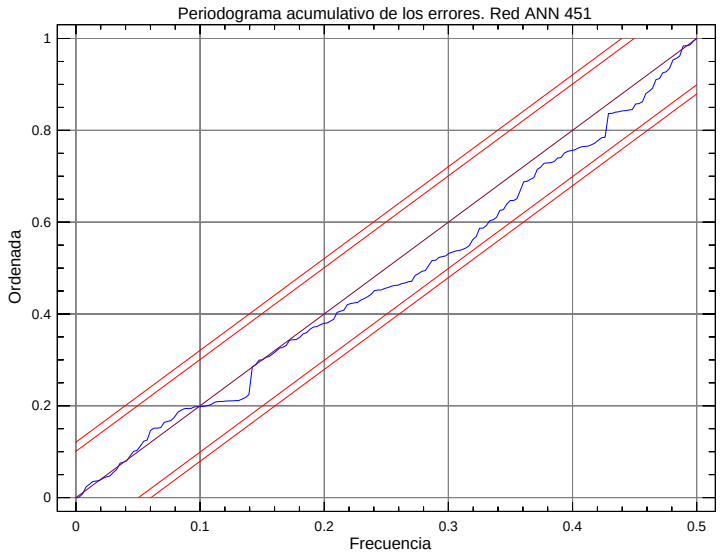


Figura 8.15: Periodograma acumulativo de los residuos, *ANN*451, predicción a 1 día

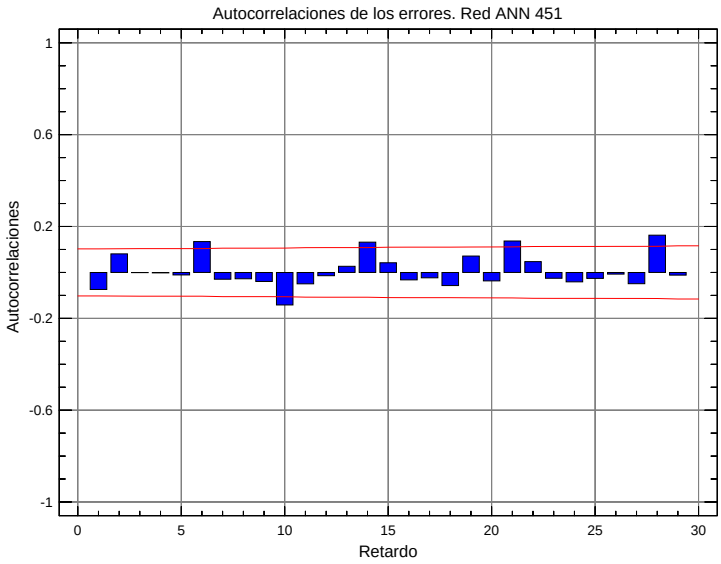


Figura 8.16: ACF de los errores, *ANN*451, predicción a 1 día

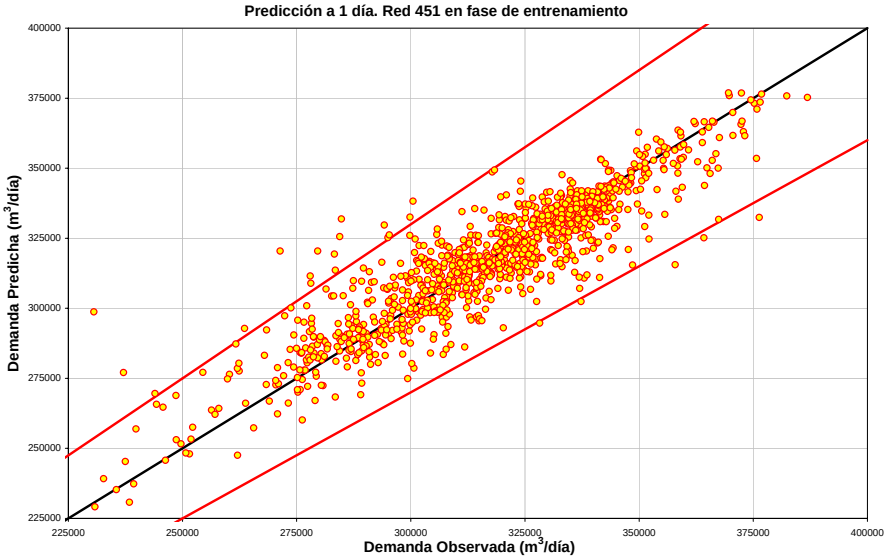


Figura 8.17: Gráfico de Demanda observada vs. Demanda Predicha en fase de entrenamiento, Red 451 y predicción a 1 día. Las líneas rojas representan los límites de errores $\pm$ 10% de la demanda observada

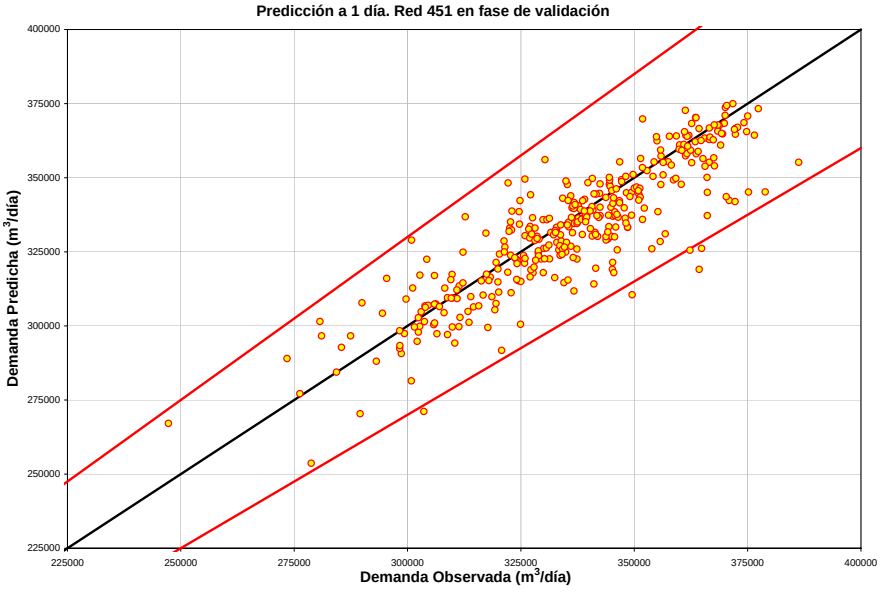


Figura 8.18: Gráfico de Demanda observada vs. Demanda Predicha en fase de Validación, Red 451 y predicción a 1 día. Las líneas rojas representan los límites de errores $\pm$ 10% de la demanda observada

8.3.4. *Red 551*

Los resultados de la red anterior (*ANN* 451) no han sido mejores que los de la red que cuenta con una neurona menos en la capa oculta. En esta nueva red que analizaremos a continuación, hemos conservado las 5 neuronas en la capa oculta y se ha incluido una nueva variable cualitativa al vector de entradas. La nueva variable intenta informar a la red, de la disminución que sufre la demanda durante los meses de agosto. Ya se ha comentado en capítulos anteriores que la demanda disminuye durante este mes principalmente por la suspensión de labores de muchas empresas por el periodo vacacional. En muchos casos los habituales residentes abandonan la ciudad hacia poblaciones aledañas. Se le ha asignado el valor 0 a los 31 días de los meses de agosto y el valor 1 al resto de los días. El cuadro 8.8 presenta las principales características de la red y la figura 8.13 presenta su topología. Se ha probado su desempeño para predecir la demanda diaria a un día.

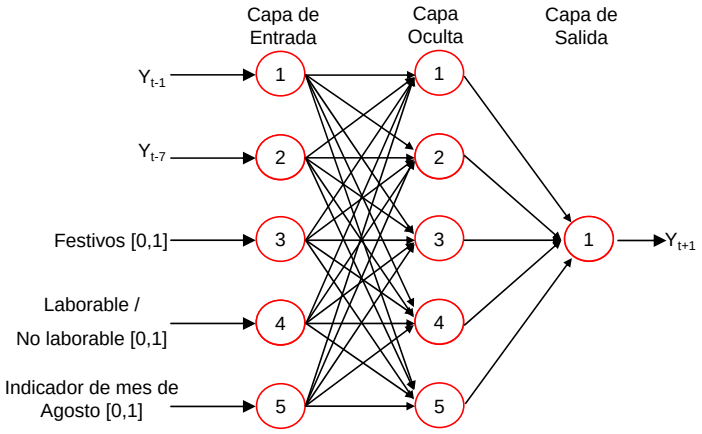


Figura 8.19: Estructura de la red neuronal 551

Dato	Característica
Entradas	5
Salidas	1
Capas	3
Neuronas capa oculta	5
Neuronas capa salida	1
Función de inicializado de red	Nguyen-Widrow
Función de transferencia	Tansigmoidal/lineal

Cuadro 8.8: Variables utilizadas en los vectores de entradas a la red neuronal *ANN*551

Red 551. Predicción a 1 día

Los resultados obtenidos se presentan en el cuadro 8.9 para las fases de entrenamiento y validación. De la comparación de los estadísticos de los errores con los de las redes *ANN*441 y 551, no se aprecian mejores resultados. Los coeficientes de correlación y de determinación tampoco han mejorado, por lo que la inclusión de una quinta neurona y la variable categórica indicativa de los meses de agosto, no parecen afectar positivamente el desempeño de la red neuronal *ANN* 551.

Estadístico	**Periodo de Entrenamiento**	**Periodo de Validación**
RMSE	10,397.26	11,556.67
MAE	7,453.43	8,381.00
MAPE	2.39	2.50
ME	0.00	3,949.43
MPE	-0.11	1.12
Error Máx. Abs.	51,855.07	61,913.09
r	0.91	0.88
R^2	0.83	0.78

Cuadro 8.9: Desempeño de la *ANN*551, estimación y validación a un día

La figura 8.20 presenta el desempeño de la red entrenada para las entradas correspondientes al año 2004. Los residuos conservan una buena apariencia, sin embargo algún valor supera los 50,000 m^3, cosa que no ocurría con las configuraciones anteriores.

Los residuos mantienen una buena apariencia, siempre en valores mínimos. En el análisis más a detalle de los errores, encontramos que la apariencia del periodograma acumulativo y la ACF de los errores (figuras 8.21 y 8.22) nos indica que los errores tienen la apariencia de un ruido blanco. La apariencia del autocorrelograma de los residuos no ha mejorado. Finalmente, las figuras 8.23 y 8.24 presentan el desempeño de la red en las fases de entrenamiento y validación respectivamente, presentando resultados muy similares a las redes anteriores.

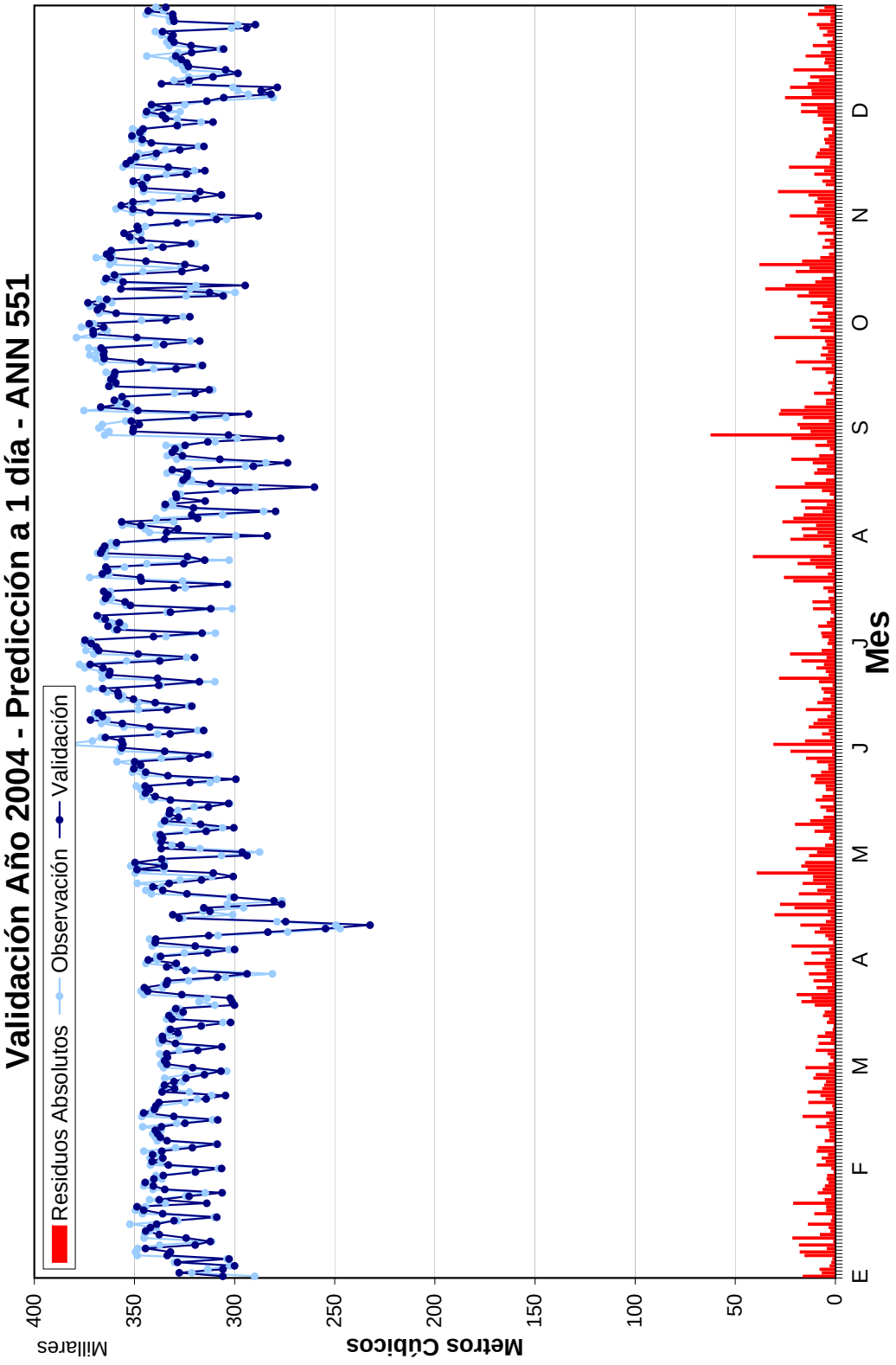


Figura 8.20: Gráfico de Validación vs. Observación, Año 2004. Predicción a 1 día Red 551

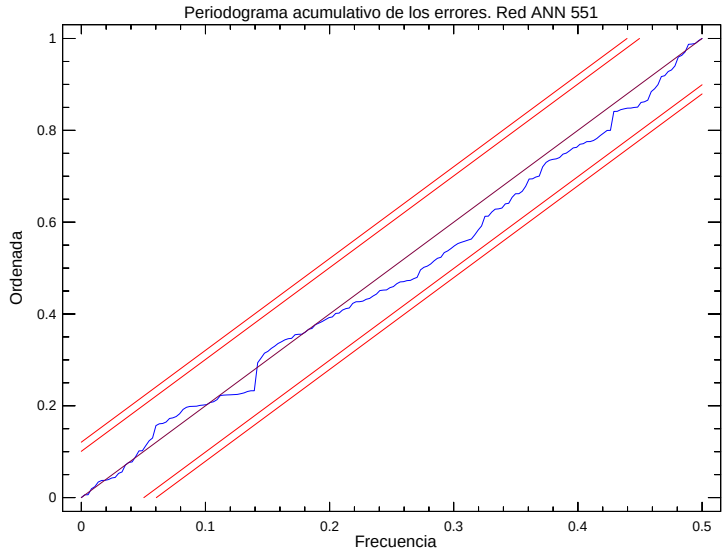


Figura 8.21: Periodograma acumulativo de los residuos, *ANN*551, predicción a 1 día

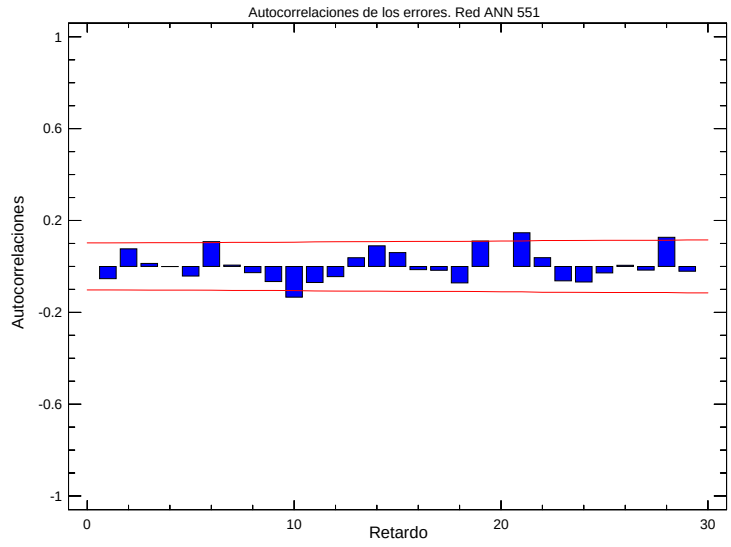


Figura 8.22: ACF de los errores, *ANN*551, predicción a 1 día

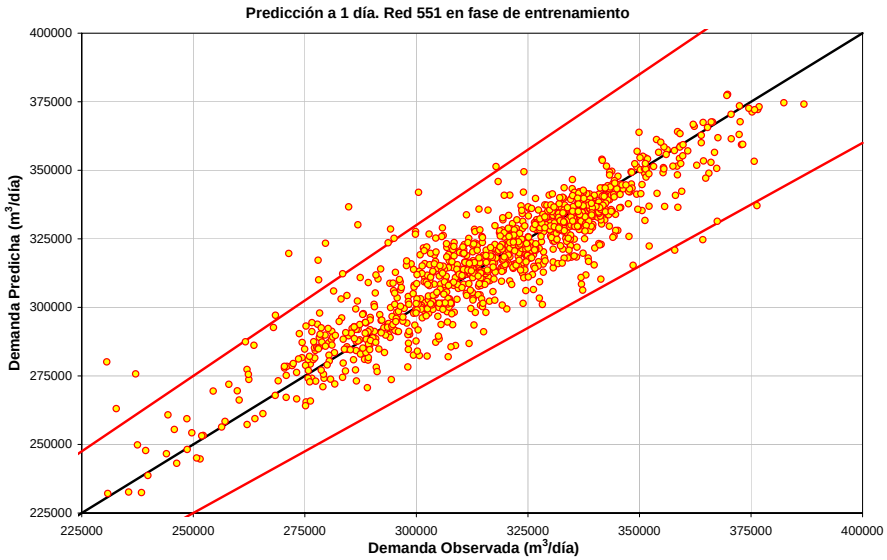


Figura 8.23: Gráfico de Demanda observada vs. Demanda Predicha en fase de entrenamiento, Red 551 y predicción a 1 día. Las líneas rojas representan los límites de errores $\pm$ 10% de la demanda observada

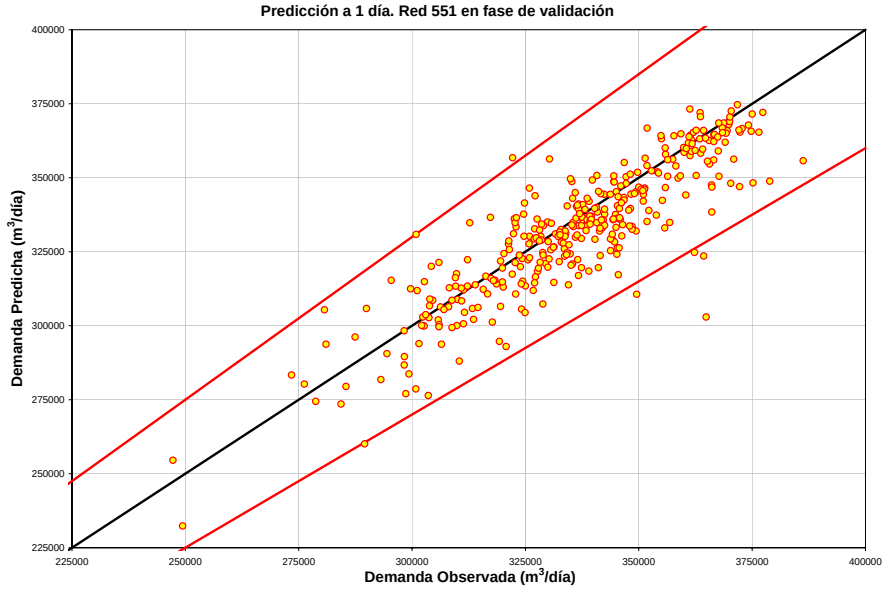


Figura 8.24: Gráfico de Demanda observada vs. Demanda Predicha en fase de Validación, Red 551 y predicción a 1 día. Las líneas rojas representan los límites de errores $\pm$ 10% de la demanda observada

8.3.5. *Red 541*

Con esta nueva arquitectura de red se pretende analizar los resultados que se obtienen utilizando el vector de entradas con 5 variables y 4 neuronas en la capa oculta. El cuadro 8.10 presenta las principales características de la red utilizada y la figura 8.25 presenta su topología. Se ha probado su desempeño para predecir la demanda diaria a un día.

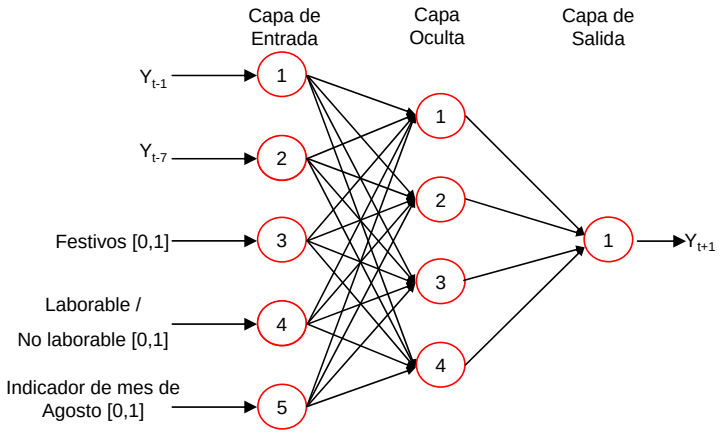


Figura 8.25: Estructura de la red neuronal 541

Dato	Característica
Entradas	5
Salidas	1
Capas	3
Neuronas capa oculta	4
Neuronas capa salida	1
Función de inicializado de red	Nguyen-Widrow
Función de transferencia	Tansigmoidal/lineal

Cuadro 8.10: Variables utilizadas en los vectores de entradas a la red neuronal *ANN*541

Red 541. Predicción a 1 día

Los resultados obtenidos se presentan en el cuadro 8.11 para las fases de entrenamiento y validación. La reducción de una neurona en la capa oculta con respecto a la red anteriormente analizada (*ANN*551) conservando las características del vector de entradas, no indica un deterioro de las predicciones más bien lo contrario. Sin embargo la mejora es muy leve y los resultados no son mejores que la red *ANN* 451 más simple y que no incluye la quinta variable indicativa del mes de agosto.

