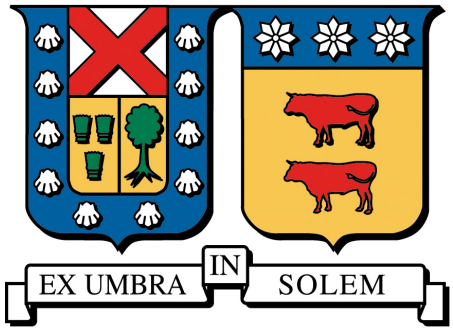
2018

PRONÓSTICO DE DEMANDA ELÉCTRICA UNIVARIADA A CORTO PLAZO MEDIANTE APROXIMACIONES ESTADÍSTICAS Y DE INTELIGENCIA ARTIFICIAL. CASO APLICADO A OPERADOR DE SISTEMA DE TRANSMISIÓN FRANCÉS

ABALLAY LEIVA, BASTIÁN ALEXIS



PRONÓSTICO DE DEMANDA ELÉCTRICA UNIVARIADA A CORTO PLAZO MEDIANTE APROXIMACIONES ESTADÍSTICAS Y DE INTELIGENCIA ARTIFICIAL. CASO APLICADO A OPERADOR DE SISTEMA DE TRANSMISIÓN FRANCÉS.

Tesis de Grado presentado por

Bastián Alexis Aballay Leiva

como requisito parcial para optar al título de
Ingeniero Civil Industrial

y al grado de
Magíster en Ciencias de la Ingeniería Industrial

Profesor Referente:
Dr. Werner Kristjanpoller Rodríguez

Profesor Coreferente Interno:
Dr. Javier Scavia Dal Pozzo

Profesor Coreferente Externo:
Dr. Hugo Garcés Hernández

NOVIEMBRE 2018

TITULO DE LA TESIS:

PRONÓSTICO DE DEMANDA ELÉCTRICA UNIVARIADA A CORTO PLAZO MEDIANTE APROXIMACIONES ESTADÍSTICAS Y DE INTELIGENCIA ARTIFICIAL. CASO APLICADO A OPERADOR DE SISTEMA DE TRANSMISIÓN FRANCÉS.

AUTOR:

Bastián Alexis Aballay Leiva

TRABAJO DE TESIS, presentado en cumplimiento parcial de los requisitos para el Grado de Magíster en Ciencias de la Ingeniería Industrial y de Ingeniero Civil Industrial de la Universidad Técnica Federico Santa María.

Dr. Werner Kristjanpoller Rodríguez .....

Dr. Javier Scavia Dal Pozzo .....

Dr. Hugo Garcés Hernández .....

VALPARAÍSO, Chile. NOVIEMBRE 2018

a Florinda Carolina y Eloísa Italma.

EX UMBRA
IN
SOLEM

AGRADECIMIENTOS

Agradezco a mi familia, Jacqueline, Marcos y Felipe, por el cariño y apoyo indiscriminado.

Agradezco a Paloma, por cambiar su corazón por el mío.

Agradezco a Alejandro, por los años de camaradería.

Agradezco a Hugo, por el buen debate.

Agradezco a todos esos amigos que siguen hasta hoy.

Y también a los que se han ido.

Por todo esto,

y por todo lo que me ha llevado a ser quien soy,

estoy agradecido.

EX UMBRA
IN
SOLEM

RESUMEN EJECUTIVO

El pronóstico de carga a corto plazo (STLF, por sus siglas en inglés) juega un papel fundamental en la planificación y operación eficiente de los sistemas de energía. Los pronósticos a corto plazo precisos ayudan con las decisiones sobre programación de unidades, transferencia energética, planes de mantenimiento y respuesta a la demanda. Diversos modelos han sido desarrollados para obtener pronósticos precisos, sin embargo, pocos se encuentran disponibles gratuitamente para cualquier practicante. En el presente trabajo se comparan enfoques estadísticos y de inteligencia artificial para el pronóstico de la demanda eléctrica cuyo horizonte es un día en adelante. Para este fin se utilizan paquetes de software gratuito que facilitan el modelamiento de series de tiempo y la especificación de modelos estadísticos así como también no-lineales. El análisis comparativo se enfoca en técnicas de pronóstico univariado que pueden establecerse como punto de referencia para modelos más complejos. Para proporcionar un análisis integrado, se realiza Análisis de Datos Exploratorio (EDA) y las visualizaciones necesarias para comprender los datos son entregadas. Los métodos son revisados y comparados por tipo de técnica utilizando la base de datos proveída de manera libre por el sistema de transmisión francés RTE. La codificación estacional determinística para la serie de carga se compara con el enfoque de diferenciación estacional. Se considera la función de autocorrelación y los procedimientos de preprocesamiento de información mutua para llevar a cabo la selección de variables a utilizar en los modelos de inteligencia artificial. En los experimentos numéricos, el promedio de la media de error absoluta de los mejores modelos por técnica revisada fue inferior al 3 %. El modelo Holt Winters con Estacionalidad Doble supera a todos los modelos considerando un año entero como período de prueba. Los modelos de inteligencia artificial logran mayor precisión cuando la doble diferenciación estacional es utilizada en las etapas de preprocesamiento. Este estudio puede ser de utilidad tanto para los operadores del sistema, así como también para los practicantes que buscan una introducción al problema de STLF, centrándose en modelos disponibles al alcance de la mano.

Palabras Clave: Pronóstico de carga a corto plazo, Escenarios diarios, Selección de Variables, Análisis de Datos Exploratorio, Precisión de pronóstico.

ABSTRACT

Short-term load forecasting (STLF) plays a fundamental role in the efficient planning and operation of power systems. Accurate short-term forecasts help with decisions regarding to unit commitment, economic dispatch, maintenance plans and demand response. Several models have been developed to obtain accurate forecasts, however, few of them are freely available to any practitioner. In this work, statistical and artificial intelligence approaches for one day-ahead electricity demand forecasting are compared. To this end, we use free software environment packages that facilitate time series modelling and non-linear model specification. We focus our comparative analysis to univariate forecast techniques that can be established as benchmark for more complex models. To provide an integrated analysis, *Exploratory Data Analysis* (EDA) is performed and the necessary visualizations to understand the data are provided. All methods are reviewed and compared among each technique using the RTE French database. Deterministic seasonal encoding for the load series is compared to the seasonal differencing approach. Autocorrelation function and mutual information preprocessing procedures are considered to perform feature selection of the artificial intelligence input variables. In the numerical experiments, the average mean absolute percent errors of the best models per technique reviewed were less than 3 %. Double seasonal Holt Winters outperforms all models considering one year as test period. Artificial intelligence models were more accurate when double seasonal differencing was used in the preprocessing stages. This study should be useful to system operators as well as practitioners looking for an introduction to the STLF problem with focus on models at hand.

Keywords: Short-term load forecasting (STLF), Day-ahead scenario, Feature Selection, Exploratory Data Analysis, Forecasting accuracy

Índice de Contenidos

Índice de Tablas

Índice de Figuras

Capítulo 1

Introducción

1.1. Motivación

Los sistemas de energía eléctrica corresponden a la elaboración más compleja fabricada por la humanidad, siendo capaces de producir y distribuir electricidad a más de 7.500 millones de personas en todo el mundo. Tal como en la mayoría de las industrias, la industria de la energía eléctrica requiere del pronóstico de niveles de oferta, demanda y precio de recursos para la correcta planificación y operación de las redes de distribución. Mientras algunas de estas industrias poseen formas de inventario para el almacenamiento y regulación de la oferta de sus productos; la industria eléctrica no posee la tecnología necesaria para llevar a cabo tales tareas. Como resultado, la electricidad debe ser generada y distribuida para su utilización inmediata por parte de los consumidores, sea para uso doméstico o industrial. En otras palabras, *los sistemas deben ser capaces de balancear la oferta y demanda en cada instante* [1]. Así, las limitaciones de almacenamiento y la dependencia de la sociedad actual con respecto a la energía eléctrica, hacen necesario que los operadores de sistemas de transmisión posean entendimiento de los patrones relacionados al comportamiento y consumo de electricidad; permitiendo llevar a cabo una correcta y precisa estimación de la demanda eléctrica futura.

El pronóstico energético en la industria de los sistemas de energía eléctrica posee diversos aspectos, como lo son el pronóstico de carga a corto y largo plazo, el pronóstico de carga espacial, de precios, de respuesta de la demanda y de generación renovable. Para llevar a cabo dichas prácticas, las técnicas de pronóstico han experimentado algunas etapas importantes de evolución en la historia de la humanidad, comenzando con aproximaciones gráficas y de tablas en la era pre-computacional, hasta los métodos computacionales más recientes [2]. La inversión en redes de medidores eléctricos inteligentes y la tecnología asociada a éstos ha traído consigo nuevos desafíos al campo del pronóstico energético eléctrico, hallando un nuevo aliciente en la era de la información y la inteligencia de sistemas [3].

El pronóstico de demanda de energía eléctrica (EED) involucra la predicción de valores horarios, semanales, mensuales y anuales del sistema, así como también de sus *peaks* o cargas máximas [4]. En la literatura asociada al pronóstico de EED, poca importancia

se ha otorgado al desarrollo completo de un esquema de análisis exploratorio de los datos (EDA), en conjunto al *preprocesamiento* de dicha información, para luego formular modelos de predicción que sean útiles para la industria. Un pronóstico preciso depende de la implementación de un modelo útil, que a su vez depende de una correcta descripción y uso de los datos disponibles [5]. La cantidad de técnicas originales para llevar a cabo tareas de pronóstico de EED sobrepasa el centenar, sin embargo, no todas son útiles o accesibles a cualquier practicante. La comprensión de que *no existe una técnica universal para obtener el mejor pronóstico* [2] hace necesario que las empresas y sus equipos de planificación entiendan primero las necesidades de su negocio, para luego analizar los datos y - mediante un proceso de prueba y error -, obtener cuál es la mejor técnica para aquel conjunto de datos en específico, en un contexto determinado. Por ende, el error de pronóstico diferirá de manera significativa para cada sistema en particular, para las zonas cubiertas por éste y para diferentes períodos de tiempo.

En la actualidad, no existe consenso para clasificar los horizontes de pronóstico energético. Sin embargo, es posible agrupar los procesos de predicción en cuatro categorías basadas en el horizonte temporal a pronosticar: pronóstico de carga a muy corto plazo (VSTLF), pronóstico de carga a corto plazo (STLF), pronóstico de carga a mediano plazo (MTLF) y pronóstico de carga a largo plazo (LTLF). El horizonte de pronóstico comprendido para estas cuatro categorías son un día, dos semanas, y tres años respectivamente [6].

Los servicios eléctricos pronostican la carga por hora de los sistemas así como también los *peaks* para llevar a cabo tareas de planificación y programación de mantención de generadores, lo que permite elegir en línea la combinación de capacidad óptima que abastecerá a la red. Como algunas instalaciones pueden ser menos eficientes que otras, es natural ponerlas en servicio sólo durante las horas en que la carga predicha será alta. Actualmente, la necesidad por pronósticos a corto plazo precisos es aún mayor. La inclusión de nuevas tecnologías disponibles para la generación y transmisión de energía han logrado que las empresas del rubro eléctrico adquieran pequeños equipos de generación, otorgando mayor flexibilidad al ajuste de capacidad que responde a los requerimientos de los consumidores. Más aún, hoy en día los excesos de generación pueden ser transados, por lo que un cálculo

cuidadoso de la demanda esperada puede conducir a contratos que aumenten la rentabilidad de la instalación y a un mejoramiento del nivel de servicio [7]. La *sobreestimación* de la demanda eléctrica conducirá a una operación conservadora, lo que provoca la utilización y encendido de muchas unidades; o bien la compra de energía en exceso, estableciendo niveles de oferta innecesarios. Por otra parte, la subestimación de la demanda eléctrica genera un estado de operación riesgoso asociado a una demanda insatisfecha. Lo anterior, cobra mayor importancia cuando se considera que un error de 1 % en el pronóstico en términos de porcentaje de error medio absoluto (MAPE) se traduce en ahorros de cientos de miles de dólares por gigawatt (GW) [8].

Debido a la importancia del pronóstico de carga, en las últimas décadas se han reportado numerosos métodos para STLF. Estos procedimientos pueden ser resumidos en aproximaciones determinísticas, estocásticas, de Sistemas Expertos basados en Conocimiento, Redes Neuronales Artificiales (ANN) e interfaces de lógica difusa [9]. Los métodos deterministicos corresponden a modelos de regresión causales clásicos de carga que consideran al clima como variable independiente. Lo anterior, incluye ajuste de curvas, extrapolación de datos y métodos de suavizamiento. Los métodos estocásticos modelan el comportamiento de la demanda en términos de un proceso estocástico. Los filtros de Kalman, médias móviles autoregresivas y aproximaciones de series de tiempo forman parte de esta categoría. Los sistemas expertos basados en conocimiento son modelos construidos a partir de conocimiento de un experto acerca del comportamiento histórico. Aquí, la palabra “técnica” es utilizada para referirse a un grupo de modelos que pertenecen a la misma familia, como lo son los Modelos de Regresión Múltiple (MLR) y ANN. Por otra parte, “metodología” corresponde a la representación de un esquema de solución general que puede ser implementando con múltiples técnicas. Es así como por ejemplo la metodología de selección de variables puede ser aplicada tanto a modelos MLR como a ANN.

La frontera entre las técnicas estadísticas con respecto a las técnicas de IA se hace cada vez más ambigua, como resultado de las colaboraciones multidisciplinarias en la comunidad científica. Un buen sistema de predicción debiese contemplar al menos un par de técnicas de cada grupo a ser implementadas luego de realizar EDA que permita vislumbrar características a explotar por dichos modelos. Las técnicas estadísticas más consideradas son

modelos MLR, modelos aditivos semi-paramétricos, modelos autorregresivos, integrados y de medias móviles (ARIMA) y suavizamientosexponenciales. Por otro lado, las técnicas de IA contemplan entre ellas ANN, modelos de regresión difusa, máquinas de vectores de soporte (SVM) y potenciación del gradiente (*Gradient Boosting*).

Para hacer frente a los desafíos emergentes asociados al pronóstico de energía eléctrica, el Instituto de Ingenieros Eléctricos y Electrónicos (IEEE) ha organizado *Global Energy Forecasting Competition* (GEFCom), una competencia de pronóstico que ha acercado la disciplina a los científicos de datos. Asimismo, operadores de transmisión como RTE [10] enfocados en la mejora del pronóstico a corto plazo para su sistema de operación.

Dado el escenario expuesto en esta sección y considerando la urgencia de realizar aportes a la comunidad científica asociada al pronóstico de demanda de energía eléctrica, la presente memoria de grado busca unificar en un esquema general el análisis exploratorio de datos para una curva de demanda y a su vez considerar un conjunto de modelos estadísticos y de inteligencia artificial para obtener predicciones provenientes de maneras distintas de abordar la misma problemática. Más aún, la disponibilidad de entornos de software libre para computación y gráficos estadísticos como R [11], permiten la utilización de librerías utilizadas por los principales investigadores del área (`tidyverse, lubridate, atsa, forecast, nnfor, nnet`, entre otras) así como también herramientas de visualización y diseño (`ggplot2, lattice`). Lo anterior permitirá aportar con un caso de estudio de investigación reproducible y replicable, siendo útil para investigadores y practicantes, y aportando a la literatura del área con un estudio completo del problema energético descrito.

1.2. Alcance

El presente estudio introduce los procedimientos esenciales para abordar la problemática del análisis exploratorio de curvas de demanda eléctrica, proceso que es necesario para la formulación de modelos predictivos capaces de pronosticar - con el mínimo error posible -, futuros valores de carga eléctrica. En particular, el caso del sistema de transmisión francés RTE es considerado debido a la disponibilidad de datos ofrecidos mediante su herramienta de acceso libre de cargo *éCO$_2$mix* [12]. No obstante, el proceso de análisis exploratorio de

datos, así como también las metodologías de formulación de modelos de pronóstico de EED tratadas en este documento pueden ser replicadas a cualquier país o empresa de servicios que busque aplicar metodologías que forman parte del estado del arte de la problemática energética del nuevo siglo y que busque obtener resultados robustos en sus predicciones.

1.3. Objetivos

1.3.1. Objetivo general

El objetivo general de esta investigación es establecer un marco de trabajo para la caracterización y el pronóstico de la demanda de energía eléctrica univariada a corto plazo, comparando el rendimiento de aproximaciones de inteligencia artificial con el desempeño obtenido a partir de modelos estadísticos relevantes, utilizando como caso de estudio la curva de demanda de un operador de sistema de transmisión eléctrico francés.