Estadístico	Periodo de Entrenamiento	Periodo de Validación
RMSE	10,577.14	11,419.78
MAE	7,605.18	8,283.65
MAPE	2.45	2.47
ME	0.00	3,751.88
MPE	-0.11	1.06
Error Máx. Abs.	53,718.16	63,132.19
r	0.91	0.88
R^2	0.83	0.78

Cuadro 8.11: Desempeño de la *ANN*541, estimación y validación a un día

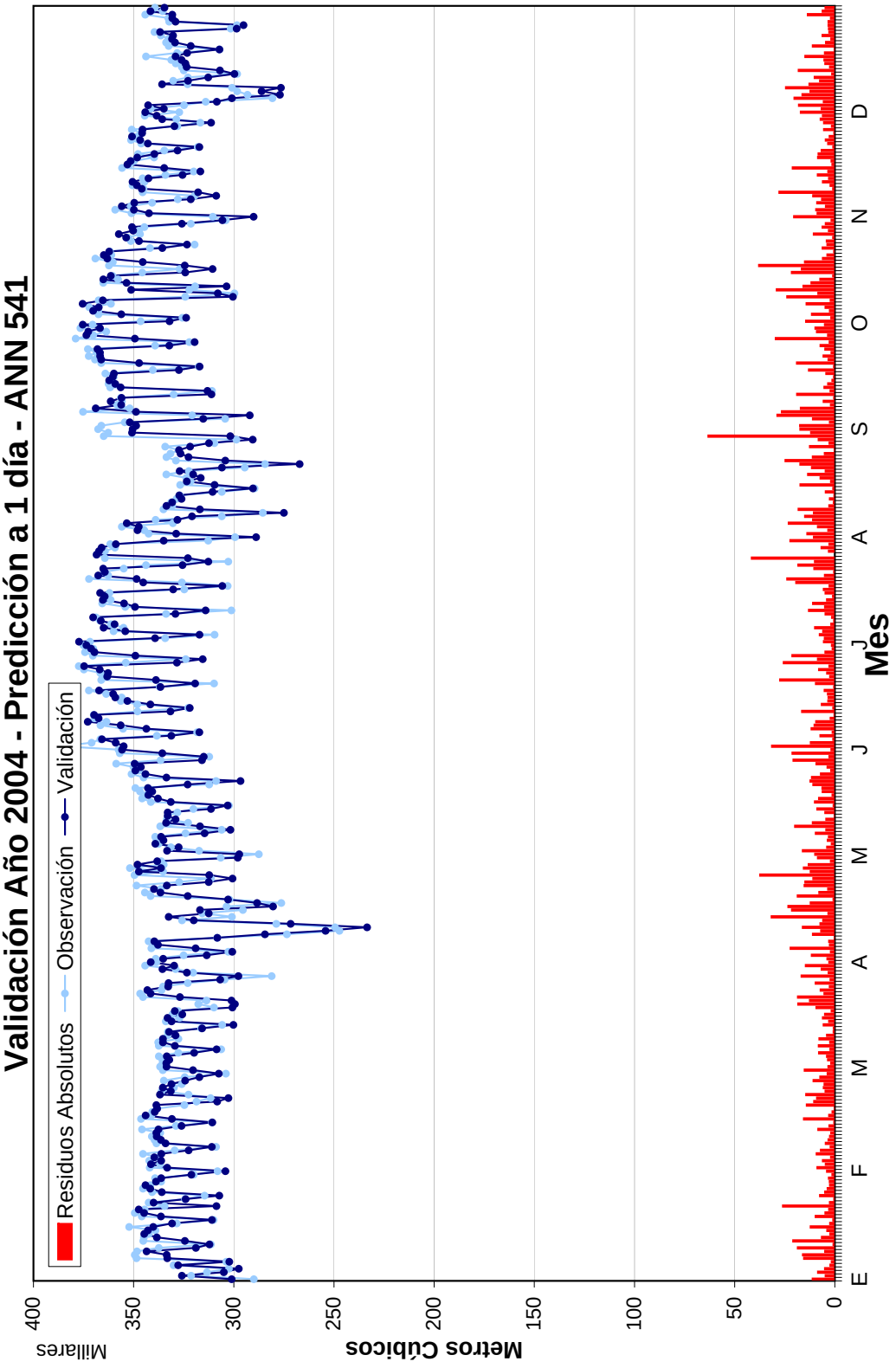


Figura 8.26: Gráfico de Validación vs. Observación, Año 2004. Predicción a 1 día Red 541

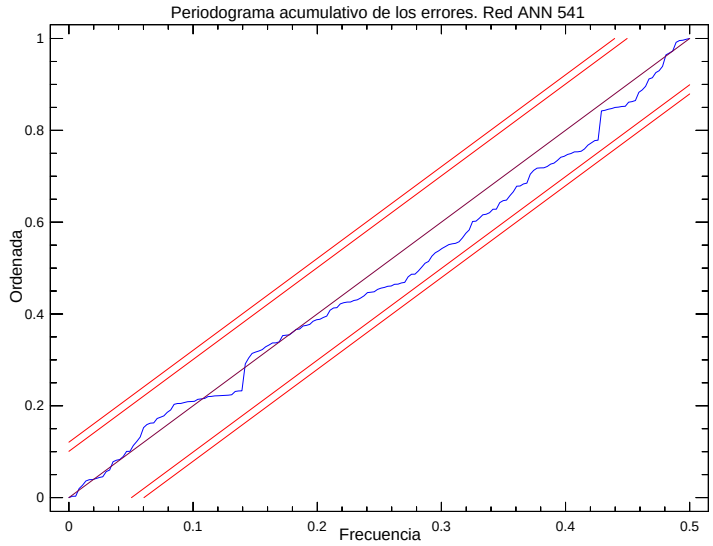


Figura 8.27: Periodograma acumulativo de los residuos, *ANN*541, predicción a 1 día

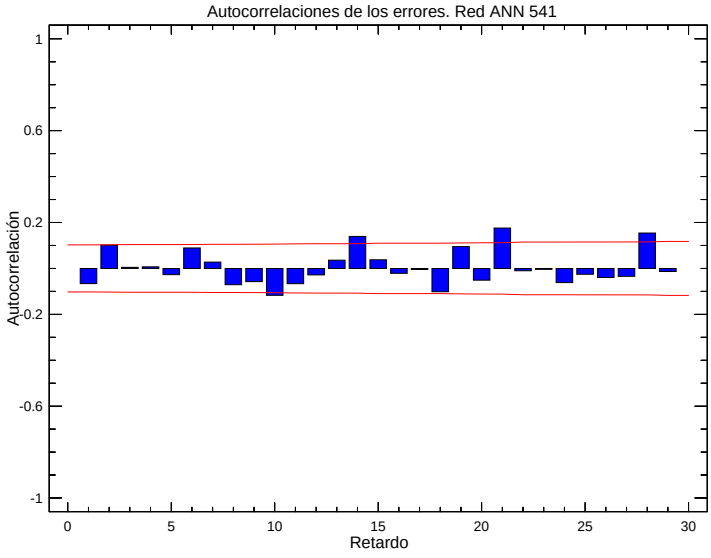


Figura 8.28: ACF de los errores, *ANN*541, predicción a 1 día

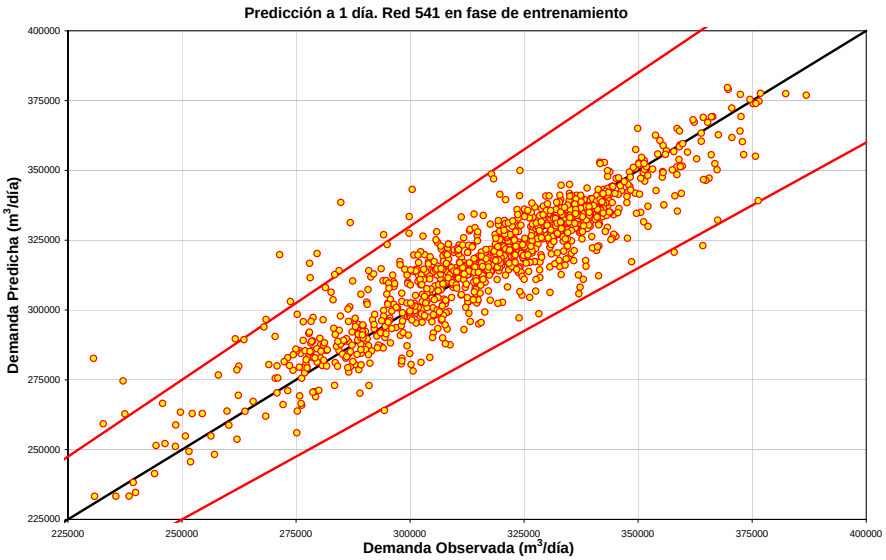


Figura 8.29: Gráfico de Demanda observada vs. Demanda Predicha en fase de entrenamiento, Red 541 y predicción a 1 día. Las líneas rojas representan los límites de errores $\pm$ 10% de la demanda observada

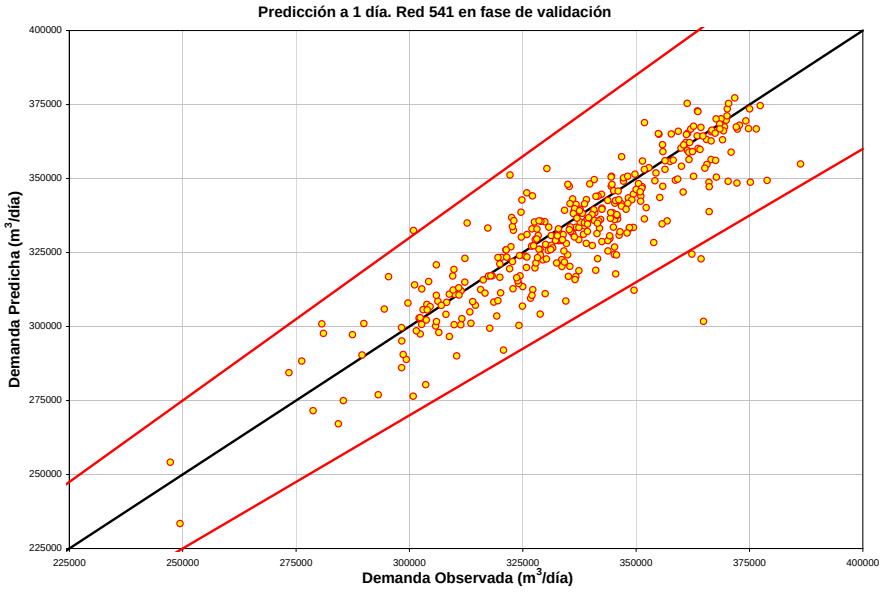


Figura 8.30: Gráfico de Demanda observada vs. Demanda Predicha en fase de Validación, Red 541 y predicción a 1 día. Las líneas rojas representan los límites de errores $\pm$ 10% de la demanda observada

8.3.6. *Red 541 incluyendo la temperatura máxima*

El análisis conjunto de la temperatura y la demanda de agua urbana, realizado en la sección 7.3.3, nos indica que la temperatura podría aportar información que resulte en una mejora del desempeño de las redes neuronales hasta ahora analizadas. En esta nueva red, repetiremos la estructura de la red anterior, reemplazando el indicador del mes de agosto que no reportó mejoras en el desempeño, por la temperatura máxima diaria observada en la ciudad de Valencia en una estación. Las principales características de esta red son idénticas a la anterior por lo que se omite el cuadro (ver 8.10). La figura 8.31 presenta su topología. Se ha probado su desempeño para predecir la demanda diaria a un día.

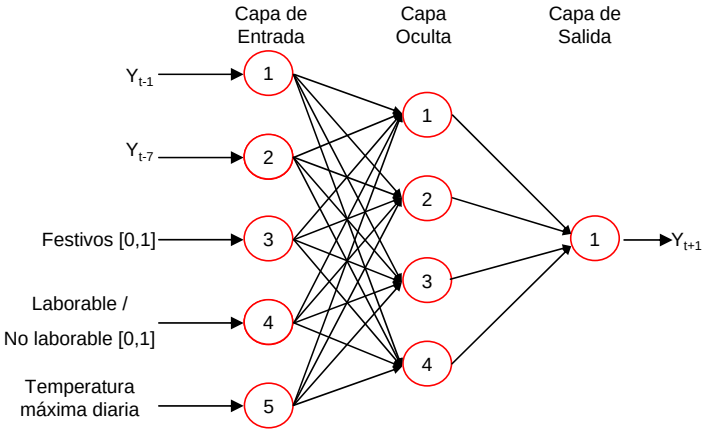


Figura 8.31: Estructura de la red neuronal 541 que incluye la temperatura máxima en el vector de entradas

Red 541 + Tmax. Predicción a 1 día

Los resultados obtenidos se presentan en el cuadro 8.12 para las fases de entrenamiento y validación. El desempeño de la red para el año de validación se observa en el gráfico 8.32. De la comparación de los estadísticos de los errores con los obtenidos con las redes anteriormente analizadas, no se observan grandes mejoras. Algunos estadísticos han mejorado levemente pero otros son peores. Por la parte de los coeficientes de correlación y de determinación no hay ninguna mejoría relevante y se mantienen como los anteriores. En cuanto a los gráficos del periodograma acumulativo de los errores, el autocorrelograma (figuras 8.33 y 8.34) así como los gráficos del desempeño de la red en fase de entrenamiento y validación (8.35 y 8.36) presentan características similares a las de las últimas redes analizadas.

Estadístico	Periodo de Entrenamiento	Periodo de Validación
RMSE	10,947.54	11,428.93
MAE	7,612.77	8,221.85
MAPE	2.44	2.44
ME	3.96	4,087.98
MPE	-0.12	1.12
Error Máx. Abs.	67,658.60	50,416.62
r	0.90	0.88
R^2	0.81	0.78

Cuadro 8.12: Desempeño de la *ANN*541 que incluye la temperatura máxima en el vector de entrada, estimación y validación a un día

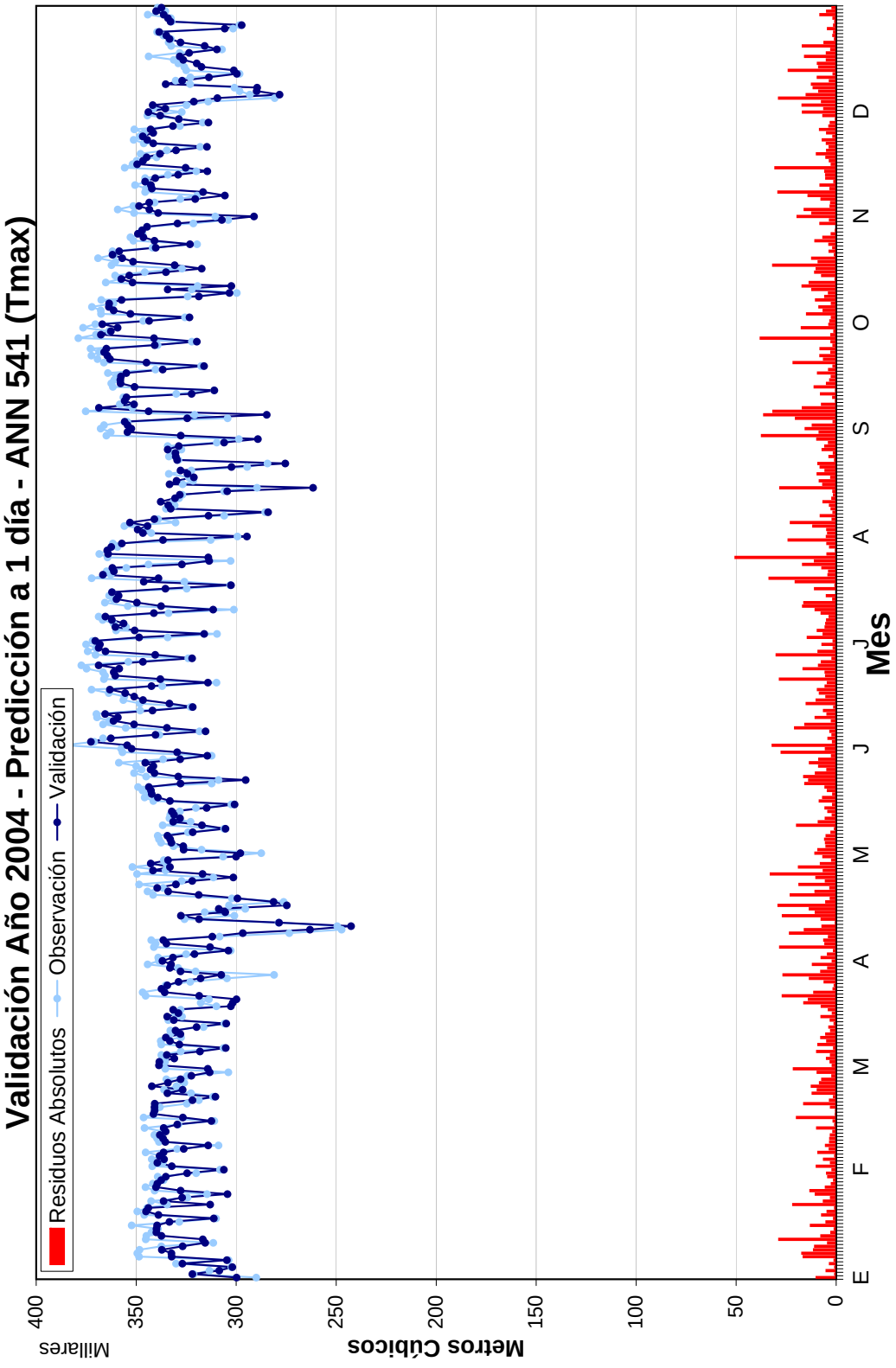


Figura 8.32: Gráfico de Validación vs. Observación, Año 2004. Predicción a 1 día Red 541 + Tmax

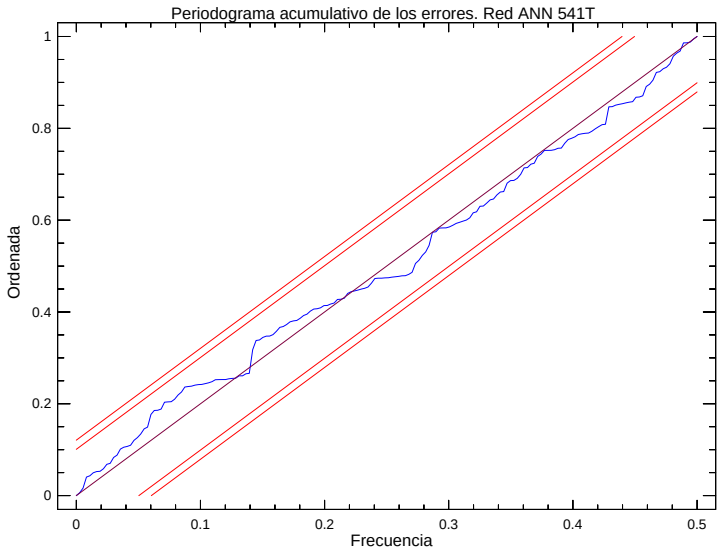


Figura 8.33: Periodograma acumulativo de los residuos, *ANN*541 + Tmax, predicción a 1 día

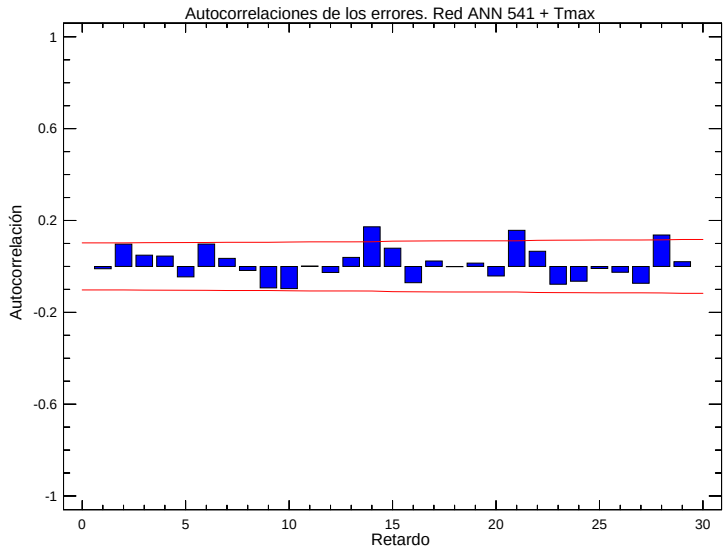


Figura 8.34: ACF de los errores, *ANN*541 + Tmax, predicción a 1 día

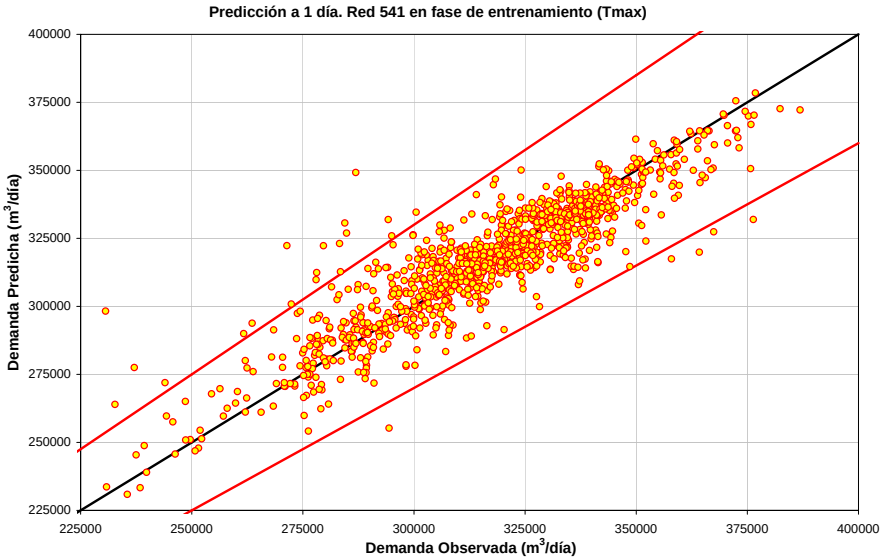


Figura 8.35: Gráfico de Demanda observada vs. Demanda Predicha en fase de entrenamiento, Red 541 + Tmax y predicción a 1 día. Las líneas rojas representan los límites de errores $\pm$ 10 % de la demanda observada

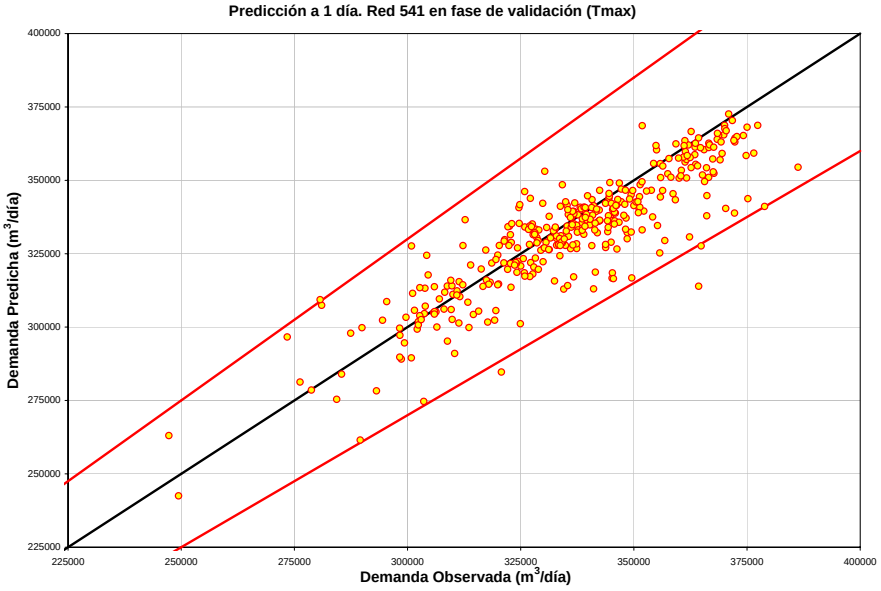


Figura 8.36: Gráfico de Demanda observada vs. Demanda Predicha en fase de Validación, Red 541 + Tmax y predicción a 1 día. Las líneas rojas representan los límites de errores $\pm$ 10 % de la demanda observada

8.3.7. *Comparación del desempeño de ANNs*

A lo largo de esta sección se han analizado 6 redes neuronales diferentes. Se diferencian entre sí tanto por la dimensión del vector de entradas, como por la cantidad de neuronas incluidas en la capa oculta. Se utilizaron vectores de entrada de dimensión 2, 4 y 5, con valores antecedentes de la demanda, variables climáticas (temperatura máxima) y valores binarios para representar variables categóricas. Los resultados obtenidos para todas las redes se presentan resumidos en el cuadro 8.13.

Se inició con la red más sencilla, y a lo largo de la sección se aumentó gradualmente su complejidad. La primera que fue entrenada y su desempeño evaluado, fue la *ANN231*, que utiliza solamente los valores pasados de la demanda de agua, utilizando el último dato de demanda disponible y el de 7 días antes. Con esta red se obtuvieron valores de $R^2 = 0{,}603$.

La siguiente red analizada fue la *ANN441* que incorporó en su vector de entradas una variable binaria indicativa de los días de lunes a viernes (1) y los de fin de semana, sábado y domingo (0), así como también otra variable binaria indicativa de los días de festividades relevantes de la ciudad (0), y días sin festividades (1). Con este nuevo vector de entradas se obtuvo un $R^2 = 0{,}78$, mejorando notablemente a la red inicial, no solo en términos del coeficiente de determinación sino también en lo que respecta al RMSE, MAE y MAPE. Igualmente la apariencia de los gráficos de autocorrelación y periodograma acumulativo de los errores mejoró, eliminando la correlación que se observaba en los retardos 1 y 2, y conteniendo dentro de los límites críticos del 95 % y 99 % a los valores acumulativos del periodograma. De la misma forma los gráficos de desempeño presentan mejoras importantes. Podemos concluir pues, que la inclusión de las variables categóricas binarias indicativas del periodo semanal, así como de las festividades locales y nacionales aportan información relevante a las *ANNs* que resultan en mejoras evidentes en las predicciones obtenidas.

Posteriormente se entrenaron redes neuronales en las cuales se aumentó una neurona en la capa oculta (*ANN451*), u otra en la que se incluyó una variable adicional indicativa de los meses de agosto (durante el cual se presenta una reducción importante de la demanda diaria) en el vector de entrada (*ANN551*), sin observar mejoras que justificasen su inclusión, más bien al contrario, los errores máximos absolutos se vuelven más grandes. De la misma forma se probó este mismo vector con una red con 4 neuronas en la capa oculta (*ANN541*) encontrando resultados muy similares a la red anteriormente mencionada. Finalmente se incluyó una variable climática, la temperatura

máxima, para proporcionar a la red (*ANN541+Tmax*) con un indicador adicional de la respuesta de la demanda de agua diaria a lo largo del tiempo.

Los desempeños de las redes neuronales –*ANN,451551,541,541+Tmax*–, considerados únicamente en términos de los coeficientes de correlación (r) y de determinación (R^2) no han logrado mejorar el que obtuvo una de las redes más simples, la *ANN441*. La inclusión del indicador de los meses de agosto resultó en errores máximos absolutos más grandes por lo que no se justifica su inclusión y lo mismo ocurre con la inclusión de la temperatura máxima en el vector de entradas.

Estadístico	**ANN 231**	**ANN 441**	**ANN 451**	**ANN 551**	**ANN 541**	**ANN 541+Tmax**
RMSE	14,395.23	11,359.23	11,459.44	11,556.67	11,419.78	11,428.93
MAE	10,379.10	8,332.94	8,234.82	8,381.00	8,283.65	8,221.85
MAPE	3.15	2.49	2.46	2.50	2.47	2.44
ME	1,103.90	3,859.20	3,565.81	3,949.43	3,751.88	4,087.98
MPE	0.146	1.09	1.00	1.12	1.06	1.12
Error Máx. Abs.	67,676.40	42,232.96	45,260.81	61,913.09	63,132.19	50,416.62
r	0.776	0.88	0.88	0.88	0.88	0.88
R^2	0.603	0.78	0.78	0.78	0.78	0.78

Cuadro 8.13: Comparación del desempeño de *ANNs*, validación a un día

8.4. Conclusiones sobre predicción de la demanda con Redes Neuronales

Con el análisis anteriormente realizado es evidente que la técnica de *ANNs* es una herramienta útil y válida para realizar predicciones a corto plazo de la demanda de agua urbana. Mediante la utilización de arquitecturas y topologías de *ANNs* simples, logran aprender las relaciones entre las entradas (valores pasados de la demanda diaria, variables categóricas) y las salidas (valores futuros de demanda), obteniendo con las redes entrenadas unas predicciones con propiedades estadísticas aceptables (ver cuadro 8.13).

Las *ANNs* fueron entrenadas para realizar predicciones a un intervalo de tiempo, es decir a un día vista, con los resultados anteriormente reportados. Sin embargo, también se realizaron varias pruebas para realizar predicciones a 7 intervalos de tiempo ó 7 días vista (Y_{t+7}), con varias arquitecturas de red simples similares a las que fueron utilizadas para predecir a un intervalo de tiempo, con el mismo conjunto de datos de entrenamiento y validación sin obtener buenos resultados. Si bien las *ANNs* entrenadas para predecir a 7 días lograron captar los ciclos existentes, no interpretaban adecuadamente que la predicción esperada era la que se presentaría dentro de 7 periodos. Los resultados escasamente superaron a los obtenidos con un modelo de persistencia, es decir que el último valor de demanda observado se repita 7 periodos de tiempo después. Se hicieron pruebas en la cuales se le proporcionaba a la red con un vector de objetivos que contuviera las predicciones esperadas a 1 (Y_{t+1}) y 7 (Y_{t+7}) días simultáneamente, e incluso con un vector que incluía también los objetivos de una predicción intermedia, a 3 días (Y_{t+3}), sin obtener mejores resultados.

El tipo de redes neuronales que hemos utilizado en este caso pertenecen a las que podríamos llamar clásicas hacia adelante, que se consideran como estáticas. Este tipo de redes neuronales no cuenta con ningún tipo de retroalimentación por lo que la salida de la red es calculada directamente desde el vector de entradas mediante la modificación de los pesos sinápticos de las conexiones hasta encontrar una solución óptima. Existen también redes neuronales más potentes –que no son objeto de este estudio–, que a su vez son más complejas y difíciles de entrenar y que se consideran dinámicas. En este tipo de redes la salida de la red depende no solo de las entradas a la red actuales sino también de los valores de entradas y objetivos pasados, así como de una memoria parcial de la historia relevante de la secuencia por medio de una representación en forma de estados. Estos valores son presentados a la red como una nueva entrada. Es muy probable que con la utilización de redes neuronales dinámicas, se pudieran obtener mejores resultados

tanto en las predicciones a 1, pero sobre todo a 7 días. Las posibles mejoras significarían mejores ajustes hasta un determinado punto. Sin embargo, ya que el proceso de demanda de agua no puede ser considerado –de hecho no lo es– como físicamente basado, sino que es la respuesta del conjunto de una población a sus requerimientos diarios de agua, el proceso tiene siempre una incertidumbre asociada esperable, que limita el grado de ajuste tanto de la metodología de redes neuronales como de cualquier otra.

8.5. Comparación de las metodologías empleadas

La primeras secciones de este capítulo, fueron empleadas para realizar un análisis preliminar de la serie temporal de demandas que estamos analizando. Con ese análisis obtuvimos información relevante sobre las autocorrelaciones existentes, las correlaciones cruzadas con las variables meteorológicas, las periodicidades de la serie así como la tendencia a largo plazo de la misma. La información obtenida fue de utilidad a la hora de seleccionar las variables empleadas para construir un grupo de modelos de series temporales ARIMA, así como también una serie de redes neuronales que reprodujeran y predijeran adecuadamente la demanda diaria.

De los análisis realizados en las secciones 7.4 y 8.3 de este capítulo, quedó demostrado que tanto los modelos del tipo ARIMA con estacionalidad, como los de redes neuronales fueron capaces de identificar y reproducir el proceso de la demanda de agua de la ciudad de Valencia con resultados muy aceptables, aún y cuando los fundamentos de las metodologías empleadas parten de postulados muy diferentes. Los desempeños de los modelos construidos, tanto del tipo ARIMA como los de *ANN* se presentan en el cuadro 8.14 y se presentan también en el cuadro 8.15 agrupados por los promedios de las metodologías y por variables empleadas.

Modelo	RMSE	MAE	MAPE	ME	MPE	Error Max.	r	R^2
ARIMA A	11,143.56	7,932.46	2.40	-17.65	-0.0861	48,593.50	0.87	0.76
ARIMA B	10,662.31	7,629.09	2.32	2.32	-0.0656	43,581.10	0.88	0.78
ARIMA C	11,101.84	7,926.92	2.41	-24.56	-0.0955	47,324.60	0.87	0.76
ARIMA D	10,603.67	7,553.24	2.29	-9.90	-0.0817	44,403.20	0.88	0.78
ANN231	14,395.23	10,379.10	3.15	1,103.90	0.146	67,676.40	0.77	0.603
ANN441	11,359.23	8,332.94	2.49	3,859.20	1.09	42,232.96	0.88	0.78
ANN451	11,459.44	8,234.94	2.46	3,565.81	1.00	45,260.81	0.88	0.78
ANN551	11,556.67	8,381.00	2.50	3,949.43	1.12	61,913.09	0.88	0.78
ANN541	11,419.78	8,283.65	2.47	3,751.88	1.06	63,132.19	0.88	0.78
ANN541+Tmax	11,428.93	8,221.85	2.44	4,087.98	1.12	50,416.62	0.88	0.78

Cuadro 8.14: Desempeño de los modelos ARIMA y ANN en fase de validación para la predicción de la demanda diaria a un día

De estos dos cuadros podemos destacar varios puntos:

- El más simple de los modelos ARIMA (línea 1 del cuadro 8.15) que solo utiliza datos del pasado de la serie de demandas para calibrar sus parámetros autoregresivos y de media móvil, presenta mejores resultados que la *ANN* más simple (línea 2 del mismo cuadro) que también utiliza solo valores del pasado de la serie de demandas para entrenarse.
- Los resultados de los modelos de ambas metodologías que incorporan a la temperatura en sus estructuras (líneas 3 y 4 del cuadro 8.15), resultan

en mejores desempeños para el modelo del tipo ARIMA, presentando los mismos resultados que la red neuronal únicamente en términos de los coeficientes r y R^2.

- Las redes neuronales con las arquitecturas más complejas de entre las *ANN* entrenadas (línea 6 del cuadro 8.15), que incorporan las diferentes variables categóricas empleadas, no fueron mejores que el más complejo de los modelos del tipo ARIMA (línea 3 del mismo cuadro). Ambas metodologías presentan valores similares únicamente en los coeficientes r y R^2.

- El promedio del desempeño de los diferentes modelos de ambas metodologías, resulta en mejores valores en todos los puntos analizados para la metodología de modelos del tipo ARIMA. Sin embargo en términos absolutos y porcentuales (MAPE) los desempeños de ambas metodologías son muy similares.

- El mejor de los modelos ARIMA construidos ((0,1,2)x(0,1,1)7 + Tmax) requiere la estimación de 4 parámetros (3 pertenecientes al modelo ARIMA más el coeficiente de regresión de la temperatura) para realizar sus predicciones. Cada uno de los parámetros nos aporta información significativa de la relación entre las predicciones y el pasado de la serie.

- La red *ANN*451 requiere de la estimación de 31 parámetros para realizar sus predicciones. No podemos obtener ninguna información de los parámetros entrenados.

- Los resultados obtenidos están acordes con los obtenidos por De~la Fuente~García et~al. (1996), donde concluye que la metodología Box-Jenkins obtiene mejores resultados que la de redes neuronales (y que la de espacio de los estados) cuando la serie que se está tratando tiene un patrón de estacionalidad, que es nuestro caso.

Modelo	RMSE	MAE	MAPE	ME	MPE	Error Max.	r	R^2
(1) ARIMA Univariada	11,143.56	7,932.46	2.40	-17.65	-0.0861	48,593.50	0.87	0.76
(2) *ANN* Univariada	14,395.23	10,379.10	3.15	1,103.90	0.146	67,676.40	0.77	0.603
(3) ARIMA + Tmax	10,632.99	7,591.16	2.30	-3.79	0.07365	43,992.15	0.88	0.78
(4) *ANN541+Tmax*	11,428.93	8,221.85	2.44	4,087.98	1.12	50,416.62	0.88	0.78
(5) *ANN+categóricas*	11,448.78	8,308.03	2.48	3,781.58	1.06	53,134.76	0.88	0.78
(6) ARIMA total	10,877.84	7,760.42	2.36	-12.44	-0.0822	45,976.35	0.875	0.77
(7) *ANN* total	11,936.54	8,638.91	2.58	3,386.36	0.92	55,105.34	0.861	0.75

Cuadro 8.15: Desempeños de los modelos ARIMA y ANN en fase de validación para la predicción de la demanda diaria a un día. Valores promedio para cada clase de modelo

8.6. Conclusiones del caso de estudio

El trabajo realizado a lo largo del presente capítulo nos lleva a encontrar que las metodologías empleadas son adecuadas para predecir la demanda de agua urbana de la ciudad de Valencia, España. Los modelos del tipo ARIMA estacionales con y sin variables exógenas nos permitieron realizar predicciones de 1 y 7 días de antelación. Con la metodología de redes neuronales obtuvimos resultados equiparables a los obtenidos con los modelos ARIMA para el caso de 1 día de antelación, sin embargo las predicciones a 7 días con esta metodología no ofrecieron buenos resultados por lo que se omitió su análisis más en detalle. Un modelo de persistencia nos aportaba resultados similares a los que se obtenían con las redes neuronales entrenadas.

Es adecuado comentar que en este caso de estudio se pretendió siempre emplear modelos parsimoniosos, tanto para los modelos ARIMA como para las redes neuronales, por lo cual los modelos y redes neuronales construidas fueron siempre modelos simples. Ya se comentó anteriormente que existen redes neuronales más potentes, como son las denominadas dinámicas o recurrentes, pero este tipo de redes lleva implícito una complejidad añadida en su topología y en la fase de entrenamiento, por lo que no fueron empleadas en este caso. Se podrían haber construido redes neuronales simples (estáticas) con una cantidad grande de entradas y neuronas en la capa oculta para buscar mejores ajustes. Tal es el caso de Griñó~C. (1991), donde se emplearon redes con arquitecturas del tipo 15-20-1 o 19-35-1, para obtener predicciones con errores del orden del 5% (MAPE) de la demanda. El mismo caso se reproduce en Zhang (2003), que utilizan una red con arquitectura 10-13-2 para predecir la demanda a 1 y 2 días en los meses de verano, incorporando también predicciones de variables meterológicas de lluvia y temperatura (con su propia incertidumbre) obteniendo predicciones con errores inferiores al 5% (MAPE). Si comparamos estos resultados con los obtenidos en el caso que estamos analizando con modelos simples, el MAPE que obtienen el conjunto de los modelos ARIMA y de redes neuronales construidos es de 2.36% y 2.58% respectivamente. Es evidente que los resultados que hemos obtenido son cuando menos prometedores, sobre todo si analizamos que la media diaria de la serie de demandas analizadas para el año 2004 reservado para validación es de 321,573.40 m^3, por lo que un error de predicción típico tendría valores entre 7,760.42-8,638.91 m^3.

El proceso que genera la demanda diaria está regido por un fenómeno sociológico y no por uno físico. Es la respuesta de un conjunto de personas demandando el agua que requieren para cubrir sus necesidades, por lo que es inherentemente cambiante y no todo el tiempo la respuesta es la misma. Más bien es variable según la época del año, la climatología imperante, el

tipo de día (laborable, festivo) etc. Los fenómenos que afectan a la demanda son complejos de identificar y por el mismo motivo su incorporación a los modelos de predicción es complicada. Cualquiera de las metodologías que se emplee para predecir la demanda se verá limitada por este fenómeno y deberá asumir que una parte del proceso de demanda no es totalmente predecible, y se deberá asumir con una incertidumbre inherente al proceso generador de la demanda, es decir una componente estocástica. Se deduce que muy probablemente, no es casual que los valores de correlación entre los valores observados y los predichos r, y de varianza explicada R^2, de 0.88 y 0.78 obtenidos con las dos metodologías empleadas en este caso, estén muy próximos a los valores que se obtendrían con un modelo óptimo de predicción que incorporara todas las variables que rigen al conjunto de usuarios de una ciudad para demandar agua.

Como conclusión final de esta sección y como tema a desarrollarse en el capítulo siguiente queda, la identificación de las variables que pudieran mejorar las predicciones de la demanda de agua en entornos urbanos, la cuantificación de sus efectos y la incorporación implícita en un modelo eficiente de predicción a corto plazo, así como la determinación del componente estocástico de la demanda de agua urbana.

Capítulo 9

Rasgos peculiares de la demanda

9.1. Identificación de rasgos peculiares de la demanda

Hasta este punto hemos identificado la estructura de un conjunto de modelos de series temporales ARIMA, que reproducen el comportamiento de una serie de demanda diaria de agua. Se han obtenido predicciones puntuales a uno y siete pasos de tiempo con sus incertidumbres asociadas. Sin embargo estas estructuras ARIMA identificadas con sus componentes regulares y estacionales, nos sirven solamente como regla general de comportamiento para la serie que estamos analizando, estas han logrado captar la porción de variabilidad sistemática que se observa a lo largo de la serie completa. Esto significa que aún utilizando una estructura ARIMA muy compleja no se habrán tomado en cuenta los fenómenos que a continuación se presentan y que son de frecuente ocurrencia en series de demanda de agua potable.

El objetivo principal de esta sección es el de identificar los componentes de variabilidad sistemática irregular y atípicos en la demanda, sus efectos tanto en la identificación de los modelos ARIMA como en la predicción con los mismos ya han sido comentados en las secciones 2.4 5.1.

Los modelos que han sido estimados previamente con la metodología de Box-Jenkins no han tomado en cuenta ningún tipo de peculiaridad. La serie temporal no ha sido analizada en la búsqueda de rasgos que nos pudieran indicar algún tipo de anomalía en los datos. Es de esperar por tanto, que los modelos estimados no sean los mejores posibles para representar el proceso de la demanda. Con esta finalidad hemos realizado un análisis de

valores atípicos, nos hemos limitado a identificar los atípicos del tipo aditivo (AO) de la serie en estudio. Se ha ajustado el modelo ARIMA(0,1,2)x(0,1,1) y se ha ejecutado simultáneamente la búsqueda de valores atípicos del tipo antes mencionado. Previamente se le indicaron al modelo los días en los que tiene lugar algún tipo de evento derivado de los patrones de festividad de la ciudad. Se ha generado un catálogo de festividades que se tienen registradas en Valencia, se presentan en el cuadro 9.1. Es de esperar que estos días sean identificados como atípicos. Los resultados se presentan en el cuadro 9.2. Se han eliminado los días clasificados como festivos pero que no fueron estadísticamente significativos, en su mayoría por que se presentan en fin de semana y su efecto en la mayoría de los casos se anula.

Los resultados presentados en el cuadro 9.2 confirman que los días en los que se presenta una festividad, las variaciones que se registran son estadísticamente significativas y pueden ser considerados como valores atípicos de variabilidad sistemática irregular. Por otra parte, se han identificado también valores que no son explicados por el calendario de festividades y que podrían considerarse como impactos puntuales, no sistemáticos que deberán ser removidos o reemplazados por valores más representativos. La presencia y combinación de las dos clases de eventos invariablemente estarán introduciendo un sesgo en la estimación de los parámetros del modelo, por lo que se puede verificar que los que se han obtenido anteriormente no son los mejores que se pueden obtener. Los que han sido obtenido tomando en cuenta las valores atípicos han considerado los 4 años de la serie, es decir que no se ha reservado un conjunto de datos para validar el desempeño del modelo estimado. Como se mencionó anteriormente, la finalidad de esta sección es la de identificar valores atípicos, su magnitud y propiedades estadísticas. En secciones posteriores se identificará –como se hizo en secciones anteriores– un modelo que considere los valores atípicos utilizando 3 años para estimar los parámetros del modelo y uno para validar el desempeño.

Fecha	**Festividad/Evento**	Día
01-ene	Año nuevo	Variable
06-ene	Día de reyes	Variable
Variable	Semana Santa	Viernes
Variable	Semana Santa	Sábado
Variable	Semana Santa	Domingo
Variable	Semana Santa	Lunes
19-mar	San José	Variable
01-may	Día del trabajo	Variable
15-ago	Virgen de la Asunción	Variable
09-oct	Día de la comunidad Valenciana	Variable
12-oct	Día de la hispanidad	Variable
01-nov	Día de todos los santos	Variable
06-dic	Día de la Constitución	Variable
08-dic	Día de la inmaculada concepción	Variable
25-dic	Navidad	Variable

Cuadro 9.1: Resumen de atípicos aditivos identificados en la serie de demandas de agua de la ciudad de Valencia. 2001-2004

Fecha	Festividad/Evento	Día	Estimación (m^3)	ET (m^3)	*t*	Sig.
01-ene-01	Año nuevo	Lun	-49,314.68	9,297.93	-5.30	0.000
19-mar-01	San José	Lun	-40,238.92	7,481.23	-5.38	0.000
13-abr-01	Semana Santa	Vie	-63,853.46	8,269.99	-7.72	0.000
14-abr-01	Semana Santa	Sáb	-44,928.66	9,032.44	-4.97	0.000
15-abr-01	Semana Santa	Dom	-38,590.32	9,034.18	-4.27	0.000
16-abr-01	Semana Santa	Lun	-41,980.66	8,277.84	-5.07	0.000
01-may-01	Día del trabajo	Mar	-37,354.38	7,500.50	-4.98	0.000
09-may-01	Atípico	Mié	-34,118.49	7,474.72	-4.56	0.000
15-ago-01	Virgen de la Asunción	Mié	-39,380.21	7,464.70	-5.28	0.000
08-oct-01	Atípico provocado por enlace	Lun	-19,084.32	8,064.45	-2.37	0.018
09-oct-01	Día de la comunidad Valenciana	Mar	-47,617.83	8,071.00	-5.90	0.000
12-oct-01	Día de la hispanidad	Vie	-24,081.81	7,486.75	-3.22	0.001
01-nov-01	Día de todos los santos	Jue	-25,672.39	7,507.65	-3.42	0.001
08-nov-01	Atípico	Jue	51,699.31	7,523.31	6.87	0.000
06-dic-01	Día de la Constitución	Jue	-23,353.19	7,484.28	-3.12	0.002
21-dic-01	Atípico	Vie	-41,109.16	7,471.35	-5.50	0.000
01-ene-02	Año nuevo	Mar	-38,810.60	7,471.16	-5.19	0.000
19-mar-02	San José	Mar	-35,096.76	7,479.22	-4.69	0.000
28-mar-02	Semana Santa	Jue	-49,531.94	8,322.34	-5.95	0.000
29-mar-02	Semana Santa	Vie	-76,464.22	9,182.00	-8.33	0.000
30-mar-02	Semana Santa	Sáb	-48,828.65	9,291.24	-5.26	0.000
31-mar-02	Semana Santa	Dom	-35,705.04	9,183.52	-3.89	0.000
01-abr-02	Semana Santa	Lun	-30,293.51	8,322.40	-3.64	0.000
01-may-02	Día del trabajo	Mié	-15,528.17	7,471.76	-2.08	0.038
29-jul-02	Atípico	Lun	-44,965.34	7,466.68	-6.02	0.000
15-ago-02	Virgen de la Asunción	Jue	-24,923.48	7,466.67	-3.34	0.001
20-ago-02	Atípico	Mar	-59,553.99	7,466.84	-7.98	0.000
09-oct-02	Día de la comunidad Valenciana	Mié	-27,597.94	7,471.98	-3.69	0.000
01-nov-02	Día de todos los santos	Vie	-29,862.84	7,467.33	-4.00	0.000
25-nov-02	Atípico	Lun	40,997.33	7,467.68	5.49	0.000
25-dic-02	Navidad	Mié	-43,365.19	7,474.82	-5.80	0.000
06-ene-03	Día de reyes	Lun	-37,662.71	7,468.54	-5.04	0.000
19-mar-03	San José	Mié	-31,075.83	7,467.53	-4.16	0.000
18-abr-03	Semana Santa	Vie	-55,897.42	8,263.57	-6.76	0.000
19-abr-03	Semana Santa	Sáb	-59,314.39	9,039.61	-6.56	0.000
20-abr-03	Semana Santa	Dom	-41,267.50	9,033.53	-4.57	0.000
21-abr-03	Semana Santa	Lun	-46,291.22	8,264.28	-5.60	0.000
15-ago-03	Virgen de la Asunción	Vie	-30,023.24	7,466.19	-4.02	0.000
09-oct-03	Día de la comunidad Valenciana	Jue	-27,691.52	8,054.70	-3.44	0.001
10-oct-03	Atípico provocado por enlace	Vie	-16,732.10	8,055.95	-2.08	0.038
08-dic-03	Día de la inmaculada concepción	Lun	-23,839.27	7,481.81	-3.19	0.001
25-dic-03	Navidad	Jue	-27,803.19	7,499.82	-3.71	0.000
01-ene-04	Año nuevo	Jue	-37,886.53	7,499.35	-5.05	0.000
06-ene-04	Día de reyes	Mar	-35,776.64	7,467.12	-4.79	0.000
19-mar-04	San José	Vie	-21,864.64	7,477.70	-2.92	0.004
09-abr-04	Semana Santa	Vie	-43,599.03	8,305.07	-5.25	0.000
10-abr-04	Semana Santa	Sáb	-55,757.41	9,036.49	-6.17	0.000
11-abr-04	Semana Santa	Dom	-33,795.65	9,034.03	-3.74	0.000
12-abr-04	Semana Santa	Lun	-39,903.60	8,262.13	-4.83	0.000
19-jul-04	Atípico	Lun	-33,934.75	7,466.26	-4.55	0.000
11-oct-04	Atípico provocado por enlace	Lun	-31,285.30	8,097.27	-3.86	0.000
12-oct-04	Día de la hispanidad	Mar	-39,903.11	8,075.51	-4.94	0.000
01-nov-04	Día de todos los santos	Lun	-35,497.11	7,505.15	-4.73	0.000
06-dic-04	Día de la constitución	Lun	-27,330.51	8,314.85	-3.29	0.001
07-dic-04	Atípico provocado por enlace	Mar	-28,012.93	8,950.43	-3.13	0.002
08-dic-04	Día de la inmaculada concepción	Mié	-25,695.99	8,291.65	-3.10	0.002
25-dic-04	Navidad	Sáb	-18,949.32	7,834.72	-2.42	0.016

Cuadro 9.2: Resumen de atípicos aditivos identificados en la serie de demandas de agua de la ciudad de Valencia. 2001-2004

9.1.1. *Caracterización de los valores atípicos*

Analizando los datos obtenidos en la sección anterior se observa que de un máximo de 60 días de festividades para los cuatro años en estudio, 45 de ellos fueron identificados como atípicos. A continuación detallaremos los eventos identificados para cada uno de los años de la serie.

Año 2001

En este año 12 de los 15 días festivos fueron detectados como atípicos. De entre los restantes, 2 ocurrieron en fin de semana y su efecto se anuló y uno más, el 25 de diciembre no presentó un descenso en la demanda suficientemente significativo. Adicionalmente, el 8 de octubre que ocurrió en un día lunes, presentó un descenso de la demanda esperada, esto se puede explicar por un efecto de *enlace* porque tanto el día 6, como el 9 de octubre son festivos (este año ocurriendo en un día martes). Si consideramos este último día como festivo, se habrán identificado 14 días de este tipo. Ver cuadro 9.3

Fecha	Festividad/Evento	Día	Estimación (m^3)	ET (m^3)	*t*	Sig.
01-ene-01	Año nuevo	Lun	-49,314.68	9,297.93	-5.30	0.000
19-mar-01	San José	Lun	-40,238.92	7,481.23	-5.38	0.000
13-abr-01	Semana Santa	Vie	-63,853.46	8,269.99	-7.72	0.000
14-abr-01	Semana Santa	Sáb	-44,928.66	9,032.44	-4.97	0.000
15-abr-01	Semana Santa	Dom	-38,590.32	9,034.18	-4.27	0.000
16-abr-01	Semana Santa	Lun	-41,980.66	8,277.84	-5.07	0.000
01-may-01	Día del trabajo	Mar	-37,354.38	7,500.50	-4.98	0.000
15-ago-01	Virgen de la Asunción	Mié	-39,380.21	7,464.70	-5.28	0.000
08-oct-01	Atípico provocado por enlace	Lun	-19,084.32	8,064.45	-2.37	0.018
09-oct-01	Día de la comunidad Valenciana	Mar	-47,617.83	8,071.00	-5.90	0.000
12-oct-01	Día de la hispanidad	Vie	-24,081.81	7,486.75	-3.22	0.001
01-nov-01	Día de todos los santos	Jue	-25,672.39	7,507.65	-3.42	0.001
06-dic-01	Día de la Constitución	Jue	-23,353.19	7,484.28	-3.12	0.002
Festivos no identificados como atípicos						
06-ene-01	Día de Reyes	Sábado	**Su efecto se anula en fin de semana**			
08-dic-01	Día de la Inmaculada concepción	Sábado	**Su efecto se anula en fin de semana**			
25-dic-01	Navidad	Martes	**El descenso no fue significativo**			

Cuadro 9.3: Resumen de festivos identificados y no identificados como atípicos para el año 2001

Año 2002

En este caso 11 de los 15 días festivos fueron detectados como atípicos, de entre los que no fueron identificados, el 12 de octubre y el 8 de diciembre ocurrieron en fin de semana. El 6 de enero y el 6 de diciembre ocurrieron en viernes, sin embargo sus descensos no fueron significativos. Finalmente el día 28 de marzo que no es festivo, que ocurrió el jueves que precede al viernes de semana santa presentó un descenso en la demanda muy significativo que se entiende como un arrastre del conjunto de días festivos que le suceden. Si contamos a éste último como festivo tendríamos 12 días identificados. Ver cuadro 9.4

Fecha	Festividad/Evento	Día	Estimación (m^3)	ET (m^3)	*t*	Sig.
01-ene-02	Año nuevo	Mar	-38,810.60	7,471.16	-5.19	0.000
19-mar-02	San José	Mar	-35,096.76	7,479.22	-4.69	0.000
28-mar-02	Semana Santa	Jue	-49,531.94	8,322.34	-5.95	0.000
29-mar-02	Semana Santa	Vie	-76,464.22	9,182.00	-8.33	0.000
30-mar-02	Semana Santa	Sáb	-48,828.65	9,291.24	-5.26	0.000
31-mar-02	Semana Santa	Dom	-35,705.04	9,183.52	-3.89	0.000
01-abr-02	Semana Santa	Lun	-30,293.51	8,322.40	-3.64	0.000
01-may-02	Día del trabajo	Mié	-15,528.17	7,471.76	-2.08	0.038
15-ago-02	Virgen de la Asunción	Jue	-24,923.48	7,466.67	-3.34	0.001
09-oct-02	Día de la comunidad Valenciana	Mié	-27,597.94	7,471.98	-3.69	0.000
01-nov-02	Día de todos los santos	Vie	-29,862.84	7,467.33	-4.00	0.000
25-dic-02	Navidad	Mié	-43,365.19	7,474.82	-5.80	0.000
Festivos no identificados como atípicos						
06-ene-02	Día de Reyes	Viernes	**El descenso no fue significativo**			
12-oct-02	Día de la Hispanidad	Sábado	**Su efecto se anula en fin de semana**			
06-dic-02	Día de la constitución	Viernes	**El descenso no fue significativo**			
08-dic-02	Día de la Inmaculada concepción	Domingo	**Su efecto se anula en fin de semana**			

Cuadro 9.4: Resumen de festivos identificados y no identificados como atípicos para el año 2002

Año 2003

En este año 10 de los 15 días festivos fueron identificados como atípicos. El día 1 de enero no fue identificado como tal, a pesar de ocurrir en un día miércoles su disminución no fue significativa. Los días 12 de octubre, 1 de noviembre y 6 de diciembre acontecieron en fin de semana y su efecto se disipó. Finalmente el día 10 de octubre (viernes), que no es festivo, presentó un descenso que es levemente significativo ($t = -2{,}08$ y $p = 0{,}038$). Este descenso se explica como un enlace entre el día festivo de jueves 9 de octubre y el fin de semana. Si tomamos en cuenta este último día, 11 de los 15 festivos fueron identificados. Los días 1 de enero y 1 de mayo, miércoles y jueves respectivamente no presentaron descensos significativos. Ver cuadro 9.5

Fecha	Festividad/Evento	Día	Estimación (m^3)	ET (m^3)	*t*	Sig.
06-ene-03	Día de reyes	Lun	-37,662.71	7,468.54	-5.04	0.000
19-mar-03	San José	Mié	-31,075.83	7,467.53	-4.16	0.000
18-abr-03	Semana Santa	Vie	-55,897.42	8,263.57	-6.76	0.000
19-abr-03	Semana Santa	Sáb	-59,314.39	9,039.61	-6.56	0.000
20-abr-03	Semana Santa	Dom	-41,267.50	9,033.53	-4.57	0.000
21-abr-03	Semana Santa	Lun	-46,291.22	8,264.28	-5.60	0.000
15-ago-03	Virgen de la Asunción	Vie	-30,023.24	7,466.19	-4.02	0.000
09-oct-03	Día de la comunidad Valenciana	Jue	-27,691.52	8,054.70	-3.44	0.001
10-oct-03	Atípico provocado por enlace	Vie	-16,732.10	8,055.95	-2.08	0.038
08-dic-03	Día de la inmaculada concepción	Lun	-23,839.27	7,481.81	-3.19	0.001
25-dic-03	Navidad	Jue	-27,803.19	7,499.82	-3.71	0.000
Festivos no identificados como atípicos						
01-ene-03	Año nuevo	Mié	**El descenso no fue significativo**			
01-may-03	Día del trabajo	Jue	**El descenso no fue significativo**			
12-oct-03	Día de la Hispanidad	Dom	**Su efecto se anula en fin de semana**			
01-nov-03	Día de todos los santos	Sáb	**Su efecto se anula en fin de semana**			
06-dic-03	Día de la constitución	Sáb	**Su efecto se anula en fin de semana**			

Cuadro 9.5: Resumen de festivos identificados y no identificados como atípicos para el año 2003

Año 2004

Para este caso 12 de los 15 días festivos fueron identificados como atípicos. Los días 1 mayo, 15 de agosto y 9 de octubre acontecieron en fin de semana y su efecto se disipó. El día 11 de octubre que no es festivo, aconteció un día lunes y presentó un descenso significativo como resultado del enlace entre los días del fin de semana que le preceden y el día martes 12 de octubre que le sucede. Un comportamiento similar se observó el día martes 7 de diciembre. Presentó descensos en la demanda esperada por estar precedido del día festivo lunes 6 de diciembre y a continuación el día miércoles 8 de diciembre. El día martes 7 de diciembre enlaza dos días festivos aconteciendo en días laborables. Ver cuadro 9.6

Fecha	Festividad/Evento	Día	Estimación (m^3)	ET (m^3)	*t*	Sig.
01-ene-04	Año nuevo	Jue	-37,886.53	7,499.35	-5.05	0.000
06-ene-04	Día de reyes	Mar	-35,776.64	7,467.12	-4.79	0.000
19-mar-04	San José	Vie	-21,864.64	7,477.70	-2.92	0.004
09-abr-04	Semana Santa	Vie	-43,599.03	8,305.07	-5.25	0.000
10-abr-04	Semana Santa	Sáb	-55,757.41	9,036.49	-6.17	0.000
11-abr-04	Semana Santa	Dom	-33,795.65	9,034.03	-3.74	0.000
12-abr-04	Semana Santa	Lun	-39,903.60	8,262.13	-4.83	0.000
11-oct-04	Atípico provocado por enlace	Lun	-31,285.30	8,097.27	-3.86	0.000
12-oct-04	Día de la hispanidad	Mar	-39,903.11	8,075.51	-4.94	0.000
01-nov-04	Día de todos los santos	Lun	-35,497.11	7,505.15	-4.73	0.000
06-dic-04	Día de la constitución	Lun	-27,330.51	8,314.85	-3.29	0.001
07-dic-04	Atípico provocado por enlace	Mar	-28,012.93	8,950.43	-3.13	0.002
08-dic-04	Día de la inmaculada concepción	Mié	-25,695.99	8,291.65	-3.10	0.002
25-dic-04	Navidad	Sáb	-18,949.32	7,834.72	-2.42	0.016
Festivos no identificados como atípicos						
01-may-04	Día del trabajo	Sáb	**Su efecto se anula en fin de semana**			
15-ago-04	Virgen de la Asunción	Dom	**Su efecto se anula en fin de semana**			
09-oct-04	Día de la comunidad Valenciana	Sáb	**Su efecto se anula en fin de semana**			

Cuadro 9.6: Resumen de festivos identificados y no identificados como atípicos para el año 2004

9.1.2. *Clasificación de los valores atípicos*

Una clasificación de los días identificados como atípicos y ordenados por el día de la semana en que han ocurrido se presenta en el cuadro 9.7 y en la figura 9.1. De estos podemos obtener las siguientes conclusiones.