1.3.2. Objetivos específicos

- Establecer una metodología integrada para el *análisis de datos exploratorio* así como también el *pronóstico de series de tiempo* para curvas de demanda de alta frecuencia.
- Determinar las bondades y limitaciones de los modelos de inteligencia artificial de pronóstico univariado así como también las de las aproximaciones estadísticas .
- Aplicar los modelos propuestos a la curva de demanda francesa y comparar su rendimiento para el set de datos en específico.
- Implementar las metodologías revisadas en un entorno de software libre (R).

Capítulo 2

Marco Teórico

2.1. Pronóstico de series de tiempo

Muchos problemas de predicción pueden involucrar componentes temporales. La dependencia temporal de éstos usualmente es tanto una restricción como también una estructura que provee de una fuente de información adicional al problema. El objetivo será distinto dependiendo de si existe interés en entender un conjunto de datos o bien realizar predicciones acerca de él. El modelamiento descriptivo o *análisis de series de tiempo* puede ayudar a realizar mejores predicciones, pero no es estrictamente requerido y puede resultar en gran inversión de tiempo cuando se tiene en mente pronosticar el futuro. Una serie puede ser modelada para determinar sus componentes en términos de patrones estacionales, tendencias, relación con factores externos, entre otros, desarrollando modelos matemáticos que otorguen descripciones plausibles a partir de datos muestrales. Por otro lado, el *pronóstico de series de tiempo* usa la información de una serie de tiempo (e incluso información adicional) para pronosticar valores futuros de ella. En este caso, el interés no radica en describir de mejor manera los datos, sino más bien en *ajustar y usar datos históricos para predecir observaciones futuras*. Por ello, la habilidad de un modelo de pronóstico de series de tiempo es determinada por su desempeño prediciendo el futuro.

En particular, los siguientes aspectos son fundamentales cuando se enfrenta un problema de modelamiento predictivo.

Entendimiento de los factores que contribuyen a la predictibilidad de los eventos

Disponibilidad de los datos y frecuencia de su obtención Más datos ofrecen una oportunidad para el análisis exploratorio de los datos, ajuste y testeo de modelos.

Horizonte de tiempo requerido para los pronósticos Definición de corto, mediano y largo plazo para el problema específico.

Frecuencia de actualización del pronóstico Posibilidad de actualizar pronósticos a medida que nueva información se encuentra disponible a menudo puede mejorar la capacidad predictiva.

Frecuencia temporal de requerimiento del pronóstico Si los pronósticos pueden afectar lo que se está tratando de pronosticar .

A modo de ejemplo, los pronósticos de demanda eléctrica pueden ser muy precisos debido a que las tres condiciones generalmente se cumplen. En general, se tiene una idea de los factores que contribuyen a la evolución de ésta. La demanda eléctrica se ve afectada en gran medida por las temperaturas, con efectos más pequeños para las variaciones en el calendario - como las feriados y vacaciones -, y las condiciones económicas. Siempre que haya un historial suficiente de datos sobre la carga y las condiciones climáticas, y que se posean las habilidades para desarrollar un buen modelo que vincule la demanda de electricidad y las variables que la afectan, los pronósticos pueden ser muy precisos.

En el pronóstico de series temporales, un paso clave es saber cuándo algo puede ser pronosticado con precisión y cuando los pronósticos no serán mejores que lanzar una moneda. Buenos pronósticos capturan patrones genuinos y relaciones que están presentes en los datos históricos, pero no replican eventos pasados que no ocurrirán otra vez. Es decir, existe una diferencia entre una fluctuación aleatoria en los datos pasados que debe ser ignorada, y un patrón genuino que debe ser modelado y extrapolado [13]. A menudo es erróneo asumir que el pronóstico no es posible de llevar a cabo en un entorno cambiante. Cada entorno cambia, y buenos modelos de pronóstico capturan la manera en que las cosas cambian. Los pronósticos raramente asumen que el entorno no cambia, lo que normalmente se asume es que la manera en que el entorno cambia continuará en el futuro.

Las organizaciones requieren del desarrollo de sistemas de pronóstico que involucren diversas aproximaciones para predecir eventos inciertos. Tales sistemas necesitan habilidad en la identificación de problemas de pronóstico, aplicando un rango de métodos y seleccionando los apropiados para cada problema, mientras se evalúan y refinan los métodos de pronóstico en el tiempo. Una vez determinados los pronósticos que serán requeridos, es necesario hallar o coleccionar los datos en los que se basarán los pronósticos. La información requerida para predecir podría existir y estar disponible. Actualmente, gran parte de la información se encuentra almacenada y es tarea del practicante identificar dónde y cómo los datos requeridos están almacenados. Así, gran parte del tiempo de estudio será utilizada en ubicar y compaginar datos disponibles previo al desarrollo de modelos de pronóstico

adecuados.

Las *variables predictivas* son a menudo útiles en pronóstico de series de tiempo. Retomando el ejemplo anterior, suponiendo que se desee estimar el pronóstico de la demanda eléctrica horaria de una región en particular. Un modelo con variables predictivas podría ser de la siguiente forma

$$\begin{aligned}\text{EED} = f(&\text{Temperatura, población, fortaleza de la economía,Hora del día,}\\ &\text{Día de la semana, error}).\end{aligned} \tag{2.1}$$

Dicha relación no es exacta, pues siempre existirán cambios en la demanda que no pueden ser considerados por las variables predictivas. El término del *error* admite variaciones aleatorias y los efectos de variables relevantes que no están incluidas en el modelo. En particular, dichos modelos son llamados *modelos explicativos* porque permiten explicar qué causa la variación en la demanda eléctrica. Dado que la demanda es también una serie de tiempo en sí misma, es posible utilizar modelos de series de tiempo para llevar a cabo pronósticos. Así, una ecuación adecuada de series de tiempo podría tener la forma

$$\text{EED}_{t+1} = f(EED_t, EED_{t-1}, EED_{t-2}, EED_{t-3}, \ldots, \text{error}). \tag{2.2}$$

En este caso, la predicción del futuro está basado en valores pasados de una variable, pero no en variables externas que podrían afectar el sistema. El término del error cumple el mismo rol de la Ec. (2.1). Finalmente, el resultado de combinar las dos aproximaciones previamente expuestas podría ser

$$\text{EED}_{t+1} = f(EED_t, \text{Temperatura, Hora del día, Día de la semana, error}), \tag{2.3}$$

lo que se denomina *modelo mixto*, y también es conocido como modelo de regresión dinámica, modelo longitudinal o de datos de panel, modelo de función de transferencia y (asumiendo que f es lineal) sistema de modelos lineales. Los modelos explicativos son útiles dado que incorporan información de otras variables, en lugar de sólo valores históricos de la variable a pronosticar. Sin embargo, existen variadas razones para que un

practicante seleccione un modelo de serie de tiempo en lugar de un modelo explicativo y mixto, entre ellas:

- El sistema puede no ser entendido o bien es extremadamente difícil medir las relaciones que se asumen gobiernan su comportamiento
- Es necesario saber o pronosticar el valor futuro de las variables predictivas para ser capaces de pronosticar la variable de interés, lo que en ciertos casos puede ser difícil.
- La principal preocupación puede ser sólo predecir lo que ocurrirá, no explicar por qué ocurre.

El modelo a ser utilizado para generar pronósticos depende de los recursos y datos disponible, la precisión de los modelos evaluados, y la manera en que el modelo de pronóstico será utilizado.

2.1.1. Etapas de las tareas de pronóstico

Un problema de pronóstico habitualmente involucra las siguientes cinco etapas:

Etapa 1: Definición del problema Requiere entendimiento de la manera en que los pronósticos serán utilizados, quién los requiere y de qué manera la función de pronóstico se ajusta dentro de la organización que requiere el pronóstico.

Etapa 2: Reunión de información Se requiere al menos de datos estadísticos y del nivel de habilidad de quienes colectan los datos y usan los pronósticos.

Etapa 3: Análisis de datos exploratorio Exposición gráfica de los datos que permite estudiar patrones consistentes (véase sección 2.4.3.1), tendencias, estacionalidad, evidencia de ciclos, *outliers* u observaciones atípicas y relaciones entre variables disponibles para analizar.

Etapa 4: Elección y ajuste de modelos El mejor modelo a usar depende de la disponibilidad de datos históricos, la fuerza de las relaciones entre la variable pronosticada y cualquier variable explicativa, y la manera en que los pronósticos serán utilizados. Es

habitual comparar el potencial de diversos modelos. Cada modelo en sí mismo es una construcción artificial basada en un conjunto de supuestos (explícitos e implícitos) e involucra uno o más parámetros que deben ser estimados usando la data histórica conocida.

Etapa 5: Uso y evaluación de modelos Una vez elegido el modelo y estimados sus parámetros, el modelo puede ser usado para producir pronósticos. El rendimiento del modelo sólo puede ser evaluado luego de que los datos para el período de pronóstico resultan disponibles, para lo cual varias metodologías han sido desarrolladas.

2.1.2. Perspectiva estadística del pronóstico

Aquello que se desea pronosticar es desconocido - pues si no lo fuese no tendría sentido pronosticarlo -, y puede ser entendido como una *variable aleatoria*. En la mayoría de las situaciones que involucran pronósticos, la variación asociada con aquello que se desea pronosticar se reducirá a medida que el evento se aproxima. Es decir, cuanto más alejado el pronóstico, más incierto será.

Es posible imaginar muchos futuros posibles, cada uno con un valor diferente para lo que se busca pronosticar. Cuando un pronóstico es obtenido, en realidad se estima la media del rango de los valores posibles que la variable aleatoria podría tomar. A menudo, el pronóstico se acompaña de un *intervalo de predicción* otorgando un rango de valores que la variable aleatoria puede tomar con relativa alta probabilidad. A modo de ejemplo, un intervalo de predicción de 95 % contiene un rango de valores que debiese incluir el valor futuro estimado con probabilidad 95 %. Así, el promedio de los posibles valores futuros es llamado *pronóstico puntual* para determinado período t en el tiempo. Suponiendo que toda la información disponible al período t se denota I y que se desea pronosticar y_t, entonces es posible escribir $y_t|I$ para referirse a "la variable aleatoria y_t dado lo que se sabe en I". El conjunto de valores que esta variable aleatoria puede tomar, junto con sus probabilidades relativas es conocido como la "distribución de probabilidad" de $y_t|I$, que corresponde a la distribución del pronóstico. Luego, el pronóstico de y_t se denota por $\hat{y}_t$, que equivale al promedio de los posibles valores que y_t puede tomar dado todo lo que se conoce hasta t.

Finalmente, es posible especificar la información que ha sido utilizada en el cálculo de un pronóstico escribiendo, por ejemplo, $\hat{y}_{t|t-1}$ para referirse al pronóstico de y_t tomando en cuenta todas las observaciones previos $(y_1, \ldots, y_{t-1})$. De manera general, $\hat{y}_{T+h|T}$ denota el pronóstico de y_{T+h} tomando en consideración $y_1, \ldots, y_T$, es decir, un pronóstico a h pasos, considerando todas las observaciones hasta el período T.

2.2. SERIES DE TIEMPO

Una *serie de tiempo* es un conjunto de observaciones obtenidas de manera secuencial en el tiempo [14]. En la mayoría de las ramas de la ciencia, la ingeniería y el comercio, existen variables medidas secuencialmente en el tiempo. Los bancos observan tasas de interés y tipo de cambio cada día. Los departamentos de estadística de entidades gubernamentales calculan el producto interno bruto del país anualmente. Periódicos y páginas web publican el clima y su temperatura para las ciudades de todo el mundo. Centros meteorológicos registran las precipitaciones en diferentes sitios con diversas resoluciones y los operadores eléctricos de sistemas de transmisión monitorean constantemente la curva de demanda eléctrica de sus clientes. Los ejemplos anteriormente mencionados corresponden a un pequeño grupo del sinnúmero de procesos en los que las observaciones temporales son requeridas para el monitoreo, control y gestión de los recursos disponibles de una entidad determinada.

Una característica intrínseca de las series de tiempo es que las observaciones adyacentes pueden además ser *dependientes* entre sí. La naturaleza de dicho fenómeno es de considerable interés práctico. El *análisis de series de tiempo* se refiere a las técnicas para llevar a cabo el estudio de dicha dependencia, desarrollando modelos matemáticos que proveen descripciones plausibles para los datos de una muestra. Con el objetivo de describir el carácter de datos que pareciesen fluctuar de manera aleatoria en el tiempo, se asume que una serie de tiempo puede ser definida como una colección de variables aleatorias indexadas de acuerdo al orden en que son obtenidas en el tiempo, $x_1, x_2, x_3, \ldots$, donde la variable aleatoria x_1 denota el valor obtenido de la serie en el primer instante de tiempo, la variable x_2 denota el valor para el segundo período de tiempo y así sucesivamente. En general, una colección de variables aleatorias, $\{x_t\}$, indexada por t es usualmente referida como un *proceso estocástico*. Los valores observados de un proceso estocástico son entendidos como *realizaciones* de un proceso estocástico. Usualmente es posible inferir el uso del término *serie de tiempo* dependiendo del contexto en el que es usado, pudiendo referirse de manera genérica al proceso, o a una realización particular, sin necesidad de hacer una distinción en la notación de ambos conceptos [15].

Es convencional exponer gráficamente una serie de tiempo con los valores obtenidos por las variables aleatorias en el eje vertical, con la escala temporal en el eje horizontal, conectando los valores en períodos adyacentes mediante una línea con el objetivo de reconstruir visual e hipotéticamente, una serie de tiempo continua que pudo haber producido los valores como una muestra discreta de ella. La aproximación de las series mediante una *serie de parámetros discretos* muestreados mediante puntos equiespaciados en el tiempo es un reconocimiento de que los datos muestreados, en su mayor parte, serán discretos debido a las restricciones inherentes en el método de colección. Más aún, las técnicas de análisis pueden ser utilizadas de manera factible mediante el uso de computadores, cuyos cálculos digitales son limitados. Asimismo, los desarrollos teóricos se basan también en la idea que una serie de tiempo de parámetros continuos debiese ser especificada en términos de una *función de distribución* de dimensiones finitas definida sobre un número finito de puntos en el tiempo. Luego, una descripción completa de una serie de tiempo, observada como una colección de n variables aleatorias en puntos temporales arbitrarios $t_1, t_2, \ldots, t_n$, para cualquier entero positivo n, es obtenida de la función de distribución conjunta, evaluada como la probabilidad de que los valores de la serie sean, de manera conjunta, menores que las n constantes, $c_1, c_2, \ldots, c_n$; por ejemplo,

$$F_{t_1,t_2,\ldots,t_n}(c_1, c_2, \ldots, c_n) = Pr(x_{t_1} \leq c_1, x_{t_2} \leq c_2, \ldots, x_{t_n} \leq c_n). \tag{2.4}$$

Sin embargo, dichas funciones no pueden se descritas fácilmente a menos que las variables aleatorias se distribuyan normal de manera conjunta. Un proceso x_t se dice proceso *Gaussiano* si los vectores n-dimensionales $x = (x_{t_1}, x_{t_2}, \ldots, x_{t_n})'$, para cada colección de distintos puntos temporales $t_1, t_2, \ldots, t_n$ y cada entero positivo n, poseen una distribución normal multivariada. Definiendo el vector $n \times 1$ de medias $E(x) \equiv \mu = (\mu_{t_1}, \mu_{t_2}, \ldots, \mu_{t_n})'$ y la matriz $n \times n$ de covarianza como var$(x) \equiv \Gamma = \{\gamma(t_i, t_j); i, j = 1, \ldots, n\}$, que se asume es positiva definida [16], la función de densidad normal multivariada puede ser escrita como

$$f(x) = (2\pi)^{-n/2}|\Gamma|^{-1/2}\exp\Big\{-\frac{1}{2}(x-\mu)'\Gamma^{-1}(x-\mu)\Big\}, \tag{2.5}$$

para $x \in \mathbb{R}^n$, donde $|\cdot|$ denota el determinante.