- Los festivos que han ocurrido en martes han resultado en la reducción media de la demanda de mayor magnitud, 39,093.22 m^3, seguida de los festivos ocurridos en día lunes que resultan en una reducción media de la demanda de 35,647.20 m^3.
- Los festivos ocurridos en jueves y viernes (descontando los de semana santa) son los que menores reducciones medias de la demanda presentan, con valores 29,829.29 m^3 y 29,458.13 m^3 respectivamente.
- La reducción media global por ocurrencia de día festivo es de 32,235.98 m^3.
- De un total de 10 festivos que ocurrieron en fin de semana, solo uno resultó ser significativo, el 25-dic-04. Por lo que se podría establecer como regla general que sus efectos se disipan si ocurren en fin de semana por los descensos propios del patrón semanal de la serie en estudio. Este patrón fue presentado en la figura 7.25 de la página 179.
- La mayor reducción registrada se produjo el jueves 28-mar-02 previo a semana santa con 49,531.94 m^3, que no es festivo pero precede a una serie de 4 días que si lo son, seguido del 01-ene-01 (lunes) y el 09-oct-01 (martes) con reducciones de 49,314.68 y 47,617.83 m^3 respectivamente.
- Los eventos atípicos provocan efectos de *enlace* y *arrastre* en los días anteriores y posteriores a su ocurrencia. Estos efectos requerirán del juicio del modelador para la desición de tomarlos en cuenta o ignorarlos.

Fecha	Festividad/Evento	Día	Estimación (m^3)
01-ene-01	Año nuevo	Lun	-49,314.68
19-mar-01	San José	Lun	-40,238.92
06-ene-03	Día de reyes	Lun	-37,662.71
08-dic-03	Día de la inmaculada concepción	Lun	-23,839.27
01-nov-04	Día de todos los santos	Lun	-35,497.11
06-dic-04	Día de la constitución	Lun	-27,330.51
Estimación media			**-35,647.20**
Estimación máxima			**-49,314.68**
Estimación mínima			**-23,839.27**
01-may-01	Día del trabajo	Mar	-37,354.38
09-oct-01	Día de la comunidad Valenciana	Mar	-47,617.83
01-ene-02	Año nuevo	Mar	-38,810.60
19-mar-02	San José	Mar	-35,096.76
06-ene-04	Día de reyes	Mar	-35,776.64
12-oct-04	Día de la hispanidad	Mar	-39,903.11
Estimación media			**-39,093.22**
Estimación máxima			**-47,617.83**
Estimación mínima			**-35,096.76**
15-ago-01	Virgen de la Asunción	Mié	-39,380.21
01-may-02	Día del trabajo	Mié	-15,528.17
09-oct-02	Día de la comunidad Valenciana	Mié	-27,597.94
25-dic-02	Navidad	Mié	-43,365.19
19-mar-03	Día de San José	Mié	-31,075.83
08-dic-04	Día de la inmaculada concepción	Mié	-25,695.99
Estimación media			**-30,440.55**
Estimación máxima			**-43,365.19**
Estimación mínima			**-15,528.17**
01-nov-01	Día de todos los santos	Jue	-25,672.39
06-dic-01	Día de la Constitución	Jue	-23,353.19
28-mar-02	Semana Santa	Jue	-49,531.94
15-ago-02	Vírgen de la Asunción	Jue	-24,923.48
09-oct-03	Día de la comunidad Valenciana	Jue	-27,691.52
25-dic-03	Navidad	Jue	-27,803.19
Estimación media			**-29,829.29**
Estimación máxima			**-49,531.94**
Estimación mínima			**-23,353.19**
12-oct-01	Día de la hispanidad	Vie	-24,081.81
01-nov-02	Día de todos los santos	Vie	-29,862.84
15-ago-03	Vírgen de la Asunción	Vie	-30,023.24
19-mar-04	San José	Vie	-21,864.64
Estimación media			**-26,458.13**
Estimación máxima			**-30,023.24**
Estimación mínima			**-21,864.64**
25-dic-04	Navidad	Sáb	-18,949.32

Cuadro 9.7: Resumen de las reducciones en la demanda producidas por festivos clasificados según el día de su ocurrencia

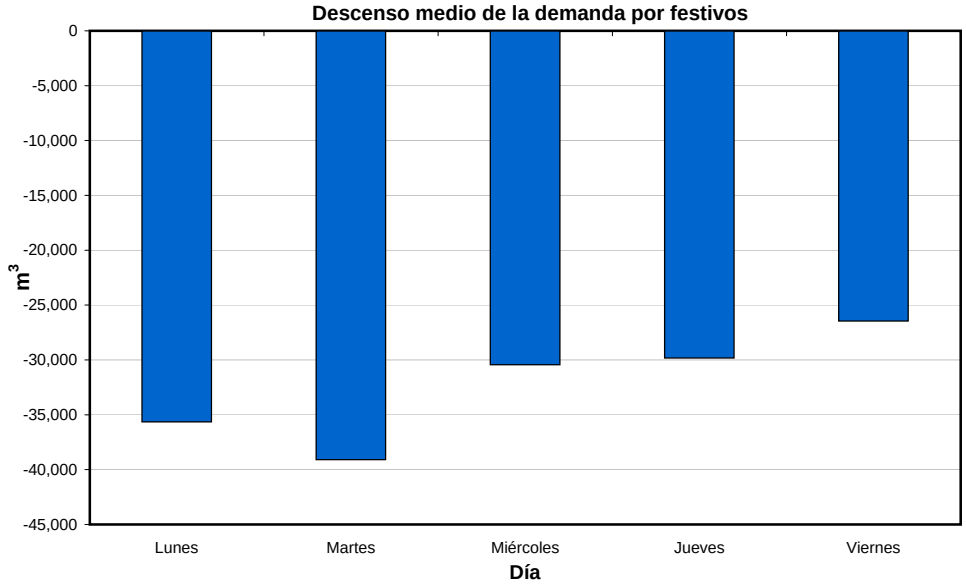


Figura 9.1: Descensos de la demanda media para los días de lunes a viernes provocados por festivos durante los años 2001 al 2004

Los festivos de semana santa han sido analizados por separado en otro cuadro (ver cuadro 9.8) ya que presentan valores muy diferentes a los del resto de los festivos que resultan en descensos globales máximos. Los descensos observados durante los días festivos de semana santa han sido graficados en la figura 9.2. Si bien es cierto que es complicado obtener un patrón de un conjunto tan limitado de eventos, el agrupamiento de 4 días festivos consecutivos nos permite obtener un patrón aproximado promediando cada uno de los días de los diferentes años (ver gráfico 9.3). De esta manera podemos observar que el descenso más importante se presenta al inicio del conjunto de días festivos consecutivos, es decir el viernes. Posteriormente los descensos se reducen paulatinamente en sábado y domingo para volver a aumentar el día lunes. Se ha ajustado una función polinómica de orden 3 que nos reproduce este patrón con un valor de R^2 cercano a 1. Una mayor cantidad de eventos nos permitiría obtener una función que resultara más representativa del conjunto de eventos.

Fecha	Festividad/Evento	Día	Estimación (m^3)
2001			
13-abr-01	Semana Santa	Vie	-63,853.46
14-abr-01	Semana Santa	Sáb	-44,928.66
15-abr-01	Semana Santa	Dom	-38,590.32
16-abr-01	Semana Santa	Lun	-41,980.66
Descenso global			**-189,353.10**
Estimación media			**-47,338.28**
Estimación máxima			**-63,853.46**
Estimación mínima			**-38,590.32**
2002			
29-mar-02	Semana Santa	Vie	-76,464.22
30-mar-02	Semana Santa	Sáb	-48,828.65
31-mar-02	Semana Santa	Dom	-35,705.04
01-abr-02	Semana Santa	Lun	-30,293.51
Descenso global			**-191,291.42**
Estimación media			**-47,822.85**
Estimación máxima			**-76,464.22**
Estimación mínima			**-30,293.51**
2003			
18-abr-03	Semana Santa	Vie	-55,897.42
19-abr-03	Semana Santa	Sáb	-59,314.39
20-abr-03	Semana Santa	Dom	-41,267.50
21-abr-03	Semana Santa	Lun	-46,291.22
Descenso global			**-202,770.53**
Estimación media			**-50,692.63**
Estimación máxima			**-59,314.39**
Estimación mínima			**-41,267.50**
2004			
09-abr-04	Semana Santa	Vie	-43,599.03
10-abr-04	Semana Santa	Sáb	-55,757.41
11-abr-04	Semana Santa	Dom	-33,795.65
12-abr-04	Semana Santa	Lun	-39,903.60
Descenso global			**-173,055.69**
Estimación media			**-43,263.92**
Estimación máxima			**-55,757.41**
Estimación mínima			**-33,795.65**
Semana Santa típica			
Descenso global típico			**-189,117.69**
Viernes típico			**-59,953.53**
Sábado típico			**-52,207.28**
Domingo típico			**-37,339.63**
Lunes típico			**-39,617.25**

Cuadro 9.8: Resumen de las reducciones en la demanda producidas por los festivos de semana santa

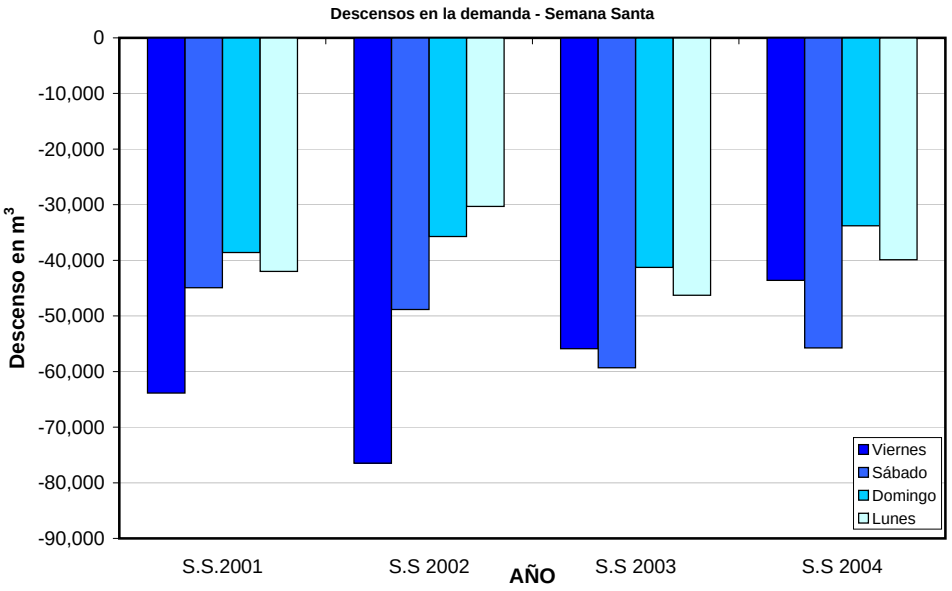


Figura 9.2: Descensos de la demanda durante semana santa observados en los años 2001 al 2004

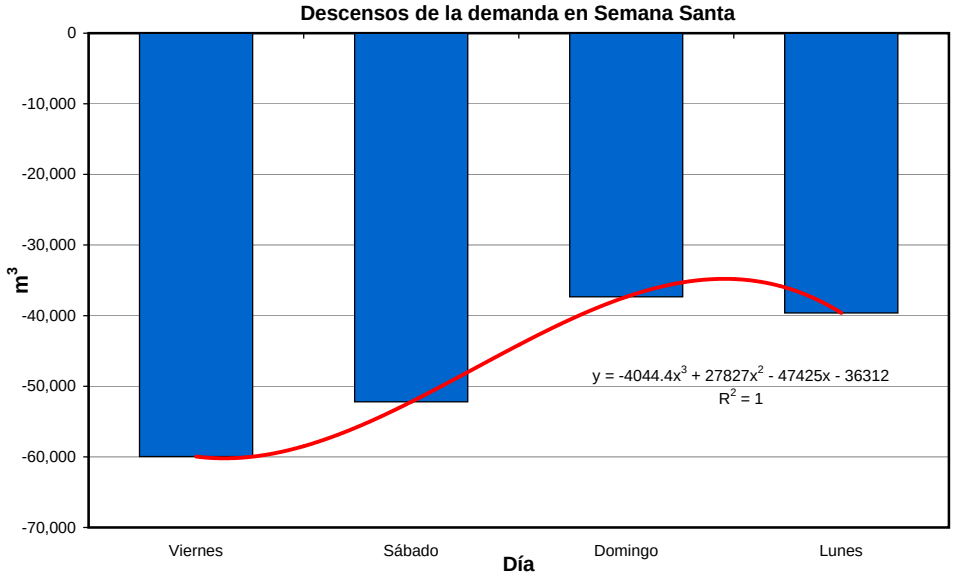


Figura 9.3: Patrón de descensos de la demanda provocados por semana santa

Parámetro	Estimado	Error Stnd.	t	P-value
MA(1)	0.369646	0.030945	11.945377	0.000000
MA(2)	0.157493	0.030945	5.089402	0.000000
SMA(1)	0.825319	0.018158	45.451470	0.000000

Cuadro 9.9: Resumen del modelo (E)

9.2. Identificación del modelo con intervenciones

Con el fin de conocer los impactos que los eventos detectados tienen sobre los parámetros de los modelos ARIMA, se ha procedido a identificar un nuevo modelo que los incluye considerando la estructura de un modelo de m intervenciones y un modelo ARIMA de la forma contemplada en capítulos anteriores. Se han utilizado los datos diarios de demanda de los 3 primeros años (2001, 2002, 2003) junto con los eventos detectados en ese mismo periodo y cuyas magnitudes se han presentado en los cuadros 9.3, 9.4, 9.5. Los eventos que fueron detectados como atípicos de variabilidad sistemática irregular y que no tienen relación con el calendario de festividades con el que contamos, han sido reemplazadas por valores representativos de la serie en estudio porque entendemos que esos valores son errores de registro o de manipulación de los registros diarios de demanda.

Con estas condicionantes se ha identificado el modelo que hemos nombrado "E", sus parámetros se presentan en el cuadro 9.9 y los estadísticos descriptivos de los residuos en el cuadro 9.18

Estadístico	Periodo de Estimación (E)
RMSE	9,599.31
MAE	7,397.67
MAPE	2.36
ME	541.19
MPE	0.13

Cuadro 9.10: Desempeño del modelo (E), estimación y validación

9.3. Mejoras del modelo con intervenciones

Es de esperar que al identificar conjuntamente los parámetros del modelo y los eventos atípicos, encontremos que los estadísticos de los residuos

que miden el ajuste del modelo sean mejores a los obtenidos para los modelos anteriores. Esto queda demostrado si comparamos los resultados de los modelos ARIMA que no han tomado en cuenta las intervenciones obtenidos en fase de validación con los del cuadro 9.18. De estos podemos observar que los estadísticos que analizan los residuos nos indican un mejor ajuste del modelo E.

9.3.1. *Mejoras en los residuos y ajuste - Estimación*

De entre los modelos que no utilizan la temperatura en su ecuación el que mejores resultados obtuvo es el modelo C. El modelo E mejora sus resultados en términos de RMSE en fase de validación, 9,599.31 contra 12,297.30, es decir un 21 %, en términos de MAE 7,077.52 contra 8,684.44, un 18 %, en términos de MAPE 2.26 contra 2.82, un 19 %, en términos de ME 41.09 contra 123.62, un 66 %, en términos de MPE 0.07 contra 0.15, un 53 %, ver cuadro 9.11.

Estadístico	**Periodo de Estimación (E)**	**Periodo de Estimación (C)**	Diferencia %
RMSE	9,599.31	12,297.30	21
MAE	7,397.67	8,684.44	14
MAPE	2.36	2.82	19
ME	541.19	-123.63	66
MPE	-0.070	-0.157	53
r	0.93	0.88	5
R^2	0.13	0.78	9

Cuadro 9.11: Comparación del desempeño de los modelos (E) y (C), fase de estimación

En lo que se refiere a los coeficientes de determinación y de correlación las mejoras en términos porcentuales han sido menores que los del resto de los coeficientes, lo cual es esperable ya que valores de cercanos a 0.9 de estos coeficientes son complicados de obtener para este tipo de series que contiene una variabilidad que no responde a un fenómeno físico.

Es evidente que la inclusión de las variables determinísticas identificadas mediante detección de valores atípicos e integradas por medio de las intervenciones caracterizadas, han logrado mejorar el desempeño de los modelos ARIMA operando por si solos. El modelo ha logrado explicar una gran parte de la varianza que contiene la serie de demandas en estudio.

9.3.2. *Mejoras en los residuos y ajuste - Validación*

Una vez obtenidos los parámetros del modelo que contempla un conjunto de intervenciones, se ha procedido a probar su desempeño para predecir la demanda de agua diaria a un día empleando la metodología propuesta.

Consideraciones

En la sección 9.1.2 de la página 273 se han identificado los eventos que han provocado una intervención estadísticamente significativa del proceso que sigue la demanda y han sido clasificados según el día de su ocurrencia. Para el conjunto de los días ocurriendo en los diferentes días de la semana se obtuvieron valores medios: estos se resumen en el cuadro 9.12. Ya se comentó anteriormente que los eventos atípicos que ocurren en fin de semana tienen efecto nulo y por lo tanto no se considerarán en las predicciones. Para el caso de los días de viernes a lunes de semana santa, ya que como se comentó anteriormente, sus efectos presentan valores marcadamente diferenciados de los del resto de las intervenciones, se utilizarán los valores obtenidos en el cuadro 9.8 de la página 276 cuyos valores medios se resumen en el cuadro 9.13.

Día	**Intervención media Estimada** m^3
Lunes	-35,647.20
Martes	-39,093.22
Miércoles	-30,440.55
Jueves	-29,829.29
Viernes	-26,458.13
Sábado	–
Domingo	–

Cuadro 9.12: Intervenciones medias producidas por atípicos según el día de su ocurrencia

Los valores medios obtenidos anteriormente se emplearán para modelar los eventos atípicos que acontecerán durante el año 2004 que utilizaremos para validar el desempeño del modelo. Del análisis que hemos realizado en secciones anteriores conocemos que ocurrirán un total de 14 eventos en días laborables (entonces $m = 14$) y que por lo tanto alterarán el proceso de demanda. Estos eventos se presentan a continuación en el cuadro 9.14. Si bien las magnitudes de las intervenciones que producen han sido estimadas, en

Día	Intervención media Estimada m^3
Viernes	-59,953.53
Sábado	-52,207.28
Domingo	-37,339.63
Lunes	-39,617.25

Cuadro 9.13: Intervenciones medias caracterizadas para los días de semana santa

un escenario de predicción hemos de suponer que solamente conocemos el momento de su ocurrencia, es decir la fecha y el día de la semana y emplearemos los valores medios (del cuadro 9.12) para considerar sus efectos e incorporarlos implícitamente al modelo de predicción. Está claro que ningún evento será idéntico a otro por más que acontezcan en el mismo día de la semana, sin embargo los valores medios estimados pueden ser una buena aproximación para mejorar las predicciones. Hemos de asumir pues la incertidumbre que esta hipótesis aporta en la fase de predicción.

El modelo de predicción consistirá de una estructura ARIMA (0,1,2)x(0,1,1)7 que utilizará los coeficientes obtenidos en la fase de estimación (del cuadro 9.9), acoplado con un conjunto de m intervenciones caracterizadas que se activarán solamente en las fechas definidas previamente (del cuadro 9.14). Durante los q periodos posteriores se corregirán las aportaciones producidas en los residuos en una proporción igual a la magnitud (del cuadro 9.12) del evento afectada por los coeficientes θ_1 y θ_2.

Finalmente, es importante aclarar que el inicializado del modelo, es decir el inicio de la obtención de las predicciones desde el punto que nos interesa conocerlas, debe comenzar con una cantidad tan larga como sea posible de periodos de antelación de donde se obtengan valores la relación observación - predicción con el fin de que el sistema obtenga valores estables para el momento de las predicciones deseadas. Si se ignorara este punto, sucederá que las predicciones obtenidas serán peores durante un cantidad de periodos de tiempo hasta que el modelo logre establizarse, una vez superado esa transición las predicciones serán las verdaderas.

Desempeño del modelo ARIMA con intervenciones en fase de validación, predicción a 1 día

Los resultados del modelo E, ARIMA con intervenciones, para predicciones a un día se presentan en el cuadro 9.15. Se observa que los estadísticos

Fecha	Festividad/Evento	Día
01-ene-04	Año nuevo	Jue
06-ene-04	Día de reyes	Mar
19-mar-04	San José	Vie
09-abr-04	Semana Santa	Vie
10-abr-04	Semana Santa	Sáb
11-abr-04	Semana Santa	Dom
12-abr-04	Semana Santa	Lun
11-oct-04	Atípico provocado por enlace	Lun
12-oct-04	Día de la hispanidad	Mar
01-nov-04	Día de todos los santos	Lun
06-dic-04	Día de la constitución	Lun
07-dic-04	Atípico provocado por enlace	Mar
08-dic-04	Día de la inmaculada concepción	Mié
Festivos no identificados como atípicos		
01-may-04	Día del trabajo	Sáb
15-ago-04	Virgen de la Asunción	Dom
09-oct-04	Día de la comunidad Valenciana	Sáb
25-dic-04	Navidad	Sáb

Cuadro 9.14: Resumen de eventos para el año 2004

de los residuos en fase de validación no han empeorado y en algunos casos incluso han sido mejores que los de la fase de estimación. Los estadísticos que evalúan el ajuste del modelo son levemente peores que los obtenidos en la fase de estimación. En el mismo cuadro se presenta también el desempeño del modelo E sin considerar en su estructura las intervenciones. Si se comparan los resultados obtenidos con este modelo con los que no han considerado las intervenciones a la hora de estimar los parámetros (modelos A,B,C,D), se evidencia que existe una mejoría al utilizar estos nuevos parámetros más característicos incluso sin considerar las intervenciones.

Estadístico	Periodo de Estimación (E)	Periodo de Validación (E)	Periodo de Validación (E) sin intervenciones
RMSE	9,507.55	9,080.00	10,880.21
MAE	7,077.52	6,842.13	7,789.81
MAPE	2.26	2.06	2.37
ME	-41.09	667.26	-7.74
MPE	-0.070	0.20	-0.07
r	0.93	0.92	0.88
R^2	0.86	0.85	0.78

Cuadro 9.15: Desempeño del modelo (E), estimación y validación.

Podemos concluir pues, que el modelo se muestra robusto a pesar de las hipótesis que hemos asumido para modelar las intervenciones y que inevitablemente esta consideración resulta en incertidumbre adicional de las predicciones. Los resultados de la fase de validación han demostrado un comportamiento estable y a la vez reproduce eficientemente los eventos que

modifican el proceso de demanda y que aportan una variabilidad al modelo predictor. Los valores del error medio (ME) y error medio porcentual (MPE) que obtiene el modelo (a pesar de entender que los residuos positivos y negativos se anulan mutuamente) son muy pequeños considerando que la desviación estándar de la serie orignal es 22,783.54 m^3, pero aún se resultan más pequeños si consideramos el almacenamiento de agua con que dispone el sistema de Valencia, de 90,000 m^3, volumen que regula las variaciones de picos máximos y mínimos que presenta la demanda. Más representativo resulta el error medio absoluto porcentual (MAPE) que nos indica que el error que comete el modelo en sus predicciones diarias es del orden del 2 %, o 6,842.13 m^3, que representa una tercera parte de la desviación estándar de la serie original. Es decir que los errores medios cometidos están contenidos dentro de una tercera parte de la desviación estándar de la serie.

Como una evaluación en términos globales se ha realizado una predicción de la misma serie pero sin la consideración de las intervenciones del modelo anterior, encontrando que la incorporación de los eventos de intervención producen una reducción del 12 % de los residuos de predicción acumulados a lo largo del año en términos absolutos que equivalen a 346,846.43 m^3. Es importante destacar que la consideración de un número pequeño de eventos de intervención, en total 14, han conseguido la mejora del 12 % antes mencionada, lo cual demuestra la gran influencia que estos eventos tienen en un escenario de predicción de demanda de agua en entornos dénsamente poblados. El cuadro 9.16 presenta los residuos mensuales que comete el modelo E que no incluye las intervenciones comparado con el que si las incluye.

Se presentan también los gráficos donde se comparan los valores observados y modelados conjuntamente con los residuos absolutos que ha producido el modelo. Se ha genera un gráfico para el modelo sin las intervenciones y otro con intervenciones para que los resultados puedan ser comparados. Los meses de Enero (Figuras 9.5, 9.6), Marzo (Figuras 9.9, 9.10), Abril (Figuras 9.11, 9.12), Octubre (Figuras 9.23, 9.24), Noviembre (Figuras 9.25, 9.26)y Diciembre (Figuras 9.27, 9.28) contienen intervenciones. En el resto de los meses no han acontecido eventos de este tipo y por lo tanto la predicción obtenida con ambos modelos es la misma.

Si analizamos los desempeños mes a mes durante el año 2004 encontramos que en aquellos meses en los que ocurrieron eventos y que fueron modelados mediante intervenciones se han obtenido valores muy superiores a los obtenidos ignorando su ocurrencia. El cuadro 9.17 presenta los valores de correlación y de determinación obtenidos mes a mes. Se han destacado en gris los meses con intervenciones.

Mes	Mod. (E) sin interv. Residuos absolutos	Mod. (E) con interv. Residuos absolutos	Diferencia m^3	%
Enero	279,510.16	197,015.79	82,494.36	29
Febrero	141,624.07	141,624.06	–	–
Marzo	214,279.91	184,039.57	30,240.34	14
Abril	411,368.26	345,564.36	65,803.89	15
Mayo	202,754.96	202,754.96	–	–
Junio	208,511.73	208,511.73	–	–
Julio	220,950.62	220,950.62	–	–
Agosto	216,822.87	216,822.874	–	–
Septiembre	191,023.64	191,023.64	–	–
Octubre	254,399.16	195,911.31	58,487.84	23
Noviembre	236,908.19	185,232.89	51,675.30	21
Diciembre	272,915.76	214,771.08	58,144.68	21
Total	**2,851,069.38**	**2,504,222.94**	**346,846.43**	12

Cuadro 9.16: Residuos mensuales acumulados del modelo (E) con y sin intervenciones

Estadístico	Mod. (E) sin interv. r - R^2	Mod. (E) con interv. r - R^2
Enero	0.68 - 0.46	0.89 - 0.79
Febrero	0.86 - 0.74	0.86 - 0.74
Marzo	0.78 - 0.61	0.84 - 0.71
Abril	0.77 - 0.59	0.88 - 0.78
Mayo	0.87 - 0.76	0.87 - 0.76
Junio	0.88 - 0.79	0.88 - 0.79
Julio	0.93 - 0.87	0.93 - 0.87
Agosto	0.92 - 0.84	0.92 - 0.84
Septiembre	0.90 - 0.82	0.90 - 0.82
Octubre	0.83 - 0.70	0.92 - 0.86
Noviembre	0.71 - 0.50	0.88 - 0.78
Diciembre	0.74 - 0.55	0.87 - 0.75
Total	0.88 - 0.78	0.92 - 0.85

Cuadro 9.17: Correlación y R^2 mensuales del modelo (E) con y sin intervenciones

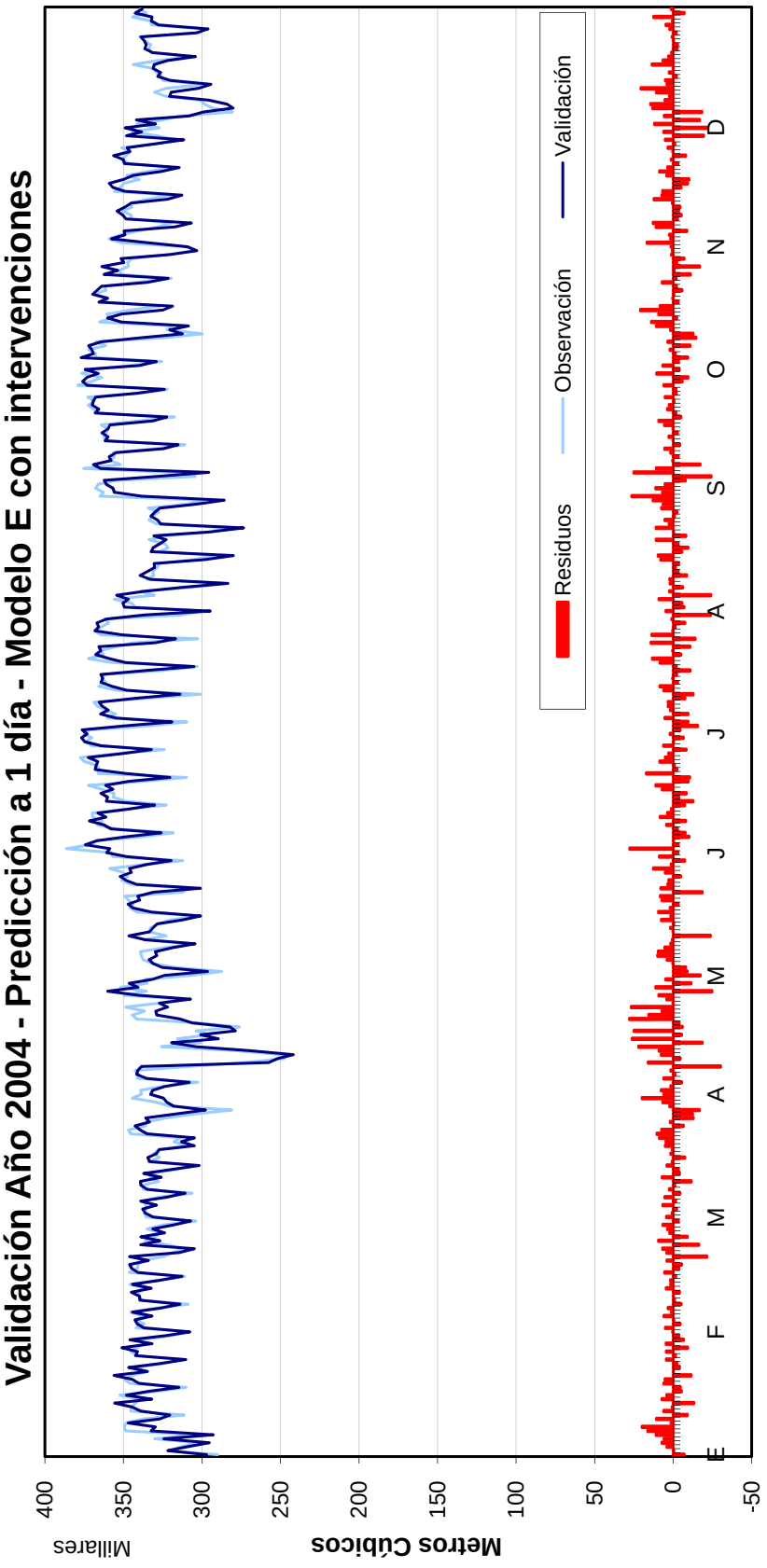


Figura 9.4: Predicción a un día. Gráfico de Validación vs. Observación del, Modelo E con intervenciones. Año 2004

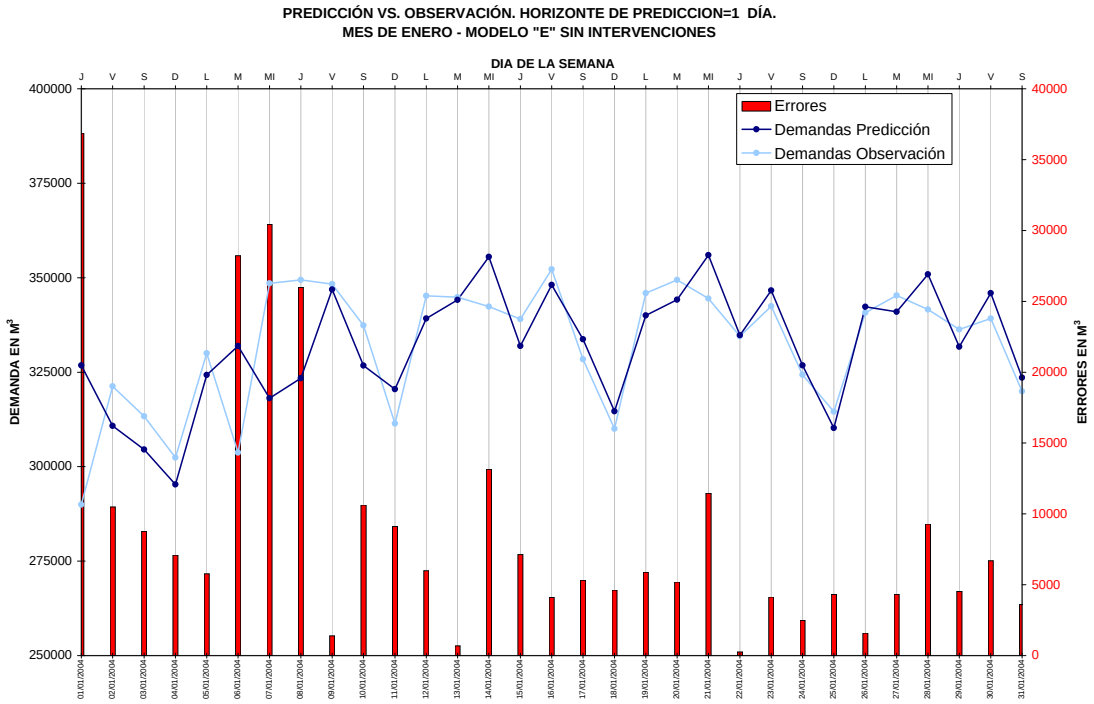


Figura 9.5: Predicción vs. Observación, mes de Enero del 2004. Modelo E sin intervenciones

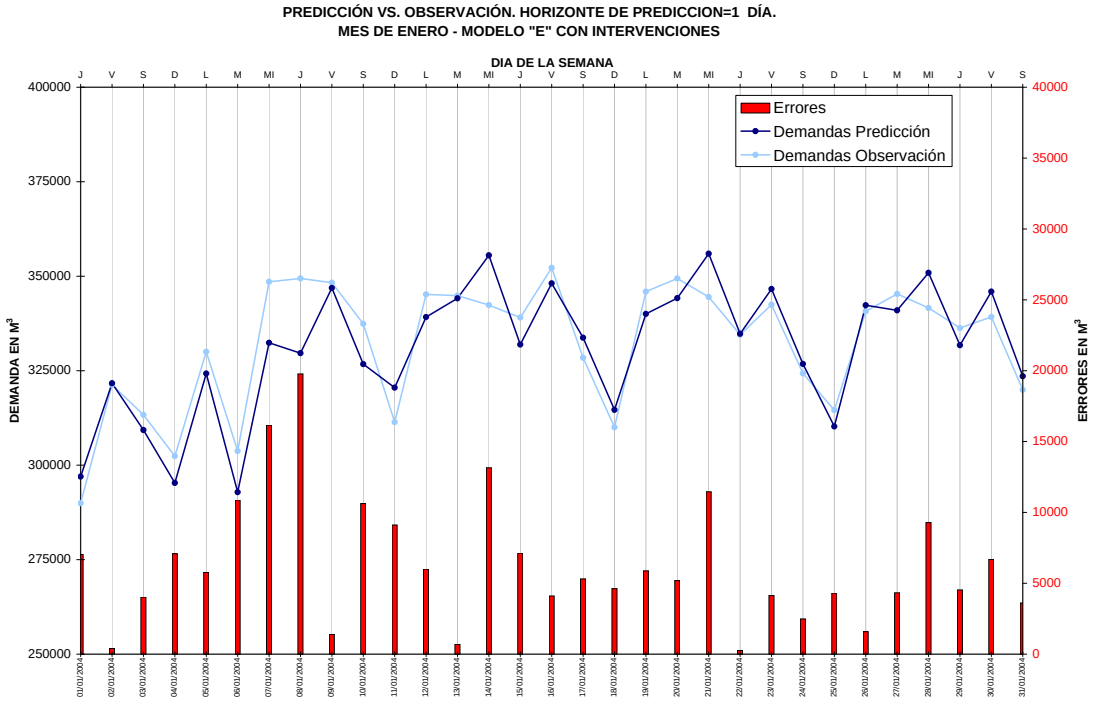


Figura 9.6: Predicción vs. Observación, mes de Enero del 2004. Modelo E con intervenciones

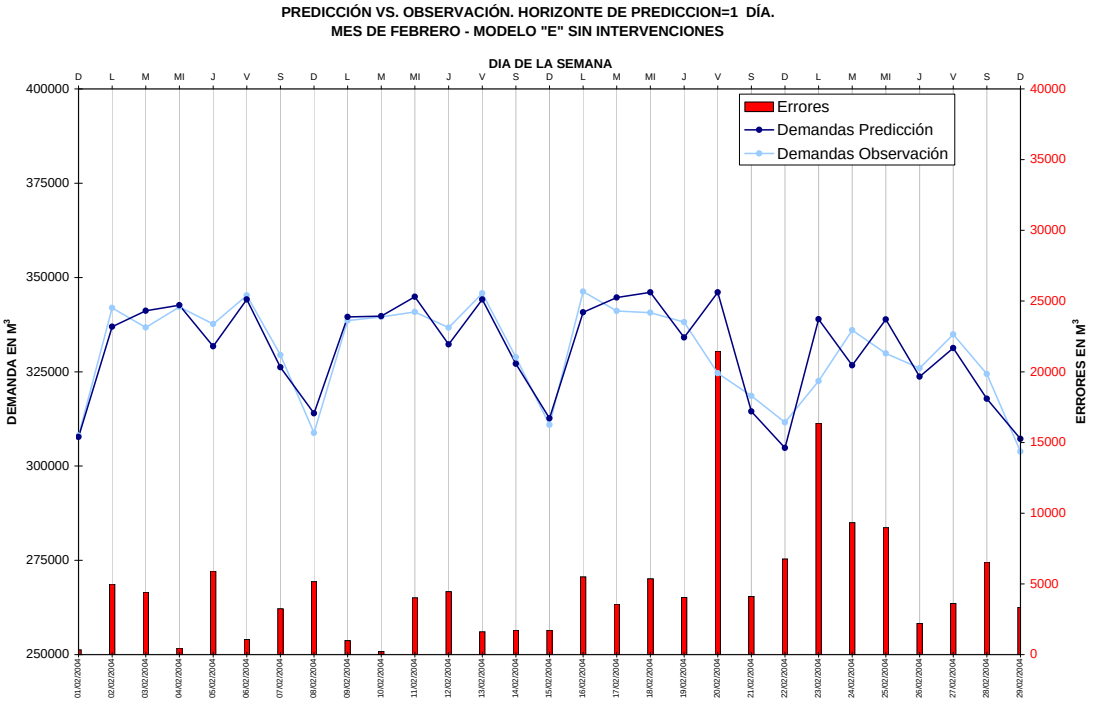


Figura 9.7: Predicción vs. Observación, mes de Febrero del 2004. Modelo E sin intervenciones

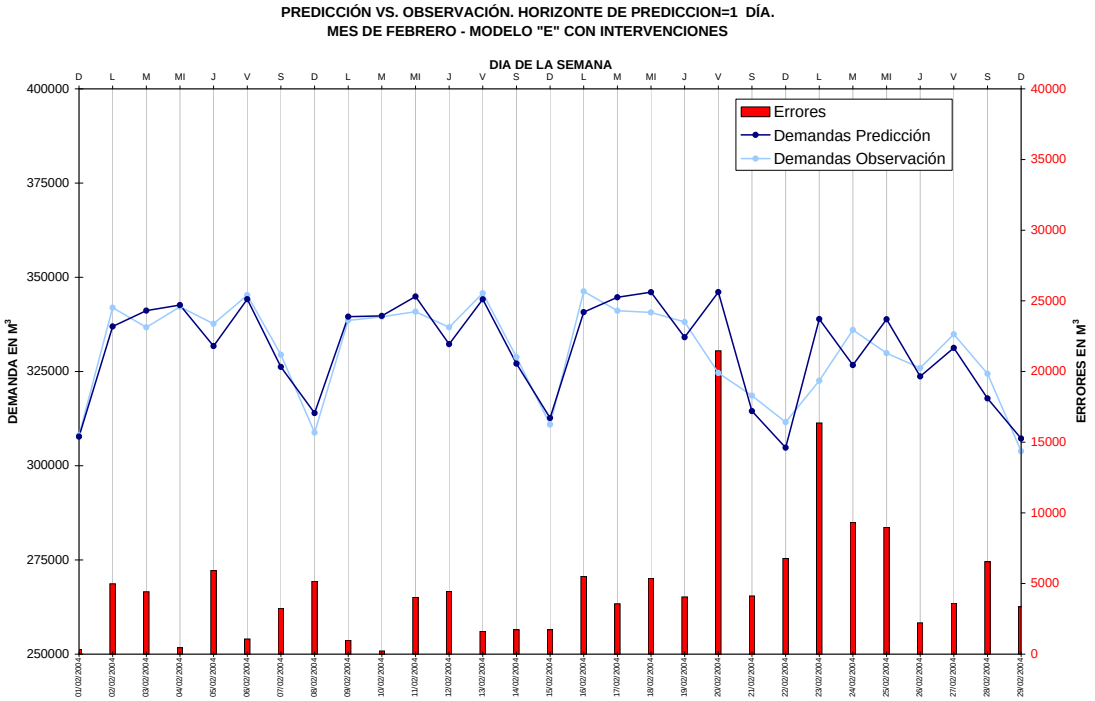


Figura 9.8: Predicción vs. Observación, mes de Febrero del 2004. Modelo E con intervenciones

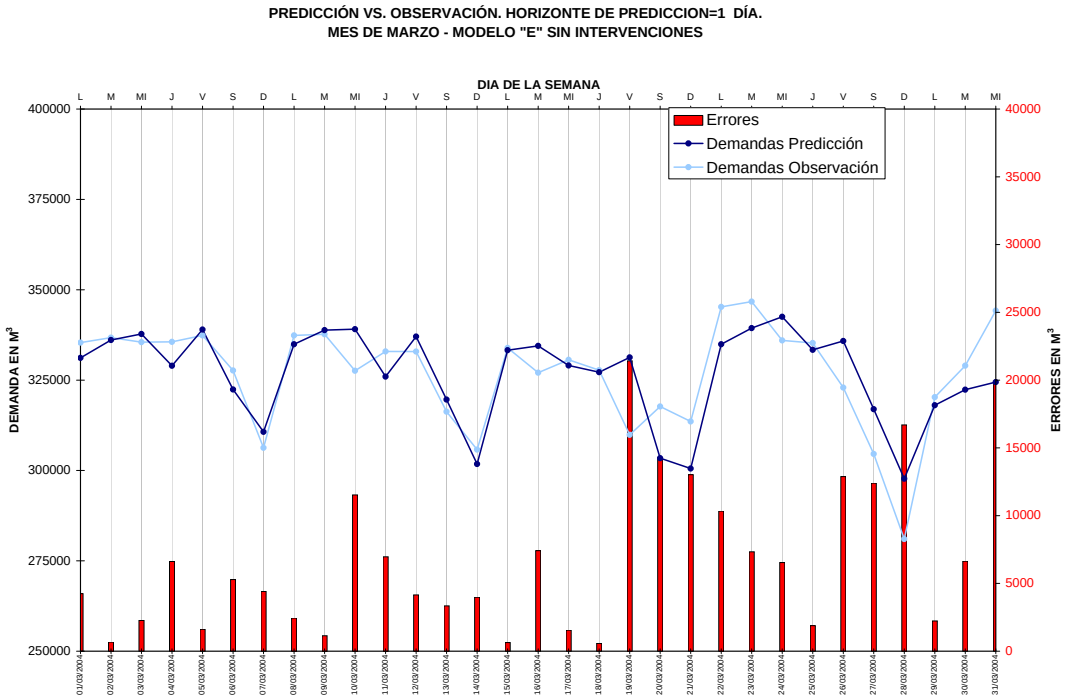


Figura 9.9: Predicción vs. Observación, mes de Marzo del 2004. Modelo E sin intervenciones

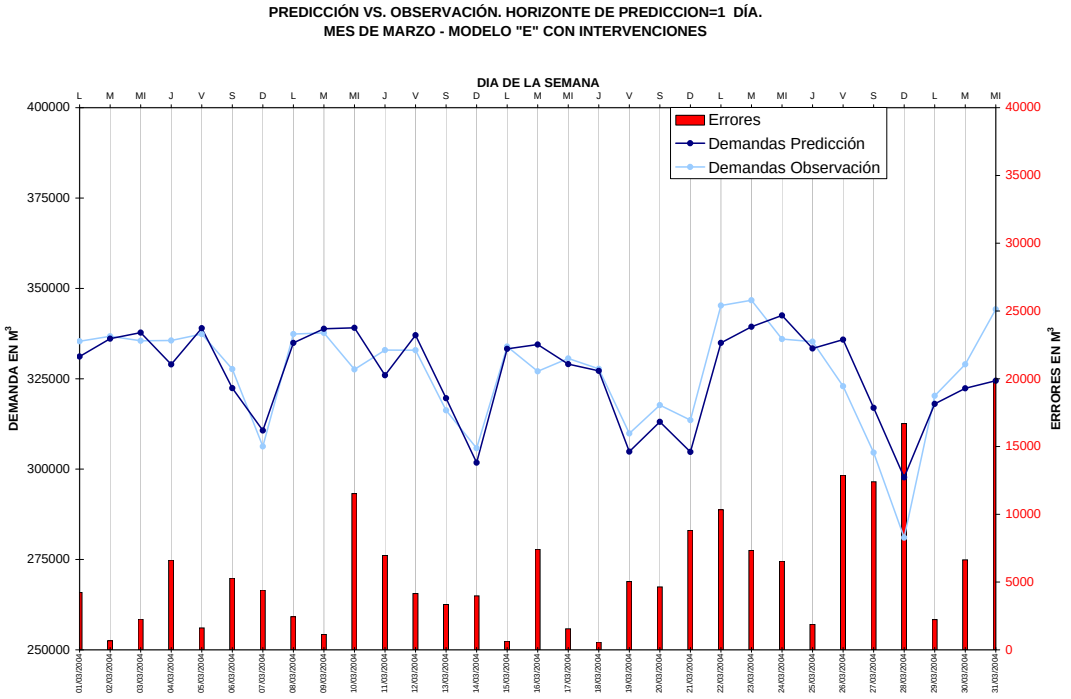


Figura 9.10: Predicción vs. Observación, mes de Marzo del 2004. Modelo E con intervenciones

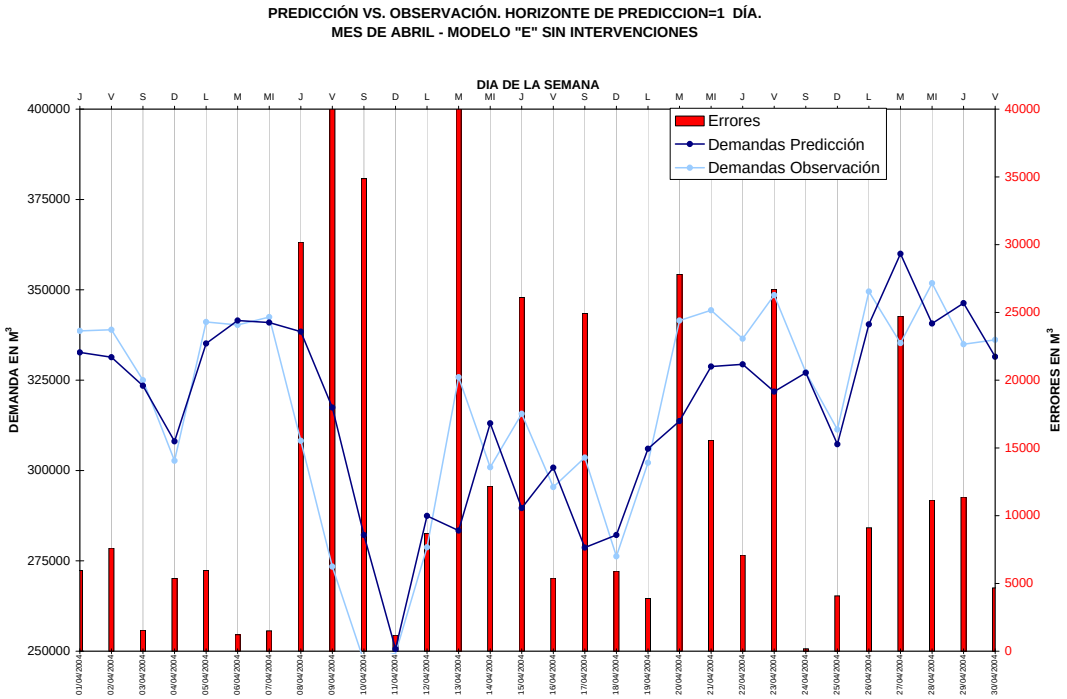


Figura 9.11: Predicción vs. Observación, mes de Abril del 2004. Modelo E sin intervenciones