Pese a que la función de distribución conjunta describe los datos completamente, es la herramienta inadecuada para mostrar y analizar series de tiempo. La función de distribución de la Ec. (2.4), debe ser evaluada en función de n argumentos, haciendo virtualmente imposible cualquier tipo de exposición visual. Las funciones de distribución marginal

$$F_t(x) = P\{x_t \leq x\}, \tag{2.6}$$

o las correspondientes funciones de densidad marginales

$$f_t(x) = \frac{\partial F_t(x)}{\partial x}, \tag{2.7}$$

cuando existen, son a menudo informativas para examinar el comportamiento marginal de una serie. Si x_t es Gaussiano con media μ_t y varianza σ_t^2, abreviado $x_t \sim N(\mu_t, \sigma_t^2)$, la densidad marginal está dada por

$$f_t(x) = \frac{1}{\sigma_t \sqrt{2\pi}} \exp\left\{ -\frac{1}{2\sigma_t^2}(x - \mu_t)^2 \right\}, x \in \mathbb{R}. \tag{2.8}$$

Asumiendo que existe, es posible definir la función *media* como

$$\mu_{xt} = E(x_t) = \int_{-\infty}^{\infty} x f_t(x) dx, \tag{2.9}$$

donde E denota el operador valor esperado o esperanza. Cuando no hay confusión acerca de la serie de tiempo a la cual se está refiriendo, es posible omitir el subíndice y escribir μ_{xt} como μ_t.

2.2.1. Autocovarianza

Luego de introducir en la sección (2.2) al operador esperanza, y considerando que la falta de independencia entre dos valores adyacentes x_s y x_t puede ser abordada numéricamente mediante el uso de las nociones de covarianza y correlación; asumiendo que la varianza de x_t es finita, es posible definir la función *autocovarianza* como el segundo

momento tal que

$$\gamma_x(s,t) = \text{cov}(x_s, x_t) = E[(x_s - \mu_s)(x_t - \mu_t)], \tag{2.10}$$

para todo s y t. Cuando no exista confusión posible acerca de cual serie se está refiriendo, es posible omitir el subíndice y escribir $\gamma_x(s,t)$ como $\gamma(s,t)$. Notar que $\gamma_x(s,t) = \gamma_x(t,s)$ para todos los períodos s y t.

La autocovarianza mide la dependencia *lineal* entre dos puntos en la misma serie observada en períodos distintos. Series muy suaves exhiben funciones de autocovarianza que se mantienen en niveles elevados incluso cuando s y t son distantes, mientras que series con cambios más abruptos tienden a tener niveles de autocovarianza cercanos a cero para observaciones distantes. Tal como en la estadística clásica, si $\gamma_x(s,t) = 0$, x_s y x_t no están relacionadas de manera lineal, aunque aún podría existir algún tipo de estructura de dependencia entre ellas. Si, sin embargo, x_s y x_t son normales bivariadas, $\gamma_x(s,t) = 0$ asegura su independencia. Es claro que, para $s = t$, la autocovarianza se reduce a la (asumida finita) *varianza*, ya que

$$\gamma_x(t,t) = E[(x_t - \mu_t)^2] = \text{var}(x_t). \tag{2.11}$$

Es usual lidiar de una manera más conveniente con medidas de asociación entre -1 y 1, las que pueden ser obtenidas usando la *función de autocorrelación* (ACF) definida como

$$\rho(s,t) = \frac{\gamma(s,t)}{\sqrt{\gamma(s,s)\gamma(t,t)}}. \tag{2.12}$$

La ACF mide la predictibilidad lineal de la serie en el período t, x_t, usando sólo el valor x_s. La desigualdad de Cauchy-Schwarz permite mostrar que $-1 \leq \rho(s,t) \leq 1$ pues $|\gamma(s,t)|^2 \leq \gamma(s,s)\gamma(t,t)$. Así, es posible obtener una métrica aproximada de la habilidad de pronóstico de la serie de tiempo en el período t a partir de su valor en el período s.

A menudo es de interés medir la capacidad de predicción de otra serie y_t con respecto a x_s. Asumiendo que ambas series poseen varianza finita, es posible introducir la *función*

de covarianza cruzada entre dos series, x_t e y_t como

$$\gamma_{xy}(s,t) = \text{cov}(x_s, y_t) = E[(x_s - \mu_{xs})(y_t - \mu_{yt})], \tag{2.13}$$

cuya versión escalada es conocida como función de *correlación cruzada* dada por

$$\rho_{xy}(s,t) = \frac{\gamma_{xy}(s,t)}{\sqrt{\gamma_x(s,s)\gamma_y(t,t)}}. \tag{2.14}$$

Finalmente, es posible extender las ideas anteriormente descritas para el caso de una *serie de tiempo multivariada*, $x_{t1}, x_{t2}, \ldots, x_{tr}$, con r componentes, por ejemplo

$$\gamma_{jk}(s,t) = E[(x_{sj} - \mu_{sj})(x_{tk} - \mu_{tk})], \qquad j,k = 1,2,\ldots,r. \tag{2.15}$$

Para las definiciones previas, las funciones de autocovarianza y covarianza cruzada podrían variar a través de la serie, pues sus valores sólo dependen de s y t, la ubicación de los puntos en el tiempo. En particular, la función de autocovarianza depende de la separación de x_s y x_t, $h = |s - t|$, y no del lugar donde los puntos están ubicados en el tiempo. Mientras los puntos estén separados h unidades, la ubicación de los dos puntos en el tiempo pierde importancia.

2.2.2. Estacionariedad

Sin realizar ningún supuesto previo acerca del comportamiento de una serie de tiempo, las definiciones expuestas en la sección (2.2.1) conducen al análisis de ciertas regularidades que podrían existir en el comportamiento de una serie en el tiempo. Es posible introducir la noción de regularidad usando el concepto de *estacionariedad*. Una serie de tiempo *estrictamente estacionaria* es aquella serie cuyo comportamiento probabilístico para cada colección de valores

$$\{x_{t_1}, x_{t_2}, \ldots, x_{t_k}\},$$

es idéntica al set trasladado en el tiempo

$$\{x_{t_1+h}, x_{t_2+h}, \ldots, x_{t_k+h}\}.$$

Esto es,

$$Pr\{x_{t_1} \leq c_1, \ldots, x_{t_k} \leq c_k\} = Pr\{x_{t_1+h} \leq c_1, \ldots, x_{t_k+h} \leq c_k\}, \tag{2.16}$$

para todo $k = 1, 2, \ldots$, todos los períodos $t_1, t_2, \ldots, t_k$, para todos los números $c_1, c_2, \ldots, c_k$, y todos los desplazamientos temporales $h = 0, \pm 1, \pm 2, \ldots$.

Si una serie es estrictamente estacionaria, todas las funciones de distribución multivariadas para los subconjuntos de variables deben concordar con sus contrapartes en el conjunto desplazado para todos los valores de traslación, separación o desplazamiento h. Para $k = 1$, si la media μ_t, de la serie existe, entonces $\mu_s = \mu_t$ para todo s y t, y por ende, u_t debe ser constante. Más aún, cuando $k = 2$, es posible escribir la Ec. (2.16)

$$Pr\{x_s \leq c_1, x_t \leq c_2\} = Pr\{x_{s+h} \leq c_1, x_{t+h} \leq c_2\}, \tag{2.17}$$

para cualquier período s y t separados por h. Entonces, si la varianza del proceso existe, la autocovarianza de x_t satisface

$$\gamma(s, t) = \gamma(s + h, t + h),$$

para todo s, t y h. Cuya interpretación nuevamente resulta en la dependencia con respecto a la diferencia entre s y t, no de los períodos actuales.

La definición anterior de estacionariedad es demasiado fuerte para la mayoría de las aplicaciones. Más aún, es difícil establecer estacionariedad a partir de un solo set de datos. En la práctica, en lugar de imponer condiciones a todas las distribuciones posibles de una serie de tiempo, una relajación de dicha condición basada en los primeros dos momentos de la serie es utilizada. Una serie de tiempo x_t *débilmente estacionaria* es un proceso de varianza finita cuya media, μ_t, definida en la Ec. (2.9) es constante y no depende del período t. De igual manera, la autocovarianza, $\gamma(s, t)$, definida en la Ec. (2.10) depende de s y t sólo a través de su diferencia $|s - t|$. Luego, el término *estacionario* se usará para referirse a

una serie débilmente estacionaria; si un proceso es estacionario en el sentido estricto de su definición, el término *estrictamente estacionario* será utilizado.

La estacionariedad requiere que las funciones de media y autocorrelación sean regulares en el sentido de que puedan ser estimadas (al menos) mediante su promedio. Claramente, una serie de tiempo estrictamente estacionaria con varianza finita es también estacionaria. Sin embargo, lo contrario no es cierto a menos que se exijan ciertas condiciones. En particular, una serie estacionaria será estrictamente estacionaria si la serie de tiempo es Gaussiana, donde todas las distribuciones finitas de la serie son Gaussianas.

Dado que la media, $E(x_t) = \mu_t$, de una serie estacionaria es independiente del período t, es posible escribir

$$\mu_t = \mu. \tag{2.18}$$

Dado que la autocovarianza, $\gamma(s,t)$, de una serie estacionaria x_t, depende de s y t sólo a través de su diferencia $|s-t|$, es posible simplificar su notación si $s = t+h$, donde h representa el desplazamiento temporal, rezago o *lag*. Entonces

$$\gamma(t+h,t) = \text{cov}(x_{t+h}, x_t) = \text{cov}(x_h, x_0) = \gamma(h,0), \tag{2.19}$$

pues la diferencia temporal entre los períodos $t+h$ y t es la misma diferencia entre h y 0. Luego, la autocovarianza de una serie de tiempo estacionaria no depende del argumento temporal t. Por conveniencia, el segundo argumento de $\gamma(h,0)$ es omitido. Luego la *función de autocovarianza de una serie de tiempo estacionaria* se escribe como

$$\gamma(h) = \text{cov}(x_{t+h}, x_t) = E[(x_{t+h}-\mu)(x_t-\mu)]. \tag{2.20}$$

La autocovarianza de un proceso estacionario posee varias propiedades especiales. En particular, $\gamma(h)$ es no-negativa definida [16], asegurando que las varianzas de las combinaciones lineales de las variables x_t nunca sean negativas. Esto es, para cualquier $n \geq 1$ y para las constantes $a_1, \ldots, a_n$,

$$0 \leq \text{var}(a_1x_1 + \cdots + a_nx_n) = \sum_{j=1}^{n}\sum_{k=1}^{n} a_j a_k \gamma(j-k) \tag{2.21}$$

También, para $h = 0$

$$\gamma(0) = E[(x_t - \mu)^2], \tag{2.22}$$

que corresponde a la varianza de una serie de tiempo. La desigualdad de Cauchy-Schwarz implica que

$$|\gamma(h)| \leq \gamma(0). \tag{2.23}$$

Finalmente, la autocovarianza de una serie estacionaria es simétrica con respecto al origen, esto es,

$$\gamma(h) = \gamma(-h), \tag{2.24}$$

para todo h, ya que

$$\gamma((t+h) - t) = \text{cov}(x_{t+h}, x_t) = \text{cov}(x_t, x_{t+h}) = \gamma(t - (t+h)). \tag{2.25}$$

Considerando la Ec.(2.12), es posible expresar la *función de autocorrelación de una serie de tiempo estacionaria* (ACF) como

$$\rho(h) = \frac{\gamma(t+h, t)}{\sqrt{\gamma(t+h, t+h)\gamma(t, t)}} = \frac{\gamma(h)}{\gamma(0)}. \tag{2.26}$$

La desigualdad de Cauchy-Schwarz permite mostrar que para la Ec. (2.26), se cumple que $-1 \leq \rho(h) \leq 1$, para todo h.

Finalmente, considérese dos series de tiempo x_t e y_t. Dichas series se dicen *conjuntamente estacionarias* si cada una es estacionaria y la función de covarianza cruzada

$$\gamma_{xy}(h) = \text{cov}(x_{t+h}, y_t) = E[(x_{t+h} - \mu_x)(y_t - \mu_y)], \tag{2.27}$$

es función solamente del lag h. Luego, la *función de correlación cruzada de series conjuntamente estacionarias*(CCF) x_t e y_t se define como

$$\rho_{xy}(h) = \frac{\gamma_{xy}(h)}{\sqrt{\gamma_x(0)\gamma_y(0)}}. \tag{2.28}$$

Nuevamente, $-1 \leq \rho_{xy}(h) \leq 1$ permite la comparación entre x_{t+h} e y_t. La función de

correlación cruzada generalmente no es simétrica con respecto a cero, es decir, $\rho_{xy}(h) \neq \rho_{yx}(-h)$, lo que es demostrable usando los mismos principios que para obtener la Ec.(2.12).

2.2.3. Ruido blanco

Uno de los usos más importantes de la función de autocorrelación es la definición de un *proceso aleatorio ideal*, que es la base de la teoría de procesos aleatorios (lineales). El más simple de los procesos de series de tiempo corresponde a una colección de variables aleatorias w_t, con media 0 y varianza finita σ_w^2. La serie de tiempo generada a partir de variables no correlacionadas es usada como modelo para el ruido en aplicaciones ingenieriles, por ello también es llamada *ruido blanco*, denotando al proceso mediante $w_t \sim \text{wn}(0, \sigma_w^2)$. La designación *blanco* se origina como analogía con la luz blanca e indica que todas las oscilaciones periódicas posibles están presentes con igual fuerza.

Usualmente es necesario que el ruido corresponda a variables aleatorias independientes e idénticamente distribuidas (iid) con media 0 y varianza σ_w^2, aquello se distingue mediante la notación $w_t \sim \text{iid}(0, \sigma_w^2)$, así como también refiriéndose como *ruido blanco independiente* o *ruido iid* [15]. Una serie de ruido blanco particularmente útil es el *ruido blanco Gaussiano*, donde w_t corresponde a variables aleatorias normales independientes, con media 0 y varianza σ_w^2, cuya notación es $w_t \sim \text{iid N}(0, \sigma_w^2)$. La Fig. (2.1) muestra tales variables aleatorias con $\sigma_w^2 = 1$ en el orden en que fueron obtenidas Se espera que una serie de ruido

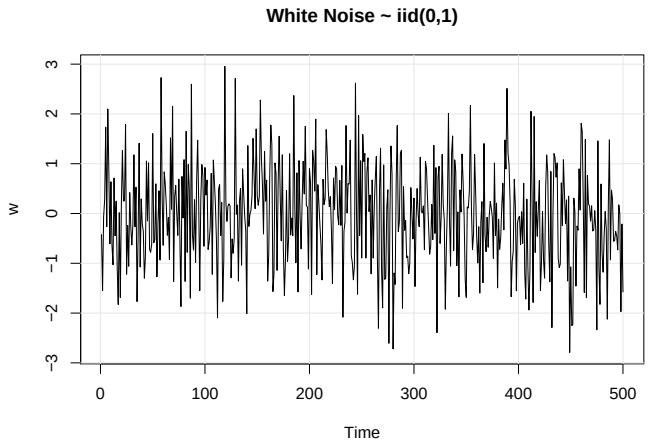


Figura 2.1: Serie ruido blanco Gaussiana

blanco tenga una autocorrelación cercana a cero. Más aún, el 95 % de los valores en la función de autocorrelación estará entre $\pm 2/\sqrt{T}$, donde T es el largo de la serie de tiempo.

2.3. Herramientas de pronóstico

2.3.1. Modelos de referencia

Antes de estudiar el rendimiento de los modelos de pronóstico, es necesario demostrar que dichos modelos son mejores que los modelos ya existentes. Cada metodología debe ser estudiada respecto de un punto de comparación o *benchmark*. En el caso de métodos univariados de series de tiempo, las comparaciones debiesen llevarse a cabo al menos considerando metodologías ingenuas de pronóstico y también metodologías estándar tales como modelos ARIMA (véase sección 2.4.2).

Para datos no estacionales, el método de pronóstico ingenuo se basa en el modelo de caminata aleatoria (véase sección 2.2.3), estableciendo que todos los pronósticos son iguales a la última observación disponible. Para datos estacinales, el mejor método ingenuo corresponde a usar la última observación de la misma estación. En el caso de modelos de serie de tiempo multivariados, los mismos modelos univariados pueden ser utilizados como puntos de comparación.

Para métodos que involucran variables independientes, una regresión lineal estándar a menudo provee de un punto de referencia básico válido. Incluso si la regresión lineal no es apropiada para los datos (ej., por relaciones no lineales o variables correlacionadas), es un método estándar y simple a vencer con modelos más complejos.

2.3.2. Transformaciones matemáticas

Si los datos muestran variaciones que incrementan o disminuyen con el nivel de la serie, entonces una transformación puede ser útil. Denotando $y_1, \ldots, y_T$ las observaciones originales y $w_1, \ldots, w_T$, entonces $w_t = log(y_t)$. Las transformaciones logarítmicas son útiles debido a su interpretabilidad: cambios en el valor logarítmico corresponden a cambios relativos (o porcentuales) en la escala original. Más aún, la transformación logarítmica restringe a los pronósticos a permanecer positivos en la escala original.

En otras ocasiones, otras transformaciones como raíces cuadradas o cúbicas son utilizadas (aunque no sean tan interpretables). Éstas son llamadas *transformaciones de potencias*,

ya que pueden ser escritas como $w_t = y_t^p$.

Una familia de transformaciones particularmente útil, que incluye logaritmos y potencias, es la familia de las *transformaciones Box-Cox* que dependen del parámetro λ y se definen como

$$w_t = \begin{cases} \log(y_t) & \text{si } \lambda = 0; \\ (y_t^\lambda - 1)/\lambda & \text{si no.} \end{cases} \tag{2.29}$$

El logaritmo en la transformación Box-Cox es siempre un logaritmo natural (base e). Así, si $\lambda = 0$, entonces el logaritmo natural es utilizado. En el caso $\lambda \neq 0$, una transformación de potencias es utilizada, seguida de un escalamiento simple. De manera particular, si $\lambda = 1$, entonces $w_t = y_t - 1$, y los datos son trasladados hacia abajo, sin un cambio en la forma de la serie de tiempo.

Una vez escogido el valor de λ, es necesario pronosticar los datos transformados para posteriormente revertir la transformación para obtener los pronósticos en la escala original, cuya reversión está dada por

$$y_t = \begin{cases} \exp(w_t) & \text{si } \lambda = 0; \\ (\lambda w_t + 1)^{(1/\lambda)} & \text{si no.} \end{cases} \tag{2.30}$$

Algunas aspectos a considerar con respecto a las transformaciones Box-Cox son:

- Si algún $y_t \leq 0$, ninguna transformación de potencias es posible a menos que todas las observaciones sean ajustadas sumando una constante a todos los valores
- Los resultados de pronóstico son insensibles a λ, sin embargo la transformación tendrá efectos mayores en cuanto a intervalos de predicción
- El pronóstico obtenido al revertir la transformación corresponde a la mediana de la distribución de pronóstico, en lugar de la media de dicha distribución. Para muchos propósitos este hecho es aceptable, sin embargo, ocasionalmente, la media del pronóstico es requerida cuando un pronóstico agregado es requerido y la suma de medianas no puede ser considerada.

En particular, para una transformación Box-Cox, la media de la reversión de la transformación está dada por

$$y_t = \begin{cases} \exp(w_t)\left[1 + \frac{\sigma_h^2}{2}\right] & \text{si } \lambda = 0; \\ (\lambda w_t + 1)^{1/\lambda}\left[1 + \frac{\sigma_h^2(1-\lambda)}{2(\lambda w_t+1)^2}\right] & \text{si no,} \end{cases} \tag{2.31}$$

donde σ_h^2 es la varianza del pronóstico a h pasos. Mientras más grande sea la varianza del pronóstico, mayor será la diferencia entre la media y la mediana.

La diferencia entre la reversión simple del pronóstico dada por la Ec. (2.30) y la media dada por la Ec. (2.31) es llamada *sesgo*. Cuando se utiliza la media en lugar de la mediana, se dice que los pronósticos puntuales han sido *ajustados al sesgo*.

De esta manera, la etapa de *preprocesamiento* de los datos es revertida a su escala original. Dichos procedimientos son llevados a cabo en el contexto de los modelos BATS/TBATS, idea que será profundizada en la sección (2.4.3.8). Asimismo, en el contexto de los modelos ARIMA (véase sección 2.4.2) se considera un preprocesamiento alternativo que también puede ser utilizado para modelos de inteligencia artificial.

2.3.3. Diagnóstico residual

Cada observación en una serie de tiempo puede ser pronosticada usando todas las observaciones previas. Dichos valores son llamados *valores ajustados* y pueden ser denotados por $\hat{y}_{t|t-1}$, el pronóstico de y_t basado en las observaciones $y_1, \ldots, y_{t-1}$. En su uso habitual, el subíndice es omitido, utilizando la notación $\hat{y}_t$ en lugar de $\hat{y}_{t|t-1}$. Los valores ajustados siempre involucran pronósticos a un paso.

Los valores ajustados no corresponden a verdaderos pronósticos, pues cualquier parámetro involucrado en el método de pronóstico es estimado utilizando todas las observaciones disponibles en la serie de tiempo, incluyendo observaciones futuras.

Los *residuales* en un modelo de series de tiempo corresponden a lo que queda luego de ajustar un modelo. En general, los residuales equivalen a la diferencia entre las

observaciones y sus correspondientes valores ajustados

$$e_t = y_t - \hat{y}_t. \tag{2.32}$$

Los residuales son útiles para diagnosticar si un modelo ha capturado la información de los datos de manera adecuada. Un buen método de pronóstico proveerá de residuales con las siguientes propiedades:

1. Los residuales no están correlacionados. De existir correlación en los residuales, entonces aún queda información en ellos que puede ser utilizada en la generación de pronósticos.

2. Los residuales poseen media cero, sino, los pronósticos serán sesgados.

Cualquier metodología que pronóstico que no satisfaga dichas propiedades puede ser mejorada. Sin embargo, aquello no significa que los métodos de pronósticos que sí las satisfacen no puedan ser mejorados. Es posible tener diferentes metodologías de pronóstico para el mismo conjunto de datos, todas satisfaciendo dichas propiedades. La revisión de dichas propiedades es importante para verificar si un método en particular está usando toda la información disponible, pero no es una buena manera de seleccionar un método de pronóstico.

Si cualquiera de estas propiedades no es satisfecha, el método de pronóstico puede ser modificado para obtener mejores pronósticos. El sesgo en el pronóstico puede ser corregido sumando una constante a todos los pronósticos. El caso de la correlación es más complejo, y será considerado y abordado en modelos de secciones posteriores.

Además de las propiedades anteriormente descritas, es útil (aunque no necesario) para los residuales tener las siguientes propiedades

3. Los residuales poseen varianza constante.

4. Los residuales poseen distribución normal.

Dichas propiedades hacen del cálculo de intervalos de predicción una tarea más fácil. Sin embargo, un método de pronóstico que no satisfaga estas propiedades no necesariamente requiere ser mejorado.

2.3.3.1. Test de Portmanteau para autocorrelación

Además de la consideración de la función de autocorrelación revisada en la sección (2.2.1) y de los gráficos que pueden ser obtenidos a partir de ella, existen pruebas formales para evaluar la presencia de autocorrelación. El *Test de Portmanteau* es un tipo de prueba de hipótesis estadística en la cual una hipótesis nula está bien especificada, y la hipótesis alternativa se especifica de manera flexible. Las pruebas construidas en este contexto pueden tener la propiedad de ser al menos moderadamente potentes contra una amplia gama de desviaciones de la hipótesis nula. Así, en la estadística aplicada, el Test de Portmanteau proporciona un modo razonable de proceder como un control general de un modelo de partida para un conjunto de datos donde hay muchas maneras diferentes en las que el modelo podrá apartarse del proceso subyacente generador de los datos. El uso de este tipo de pruebas evita tener que ser muy específico sobre el tipo de alternativas que se está probando.

Considerando ρ_k, la autocorrelación para el rezago k, es posible probar si las primeras h autocorrelaciones son significativamente diferentes de lo que se espera de un proceso ruido blanco. Una de las pruebas anteriormente descritas corresponde al *Test Box-Pierce*, basado en el siguiente estadístico

$$Q = T \sum_{k=1}^{h} \rho_k^2, \tag{2.33}$$

donde h es el rezago máximo a considerar y T es el número de observaciones. Si cada ρ_k es cercano a cero, entonces Q será pequeño. Si algunos valores ρ_k son grandes (positivos o negativos), entonces Q será grande. Lamentablemente, el Test Box-Pierce no es útil cuando h es grande.

Un test relacionado (y más preciso) es el *Test de Ljung-Box*, basado en

$$Q^* = T(T+2) \sum_{k=1}^{h} (T-k)^{-1} \rho_k^2. \tag{2.34}$$

Nuevamente, valores grandes de Q^* sugieren que las autocorrelaciones no provienen de una serie ruido blanco.

Si las autocorrelaciones provienen de una serie ruido blanco, entonces Q y Q^* tendrán una distribución χ^2 con $(h - K)$ grados de libertad, donde K es el número de parámetros del modelo.

2.3.3.2. Test Jarque-Bera para normalidad

El *Test Jarque-Bera* [17] es una prueba de bondad de ajuste que busca entregar evidencia acerca de la correspondencia de la simetría y la curtosis con respecto a una distribución normal. Asumiendo datos iid (véase sección 2.2.3), el estadístico está definido como

$$JB = \frac{n}{6}\left(S^2 + \frac{(C-3)^2}{4}\right), \tag{2.35}$$

donde S y C corresponden a la simetría y curtosis muestral, respectivamente. Acá,

$$S = \frac{\hat{\mu}_3}{\hat{\sigma}^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{3/2}}, \tag{2.36}$$

$$C = \frac{\hat{\mu}_4}{\hat{\sigma}^4} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{2}}, \tag{2.37}$$

donde $\hat{\mu}_3$ y $\hat{\mu}_4$ corresponden a los estimadores del tercer y cuarto momento, $\bar{x}$ es la media y $\hat{\sigma}^2$ el estimador de la varianza.

Para S, se considera que:

- Si $S = 0$ la distribución es simétrica
- Si $S < 0$ la distribución es sesgada a la izquierda, asimétrica negativa, o de colas pesadas a la izquierda
- Si $S > 0$ la distribución es sesgada a la derecha, asimétrica positiva o de colas pesadas a la derecha

mientras que para C:

- Si $C = 3$, la estructura de variabilidad es igual a la Gaussiana (Mesocúrtica)

- Si $C < 3$, la estructura de variabilidad es mayor a la ideal (Platicúrtica)
- Si $C > 3$, la estructura de variabilidad es menor a la ideal (Leptocúrtica)

Bajo la hipótesis nula de normalidad, el estadístico Jarque-Bera está aproximadamente distribuido χ^2 con dos grados de libertad. El test rechaza el supuesto de normalidad si el estadístico es lo suficientemente grande.

2.3.4. Evaluación de precisión del pronóstico

2.3.4.1. Sets de entrenamiento y evaluación

Es importante evaluar la precisión de los pronósticos usando pronósticos genuinos. En consecuencia, el tamaño de los residuales no es un indicador confiable de cuan grande podrán ser los errores de pronóstico. La precisión de los pronósticos sólo puede ser determinada considerando el nivel de desempeño de un modelo en datos nuevos que no han sido utilizados para ajustar el modelo.

Al momento de escoger los modelos, una práctica común es separar los datos disponibles en dos porciones, los datos de entrenamiento (*train set*) y de evaluación (*test set*). Los primeros son utilizados para determinar cualquier parámetro que sea requerido por la metodología de pronóstico, mientras que los últimos son utilizados para evaluar su rendimiento. Debido a que los datos de evaluación no son utilizados para determinar el pronóstico, éstos debiesen proveer de un indicador confiable para medir qué tan bien el modelo es capaz de pronosticar nuevos datos.

Pese a que no existe una regla general para la determinación de los porcentajes de datos por conjunto, en general el tamaño del set de evaluación es de un 20 % del total de la muestra. Sin embargo, los porcentajes dependen del tamaño de la muestra y de qué tan a futuro se desea pronosticar. Idealmente, el set de evaluación debería ser al menos tan grande como el horizonte de pronóstico máximo requerido. Algunos puntos importantes a considerar son

- Un modelo que se ajusta bien a los datos de entrenamiento no necesariamente pronosticará bien.

- Siempre puede obtenerse un ajuste perfecto con un modelo con suficientes parámetros.
- Sobreajustar un modelo a los datos (*overfitting*) es tan malo como fallar en identificar los patrones sistemáticos en los datos.

2.3.4.2. Métricas de error

El *error de pronóstico* corresponde a la diferencia entre el valor observado y su pronóstico. Aquí, *error* no significa una desviación o equivocación, sino la parte impredecible de una observación [18]. Es posible expresar el error según

$$e_{T+h} = y_{T+h} - \hat{y}_{T+h|T} \tag{2.38}$$

donde los datos de entrenamiento están dados por $\{y_1, \ldots, y_T\}$ y los datos de testeo por $\{y_{T+1}, y_{T+2}, \ldots\}$.

Notar que el error de pronóstico es diferente a los residuales en dos maneras

1. Los residuales son calculados con respecto a los datos de entrenamiento mientras que los errores de pronósticos son obtenidos a partir del set de evaluación.
2. Los residuales se basan en el pronóstico a un paso (*one-step forecast*) , mientras que los errores de pronósticos involucran pronósticos a múltiples pasos (*multi-step forecasts*).

Es posible medir la precisión del pronóstico resumiéndolo con los siguientes indicadores

Error escala-dependiente

El error de pronóstico está en la misma escala que los datos. Las medidas de precisión que se basan sólo en e_t son dependientes de la escala y por ello no pueden ser utilizados para hacer comparaciones entre series que involucran diferentes unidades.

Las dos métricas escala-dependientes más utilizadas son el *error absoluto medio* (MAE) y la *raíz del error cuadrático medio* (RMSE)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_t - \hat{y}_t| \tag{2.39}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_t - \hat{y}_t)^2} \tag{2.40}$$

Al momento de comparar métodos de pronóstico a una serie de tiempo o varias series con las mismas unidades, MAE se presenta como una alternativa fácil de calcular y entender. Un método que minimice MAE conducirá a pronósticos de la mediana, mientras que minimizar RMSE llevará a pronósticos de la media. Por ello, RMSE es ampliamente utilizado, pese a que sea más difícil de interpretar.

Error porcentual

Los errores porcentuales tienen la ventaja de ser libre de unidades, y por ello son frecuentemente utilizados para evaluar el desempeño de pronósticos entre set de datos. La medida más utilizada es el *porcentaje de error absoluto medio* (MAPE) definido como

$$MAPE = 100 \times \frac{1}{n}\sum_{t=1}^{n}\left|\frac{y_t - \hat{y}_t}{y_t}\right| \tag{2.41}$$

Las métricas basadas en error porcentual tienen la desventaja de ser infinitas o indefinidas si $y_t = 0$ para cualquier t en el período de interés, y poseer valores extremos si cualquier y_t es cercano a cero. Además, en el caso de una serie que posea un cero arbitrario como realización válida de la observación, el indicador provee de mayor significancia. Finalmente, el índice penaliza de mayor manera errores negativos que positivos, lo que ha conducido al uso de métricas simétricas del error, como por ejemplo *MAPE simétrico* (sMAPE) .

2.3.4.3. Validación cruzada para series de tiempo

Una versión más sofisticada de los sets de entrenamiento/evaluación descritos en la sección (2.3.4.1) es la *validación cruzada para series de tiempo* o *evaluación de pronóstico en ventana rodante*. En este procedimiento, se considera una serie de sets de evaluación consistentes en una sola observación cada uno. El set de entrenamiento correspondiente consiste solo en las observaciones que ocurrieron *previo* a la observación que forma el set de evaluación. Así, ninguna observación futura es usada en construir el pronóstico. En general tres parámetros son utilizados para este tipo de partición de los datos

Ventana inicial Número inicial de valores consecutivos en cada muestra del set de entrenamiento.

Horizonte Número de valores consecutivos a utilizar en el set de evaluación.

Ventana fija Corresponde a la consideración lógica con respecto al tamaño de la ventana rodante en el tiempo. Si la ventana no es fija, el set de entrenamiento siempre comenzará con la primera obsercación y su tamaño variará a través de las particiones.

La Fig. (2.2) presenta el esquema de ventana rodante para una serie de 20 puntos. Las filas de cada panel corresponden a diferentes particiones de datos y las columnas a diferentes puntos de datos. El color rojo indica las observaciones que están siendo incluidas en el set de entrenamiento y el color azul las observaciones del set de evaluación.