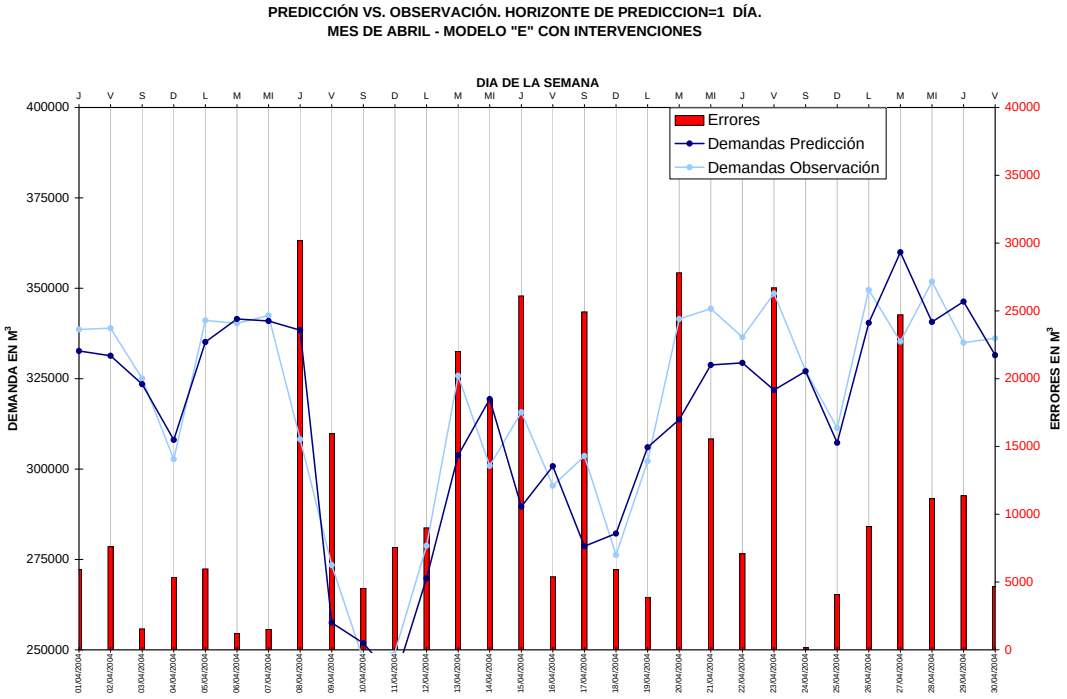


Figura 9.12: Predicción vs. Observación, mes de Abril del 2004. Modelo E con intervenciones

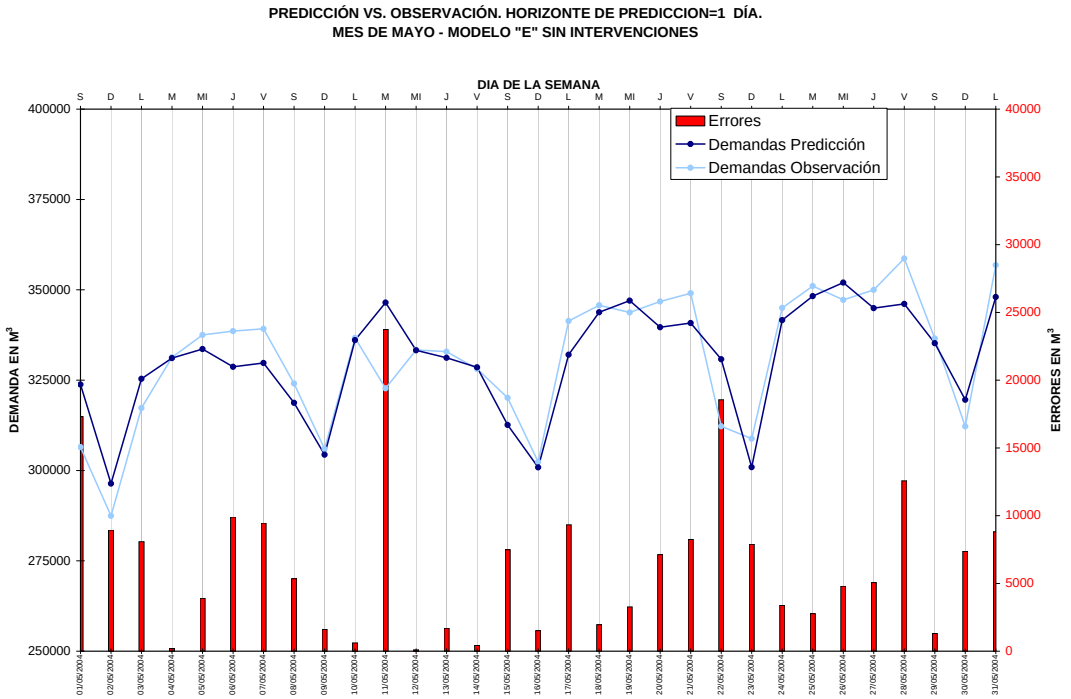


Figura 9.13: Predicción vs. Observación, mes de Mayo del 2004. Modelo E sin intervenciones

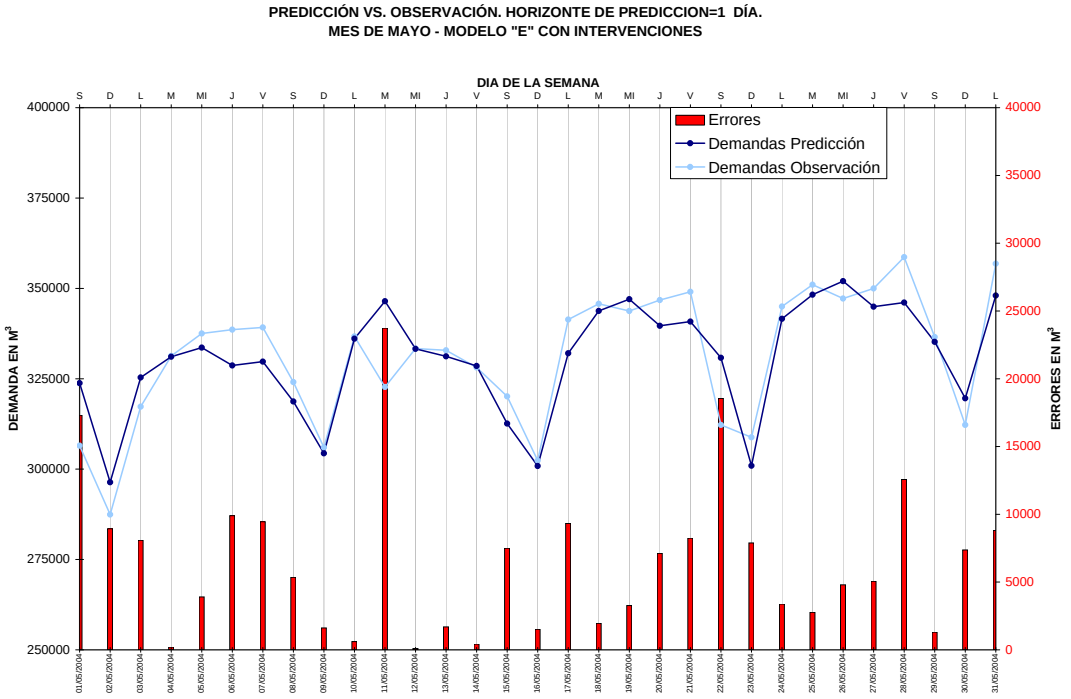


Figura 9.14: Predicción vs. Observación, mes de Mayo del 2004. Modelo E con intervenciones

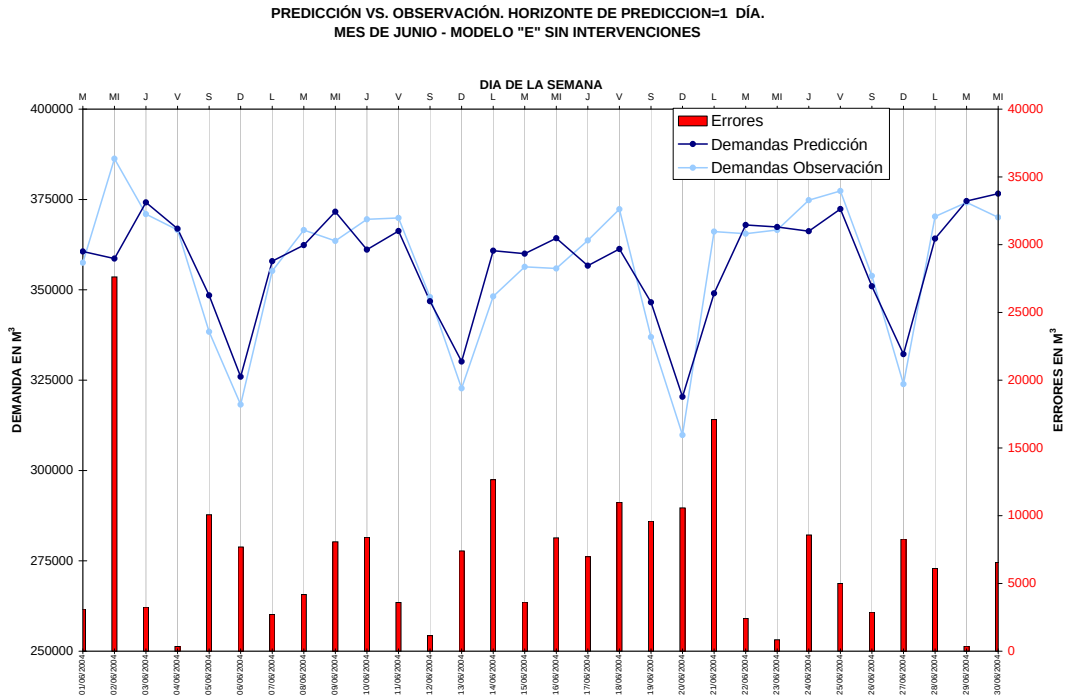


Figura 9.15: Predicción vs. Observación, mes de Junio del 2004. Modelo E sin intervenciones

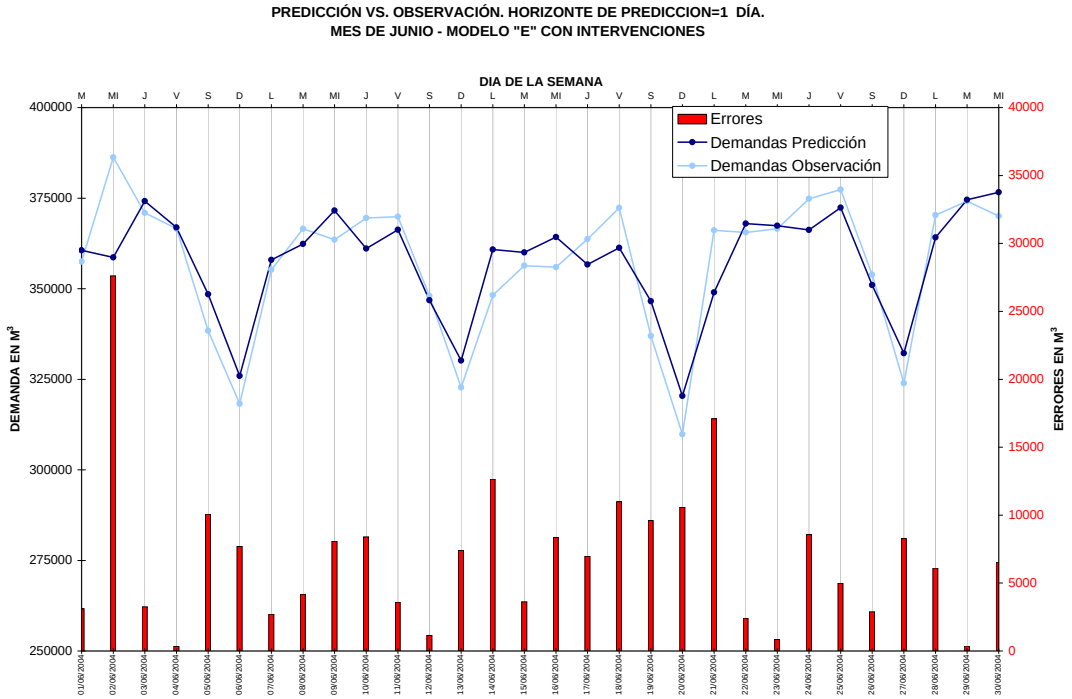


Figura 9.16: Predicción vs. Observación, mes de Junio del 2004. Modelo E con intervenciones

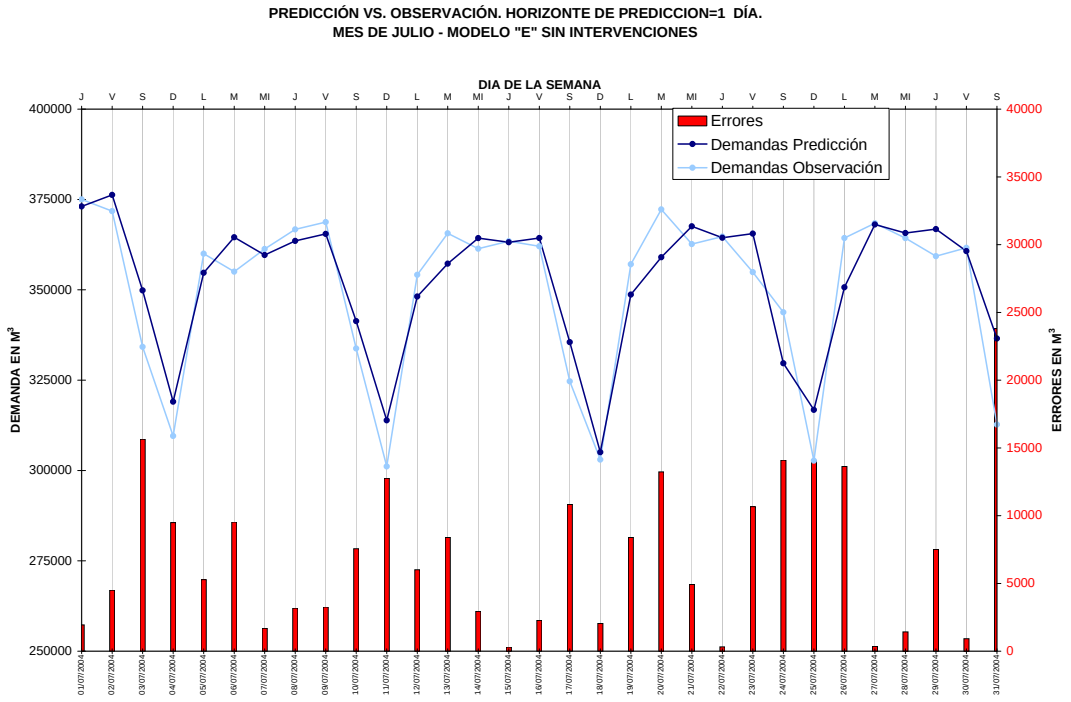


Figura 9.17: Predicción vs. Observación, mes de Julio del 2004. Modelo E sin intervenciones

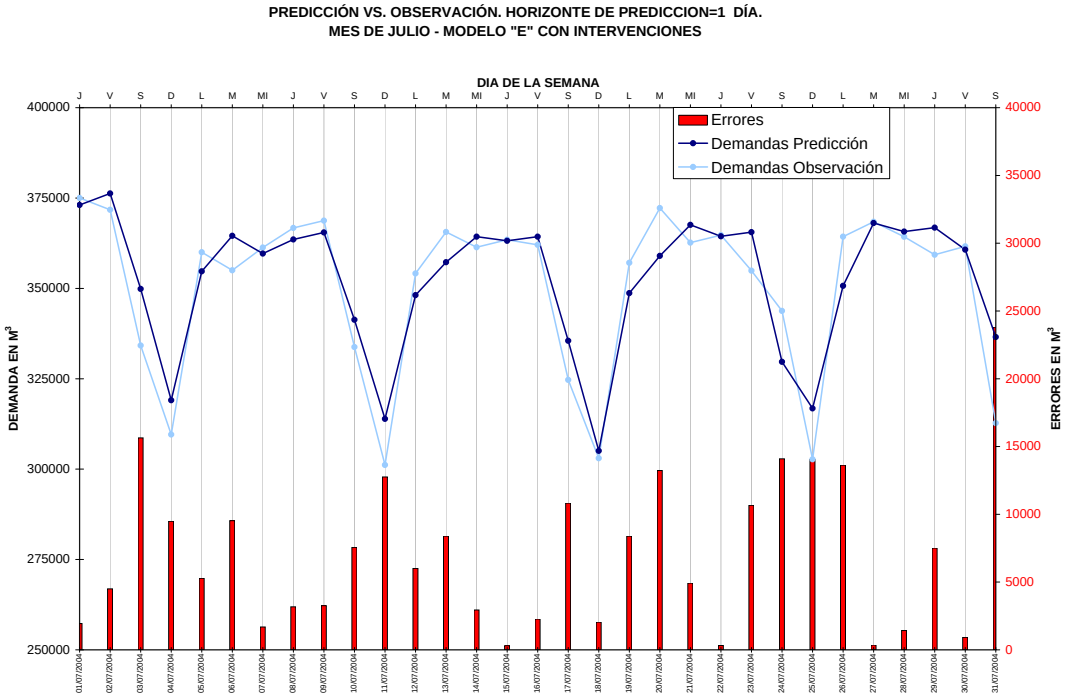


Figura 9.18: Predicción vs. Observación, mes de Julio del 2004. Modelo E con intervenciones

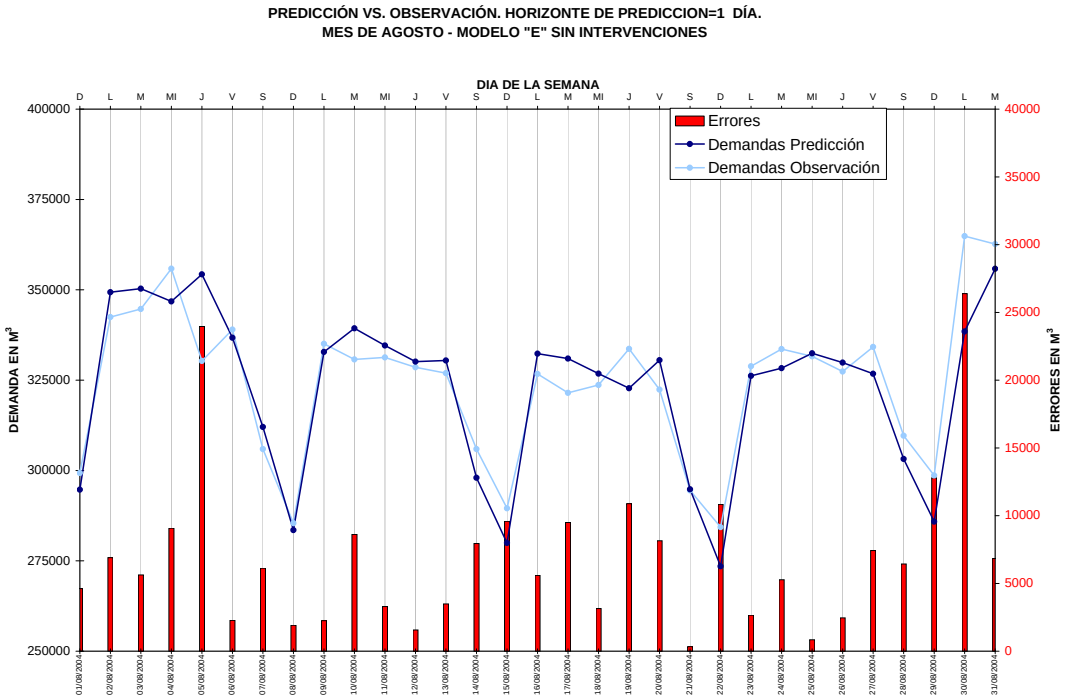


Figura 9.19: Predicción vs. Observación, mes de Agosto del 2004. Modelo E sin intervenciones

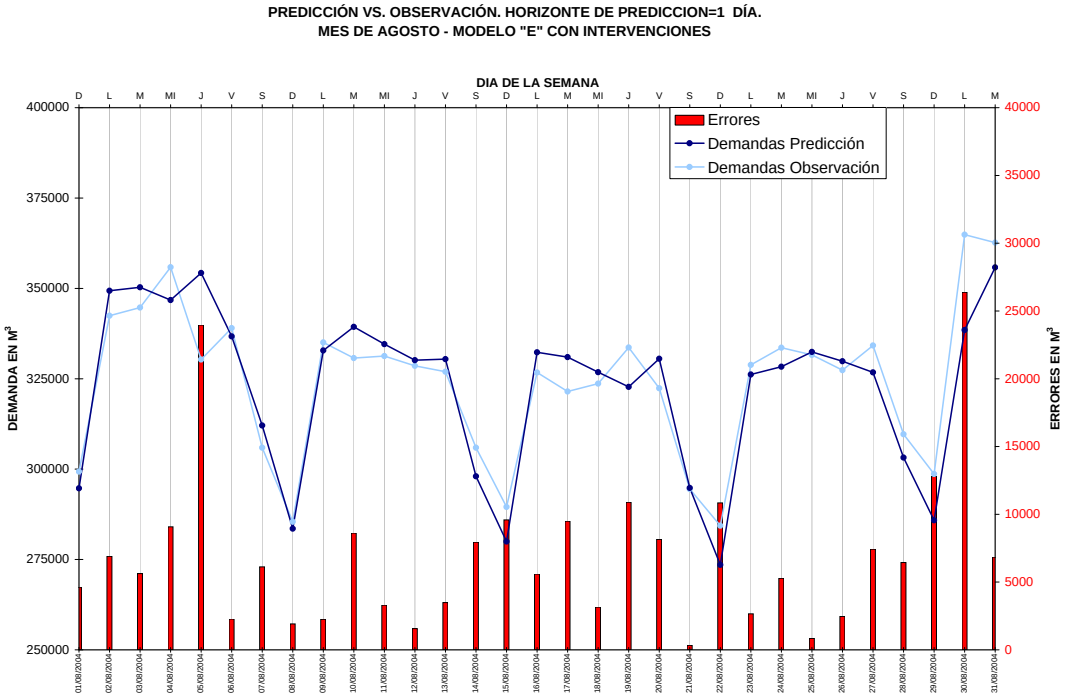


Figura 9.20: Predicción vs. Observación, mes de Agosto del 2004. Modelo E con intervenciones

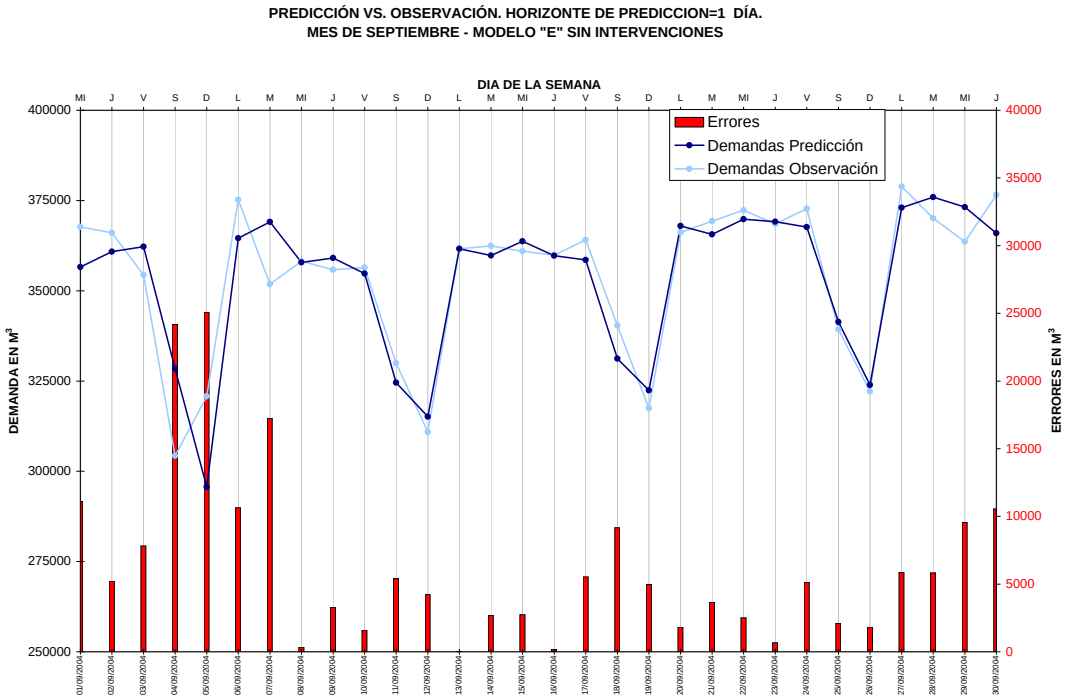


Figura 9.21: Predicción vs. Observación, mes de Septiembre del 2004. Modelo E sin intervenciones

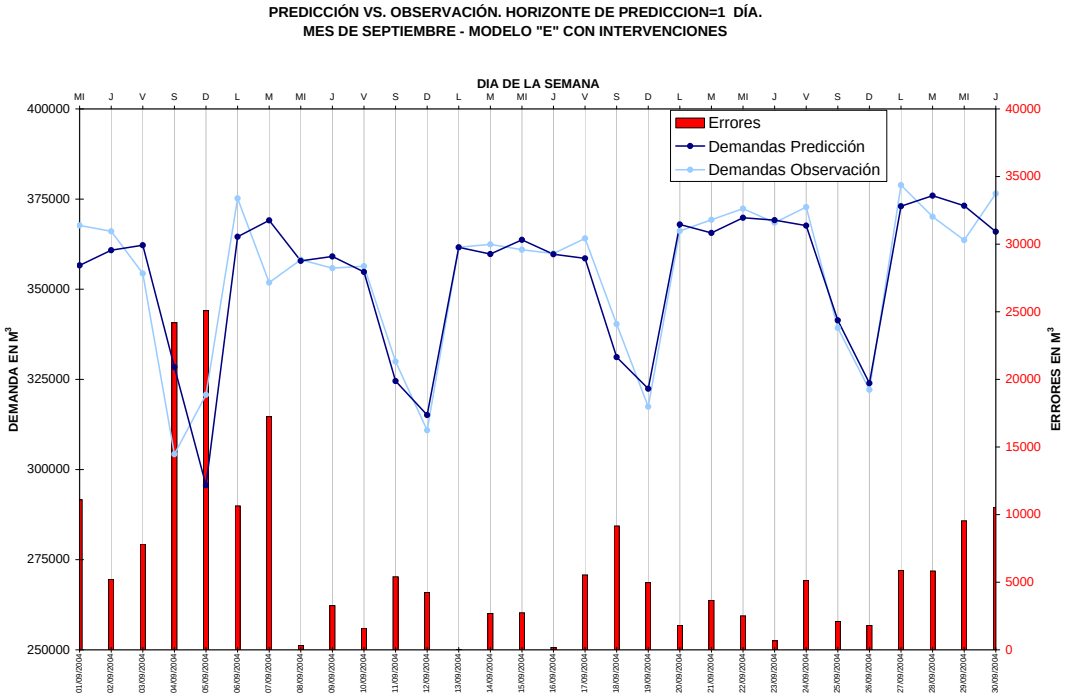


Figura 9.22: Predicción vs. Observación, mes de Septiembre del 2004. Modelo E con intervenciones