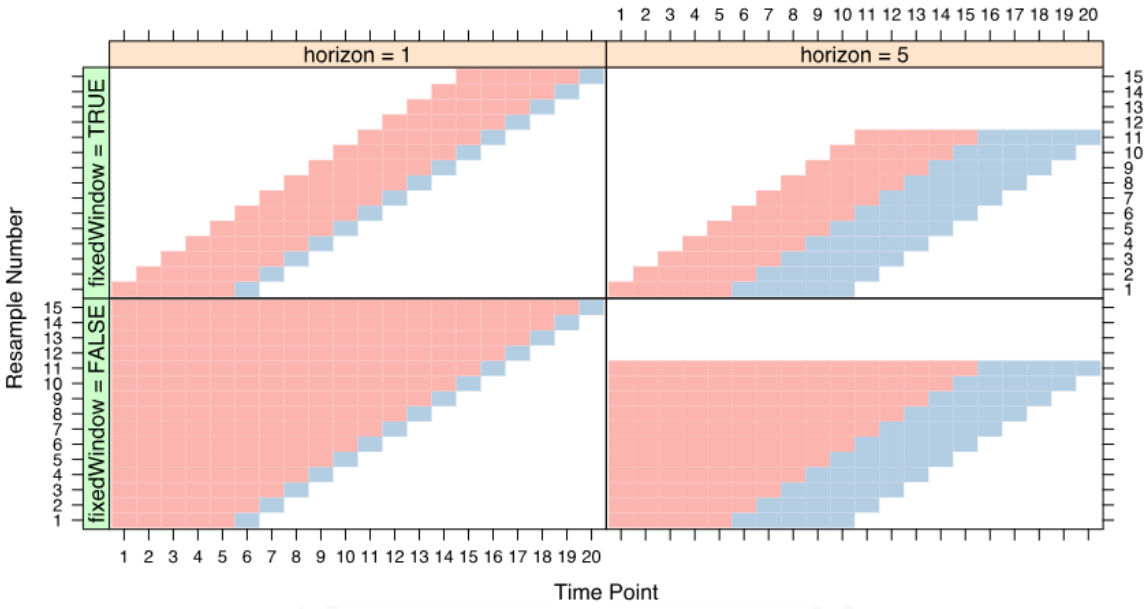


Figura 2.2: Esquema de pronóstico con horizonte rodante.
Fuente: [19]

2.4. Metodologías estadísticas de pronóstico

2.4.1. Regresión clásica en el contexto de series de tiempo

Suponiendo que la serie de tiempo, y_t, para $t = 1, \ldots, n$, está siendo influenciada por una colección de posibles series independientes, $x_{t1}, x_{t2}, \ldots, x_{tq}$, las cuales se asumen como entradas fijas y conocidas; es posible expresar esta relación a través de un *modelo de regresión lineal* como

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \cdots + \beta_q x_{tq} + \varepsilon_t, \tag{2.42}$$

donde $\beta_0, \beta_1, \ldots, \beta_q$ son coeficientes de regresión desconocidos y ε_t es un proceso aleatorio o ruido (iid) con media cero y varianza σ_ε^2. Para regresión de series de tiempo, es poco usual que el ruido sea blanco, y este supuesto usualmente es relajado.

El modelo de regresión lineal múltiple puede ser escrito en una notación más general definiendo los vectores columna $x_t = (1, x_{t1}, x_{t2}, \ldots, x_{tq})'$ y $\beta = (\beta_0, \beta_1, \ldots, \beta_q)'$. Luego la

Ec. (2.42) equivale a

$$y_t = \beta' x_t + \varepsilon_t, \tag{2.43}$$

donde $\varepsilon_t \sim iidN(0, \sigma_\varepsilon^2)$. La estimación por *mínimos cuadrados ordinarios* (OLS) permite halla el vector β que minimiza la suma de los cuadrados del error

$$Q = \sum_{t=1}^{n} \varepsilon_t^2 = \sum_{t=1}^{n} (y_t - \beta' x_t)^2, \tag{2.44}$$

con respecto a $\beta_0, \beta_1, \ldots, \beta_q$. Esta minimización puede ser llevada a cabo mediante la diferenciación de la Ec. (2.44) con respecto al vector β o mediante el uso de las propiedades de las proyecciones (usando espacios de Hilbert y el Teorema de la Proyección). En cualquier caso, la solución debe satisfacer $\sum_{t=1}^{n} (y_t - \hat{\beta}' x_t) x_t' = 0$. Este procedimiento permite obtener las *ecuaciones normales*

$$\left(\sum_{t=1}^{n} x_t x_t' \right) \hat{\beta} = \sum_{t=1}^{n} x_t y_t. \tag{2.45}$$

Si $\sum_{t=1}^{n} x_t x_t'$ es no singular, el estimador por mínimos cuadrados de β es

$$\hat{\beta} = \left(\sum_{t=1}^{n} x_t x_t' \right)^{-1} \sum_{t=1}^{n} x_t y_t. \tag{2.46}$$

La *suma de los cuadrados del error* (SSE) de la Ec. (2.44), puede ser escrita como

$$SSE = \sum_{t=1}^{n} (y_t - \hat{\beta}' x_t)^2. \tag{2.47}$$

Los estimadores OLS son insesgados ($E(\hat{\beta}) = \beta$), y tienen la menor varianza dentro de la clase de los estimadores lineales insesgados.

Si los errores ε_t se distribuyen normal, $\hat{\beta}$ es además el *estimador de máxima verosimilitud* para β y se distribuye normal con

$$\text{cov}(\hat{\beta}) = \sigma_\varepsilon^2 C, \tag{2.48}$$

donde

$$C = \left(\sum_{t=1}^{n} x_t x_t' \right)^{-1}, \tag{2.49}$$

es una notación conveniente. El estimador insesgado para la varianza σ_ε^2 es

$$s_\varepsilon^2 = MSE = \frac{SSE}{n-(q+1)}. \tag{2.50}$$

Bajo el supuesto de normalidad,

$$t = \frac{(\hat{\beta}_t - \beta_t)}{s_\varepsilon \sqrt{c_{ii}}}, \tag{2.51}$$

tiene una distribución t con $n-(q+1)$ grados de libertad; c_{ii} denota el i-ésimo elemento diagonal de C. Este resultado a menudo es utilizado para pruebas individuales de la hipótesis nula $H_0 : \beta_i = 0$ para $i = 1, \ldots, q$.

La comparación entre varios modelos a menudo es de interés para aislar o seleccionar el mejor subconjunto de variables independientes. Suponiendo que un modelo propuesto especifica que solo el subconjunto $r < q$ de variables independientes, por ejemplo, $x_{t,1:r} = x_{t1}, x_{t2}, \ldots, x_{tr}$ está influenciando la variable dependiente y_t. El modelo reducido es

$$y_t = \beta_0 + \beta_1 x_{t1} + \cdots + \beta_q x_{tr} + \varepsilon_t, \tag{2.52}$$

donde $\beta_1, \beta_2, \ldots, \beta_r$ es un subconjunto de coeficientes de las q variables originales.

La hipótesis nula en este caso es $H_0 : \beta_{r+1} = \cdots = \beta_q = 0$. Es posible evaluar el modelo reducido de la Ec. (2.52) con respecto a el modelo completo de la Ec. (2.43) mediante la comparación de la suma del error al cuadrado bajo los dos modelos usando el estadístico F definido como

$$F = \frac{(SSE_r - SSE)/(q-r)}{SSE/(n-q-1)} = \frac{MSR}{,} \tag{2.53}$$

donde SSE_r es la suma del error al cuadrado bajo el modelo reducido. Notar que $\text{SSE}_r \geq$ SSE ya que el modelo completo posee más parámetros. Si $H_0 : \beta_{r+1} = \cdots = \beta_q = 0$ es verdadera, entonces $\text{SSE}_r \approx$ SSE ya que los estimadores de aquellos βs serán cercanos a 0. Por lo tanto, no se cree H_0 si $\text{SSR} = \text{SSE}_r - \text{SSE}$ es grande. Bajo la hipótesis nula, el estadístico tiene una distribución F centrada con $q-r$ y $n-q-1$ grados de libertad cuando

el modelo de la Ec. (2.52) es el correcto.

Los resultados anteriores son a menudo resumidos en la Tabla (2.1) de *Análisis de Varianza* (ANOVA). La diferencia en el numerador es llamada *suma de cuadrados de la regresión* (SSR). La hipótesis nula es rechazada al nivel *alpha* si $F > F^{q-r}_{n-q-1}(\alpha)$, el percentil $1-\alpha$ de la distribución F con $q-r$ en el numerador y $n-q-1$ grados de libertad.

Tabla 2.1: Análisis de varianza para regresión

Origen	Grados de libertad	Suma de cuadrados	Cuadrado medio	F
$x_{t,r+1:q}$	$q-r$	$SSR = SSE_r - SSE$	$MSR = SSR/(q-r)$	F
Error	$n-(q+1)$	SSE	$MSE = SSE/(n-q-1)$	

Un caso de especial interés es la hipótesis nula $H_0 : \beta_1 = \cdots = \beta_q = 0$. En este caso $r = 0$, y el modelo de la Ec. (2.52) se reduce

$$y_t = \beta_0 + \varepsilon_t. \tag{2.54}$$

Luego es posible medir la proporción de variación explicada por todas las variables usando

$$R^2 = \frac{SSE_0 - SSE}{SSE_0}, \tag{2.55}$$

donde la suma de los cuadrados de los residuales bajo el modelo reducido es

$$SSE_0 = \sum_{t=1}^{n} (y_t - \bar{y})^2 \, . \tag{2.56}$$

En este caso SSE_0 es la suma al cuadrado de las desviaciones con respecto a la media $\bar{y}$ y es conocida como la suma total de cuadrados ajustada. El índice R^2 es llamado *coeficiente de determinación*.

Las técnicas discutidas anteriormente pueden ser utilizadas para evaluar múltiples modelos entre ellos usando el test F de la Ec. (2.53). Estos tests han sido usados previamente en el procedimiento llamado *regresión múltiple paso a paso*, cuya utilidad radica en el hallazgo de un conjunto de variables útiles para el modelo. Una alternativa es enfocarse en un procedimiento de *selección de modelos* que no proceda de manera secuencial, sino simplemente evalúe cada modelo en sus propios méritos. Suponiendo que se considera

un modelo de regresión normal con k coeficientes y denotando al *estimador de máxima verosimilitud* para la varianza como

$$\hat{\sigma}_k^2 = \frac{SSE(k)}{n}, \tag{2.57}$$

donde $SSE(k)$ denota la suma de los residuales al cuadrado bajo el modelo con k coeficientes de regresión. Akaike [20] sugiere medir la bondad de ajuste para este modelo en particular balanceando el error del ajuste contra el número de parámetros en el modelo, logrando *parsimonía* (en igualdad de condiciones, la explicación más sencilla suele ser la más probable).

2.4.1.1. Criterios de Información

De manera formal, el *criterio de información de Akaike* (AIC) se define como

$$\text{AIC} = -2log\ L_k + 2k, \tag{2.58}$$

donde L_k es la verosimilitud maximizada y k es el número de parámetros en el modelo. AIC es un estimado de la *discrepancia de Kullback-Leibler* [21] entre el verdadero modelo y un modelo candidato.

Para un modelo de regresión normal, AIC puede ser reducido a

$$\text{AIC} = log\ \hat{\sigma}_k^2 + \frac{n + 2k}{n} \tag{2.59}$$

donde $\hat{\sigma}_k^2$ está dado por la Ec. (2.57), k es el número de parámetros del modelo y n es el tamaño muestral.

El valor de k que logre el AIC mínimo especifica el mejor modelo. La idea es aproximadamente que la minimización de $\hat{\sigma}_k^2$ sería un objetivo razonable, excepto porque decrece de manera monotónica a medida que k crece. Por ello, se debe penalizar la varianza del error por un término proporcional al número de parámetros. La elección de la penalización utilizada en la definición (2.58) no es única, existiendo literatura considerable referente a distintos términos de penalización. Por ejemplo, el *AIC corregido por sesgo* (AICc)

definido como

$$\text{AICc} = log\ \hat{\sigma}_k^2 + \frac{n+k}{n-k-2} \tag{2.60}$$

A su vez, es posible derivar un término de corrección basado en argumentos Bayesianos, como lo sugiere el *coeficiente de información de Schwarz* (BIC) definido como

$$\text{BIC} = log\ \hat{\sigma}_k^2 + \frac{k\ log\ n}{n} \tag{2.61}$$

Notar que el término de penalización en BIC es mayor que en AIC. En consecuencia, BIC tiende a escoger modelos más pequeños. Estudios de simulación [22] verifican que BIC se desempeña bien en muestras grandes, mientras AICc tiende a ser superior en muestras pequeñas, donde el número relativo de parámetros es grande.

2.4.2. Modelos de Series de Tiempo

La regresión clásica restringe a la variable dependiente a ser influenciada por valores actuales de variables independientes, lo que a menudo insuficiente para explicar toda la dinámica de una serie de tiempo. Usualmente, la ACF de los residuales de una regresión lineal simple revelará estructuras adicionales que una regresión no puede capturar. En el caso de la series de tiempo, es deseable permitir que la variable dependiente se vea influenciada por los valores previos de variables independientes y posiblemente de sus propios valores pasados. La introducción de correlación que pueda ser generada a través de relaciones lineales rezagadas conduce a la propuesta de modelos *autoregresivos* (AR) y modelos *autoregresivos de media móvil* (ARMA). La adición de modelos no estacionarios a la combinación conduce al modelo *autoregresivo integrado de media móvil* (ARIMA), popularizado por Box y Jenkins [14] y su método de identificación de modelos ARIMA.

2.4.2.1. Modelo Autoregresivo (AR)

Los modelos autoregresivos están basados en la idea de que el valor actual de la serie, y_t, puede ser explicado como una función de p valores pasados, $y_{t-1}, y_{t-2}, \ldots, y_{t-p}$, donde p determina el número de pasos en el pasado requeridos para pronosticar el valor actual. Un

modelo autoregresivo de orden p, abreviado AR(p), posee la forma

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t, \tag{2.62}$$

donde y_t es estacionaria, $\varepsilon_t \sim \text{wn}(0, \sigma_\varepsilon^2)$, y $\phi_1, \phi_2, \ldots, \phi_p$ son constantes ($\phi_p \neq 0$). La media de y_t, μ, es cero, y en caso contrario, es posible reemplazar y_t por $y_t - \mu$ en la Ec. (2.62), obteniendo

$$y_t - \mu = \phi_1(y_{t-1} - \mu) + \phi_2(y_{t-2} - \mu) + \cdots + \phi_p(y_{t-p} - \mu) + \varepsilon_t, \tag{2.63}$$

o escribir

$$y_t = \alpha + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t, \tag{2.64}$$

donde $\alpha = \mu(1 - \phi_1 - \cdots - \phi_p)$.

Notar que la Ec. (2.64) es similar al modelo de regresión de la sección (2.4.1), y por ello el término “auto” regresión. Algunas dificultades técnicas surgen aplicar dicho modelo debido a que las variables regresoras $y_{t-1}, y_{t-2}, \ldots, y_{t-p}$, son componentes aleatorios, mientras que x_t se asumía fijo.

Una formulación útil sigue al usar el *operador de rezagos* definido como

$$By_t = y_{t-1} \tag{2.65}$$

extendiéndose a potencias $B^2 y_t = B(By_t) = By_{t-1} = y_{t-2}$ y así sucesivamente. Por lo tanto

$$B^k y_t = y_{t-k}. \tag{2.66}$$

La idea de un operador inverso puede ser concretada si se requiere que $B^{-1}B = 1$, así

$$y_t = B^{-1}By_t = B^{-1}y_{t-1}. \tag{2.67}$$

Esto es, B^{-1} es el *operador de traslado hacia adelante*. Además, dado que la *diferenciación* juega un rol central en el análisis de series de tiempo, recibe su propia notación. La primera

diferencia puede ser denotada como

$$\nabla y_t = y_t - y_{t-1} = (1 - B)y_t \tag{2.68}$$

y permite eliminar tendencias lineales. Una segunda diferencia permitirá eliminar tendencias cuadráticas y así en adelante. En general, la *diferencia de orden d* puede ser definida como

$$\nabla^d = (1 - B)^d \tag{2.69}$$

donde el operador $(1 - B)^d$ puede ser expandido algebraicamente para evaluar para valores altos de d.

Luego, usando la notación previa, es posible escribir el modelo AR(p) de la Ec. (2.62) como

$$(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p)y_t = \varepsilon_t \tag{2.70}$$

o de manera más concisa

$$\phi(B)y_t = \varepsilon_t \tag{2.71}$$

Las propiedades de $\phi(B)$ son importantes para resolver la Ec. (2.71) para y_t, lo que conduce a la definición del *operador autoregresivo* como

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p \tag{2.72}$$

2.4.2.2. Modelo de Media Móvil (MA)

Como una alternativa a la representación autoregresiva en la cual y_t, a la izquierda de la ecuación es asumida como una combinación lineal; el modelo de *media móvil de orden q*, MA(q), asume que el ruido blanco ε_t al lado derecho de la ecuación es combinado de forma lineal para formar los datos observados. Eso es

$$y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}, \tag{2.73}$$

donde $\varepsilon_t \sim wn(0, \sigma_\varepsilon^2)$, y $\theta_1, \theta_2, \ldots, \theta_q$ ($\theta_q \neq 0$) son parámetros.

Es posible escribir el proceso MA(q) en su forma equivalente

$$y_t = \theta(B)\varepsilon_t \tag{2.74}$$

usando la siguiente definición para el *operador de media móvil*

$$\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \theta_q B^q \tag{2.75}$$

A diferencia de un proceso autoregresivo, el proceso de media móvil es estacionario para cualquier valor de los parámetros $\theta_1, \ldots, \theta_q$

2.4.2.3. Modelo de Autoregresivo de Media Móvil (ARMA)

Una serie de tiempo $\{y_t; t = 0, \pm 1, \pm 2, \ldots\}$ es ARMA(p, q) si es estacionaria y

$$y_t = \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} \tag{2.76}$$

con $\phi_p \neq 0$, $\theta_q \neq 0$, y $\sigma_\varepsilon^2 > 0$. Los parámetros p y q determinan el orden autoregresivo y de media móvil, respectivamente. Si y_t tiene media μ distinta a cero, es posible establecer $\alpha = \mu(1 - \phi_1 - \cdots - \phi_p)$ y escribir el modelo como

$$y_t = \alpha + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} \tag{2.77}$$

donde $\varepsilon_t \sim \text{wn}(0, \sigma_\varepsilon^2)$. En particular, el modelo ARMA(p, q) puede ser escrito de manera concisa como

$$\phi(B) y_t = \theta(B)\varepsilon_t \tag{2.78}$$

2.4.2.4. Modelo de Autoregresivo Integrado de Media Móvil (ARIMA)

Un proceso y_t se denomina ARIMA(p, d, q) si

$$\nabla^d y_t = (1 - B)^d y_t \tag{2.79}$$

es ARMA(p, q). En general, se escribirá el modelo como

$$\phi(B)(1 - B)^d y_t = \theta(B)\varepsilon_t. \tag{2.80}$$

Si $E(\nabla^d y_t) = \mu$, el modelo puede ser escrito

$$\phi(B)(1 - B)^d y_t = \delta + \theta(B)\varepsilon_t. \tag{2.81}$$

donde $\delta = \mu(1 - \phi_1 - \cdots - \phi_p)$.

2.4.2.5. Funciones de autocorrelación y autocorrelación parcial

Usualmente es difícil saber a partir de un gráfico temporal si los valores de p y q elegidos en el contexto de los modelos ARIMA son los adecuados para los datos. Sin embargo, a veces es posible usar un gráfico obtenido a partir de la función de autocorrelación (véase sección 2.2.1), así como también el gráfico de la *función de autocorrelación parcial* (PACF) para determinar el valor de p y q.

El gráfico ACF muestra las autocorrelaciones que miden la relación entre y_t e y_{t-k}, para diferentes valores de k. El problema surge cuando se considera que y_t e y_{t-k} están correlacionadas y entonces y_{t-1} e y_{t-2} también lo estarán. Sin embargo, y_t e $y_t - 2$ podrían estar correlacionadas simplemente porque están conectadas a través de y_{t-1}, en lugar de que se deba a nueva información contenida en y_{t-2} que pudiese ser usada para pronosticar y_t.

Para superar este problema, es posible utilizar autocorrelaciones parciales, que permiten medir el nivel de relación entre y_t e y_{t-k} luego de remover los efectos de los rezagos $1, 2, \ldots, k - 1$. Así, la primera autocorrelación parcial será idéntica a la primera autocorrelación. Cada autocorrelación parcial puede ser estimada como el último coeficiente en un modelo autoregresivo. De manera más específica, α_k, el k-ésimo coeficiente de autocorrelación parcial, es equivalente al estimador de ϕ_k en un modelo AR(k).

Tabla 2.2: Comportamiento de ACF y PACF para modelos ARMA

	AR(p)	MA(q)	ARMA(p,q)
ACF	Disminución	Interrupción luego del lag q	Disminución
PACF	Interrupción luego del lag p	Disminución	Disminución

2.4.3. Suavizamiento Exponencial

Históricamente, el *suavizamiento exponencial* describe una clase de métodos de pronóstico. Existe una variedad de métodos que forman parte de la familia del suavizamiento exponencial, cada uno con la propiedad de que los pronósticos son combinaciones ponderadas de observaciones pasadas, con observaciones recientes que reciben más ponderación con respecto a las observaciones pasadas. El nombre "suavizado exponencial" refleja el hecho de que los pesos disminuyen exponencialmente a medida que las observaciones se hacen distantes en el tiempo. En otras palabras, cuanto más reciente sea la observación, mayor será la ponderación asociada a ella. Este esquema permite generar pronósticos confiables de manera rápida y para una amplia gama de series de tiempo, permitiendo pronosticar datos que no poseen tendencias claras o patrones estacionales [23].

2.4.3.1. Descomposición de Series de tiempo

En la sección (2.1) se ha caracterizado a las series de tiempo utilizando palabras como *tendencia*, *estacionalidad* y *ciclicidad*. En economía y negocios es común pensar en las series de tiempo como una combinación de varios componentes tales como

Tendencia (T) Dirección a largo plazo de la serie. No es necesario que sea lineal. También se considera cuando existe un cambio de dirección en el crecimiento de la serie.

Estacionalidad (S) Patrón que se repite con periodicidad conocida (ej. 12 meses por año, 7 días a la semana).

Ciclicidad (C) Patrón que se repite con cierta regularidad pero con periodicidad desconocida y cambiante (ej. ciclos económicos).

Error irregular (E) Componente impredecible de una serie.

Muchas series de tiempo incluyen tendencias, ciclos y estacionalidad. Por ello, al momento de escoger un método de pronóstico, es necesario identificar en primera instancia los patrones presentes en los datos, y luego escoger un método que sea capaz de capturar los patrones de manera adecuada.

Considerando que cualquier elemento cíclico será integrado dentro del componente de tendencia a menos que sea indicado lo contrario, es posible combinar los tres componentes restantes de diversas maneras. Un modelo puramente aditivo puede ser expresado como

$$y = T + S + E, \tag{2.82}$$

donde los tres componentes son sumados para formar la serie observada. De manera análoga, un modelo puramente multiplicativo puede ser escrito como

$$y = T \times S \times E, \tag{2.83}$$

donde los tres componentes son multiplicados para formar la serie observada. Una serie *ajustada estacionalmente* se forma mediante la extracción del componente estacional de los datos, dejando solamente los componentes de tendencia y error. En el modelo aditivo, la serie ajustada estacionalmente es $y - S$, mientras que en el modelo multiplicativo corresponde a y/S.

2.4.3.2. Clasificación de métodos de suavizamiento exponencial

Es usual comenzar considerando el componente de tendencia, el cual en sí mismo es una combinación del término del nivel (l) y un término de crecimiento (b). El nivel y el crecimiento pueden ser combinados en numerosas maneras, obteniéndose cinco tipos de tendencia futura. Sea T_h el pronóstico de tendencia para los h períodos siguientes, y sea ϕ el parámetro de amortiguación ($0 < \phi < 1$). Entonces los cinco tipos de tendencia o patrones de crecimiento son

$$\begin{aligned}
\text{Ninguno:} \quad & T_h = l \\
\text{Aditivo:} \quad & T_h = l + bh \\
\text{Aditivo amortiguado:} \quad & T_h = l + (\phi + \phi^2 + \cdots + \phi^h)b \\
\text{Multiplicativo:} \quad & T_h = lb^h \\
\text{Multiplicativo amortiguado:} \quad & T_h = lb^{(\phi+\phi^2+\cdots+\phi^h)}
\end{aligned}$$

Un método de tendencia amortiguada es apropiado cuando existe tendencia en la serie de tiempo, pero se cree que la tasa de crecimiento al final de los datos históricos es improbable que continúe por más de un período corto de tiempo en el futuro.

Habiendo escogido el componente de tendencia, es posible introducir un componente estacional de manera aditiva o multiplicativa. Finalmente, se incluye el error de la misma manera. Ignorando el componente del error, es posible definir los quince métodos de suavizamiento exponencial que se encuentran en la Tabla (2.3).

TABLA 2.3**:** Clasificación bidireccional para métodos de suavizamiento exponencial

Componente de tendencia	Componente estacional		
	N	A	M
	(Ninguno)	(Aditivo)	(Multiplicativo)
N (Ninguno)	N,N	N,A	N,M
A (Aditivo)	A,N	A,A	A,M
A_d (Aditivo amortiguado)	A_d,N	A_d,A	A_d,M
M (Multiplicativo)	M,N	M,A	M,M
M_d (Multiplicativo amortiguado)	M_d,N	M_d,A	M_d,M

Fuente: [24]

En particular, algunos de estos métodos son conocidos con otros nombres

- **(N,N):** Suavizamiento exponencial simple (SES)
- **(A,N):** Método lineal de Holt
- **(A_d,N):** Método de tendencia aditiva amortiguada
- **(A,A):** Método aditivo de Holt-Winters

- **(A,M):** Método multiplicativo de Holt-Winters
- **(A_d,M):** Método amortiguado de Holt-Winters

Denotando la serie observada por $y_1, y_2, \ldots, y_n$, el pronóstico de y_{t+h} basado en toda la información hasta el período t es denotado por $\hat{y}_{t+h|t}$. Para pronósticos a un paso, es útil la notación $\hat{y}_{t+1} \equiv \hat{y}_{t+1|t}$. Asumiendo que los parámetros relevantes requeridos para pronosticar han sido estimados, a continuación se introducen los métodos de suavizamiento exponencial más conocidos.

2.4.3.3. Suavizamiento Exponencial Simple (SES)

Suponiendo que se han observado dato hasta e incluyendo el período $t-1$, y se desea pronosticar el valor siguiente de la serie de tiempo, y_t. Denotando al pronóstico por $\hat{y}_t$, cuando la observación y_t se encuentra disponible, el error de pronóstico corresponde a $y_t - \hat{y}_t$. El Suavizamiento Exponencial Simple (SES) [25], toma el pronóstico para el período previo y lo ajusta usando el error de pronóstico según

$$\hat{y}_{t+1} = \hat{y}_t + \alpha(y_t - \hat{y}_t), \tag{2.84}$$

donde $0 < \alpha < 1$. El nuevo pronóstico será simplemente el antiguo pronóstico más un ajuste por el error cometido en el último pronóstico. Si α es cercano a 1, el nuevo pronóstico incluirá un ajuste sustancial para el error en el pronóstico previo. En caso contrario, el nuevo pronóstico se verá ajustado levemente.

$$\hat{y}_{t+1} = \alpha y_t + (1-\alpha)\hat{y}_t, \tag{2.85}$$

Así, el pronóstico $\hat{y}_{t+1}$ puede ser interpretado como un promedio ponderado entre el pronóstico más reciente y la observación más reciente. Es posible estudiar el efecto del suavizamiento exponencial si la Ec. (2.85) es expandida reemplazando $\hat{y}_t$ con sus componentes

como sigue

$$\begin{aligned}\hat{y}_{t+1} =&\alpha y_t + \alpha(1-\alpha)\hat{y}_{t-1} + \alpha(1-\alpha)^2 y_{t-2} + \alpha(1-\alpha)^3\hat{y}_{t-3} \\ &+ \alpha(1-\alpha)^4 y_{t-4} + \cdots + \alpha(1-\alpha)^{t-1} y_1 + (1-\alpha)^t \hat{y}_1\end{aligned} \tag{2.86}$$

Así $\hat{y}_{t+1}$ representa una media móvil ponderada de las observaciones pasadas con pesos que decaen de manera exponencial. La elección del valor inicial es particularmente importante y se conoce como el "problema de inicialización". En la era pre-computacional, dicho valor era establecido con la primera observación, sin embargo, hoy en día existen mejores maneras de establecer dichos parámetros.

Una representación alternativa es la definición de los componentes de un suavizamiento exponencial. La representación en componentes de un suavizamiento exponencial involucra ecuación de pronóstico y una ecuación de suavizamiento para cada uno de los componentes incluidos en el método. En el caso de SES, el único componente a considerar es el nivel, l_t, luego su formulación en componentes es

$$\text{Ecuación de pronóstico:} \qquad \hat{y}_{t+h|t} = l_t \tag{2.87a}$$

$$\text{Ecuación de suavizamiento:} \qquad l_t = \alpha y_t + (1-\alpha) l_{t-1}, \tag{2.87b}$$

estableciendo $h = 1$ se obtienen los valores ajustados, mientras que para $t = T$ se obtienen los verdaderos pronósticos fuera de los datos de entrenamiento. La ecuación de pronóstico muestra que el valor pronosticado en el período $t + 1$ es el nivel estimado al período t. La ecuación de suavizamiento para el nivel permite obtener el nivel estimado para la serie en cada período t.

Reemplazando l_t con $\hat{y}_{t+1|t}$ y l_{t-1} con $\hat{y}_{t|t-1}$ en la ecuación de suavizamiento, es posible recuperar la forma de medias ponderadas de SES.

Finalmente, la forma de componentes de SES no es particularmente útil en este escenario, sin embargo, permite establecer las bases para agregar más componentes que permitan generalizar para otros tipos de tendencias y estacionalidad.

2.4.3.4. Método Lineal de Holt

Holt [26] extendió SES a suavizamiento exponencial lineal para poder pronosticar datos con tendencias. El pronóstico para el Método de suavizamiento exponencial lineal de Holt es hallado usando dos constantes de suavizamiento, α y β^* (con valores entre 0 y 1), y tres ecuaciones

$$\text{Ecuación de Nivel:} \quad l_t = \alpha y_t + (1-\alpha)(l_{t-1} + b_{t-1}) \tag{2.88a}$$

$$\text{Ecuación de Crecimiento:} \quad b_t = \beta^*(l_t - l_{t-1}) + (1-\beta^*)b_{t-1} \tag{2.88b}$$

$$\text{Ecuación de Pronóstico:} \quad \hat{y}_{t+h|t} = l_t + b_t h. \tag{2.88c}$$

Aquí l_t denota un estimado del nivel de la serie en el período t y b_t corresponde a un estimado de la pendiente o crecimiento de la serie en el período t. Notar que b_t es una media ponderada del crecimiento previo b_{t-1} y un estimado del crecimiento basado en la diferencia entre niveles sucesivos. Además, si $\alpha = \beta^*$, el Método Lineal de Holt es equivalente a SES, y en el caso particular $\beta^* = 0$, se define el SES con tendencia (*drift*) como sigue

$$\text{Ecuación de Nivel:} \quad l_t = \alpha y_t + (1-\alpha)(l_{t-1} + b) \tag{2.89a}$$

$$\text{Ecuación de Pronóstico:} \quad \hat{y}_{t+h|t} = l_t + bh. \tag{2.89b}$$

2.4.3.5. Método de Tendencia aditiva amortiguada

Gardner y McKenzie (1985) [27] propusieron una modificación al Método lineal de Holt para permitir "amortiguación" de las tendencias. Las ecuaciones para este método son las siguientes

$$\text{Ecuación de Nivel:} \quad l_t = \alpha y_t + (1-\alpha)(l_{t-1} + \phi b_{t-1}) \tag{2.90a}$$

$$\text{Ecuación de Crecimiento:} \quad b_t = \beta^*(l_t - l_{t-1}) + (1-\beta^*)\phi b_{t-1} \tag{2.90b}$$

$$\text{Ecuación de Pronóstico:} \quad \hat{y}_{t+h|t} = l_t + (\phi + \phi^2 + \cdots + \phi^h)b_t. \tag{2.90c}$$

Así, el crecimiento para el pronóstico a un paso de y_{t+1} es ϕb_t, y el crecimiento es amortiguado por un factor de ϕ para cada periodo futuro adicional. Notar que si $\phi = 1$, el método se reduce al Método Lineal de Holt. Para $0 < \phi < 1$, a medida que $h \to \infty$ los pronósticos se aproximan a una asíntota dada por $l_t + \phi b_t/(1-\phi)$. Usualmente se restringe $\phi > 0$ para evitar la aplicación de un coeficiente negativo a b_{t-1} en la Ec. (2.90b), y $\phi \leq 1$ para evitar que b_t crezca de manera exponencial.