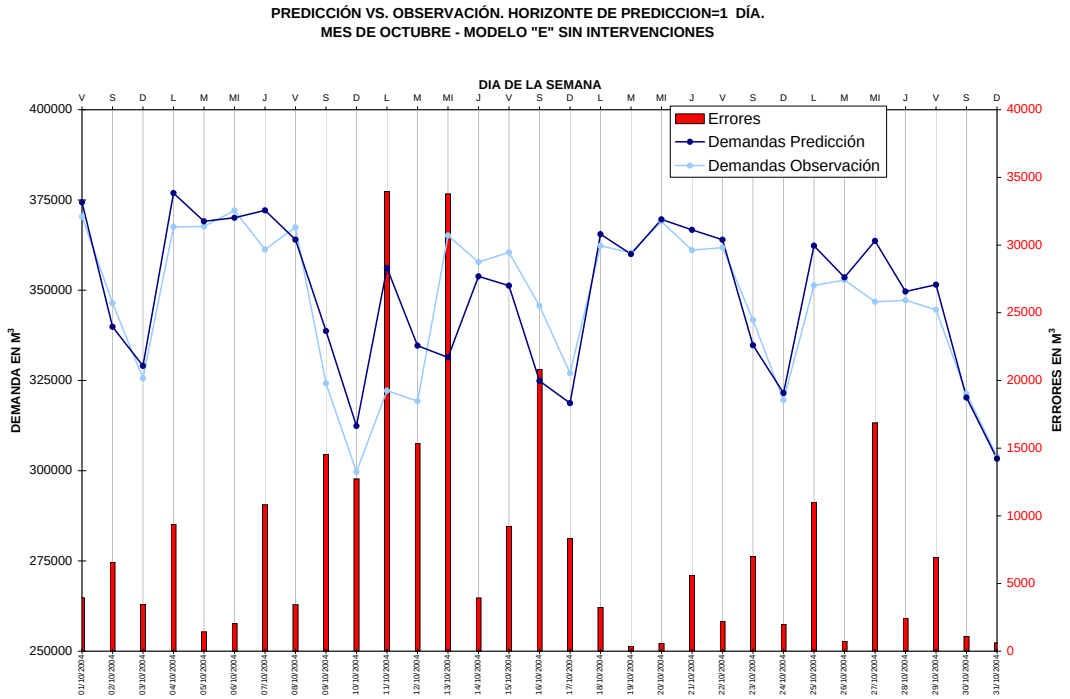


Figura 9.23: Predicción vs. Observación, mes de Octubre del 2004. Modelo E sin intervenciones

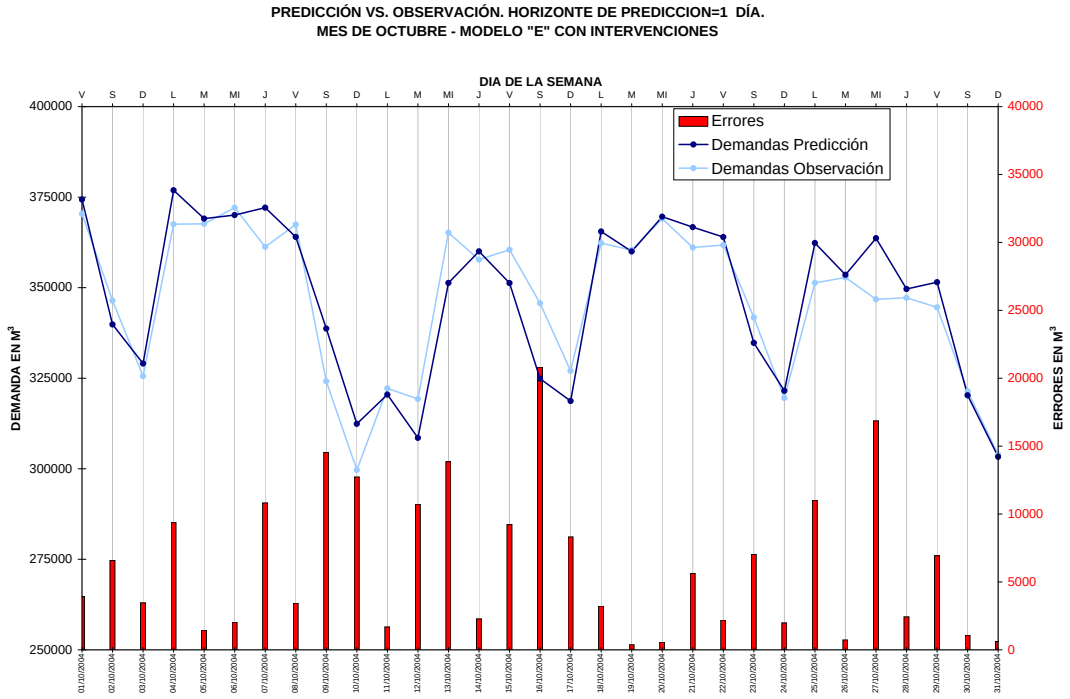


Figura 9.24: Predicción vs. Observación, mes de Octubre del 2004. Modelo E con intervenciones

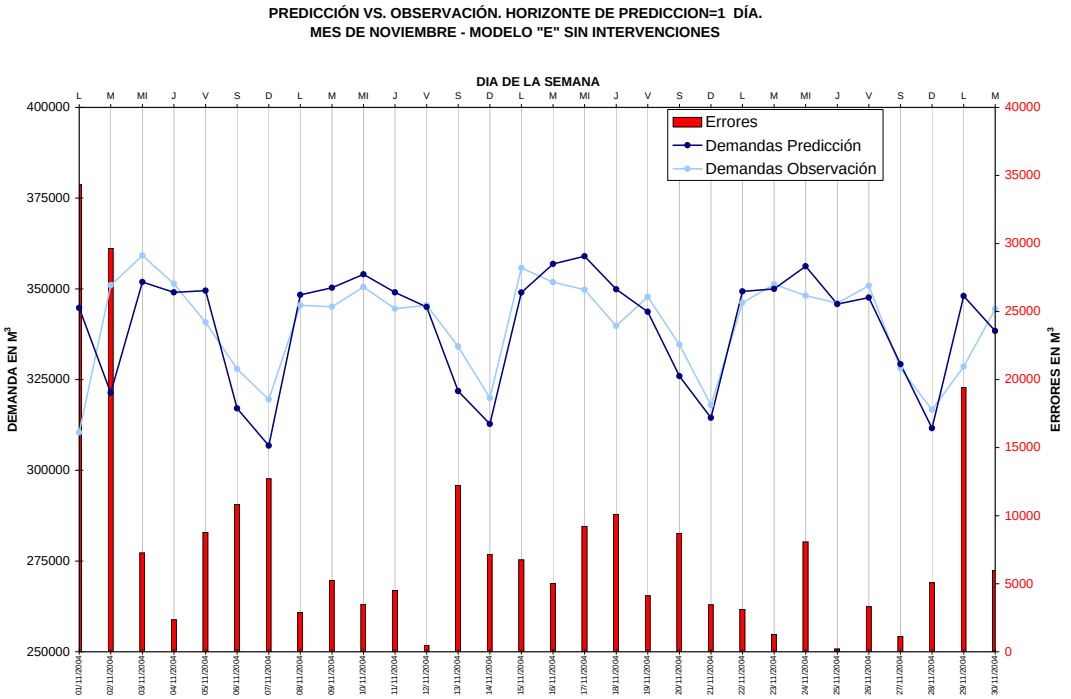


Figura 9.25: Predicción vs. Observación, mes de Noviembre del 2004. Modelo E sin intervenciones

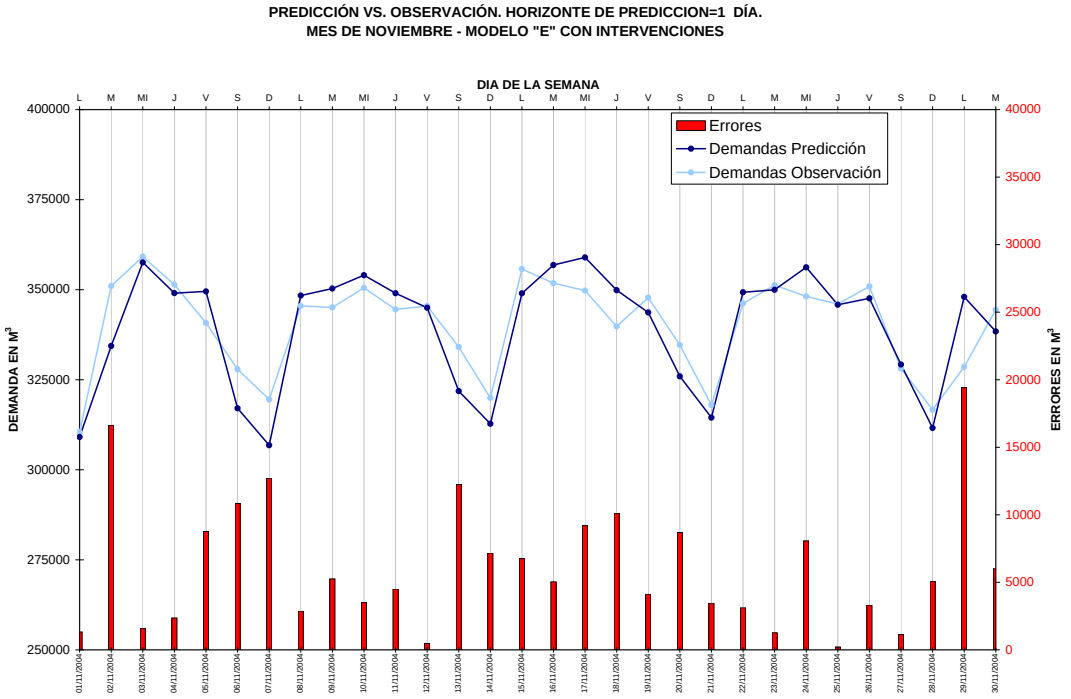


Figura 9.26: Predicción vs. Observación, mes de Noviembre del 2004. Modelo E con intervenciones

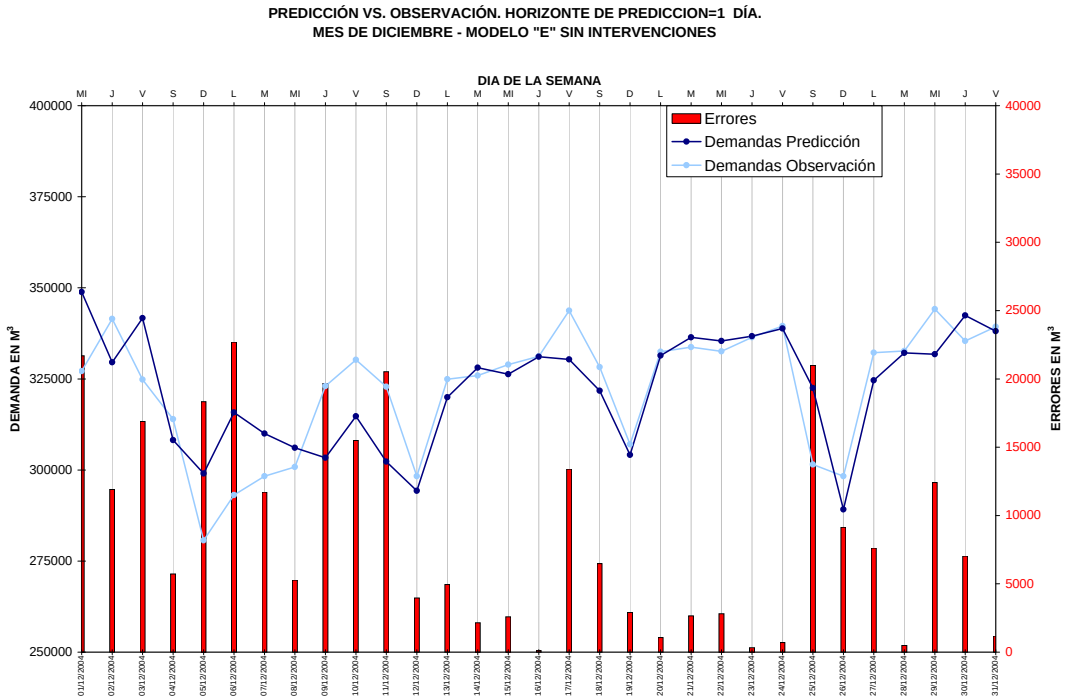


Figura 9.27: Predicción vs. Observación, mes de Diciembre del 2004. Modelo E sin intervenciones

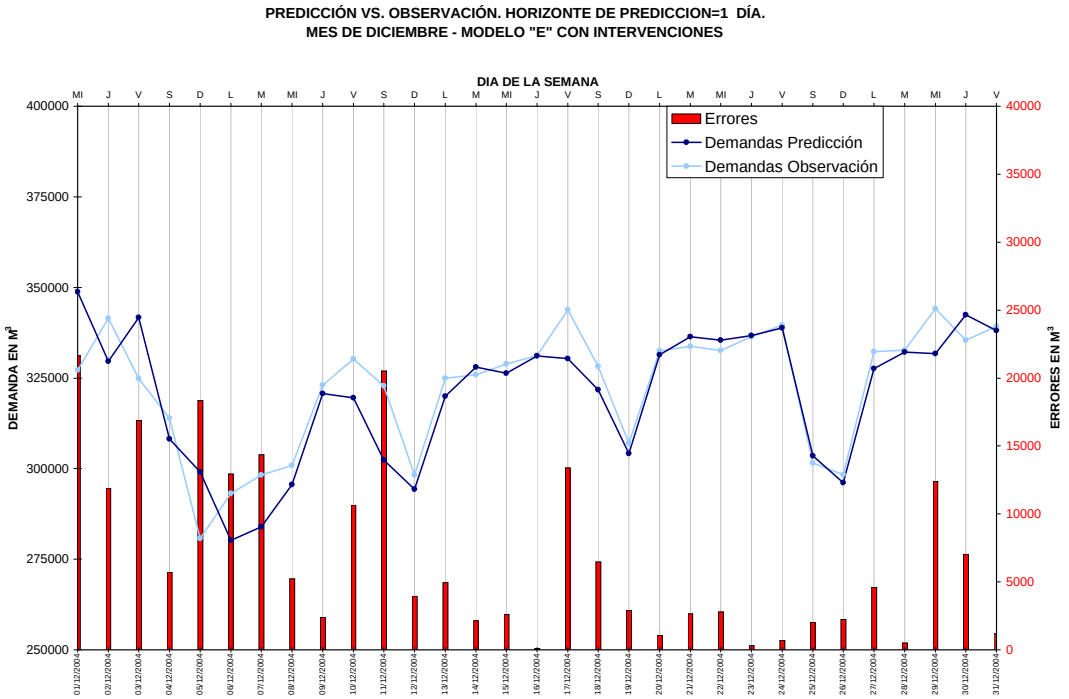


Figura 9.28: Predicción vs. Observación, mes de Diciembre del 2004. Modelo E con intervenciones

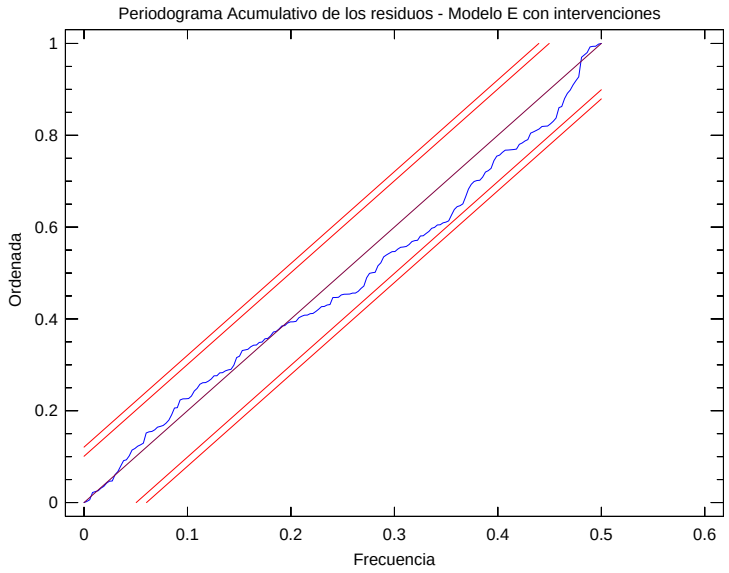


Figura 9.29: Periodograma acumulativo de los residuos de las predicciones, Año 2004. Modelo E con intervenciones

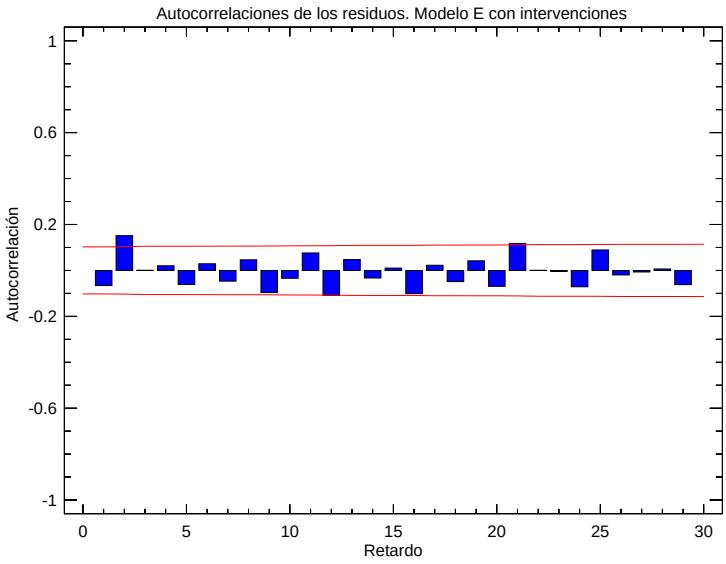


Figura 9.30: ACF de los errores, Año 2004. Modelo E con intervenciones

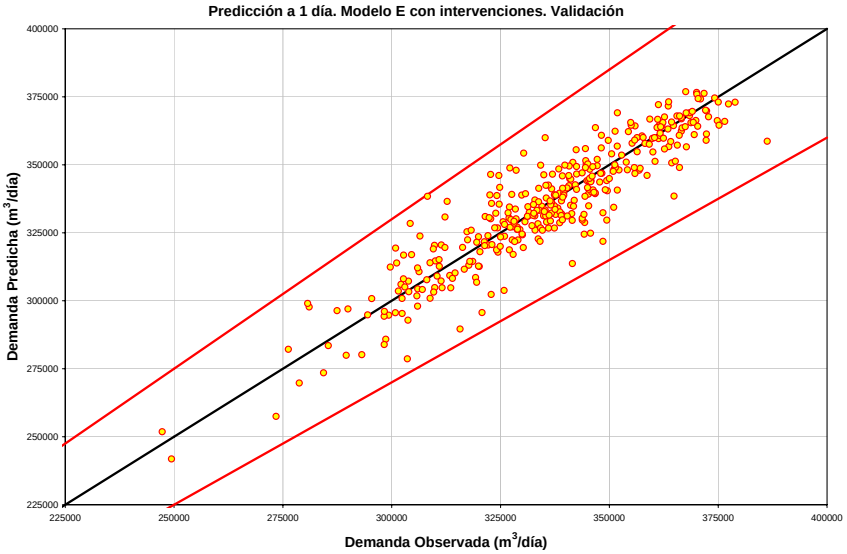


Figura 9.31: Gráfico de Demanda observada vs. Demanda Predicha en fase de validación, Modelo E con intervenciones, predicción a 1 día. Las líneas rojas representan los límites de errores ± 10 % de la demanda observada

Desempeño del modelo ARIMA con intervenciones en fase de validación, predicción a 7 días

Para efectos de planificacion de la operación y la gestión en un sistema operador de agua potable, un horizonte de predicción más grande cobra relevancia. La estructura del modelo que venimos empleando es capaz de ampliar el horizonte de predicción más allá del de un día y hasta un máximo de siete.

Con el mismo conjunto de parámetros estimados para el modelo de predicción a un día, se probará el desempeño para predecir la demanda de agua diaria con un horizonte de predicción de 7 días empleando la metodología propuesta. El mismo conjunto de consideraciones para la predicción a un día se aplican para este nuevo horizonte de predicción. Si acaso es conveniente comentar que para predecir la demanda diaria que se espera dentro de 7 días es necesario predecir previamente la demanda a 1, 2, 3, 4, 5, y 6 días, asumiendo que cada una de las predicciones realizadas es 100% correcta, lo cual en todo momento es falso salvo por coincidencia del azar. Lo anterior quiere decir que supondremos que los residuos de las predicciones serán 0 para toda predicción de más allá de un día, por lo que paulatinamente los coeficientes de media móvil irán reduciendo su impacto y los coeficientes de media movil estacional y los valores de demanda observados siete y ocho periodos antes conjuntamente con la componente estacional, pasarán a guiar el proceso. Por este motivo, para realizar las predicciones, únicamente se han considerado las disminuciones de las intervenciones cuando el día que se desea predecir está afectado por una intervención y cuando ese día es el que aporta los residuos de la componente estacional. Es evidente también que se esperarán mejores desempeños para horizontes de predicción más cortos ya que la incertidumbre es más reducida.

La predicción a 7 días desde el 1 de enero al 31 de diciembre de 2004 se inicia seis periodos antes, es decir que para obtener esta predicción se han obtenido predicciones desde el 26 de diciembre del 2003, siendo ésta la predicción a 1 día, el 27 de diciembre es la predicción a dos días y así sucesivamente hasta el 1 de enero de 2004. Previamente y para asegurar la estabilidad en el inicio de las predicciones, se han obtenido predicciones a 1 día desde el 1 de diciembre de 2003, con lo cual contamos con unos residuos válidos y representativos del comportamiento del proceso.

El modelo que contempla las intervenciones ha obtenido unos coeficientes de correlación r y de determinación R^2 de 0.80 y 0.65 respectivamente para la predicción a 7 días a lo largo del año 2004. Una comparación

de estos valores contra los que obtendría una predicción simple como puede ser un modelo de persistencia, es decir que el valor predicho sea igual al observado 7 periodos antes, se puede presenta mes a mes en el cuadro 9.19.

Si analizamos los desempeños mes a mes durante el año 2004 encontramos que en aquellos meses en los que ocurrieron eventos y que fueron modelados mediante intervenciones se han obtenido valores superiores a los obtenidos con el modelo de persistencia. El cuadro 9.19 presenta los valores de correlación y de determinación obtenidos mes a mes. Se han destacado en gris los meses con intervenciones.

Si bien la predicción es peor que la obtenida a un día, los resultados nos indican que es más válida una predicción utilizando la metodología propuesta, a pesar de las asunciones que se han considerado, que un modelo de persistencia, es decir que el valor de la demanda sea la misma que el de la semana anterior. La predicción obtenida con el modelo servirá no solo como una estimación puntual de la demanda esperada dentro de siete días, sino también como una estimación del volúmen global acumulado a lo largo de los siete días.

Estadístico	**Periodo de Validación (E)**
RMSE	14,253.10
MAE	10,344.19
MAPE	3.11
ME	90.50
MPE	0.06

Cuadro 9.18: Desempeño del modelo (E). Validación a 7 días

Estadístico	Modelo Persistencia r - R^2	Mod. (E) con interv. r - R^2
Enero	0.53 - 0.28	0.70 - 0.50
Febrero	0.79 - 0.62	0.82 - 0.67
Marzo	0.59 - 0.34	0.73 - 0.53
Abril	0.26 - 0.07	0.50 - 0.25
Mayo	0.65 - 0.43	0.72 - 0.53
Junio	0.80 - 0.64	0.79 - 0.63
Julio	0.90 - 0.82	0.89 - 0.80
Agosto	0.78 - 0.61	0.74 - 0.55
Septiembre	0.82 - 0.68	0.82 - 0.68
Octubre	0.55 - 0.31	0.81 - 0.65
Noviembre	0.60 - 0.37	0.87 - 0.76
Diciembre	0.35 - 0.12	0.63 - 4.40
Total	0.63 - 0.44	0.80 - 0.65

Cuadro 9.19: Correlación y R^2 mensuales del modelo persistencia y del modelo (E). Predicción a 7 días

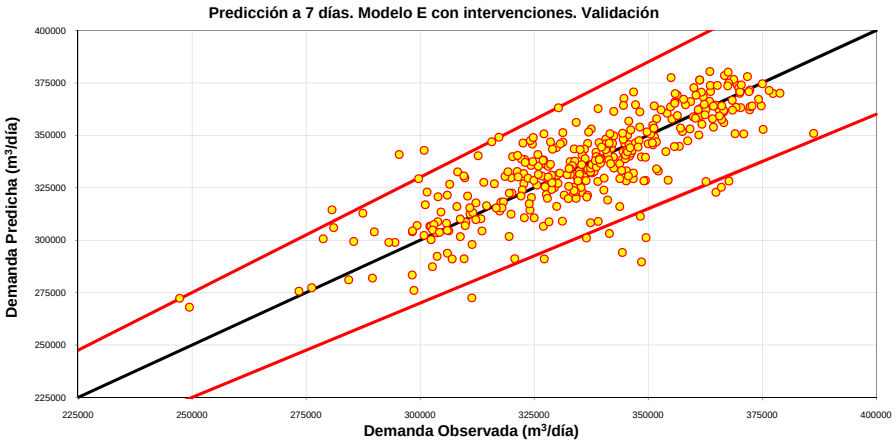


Figura 9.32: Gráfico de Demanda observada vs. Demanda Predicha en fase de validación, Modelo E con intervenciones, predicción a 1 día. Las líneas rojas representan los límites de errores ± 10 % de la demanda observada

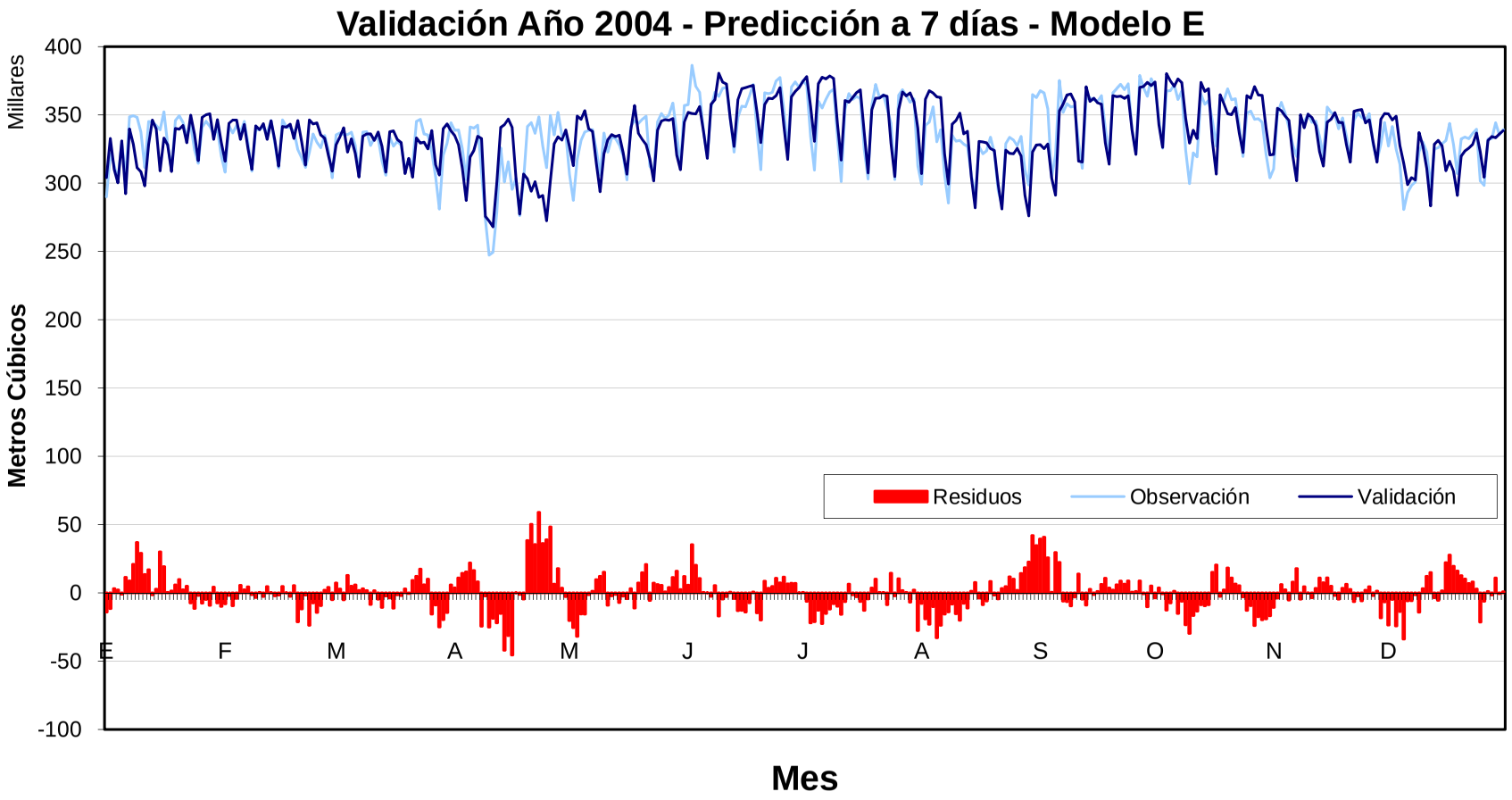


Figura 9.33: Predicción a un día. Gráfico de Validación vs. Observación del, Modelo E con intervenciones. Año 2004

Intervalos de confianza de las predicciones

Un modelo de predicción eficiente deberá entregar no solo las predicciones puntuales, sino también unos intervalos de confianza para una determinada probabilidad. En este caso, una vez analizados los residuos de las predicciones obtenidas y verificado que estos siguen una distribución muy cercana a la normal, podremos estimar los intervalos de confianza de las predicciones. En la sección 5.1.3 se definió que mediante la ecuación 5.10 es posible obtener los intervalos de confianza de las predicciones. Tan solo es necesario conocer la desviación estándar de los residuos que entrega el modelo. Para el modelo E sin intervenciones, $\sigma = 10,895\ m^3$ mientras que para el modelo E con intervenciones, $\sigma = 9,067\ m^3$ y para el caso del modelo C analizado en secciones anteriores este valor es de $\sigma = 15,541{,}3\ m^3$. Con estos datos se han calculado los intervalos de confianza para una probabilidad del 95 % y se presentan en el cuadro 9.20.