2.4.3.6. Método de Tendencia y Estacionalidad de Holt-Winters

El Método de Holt-Winters [28] se basa en tres ecuaciones de suavizamiento, agregando una ecuación para la estacionalidad al Método Lineal de Holt. Existen dos Métodos de Holt-Winters, dependiendo si la estacionalidad es modelada de manera aditiva o multiplicativa.

Estacionalidad Multiplicativa *(Método A,M)*

Las ecuaciones básicas para el método multiplicativo de Holt-Winters son

$$\text{Ecuación de Nivel:} \quad l_t = \alpha \frac{y_t}{s_{t-m}} + (1-\alpha)(l_{t-1} + \phi b_{t-1}) \tag{2.91a}$$

$$\text{Ecuación de Crecimiento:} \quad b_t = \beta^*(l_t - l_{t-1}) + (1-\beta^*)\phi b_{t-1} \tag{2.91b}$$

$$\text{Ecuación de Estacionalidad:} \quad s_t = \gamma \frac{y_t}{(l_{t-1} + b_{t-1})} + (1-\gamma) s_{t-m} \tag{2.91c}$$

$$\text{Ecuación de Pronóstico:} \quad \hat{y}_{t+h|t} = (l_t + b_t h) s_{t-m+h_m^+}, \tag{2.91d}$$

donde m es la longitud de la estacionalidad (ej., cantidad de meses o trimestres en un año), l_t representa el nivel de la serie, b_t denota el crecimiento, s_t es el componente estacional, $\hat{y}_{t+h|t}$ es el pronóstico para h períodos en adelante, y $h_m^+ = [(h-1) \bmod m] + 1$. Los parámetros ($\alpha$, β^* y γ) usualmente están restringidos entre 0 y 1. En este caso, la estacionalidad es multiplicativa en el sentido de que el nivel reinante de la serie es multiplicado por un índice estacional.

Estacionalidad Aditiva *(Método A,A)*

El componente estacional del método de Holt-Winters puede ser tratado de manera aditiva, aunque esta practica es menos frecuente. Las ecuaciones básicas para el método aditivo de Holt-Winters son

$$\text{Ecuación de Nivel:} \quad l_t = \alpha(y_t - s_{t-m}) + (1-\alpha)(l_{t-1} + \phi b_{t-1}) \tag{2.92a}$$

$$\text{Ecuación de Crecimiento:} \quad b_t = \beta^*(l_t - l_{t-1}) + (1-\beta^*)\phi b_{t-1} \tag{2.92b}$$

$$\text{Ecuación de Estacionalidad:} \quad s_t = \gamma(y_t - l_{t-1} - b_{t-1}) + (1-\gamma)s_{t-m} \tag{2.92c}$$

$$\text{Ecuación de Pronóstico:} \quad \hat{y}_{t+h|t} = l_t + b_t h + s_{t-m+h_m^+}. \tag{2.92d}$$

La única diferencia entre las ecuaciones del método con estacionalidad multiplicativa es que los índices estacionales son sumados y restados en lugar de ser productos y ratios.

2.4.3.7. Método de Holt-Winters con Estacionalidad doble (DSHW)

Taylor [29] propone una nueva formulación para el Método de Holt-Winters para que sea capaz de acomodar más de un patrón estacional. El Método de Holt-Winters para estacionalidad doble multiplicativa está dado por las Ecs. (2.93)

$$\text{Ecuación de Nivel:} \quad l_t = \alpha\left(y_t - s^{(1)}_{t-m_1} - s^{(2)}_{t-m_2}\right) + (1-\alpha)(l_{t-1} + b_{t-1}) \tag{2.93a}$$

$$\text{Ecuación de Crecimiento:} \quad b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1} \tag{2.93b}$$

$$\text{Ecuación de Estacionalidad:} \quad s^{(1)}_t = \gamma\left(y_t - l_t - s^{(2)}_{t-m_2}\right) + (1-\gamma)s^{(1)}_{t-m_1} \tag{2.93c}$$

$$\text{Ecuación de Estacionalidad:} \quad s^{(2)}_t = \delta\left(y_t - l_t - s^{(1)}_{t-m_1}\right) + (1-\delta)s^{(2)}_{t-m_2} \tag{2.93d}$$

$$\text{Ecuación de Pronóstico:} \quad \hat{y}_{t+h|t} = l_t + b_t h + s^{(1)}_{t-m_1+h} + s^{(2)}_{t-m_2+h} + \phi^h\left[y_t - \left(l_{t-1} + b_{t-1} + s^{(1)}_{t-m_1} + s^{(2)}_{t-m_2}\right)\right] \tag{2.93e}$$

donde α, β, γ, δ son los parámetros de suavizamiento. El término que involucra el parámetro ϕ corresponde a un ajuste de autocorrelación de primer orden, para el cual ha sido probado que produce mejores resultados fuera de la muestra [30]. Vale la pena mencionar que las expresiones de las Ecs. (2.93) pueden ser expandidas para incluir más patrones estacionales mediante la introducción de un índice de estacionalidad extra y su ecuación de

suavizamiento respectiva.

2.4.3.8. Modelos de espacio de estado de innovaciones (BATS - TBATS)

Una desventaja de los modelos de suavizamiento exponencial revisados en la secciones anteriores es la carencia de un esquema estadístico para producir intervalos de predicción y pronósticos puntuales. La aproximación de espacio de estado de innovaciones provee un esquema mientras conserva la naturaleza intuitiva del suavizamiento exponencial en sus ecuaciones. Dichos modelos permiten obtener intervalos de predicción, estimación por máxima verosimilitud, procedimientos de selección de modelos y otros [24].

Considerando los modelos revisados en secciones previas, es posible introducir los modelos de espacio de estado que subyacen los métodos de suavizamiento exponencial. Los modelos de espacio de estado proveen de una flexibilidad considerable en la especificación de una estructura paramétrica. Sea y_t la observación en el período t, y sea $\boldsymbol{x}_t$ el "vector de estados" que contiene componentes no observados que describen el nivel, la tendencia y estacionalidad de la serie. Entonces un modelo de espacio de estado de innovaciones puede ser escrito como

$$y_t = \boldsymbol{w}'\boldsymbol{x}_{t-1} + \varepsilon_t \tag{2.94a}$$

$$\boldsymbol{x}_t = \boldsymbol{F}\boldsymbol{x}_{t-1} + \boldsymbol{g}\varepsilon_t \tag{2.94b}$$

donde $\{\varepsilon_t\}$ es una serie ruido blanco y $\boldsymbol{F}$, $\boldsymbol{g}$ y $\boldsymbol{w}$ son coeficientes. La Ec. (2.94a) es conocida como la ecuación de la *medición* u observación y describe la relación entre los estados no observados $\boldsymbol{x}_{t-1}$ y la observación y_t. Las Ec. (2.94b) corresponde a la ecuación de *transición* o estado; describiendo la evolución de los estados en el tiempo. El uso de errores idénticos (o innovaciones) en ambas ecuaciones lo hace un modelo de espacio de estado de "innovaciones". Varios métodos de suavizamiento exponencial revisados en esta sección son equivalentes a pronósticos puntuales de casos especiales del modelo descrito por las Ecs. (2.94).

Los modelos de espacio de estados se ajustan bien con las aproximaciones de suavizamiento exponencial ya que el nivel, la tendencia y los componentes estacionales quedan

establecidos de manera explícita en los modelos.

Es posible definir modelos de espacio de estado no lineales, por ejemplo

$$y_t = w(\boldsymbol{x}_{t-1}) + r(\boldsymbol{x}_{t-1})\varepsilon_t \tag{2.95a}$$

$$\boldsymbol{x}_t = f(\boldsymbol{x}_{t-1}) + g(\boldsymbol{x}_{t-1})\varepsilon_t. \tag{2.95b}$$

Una alternativa y especificación más común es asumir que los errores en ambas ecuaciones es mutuamente independiente. Esto es, que $\boldsymbol{g}\varepsilon_t$ en la Ec. (2.94b) es reemplazado por z_t, cuando z_t consiste en una serie de ruido blanco independiente que también es independiente de ε_t , el error en la ecuación de la observación. El supuesto de que $\boldsymbol{z}_t$ y ε_t son independientes provee de restricciones para asegurar que los parámetros restantes sean *estimables* o *identificables*.

Para cada método de suavizamiento exponencial existen dos modelos, cada uno con errores aditivos o multiplicativos. Los pronósticos puntuales para los dos modelos son idénticos (si los mismos parámetros usados), pero sus intervalos de predicción podrían diferir. La notación (E,T,S) permite identificar los componentes de error (E), tendencia (T), y estacionalidad (S). Así, por ejemplo, el modelo ETS(A,A,N) tiene errores aditivos, tendencia aditiva y no posee estacionalidad; en otras palabras, corresponde al método lineal de Holt con error aditivo. Una vez el modelo ha sido especificado, es posible estudiar la distribución de probabilidad de los valores futuros de una serie y hallar, por ejemplo, la media condicional de una observación futura dado el conocimiento que se posee del pasado. Es posible de notar $\mu_{t+h|t} = E(y_{t+h|\boldsymbol{x}_t})$, donde $\boldsymbol{x}_t$ contiene los componentes no observados tales como l_t, b_t y s_t. Para $h = 1$ es posible usar $\mu_{t+1} \equiv \mu_{t+1|t}$. Para la mayoría de modelos, las medias condicionales serán idénticas a los pronósticos puntuales dados por la Tabla (2.4), así $\mu_{t+h|t} = \hat{y}_{t+h|t}$. El modelo general involucra un vector de estados $\boldsymbol{x}_t = (l_t, b_t, s_t, s_{t-1}, \ldots, s_{t-m+1})'$ y la forma de las ecuaciones presentadas en la Ec. (2.95), donde $\{\varepsilon_t\}$ es un proceso de ruido blanco Gaussiano con varianza σ^2, y $\mu_t = w(\boldsymbol{x}_{t-1})$. El modelo con errores aditivos presenta $r(\boldsymbol{x}_{t-1}) = 1$, así $y_t = \mu_t + \varepsilon_t$. El modelo con errores multiplicativos tiene $r(\boldsymbol{x}_{t-1}) = \mu_t$, así $y_t = \mu_t(1 + \varepsilon_t)$. Entonces, $\varepsilon_t = (y_t - \mu_t)/\mu_t$ es el error relativo para el modelo multiplicativo. Los modelos no son únicos. Cualquier valor de

TABLA 2.4: Fórmulas para cálculos recursivos y pronósticos puntuales

Trend	Seasonal		
	N	A	M
N	$\ell_t = \alpha y_t + (1-\alpha)\ell_{t-1}$ $\hat{y}_{t+h\|t} = \ell_t$	$\ell_t = \alpha(y_t - s_{t-m}) + (1-\alpha)\ell_{t-1}$ $s_t = \gamma(y_t - \ell_{t-1}) + (1-\gamma)s_{t-m}$ $\hat{y}_{t+h\|t} = \ell_t + s_{t-m+h_m^+}$	$\ell_t = \alpha(y_t / s_{t-m}) + (1-\alpha)\ell_{t-1}$ $s_t = \gamma(y_t / \ell_{t-1}) + (1-\gamma)s_{t-m}$ $\hat{y}_{t+h\|t} = \ell_t s_{t-m+h_m^+}$
A	$\ell_t = \alpha y_t + (1-\alpha)(\ell_{t-1} + b_{t-1})$ $b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1}$ $\hat{y}_{t+h\|t} = \ell_t + hb_t$	$\ell_t = \alpha(y_t - s_{t-m}) + (1-\alpha)(\ell_{t-1} + b_{t-1})$ $b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1}$ $s_t = \gamma(y_t - \ell_{t-1} - b_{t-1}) + (1-\gamma)s_{t-m}$ $\hat{y}_{t+h\|t} = \ell_t + hb_t + s_{t-m+h_m^+}$	$\ell_t = \alpha(y_t / s_{t-m}) + (1-\alpha)(\ell_{t-1} + b_{t-1})$ $b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)b_{t-1}$ $s_t = \gamma(y_t / (\ell_{t-1} + b_{t-1})) + (1-\gamma)s_{t-m}$ $\hat{y}_{t+h\|t} = (\ell_t + hb_t)s_{t-m+h_m^+}$
A_d	$\ell_t = \alpha y_t + (1-\alpha)(\ell_{t-1} + \phi b_{t-1})$ $b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)\phi b_{t-1}$ $\hat{y}_{t+h\|t} = \ell_t + \phi_h b_t$	$\ell_t = \alpha(y_t - s_{t-m}) + (1-\alpha)(\ell_{t-1} + \phi b_{t-1})$ $b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)\phi b_{t-1}$ $s_t = \gamma(y_t - \ell_{t-1} - \phi b_{t-1}) + (1-\gamma)s_{t-m}$ $\hat{y}_{t+h\|t} = \ell_t + \phi_h b_t + s_{t-m+h_m^+}$	$\ell_t = \alpha(y_t / s_{t-m}) + (1-\alpha)(\ell_{t-1} + \phi b_{t-1})$ $b_t = \beta^*(\ell_t - \ell_{t-1}) + (1-\beta^*)\phi b_{t-1}$ $s_t = \gamma(y_t / (\ell_{t-1} + \phi b_{t-1})) + (1-\gamma)s_{t-m}$ $\hat{y}_{t+h\|t} = (\ell_t + \phi_h b_t)s_{t-m+h_m^+}$
M	$\ell_t = \alpha y_t + (1-\alpha)\ell_{t-1}b_{t-1}$ $b_t = \beta^*(\ell_t / \ell_{t-1}) + (1-\beta^*)b_{t-1}$ $\hat{y}_{t+h\|t} = \ell_t b_t^h$	$\ell_t = \alpha(y_t - s_{t-m}) + (1-\alpha)\ell_{t-1}b_{t-1}$ $b_t = \beta^*(\ell_t / \ell_{t-1}) + (1-\beta^*)b_{t-1}$ $s_t = \gamma(y_t - \ell_{t-1}b_{t-1}) + (1-\gamma)s_{t-m}$ $\hat{y}_{t+h\|t} = \ell_t b_t^h + s_{t-m+h_m^+}$	$\ell_t = \alpha(y_t / s_{t-m}) + (1-\alpha)\ell_{t-1}b_{t-1}$ $b_t = \beta^*(\ell_t / \ell_{t-1}) + (1-\beta^*)b_{t-1}$ $s_t = \gamma(y_t / (\ell_{t-1}b_{t-1})) + (1-\gamma)s_{t-m}$ $\hat{y}_{t+h\|t} = \ell_t b_t^h s_{t-m+h_m^+}$
M_d	$\ell_t = \alpha y_t + (1-\alpha)\ell_{t-1}b_{t-1}^{\phi}$ $b_t = \beta^*(\ell_t / \ell_{t-1}) + (1-\beta^*)b_{t-1}^{\phi}$ $\hat{y}_{t+h\|t} = \ell_t b_t^{\phi_h}$	$\ell_t = \alpha(y_t - s_{t-m}) + (1-\alpha)\ell_{t-1}b_{t-1}^{\phi}$ $b_t = \beta^*(\ell_t / \ell_{t-1}) + (1-\beta^*)b_{t-1}^{\phi}$ $s_t = \gamma(y_t - \ell_{t-1}b_{t-1}^{\phi}) + (1-\gamma)s_{t-m}$ $\hat{y}_{t+h\|t} = \ell_t b_t^{\phi_h} + s_{t-m+h_m^+}$	$\ell_t = \alpha(y_t / s_{t-m}) + (1-\alpha)\ell_{t-1}b_{t-1}^{\phi}$ $b_t = \beta^*(\ell_t / \ell_{t-1}) + (1-\beta^*)b_{t-1}^{\phi}$ $s_t = \gamma(y_t / (\ell_{t-1}b_{t-1}^{\phi})) + (1-\gamma)s_{t-m}$ $\hat{y}_{t+h\|t} = \ell_t b_t^{\phi_h} s_{t-m+h_m^+}$

Fuente: [24]

$r(\boldsymbol{x}_{t-1})$ conducirá a valores idénticos para los pronósticos puntuales de y_t.