Modelo	$\pm 1{,}95\sigma$
Modelo C	30,305
Modelo E sin interv	21,245
Modelo E con interv	17,680

Cuadro 9.20: Intervalos de confianza para el 95 % de probabilidad. En m^3

Es evidente, primero que el modelo C que no contempla de ninguna forma ni los valores atípicos ni los que hemos denominado eventos de variabilidad sistemática irregular obtiene un valor de la desviación estándar (σ) superior y en consecuencia unos intervalos de confianza más grandes. El modelo E sin intervenciones, que ha sido estimado tomando en cuenta tanto atípicos como eventos de variabilidad sistemática irregular en la estimación pero no los ha tomado en cuenta en la fase de validación reduce el valor de sigma y obtiene un valor para el cálculo de los intervalos de confianza 30 % inferior. Finalmente el modelo E con intervenciones que considera todos los eventos tanto en la fase de estimación como en la de validación presenta los valores de σ más pequeños y un intervalo de confianza un 41 % menor que el modelo C y un 17 % menor que el modelo E sin intervenciones.

Parte VII

Conclusiones

Capítulo 10

Conclusiones y líneas futuras de investigación

10.1. *Conclusiones*

El trabajo desarrollado a lo largo de esta tesis estuvo encaminado a proponer un modelo de predicción a corto plazo de la demanda de agua con un horizonte de hasta 7 días. El modelo se basa en la metodología Box-Jenkins y es complementado con técnicas econométricas mediante las cuales se han incorporado variables determinísticas que alteran el proceso de demanda.

El trabajo ha iniciado con una justificación de la necesidad de un modelo predictivo de la demanda de agua urbana, encontrando que se requieren modelos eficientes de predicción que resulten en mejoras en la operación y gestión de sistemas de abastecimiento y distribución de agua potable con las cuales se consiga reducir los costes económicos y medioambientales asociados. Posteriormente se han presentado el conjunto de teorías y metodologías en las que se fundamentan las distintas técnicas que se han utilizado en la modelación y predicción de la demanda de agua potable, introduciendo conceptos y proposiciones que nos han permitido abordar el problema desde una base metodológica sólida.

El trabajo de tesis continúa con una revisión exhaustiva del conjunto de metodologías empleadas para la predicción de la demanda de agua urbana aplicadas para ciudades de diferentes países. Se identificó que la línea de investigación de los modelos estocásticos es la base de la mayoría de las metodologías empleadas, mejoradas y particularizadas para cada región.

Otras líneas de investigación basan sus procedimientos en las técnicas de inteligencia artificial como son las redes neuronales y la lógica difusa. Los resultados obtenidos con ambas líneas no son muy diferentes, sin embargo, se puede destacar que los modelos estocásticos nos permiten inferir las características de la estructura subyacente, es decir el proceso generador de los datos, mientras que por la parte de los modelos de redes neuronales destaca la facilidad de la implementación de un modelo predictivo sin necesidad de asumir hipótesis de normalidad, linealidad, independencia, etc.

De esta revisión se evidenció que las metodologías de uso más común han sido desarrolladas para explicar y predecir el comportamiento de la demanda de ciudades del tipo del sur, sureste de los Estados Unidos de Norteamérica principalmente. Este tipo de ciudades presentan patrones de consumo marcadamente diferenciados a los que se presentan en muchas de las ciudades españolas y mediterráneas, siendo la componente climática la que guía el proceso de demanda a lo largo del año y la meteorológica la que llega a perturbarla temporalmente.

El concepto de demanda base desde donde parten estas metodologías, entendiéndose por éste a la demanda mínima registrada a lo largo del año –típicamente en invierno–, del conjunto de usuarios independientemente de las condiciones climáticas y meteorológicas imperantes, y que está ligado a las necesidades de riego de zonas privadas ajardinadas de los distintos usuarios del servicio, no puede ser extrapolado a las ciudades españolas.

Estas ciudades se han urbanizado de tal forma que el hacinamiento es mucho mayor ya que el recurso suelo es más limitado, resultando en áreas densamente pobladas. En este caso las zonas particulares ajardinadas son limitadas por la disponibilidad de suelo, existiendo principalmente zonas ajardinadas comunes o parques de uso común, por lo que las necesidades de riego son mucho menores en este tipo de ciudades. Es así que las componentes climáticas y meteorológicas pierden relevancia en la composición de la demanda de agua de una ciudad. Por la parte de los valores mínimos, estos se presentan ya sea en los periodos vacacionales, típicamente en agosto –cuando cabría esperar valores de demanda máximos por ser uno de los meses más calurosos– donde se produce una disminución muy importante de la actividad productiva y comercial, o en los periodos de las festividades más importantes de las comunidades. En ambos casos la población de las ciudades tiende a desplazarse a sitios turísticos produciendo una gran cantidad de viviendas registrando demanda nula, por lo que la componente climática y meteorológica es superada por los patrones sociológicos de las comunidades.

Resumiendo las principales diferencias entre los dos tipo de ciudades, destacamos que no existe coincidencia temporal del patrón anual de demandas, ni tampoco en la composición de la demanda de las ciudades.

Los motivos y diferencias mencionados en los párrafos anteriores nos llevaron proponer una metodología que se ajustara a las condiciones de las ciudades españolas densamente pobladas y en las cuales la componente sociológica es la principal fuente de variabilidad. Se inició por probar el desempeño de la metodología Box-Jenkins para identificar y predecir el proceso de demanda en un caso real para la ciudad de Valencia, posteriormente se comparó su desempeño con las metodologías de redes neuronales como una forma de determinar si los resultados eran suficientemente buenos.

De los resultados obtenidos con la metodología de Box-Jenkins se identificó que los modelos ARIMA fueron capaces de capturar y reproducir la variabilidad sistemática regular de la serie temporal en estudio, es decir, las tendencias a largo plazo, las periodicidades de ciclo anual y finalmente las periodicidades semanales repetitivas. Sin embargo, no fueron capaces de capturar ni reproducir los eventos de variabilidad sistemática irregular. Estos eventos representaron el principal aporte de variabilidad puntual del proceso de demanda y están directamente relacionadas con el calendario de festividades de la ciudad. En un escenario de predicción con modelos ARIMA, que utilizan los valores del pasado de la serie para realizar sus predicciones, sus efectos no se limitan únicamente al momento de su ocurrencia, sino que se propagan en las predicciones de varios periodos posteriores dependiendo de la estructura del modelo utilizado.

La incorporación de los eventos de variabilidad sistemática irregular se realizó mediante técnicas que tienen su origen en la econometría como son la identificación de valores atípicos y los modelos de regresión dinámica en su modalidad de análisis de intervención. Con estas técnicas se consiguió en un primer paso identificar las componentes de variabilidad sistemática irregular en la fase de estimación de los parámetros, con lo cual se obtuvieron estimaciones más robustas. En un segundo paso se realizó una caracterización de estos eventos según el día de su ocurrencia para su posterior utilización en un caso de predicción a corto plazo de la demanda.

Los resultados que se obtienen aplicando esta metodología para el caso de validación, nos indican que el modelo ajustado se muestra robusto y estable a pesar de las hipótesis que se asumen para modelar las intervenciones. Los valores de los estadísticos de los errores se reducen en unos porcentajes significativos que justifican la utilización la metodología propuesta. Se obtienen errores absolutos porcentuales medios (MAPE) del orden del 2 % ó 6,842.13 m^3, valores muy pequeños si se considera como ejemplo el caso

del sistema de Valencia que cuenta con un volumen de regularizacion de 90,000 m^3 y que la serie de demandas presenta una desviación estandar de 22,783,54 m^3. Finalmente, los coeficientes de correlación y de varianza explicada que se obtienen son del orden de 0.92 y 0.85 respectivamente. No debemos olvidar que el proceso que estamos modelando y prediciendo es el resultado de un proceso sociológico y no de uno físico como sucede por ejemplo en la modelación hidrológica. Es la respuesta de un conjunto de individuos demandando el agua que requieren para cubrir sus necesidades de consumo y/ó producción. Es esperable entonces, que el proceso sea inherantemente cambiante según la época del año, la climatología imperante, el tipo de día o incluso podría llegar a considerarse la influencia de la situación económica de una comunidad. Se entiende así que siempre existirá una porción de la demanda que no podrá ser predecible por lo que deberemos asumir las predicciones con una incertidumbre inherente al proceso generador de la demanda y con una varianza residual que quedará sin ser explicada.

10.2. *Aportaciones más relevantes*

Las aportaciones más relevantes del presente trabajo de tesis se mencionan a continuación:

- Se ha desarrollado una metodología de predicción adecuada para modelar y predecir el proceso la demanda de ciudades del tipo españolas mediterráneas en áreas densamente pobladas. La metodología combina los modelos del tipo ARIMA con las metodologías del análisis de intervención e identificación de valores atípicos para obtener un modelo de predicción robusto que contempla en su estructura las efectos de las peculiaridades de la demanda propias de cada comunidad.

- La metodología consigue identificar y determinar la magnitud de las peculiaridades de la demanda consideradas como componentes de variabilidad sistemática irregular. Este tipo de peculiaridades son la principal fuente de varianza de la serie de demandas y en consecuencia el origen de los mayores errores de predicción cuando no son tomados en cuenta.

- La metodología propone una caracterización simple de los componentes de variabilidad sistemática irregular, con lo cual se obtiene un catalogo de eventos clasificados según el día de su ocurrencia.

- El modelo que se obtiene incorpora implícitamente en su estructura, al conjunto de eventos que modifican el proceso de demanda y los reproduce de una manera eficiente, produciendo predicciones más precisas.
- La metodología propuesta obtiene unos intervalos de confianza de las predicciones más estrechos que los que se obtienen aplicando un modelo ARIMA sin ningún tipo de intervención. En el caso analizado en esta tesis, los intervalos de confianza resultaron ser un 40 % más estrechos.

10.3. *Sugerencias para futuros desarrollos*

La modelación de la demanda de agua urbana es cada vez más viable, y a la vez ha sido una línea de investigación poco desarrollada hasta estos días. Los avances tecnológicos permiten que los datos de sistemas de distribución de ciudades de diversas características sean registrados y estén disponibles para su análisis.

En algunos capítulos de esta tesis se han analizado las variables climáticas y meteorológicas conjuntamente con la demanda de agua. La revisión bibliográfica nos indica que para algunas ciudades, estas variables llegan a ser muy importantes. En el caso de las ciudades del tipo analizadas en este trabajo, ha quedado de manifiesto que la relación demanda-temperatura no es lineal y además es variable a lo largo del tiempo y del rango de temperaturas del aire registradas. Las características de esta relación requerirá probablemente su incorporación mediante distintos modelos según la época del año, y/o mediante un planteamiento que contenga el concepto de función de calor propuesto por Maidment et~al. (1985) que fue presentado en la sección 3.4.1 de este documento.

Al existir una relación compleja entre las variables predictoras y la variable predicha, es esperable que los modelos que se obtengan para explicar esta relación contengan un alto numero de parámetros. Muy probablemente se requerirá de los modelos de regresión dinámica para modelar la relación. Sin embargo la complejidad de un determinado modelo tendrá sentido, siempre y cuando de el se obtenga información mediante la cual se consiga comprender la relación antes mencionada. Queda este tema como una línea de investigación abierta para futuros desarrollos.

Parte VIII

Apéndices

Apéndice A

Revistas científicas sobre modelación y predicción de la demanda de agua

1. Water resources research
2. Water resources Bulletin
3. Journal of water resources planning and management. ASCE
4. Journal of the american water works association. Management and operations
5. Journal of hydrologic engineering
6. Journal of hydrology
7. Water science and technology
8. Hydrological processes
9. Journal of hydroinformatics. International water association
10. Journal of the american statistical association
11. Journal of the Royal statistical society
12. Simulation
13. Neurocomputing
14. Fuzzy sets and systems

Apéndice B

Días festivos no laborables en Valencia

Mes	Día	Festividad
Enero	1,6,22	Año Nuevo, Reyes, San Vicente Mártir
Febrero		
Marzo	19	San José
Abril		
Mayo	1	Día del Trabajo
Junio		
Julio		
Agosto	15	Virgen de la Asunción
Septiembre		
Octubre	9,12	Día de la Comunidad Valenciana, Día de la Hispanidad
Noviembre	1	Día de todos los santos
Diciembre	6,8,25	Día de la Constitución, Día de la Inmaculada Concepción, Navidad

- Semana Santa varía su ocurrencia
- San Vicente Ferrer es el Lunes posterior a Semana Santa

Cuadro B.1: Días festivos no laborables en la ciudad de Valencia

Bibliografía

Adamowski, J. F. (2008, March). Peak daily demand forecast modelling using artificial neural networks. *Journal of water resources planning and management 134, 119–128.*

Akaike, H. (1974, December). A new look at statistical model identification. IEEE transaction of automatic control 19, 716–723.

Alvisi, S., M. Franchini, and A. Marinelli (2007). A short-term, pattern-based model for water-demand forecasting. *Journal of Hydroinformatics 9, 39–50.*

Aly, A. H. and N. Wanakule (2004, September/October). Short-term forecasting for urban water consumption. Journal of water resources planning and management 130, 405–410.

Anderson, R. L., T. A. Miller, and M. C. Washburn (1980). Water savings from lawn watering restrictions during a drought year, fort collins, colorado. *Water Resources Bulletin 16, 642–645.*

ASCE, T. C. (2000a, April). Artificial neural networks in hydrology. i: Preliminary concepts. Journal of Hydrologic Engineering 5, 115–123. Govindaraju, Rao S.

ASCE, T. C. (2000b, April). Artificial neural networks in hydrology. ii: Hydrologic applications. *Journal fo hydrologic engineering 5, 124–137.*

Bougadis, J., K. Adamowski, and R. Diduch (2005, January). Short-term municipal water demand forecasting. Hydrological Processes 19, 137–148.

Box, G. and G. M. Jenkins (1970). *Time-Series Analysis, Forecasting and control.* San Francisco, California: Holden-Day.

Box, G., G. M. Jenkins, and G. Reinsel (1976). *Time Series Analysis: Forecasting and Control.* San Francisco, California: Holden-Day. Revisión de la version de 1970.

Box, G. and G. C. Tiao (1975, March). Intervention analysis with applications to economic and environmental problems. *Journal of the american statistical association 70, 70–79.*

Brown, R. G. (1963). Smoothing, Forecasting and Prediction*. Englewood Cliffs, NJ: Prentice-Hall.*

Chang, I., G. C. Tiao, and C. Chen (1988, May). Estimation of time series parameters in the presence of outliers. Technometrics 30, 193–204.

Chatfield, C. (1993). Editorial: Neural networks: forecasting breakthrough or passing fad? *International Journal of forecasting 9, 1–3.*

Chatfield, C. (2001). Time-series Forecasting*. Chapman and Hall/CRC.*

Chatfield, C. (2004). The Analysis of time series, An Introduction *(Sixth ed.). Chapman and Hall/CRC.*

Chen, C. and L.-M. Liu (1991). Recent developments of time series analysis in environmental impact studies. Journal of environmental science and health 26, 1217–1252.

Chen, C. and L.-M. Liu (1993a, January). Forecasting time series with outliers. *Journal of forecasting 12, 13–35.*

Chen, C. and L.-M. Liu (1993b, March). Joint estimation of model parameters and outliers effects in time series. Journal of the american statistical association 88, 284–297.

Cutore, P., A. Campisano, C. Modica, Z. Kapelan, and D. Savic (2008). Stochastic forecasting of urban water consumption using neural networks and the scem-ua algorithm. *Urban Water Journal 5, 125–132.*

De la Fuente García, D., R. Pino Diez, C. Suárez Riestra, and J. L. Mayo Rodríguez (1996). Análisis comparativo de los métodos de previsión univariante, box-jenkins, redes neuronales artificiales y espacios de estado. Estudios de Economía Aplicada *(5), 5–33.*

Demuth, H., M. Beale, and M. Hagan (2009, March). Neural Networks toolbox 6. Users Guide. Matlab. *The MathWorks.*

Domokos, M., J. Weber, and L. Duckstein (1976, April). Problems in forecasting water requirements. Water Resources Bulletin 12, 263–275.

Faraway, J. and C. Chatfield (1998, Jun). Time series forecasting with neural networks: a comparative study using the airline data. *Journal of the Royal Statistical Society 47, 231–250.*

Franklin, S. L. and D. R. Maidment (1986, August). An evaluation of weekly and monthly time series forecast of municipal wateruse. Water Resources Bulletin - American Water Resources Association 22, 611–621.

García Bartual, R. (2005, December). Redes neuronales artificiales en ingeniería hidráulica y medio ambiental, fundamentos. Libro de apuntes de la asignatura de Redes Neuronales Artificiales en ingeniería hidráulica y medio ambiental. Doctorado. Universidad Politécnica de Valencia.

Gato, S., N. Jayasuriya, and P. Roberts (2007a, July). Forecasting residential water demand: Case study. *Journal of water resources planning and management 133, 309–319.*

Gato, S., N. Jayasuriya, and P. Roberts (2007b, April). Temperature and rainfall thresholds for base use urban water demand modelling. Journal of Hydrology 337, 364–376.

Ghiassi, M., D. K. Zimbra, and S. H. (2008, March). Urban water demand forecasting with a dynamic neural network model. *Journal of water resources planning and management 134, 138–146.*

Griñó C., R. (1991). Neural network for water demand time series forecasting. In S. B. Heidelberg (Ed.), Artificial Neural Networks*, Volume 540/1991 of* Lecture Notes in Computer Science*, pp. 453 – 460. Springer-Verlag.*

Hansen, R. D. and R. Narayanan (1978, August). A monthly time series model of municipal water demand. Water Resources Bulletin 17, 578–585.

Haykin, S. (1999). *Neural networks. A comprehensive foundation*. Prentice-Hall - Tom Robbins.

Hilera, J. R. and V. J. Martínez (1995). *Redes neuronales artificiales. Fundamentos, modelos y aplicaciones*. Paradigma. RA-MA.

Hipel, K. W., W. Lennox, T. Unny, and A. McLeod (1975, December). Intervention analysis in water resources. *Water Resources Research 11, 855–861.*

Homwongs, C., T. Sastri, and J. W. Foster (1994, November/December). Adaptive forecasting of hourly municipal water consumption. Journal of water resources planning and management 120, 888–905.

Jain, A. and E. Ormsbee (2002, July). Short-term water demand forecast modeling techniques – conventional methodsversus ai. *Journal of American Water Works Association 94, 64–72.*

Joo, C., J. Koo, and M. Yu (2002). Applications of short-term water demand prediction model to seoul. Water Science and Technology 46, 255 – 261.

Koyck, L. M. (1954). Distributed lags and investment analysis. *North Holland Publishing Company, Amsterdam*, 21–50.

Ledolter, J. (1989). The effect of additive outliers on the forecasts from arima models. *International Journal of forecasting 5, 231–240.*

Lingireddy, S. and G. M. Brion (2005). Artificial neural networks in water supply engineering. *American Society of Civil Engineers.*

Liu, L.-M. (2006, October). Time series analysis and forecasting *(Second ed.). Scientific computing associates.*

Maidment, D. R. and S.-P. Miaou (1986, June). Daily water use in nine cities. Water Resources Research 22, 845–851.

Maidment, D. R., S.-P. Miaou, and M. M. Crawford (1985, April). Transfer functions models of daily urban water use. *Water Resources Research 21, 452–432.*

Maidment, D. R. and E. Parzen (1984a, January). Cascade model of monthly municipal water use. Water Resources Research 20, 15–23.

Maidment, D. R. and E. Parzen (1984b, January). Time patterns of water use in six texas cities. *Journal of Water Resources Planning and Management 110, 90 – 107.*

Maier, H. R. and G. C. Dandy (2000, January). Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. Environmental Modelling & Software 15, 101–124.

Makridakis, S. G., S. C. Wheelwright, and R. J. Hyndman (1997). *Forecasting: Methods and Applications* (Third ed.). John Wiley & Sons.

McCulloch, W. S. and W. Pitts (1943). A logical calculus for the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics 5, 115–133.*

Miaou, S.-P. (1990, February). A class of time series urban water demand models with nonlinear climatic effects. Water Resources Research 26, 169–178.

Oh, H.-S. and H. Yamauchi (1974). An economic analysis of the patterns and trends of water consumption within the service areas of the honolulu board of water supply. Technical Report 84, Water Resources Research Center, University of Hawaii. 95 p.

Pankratz, A. (1991, November). *Forecasting with dynamic regression models.* Wiley Series in probability and mathematical statistics. ISBN: 978-0-471-61528-6.

Parzen, E. (1979, March). Nonparametric statistical data modeling. *Journal of the American Statistical Association 74, 105–131.*

Peña, D. (2008). Fundamentos de estadística. *Alianza Editorial S.A.*

Peña, D. (1990). Influential observations in time series. Journal of business & economic statistics 8, 235–241.

Peña, D. (2005). *Análisis de series temporales.* Alianza Editorial S.A.

Protopapas, A. L., S. Katchamart, and A. Platanova (2000, July). Weather effects on daily water use in new york city. *Journal of Hydrologic Engineering 5, 332–338.*

Rojas, I., O. Valenzuela, F. Rojas, A. Guillen, Luis Javier Herrera, H. Pomares, L. Marquez, and M. Pasadas (2008, January). Soft-computing techniques and arma model for time series prediction. Neurocomputing 71, 519–537.

Rumelhart, D. E., G. Hinton, and R. J. Williams (1986). Learning internal representations by error propagation. *MIT Press 1, 318–362.*

Salas LaCruz, J. and V. Yevjevich (1972). Stochastic structure of water use time series. Hydrology paper 52, Colorado State University, Fort Collins. 71 p.

Sastri, T. and J. B. Valdes (1989, July). Rainfall intervention analysis for on-line applications. Journal of Water Resources Planning and Management 115, 397–415.

Seidel, H. (1978). A statistical analysis of water utility operating data for 1965 and 1970. *Journal of American Water Works Association 70, 315–323.*

Shaw, D. T. and D. R. Maidment (1987, December). Intervention analysis of water use restrictions, austin, texas. American water resources association 23, 1037–1046.

Shvartser, L., U. Shamir, and M. Feldman (1993, November/December). Forecasting hourly water demand by pattern recognition approach. *Journal of water resources planning an management 119, 611–627.*

Tang, Z., C. de Almeida, and P. A. Fishwich (1991, November). Time series forecasting using neural networks vs. box-jenkins methodology. Simulation 57, 303–310.

Trivez, J. F. (1994). Efectos de los distintos tipos de outliers en las predicciones de los modelos arima. *Estadística Española 36, 21–58.*

Valencia-Ayuntamiento (2006a). Anuario estadístico de la ciudad de valencia. Technical report, Oficina de Estadística, Plaza del Ayuntamiento 1, 2da planta.

Valencia-Ayuntamiento (2006b). Recull estadístic. Technical report, Oficina de Estadística, Plaza del Ayuntamiento 1, 2da planta.

Valenzuela, O., I. Rojas, F. Rojas, H. Pomares, L. Herrera, A. Guillen, L. Marquez, and M. Pasadas (2008). Hybridization of intelligent techniques and arima models for time series prediction. Fuzzy Sets and Systems 159, 821–845.

Weeks, C. R. and T. A. McMahon (1973, April). A comparison of urban water use in australia and the us. *Journal of American Water Works Association 65, 232–237.*

Winters, P. R. (1960, April). Forecasting sales by exponentially weighted moving average. Management Science 6, 324–342.

Zhang, G., B. E. Patuwo, and M. Y. Hu (1998, March). Forecasting with artificial neural networks: The state of the art. *International Journal of forecasting 14, 35–62.*

Zhang, G. P. (2003). Timer series forecasting using a hybrid arima and neural network model. Neurocomputing 50, 159–175.

Zhang, J., R. Song, N. R. Bhaskar, and M. N. French (2006, August). Short-term water demand forecasting: A case of study. In *8th Annual Water Distribution Systems Analysis Symposium, Cincinnati, Ohio, USA, August 27-30*.

Zhou, S. L., T. A. McMahon, A. Walton, and J. Lewis (2000, September). Forecasting daily urban water demand: a case of study of melbourne. *Journal of Hydrology 236, 153–164.*

Zhou, S. L., T. A. McMahon, A. Walton, and J. Lewis (2002, March). Forecasting operational demand for an urban water supply zone. Journal of Hydrology 259, 189–202.