Cada método presente en la Tabla (2.4) puede ser escrito en forma dada por las Ecs. (2.95a) y (2.95b). Las ecuaciones que subyacen a los modelos de error aditivo y multiplicativo están dadas por las Tablas (2.5) y (2.6), respectivamente. En este caso, se utiliza $\beta = \alpha\beta^*$ para simplificar la notación. El error de los modelos multiplicativos es obtenido mediante el reemplazo de ε_t con $\mu_t\varepsilon_t$ en las ecuaciones de la Tabla (2.5). Es sabido que algunas combinaciones de tendencia, estacionalidad y error pueden conducir en ocasiones a dificultades numéricas; de manera más específica, cualquier ecuación de un modelo que requiera división por un componente de estado involucraría una división por cero. Los pronósticos puntuales se obtienen iterando el modelo de la Ec.(2.95) para $t = n+1, n+2, \ldots, n+h$, y estableciendo $\varepsilon_{n+j} = 0$ para $j = 1, \ldots, h$. En la mayoría de los casos, se puede mostrar que el pronóstico puntual es igual a $\mu_{t+h} = E(y_{t+h}|\boldsymbol{x}_t)$, la esperanza condicional del modelo de espacio de estados.

Estos modelos también proveen de medios para la obtención de intervalos de predicción. En el caso de los modelos lineales, donde las distribuciones de predicción son Gaussianas, es posible derivar la varianza condicional $v_{t+h|t} = \mathrm{Var}(y_{t+h}|\boldsymbol{x}_t)$ y obtener de manera acorde intervalos de predicción. Por otro lado, una aproximación más directa es simular múltiples situaciones futuras condicionadas en el último estimador del vector de estado $\boldsymbol{x}_t$. Tradicionalmente, los valores iniciales de $\boldsymbol{x}_0$ son especificados usando valores ad hoc, o mediante heurísticas como la propuesta por [23]. Los estimadores de máxima verosimilitud son obtenidos minimizando

$$\mathcal{L}^* = n\log(\sum_{t=1}^{n}\varepsilon_t^2) + 2\sum_{t=1}^{n}\log|r(\boldsymbol{x}_{t-1})| \tag{2.96}$$

que es equivalente al logaritmo negativo de la función de verosimilitud (con términos constantes eliminados), condicional a los parámetros $\boldsymbol{\theta} = (\alpha, \beta, \gamma, \phi)'$ y los estados iniciales $\boldsymbol{x}_0 = (l_0, b_0, s_0, s_{-1}, \ldots, s_{-m+1})$, donde n es el número de observaciones. Dicho valor puede ser fácilmente calculado utilizando las ecuaciones recursivas de la Tabla (2.4). Además, de manera alternativa, los estimadores pueden ser obtenidos minimizando el MSE a un paso, minimizando la varianza de los residuales σ^2, o por otro criterio de error de pronóstico.

TABLA 2.5: Ecuaciones de espacio de estado para cada modelo de error aditivo

Trend	Seasonal		
	N	A	M
N	$\mu_t = \ell_{t-1}$ $\ell_t = \ell_{t-1} + \alpha\varepsilon_t$	$\mu_t = \ell_{t-1} + s_{t-m}$ $\ell_t = \ell_{t-1} + \alpha\varepsilon_t$ $s_t = s_{t-m} + \gamma\varepsilon_t$	$\mu_t = \ell_{t-1}s_{t-m}$ $\ell_t = \ell_{t-1} + \alpha\varepsilon_t/s_{t-m}$ $s_t = s_{t-m} + \gamma\varepsilon_t/\ell_{t-1}$
A	$\mu_t = \ell_{t-1} + b_{t-1}$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha\varepsilon_t$ $b_t = b_{t-1} + \beta\varepsilon_t$	$\mu_t = \ell_{t-1} + b_{t-1} + s_{t-m}$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha\varepsilon_t$ $b_t = b_{t-1} + \beta\varepsilon_t$ $s_t = s_{t-m} + \gamma\varepsilon_t$	$\mu_t = (\ell_{t-1} + b_{t-1})s_{t-m}$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha\varepsilon_t/s_{t-m}$ $b_t = b_{t-1} + \beta\varepsilon_t/s_{t-m}$ $s_t = s_{t-m} + \gamma\varepsilon_t/(\ell_{t-1} + b_{t-1})$
A_d	$\mu_t = \ell_{t-1} + \phi b_{t-1}$ $\ell_t = \ell_{t-1} + \phi b_{t-1} + \alpha\varepsilon_t$ $b_t = \phi b_{t-1} + \beta\varepsilon_t$	$\mu_t = \ell_{t-1} + \phi b_{t-1} + s_{t-m}$ $\ell_t = \ell_{t-1} + \phi b_{t-1} + \alpha\varepsilon_t$ $b_t = \phi b_{t-1} + \beta\varepsilon_t$ $s_t = s_{t-m} + \gamma\varepsilon_t$	$\mu_t = (\ell_{t-1} + \phi b_{t-1})s_{t-m}$ $\ell_t = \ell_{t-1} + \phi b_{t-1} + \alpha\varepsilon_t/s_{t-m}$ $b_t = \phi b_{t-1} + \beta\varepsilon_t/s_{t-m}$ $s_t = s_{t-m} + \gamma\varepsilon_t/(\ell_{t-1} + \phi b_{t-1})$
M	$\mu_t = \ell_{t-1}b_{t-1}$ $\ell_t = \ell_{t-1}b_{t-1} + \alpha\varepsilon_t$ $b_t = b_{t-1} + \beta\varepsilon_t/\ell_{t-1}$	$\mu_t = \ell_{t-1}b_{t-1} + s_{t-m}$ $\ell_t = \ell_{t-1}b_{t-1} + \alpha\varepsilon_t$ $b_t = b_{t-1} + \beta\varepsilon_t/\ell_{t-1}$ $s_t = s_{t-m} + \gamma\varepsilon_t$	$\mu_t = \ell_{t-1}b_{t-1}s_{t-m}$ $\ell_t = \ell_{t-1}b_{t-1} + \alpha\varepsilon_t/s_{t-m}$ $b_t = b_{t-1} + \beta\varepsilon_t/(s_{t-m}\ell_{t-1})$ $s_t = s_{t-m} + \gamma\varepsilon_t/(\ell_{t-1}b_{t-1})$
M_d	$\mu_t = \ell_{t-1}b_{t-1}^{\phi}$ $\ell_t = \ell_{t-1}b_{t-1}^{\phi} + \alpha\varepsilon_t$ $b_t = b_{t-1}^{\phi} + \beta\varepsilon_t/\ell_{t-1}$	$\mu_t = \ell_{t-1}b_{t-1}^{\phi} + s_{t-m}$ $\ell_t = \ell_{t-1}b_{t-1}^{\phi} + \alpha\varepsilon_t$ $b_t = b_{t-1}^{\phi} + \beta\varepsilon_t/\ell_{t-1}$ $s_t = s_{t-m} + \gamma\varepsilon_t$	$\mu_t = \ell_{t-1}b_{t-1}^{\phi}s_{t-m}$ $\ell_t = \ell_{t-1}b_{t-1}^{\phi} + \alpha\varepsilon_t/s_{t-m}$ $b_t = b_{t-1}^{\phi} + \beta\varepsilon_t/(s_{t-m}\ell_{t-1})$ $s_t = s_{t-m} + \gamma\varepsilon_t/(\ell_{t-1}b_{t-1}^{\phi})$

Fuente: [24]

TABLA 2.6: Ecuaciones de espacio de estado para cada modelo de error multiplicativo

Trend	Seasonal		
	N	A	M
N	$\mu_t = \ell_{t-1}$ $\ell_t = \ell_{t-1}(1+\alpha\varepsilon_t)$	$\mu_t = \ell_{t-1} + s_{t-m}$ $\ell_t = \ell_{t-1} + \alpha(\ell_{t-1} + s_{t-m})\varepsilon_t$ $s_t = s_{t-m} + \gamma(\ell_{t-1} + s_{t-m})\varepsilon_t$	$\mu_t = \ell_{t-1}s_{t-m}$ $\ell_t = \ell_{t-1}(1+\alpha\varepsilon_t)$ $s_t = s_{t-m}(1+\gamma\varepsilon_t)$
A	$\mu_t = \ell_{t-1} + b_{t-1}$ $\ell_t = (\ell_{t-1} + b_{t-1})(1+\alpha\varepsilon_t)$ $b_t = b_{t-1} + \beta(\ell_{t-1} + b_{t-1})\varepsilon_t$	$\mu_t = \ell_{t-1} + b_{t-1} + s_{t-m}$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha(\ell_{t-1} + b_{t-1} + s_{t-m})\varepsilon_t$ $b_t = b_{t-1} + \beta(\ell_{t-1} + b_{t-1} + s_{t-m})\varepsilon_t$ $s_t = s_{t-m} + \gamma(\ell_{t-1} + b_{t-1} + s_{t-m})\varepsilon_t$	$\mu_t = (\ell_{t-1} + b_{t-1})s_{t-m}$ $\ell_t = (\ell_{t-1} + b_{t-1})(1+\alpha\varepsilon_t)$ $b_t = b_{t-1} + \beta(\ell_{t-1} + b_{t-1})\varepsilon_t$ $s_t = s_{t-m}(1+\gamma\varepsilon_t)$
A_d	$\mu_t = \ell_{t-1} + \phi b_{t-1}$ $\ell_t = (\ell_{t-1} + \phi b_{t-1})(1+\alpha\varepsilon_t)$ $b_t = \phi b_{t-1} + \beta(\ell_{t-1} + \phi b_{t-1})\varepsilon_t$	$\mu_t = \ell_{t-1} + \phi b_{t-1} + s_{t-m}$ $\ell_t = \ell_{t-1} + \phi b_{t-1} + \alpha(\ell_{t-1} + \phi b_{t-1} + s_{t-m})\varepsilon_t$ $b_t = \phi b_{t-1} + \beta(\ell_{t-1} + \phi b_{t-1} + s_{t-m})\varepsilon_t$ $s_t = s_{t-m} + \gamma(\ell_{t-1} + \phi b_{t-1} + s_{t-m})\varepsilon_t$	$\mu_t = (\ell_{t-1} + \phi b_{t-1})s_{t-m}$ $\ell_t = (\ell_{t-1} + \phi b_{t-1})(1+\alpha\varepsilon_t)$ $b_t = \phi b_{t-1} + \beta(\ell_{t-1} + \phi b_{t-1})\varepsilon_t$ $s_t = s_{t-m}(1+\gamma\varepsilon_t)$
M	$\mu_t = \ell_{t-1}b_{t-1}$ $\ell_t = \ell_{t-1}b_{t-1}(1+\alpha\varepsilon_t)$ $b_t = b_{t-1}(1+\beta\varepsilon_t)$	$\mu_t = \ell_{t-1}b_{t-1} + s_{t-m}$ $\ell_t = \ell_{t-1}b_{t-1} + \alpha(\ell_{t-1}b_{t-1} + s_{t-m})\varepsilon_t$ $b_t = b_{t-1} + \beta(\ell_{t-1}b_{t-1} + s_{t-m})\varepsilon_t/\ell_{t-1}$ $s_t = s_{t-m} + \gamma(\ell_{t-1}b_{t-1} + s_{t-m})\varepsilon_t$	$\mu_t = \ell_{t-1}b_{t-1}s_{t-m}$ $\ell_t = \ell_{t-1}b_{t-1}(1+\alpha\varepsilon_t)$ $b_t = b_{t-1}(1+\beta\varepsilon_t)$ $s_t = s_{t-m}(1+\gamma\varepsilon_t)$
M_d	$\mu_t = \ell_{t-1}b_{t-1}^{\phi}$ $\ell_t = \ell_{t-1}b_{t-1}^{\phi}(1+\alpha\varepsilon_t)$ $b_t = b_{t-1}^{\phi}(1+\beta\varepsilon_t)$	$\mu_t = \ell_{t-1}b_{t-1}^{\phi} + s_{t-m}$ $\ell_t = \ell_{t-1}b_{t-1}^{\phi} + \alpha(\ell_{t-1}b_{t-1}^{\phi} + s_{t-m})\varepsilon_t$ $b_t = b_{t-1}^{\phi} + \beta(\ell_{t-1}b_{t-1}^{\phi} + s_{t-m})\varepsilon_t/\ell_{t-1}$ $s_t = s_{t-m} + \gamma(\ell_{t-1}b_{t-1}^{\phi} + s_{t-m})\varepsilon_t$	$\mu_t = \ell_{t-1}b_{t-1}^{\phi}s_{t-m}$ $\ell_t = \ell_{t-1}b_{t-1}^{\phi}(1+\alpha\varepsilon_t)$ $b_t = b_{t-1}^{\phi}(1+\beta\varepsilon_t)$ $s_t = s_{t-m}(1+\gamma\varepsilon_t)$

Fuente: [24]

Las medidas de pronóstico revisadas en la sección (2.3.4.2) pueden ser usadas para seleccionar el modelo dado un set de datos, siempre que los errores sean calculados a partir de un set de prueba y no partir de los mismos datos que fueron utilizados en el proceso de estimación del modelo. Sin embargo, usualmente no existe la cantidad de errores fuera de muestra para obtener conclusiones confiables. Por ello, se penaliza la verosimilitud según AIC = $\mathcal{L}^*(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{x}}_0) + 2q$ (véase sección 2.4.3.1), donde q es el número de parámetros en $\boldsymbol{\theta}$ más el número de estados libres en $\boldsymbol{x}_0$, y $\hat{\boldsymbol{\theta}}$ y $\hat{\boldsymbol{x}}_0$ denota los estimadores de $\boldsymbol{\theta}$ y $\boldsymbol{x}_0$. El modelo que minimiza AIC es seleccionado de entre todos los modelos apropiados para los datos, sea aditivo o multiplicativo.

El algoritmo de pronósticos propuesto por involucra los siguientes pasos

1. Para cada serie, aplicar todos los modelos que sean apropiados, optimizando los parámetros del modelo para cada caso
2. Seleccionar los mejores modelos según AIC
3. Producir pronósticos puntuales usando el mejor modelo (con parámetros optimizados) para el horizonte requerido
4. Obtener intervalos de predicción para el mejor modelo vía resultados analíticos o simulación, hallando los percentiles respectivos a cada horizonte.

Patrones estacionales complejos

De livera et al. [31] introduce un esquema de modelamiento de espacios de estado de innovaciones para el pronóstico de series de tiempo con estacionalidades complejas tales como períodos de estacionalidad múltiple, estacionalidad de alta frecuencia, estacionalidad no-entera y efectos de calendario duales. El nuevo marco de trabajo incorpora transformaciones Box-Cox, representaciones de Fourier para los coeficientes que varían con el tiempo y una corrección ARMA para el error (véase sección 2.4.3.7). El acrónimo BATS $(p, q, m_1, m_2, \ldots, m_T)$ es utilizado para identificar las principales características del modelo que incluye transformación Box-Cox(B), errores ARMA (A), componente de tendencia (T)

y estacionalidad (S) dado por

$$y_t^{(\omega)} = \begin{cases} \frac{y_t^{\omega}-1}{\omega}; & \omega \neq 0 \\ log\ y_t; & \omega = 0 \end{cases} \tag{2.97a}$$

$$y_t^{(\omega)} = l_{t-1} + \phi b_{t-1} + \sum_{i=1}^{T} s_{t-m_i}^{(i)} + d_t \tag{2.97b}$$

$$l_t = l_{t-1} + \phi b_{t-1} + \alpha d_t \tag{2.97c}$$

$$b_t = \phi b_{t-1} + \beta d_t \tag{2.97d}$$

$$s_t^{(i)} = s_{t-m_i}^{(i)} + \gamma_i d_t \tag{2.97e}$$

$$d_t = \sum_{i=1}^{p} \rho_i d_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t \tag{2.97f}$$

donde $m_1, \ldots, m_T$ denotan los períodos estacionales, l_t y b_t representan los componentes de nivel y tendencia de la serie en el período t, respectivamente. $s_t^{(i)}$ represente el componente estacional i-ésimo en el período t, d_t denota un proceso ARMA(p, q), y ε_t es un proceso de ruido blanco Gaussiano con media cero y varianza constante σ^2. Los parámetros de suavizamiento están dados por α, β, γ_i para $i = 1, \ldots, T$, y ϕ es el parámetro de amortiguamiento. De acá, el modelo de Holt-Winters con Estacionalidad Doble (DSHW) con $\phi = 1$, $\omega = 1$ y el ajuste de residual AR(1) considerado en [29], está dado por el modelo BATS (1, 0, m_1, m_2).

Una reparametrización de los componentes estacionales basada en series de Fourier también es propuesta por [31] según:

$$s_t^{(i)} = \sum_{j=1}^{k_i} s_{j,t}^{(i)} \tag{2.98a}$$

$$s_{j,t}^{(i)} = s_{j,t-1}^{(i)} \cos\lambda_j^{(i)} + s_{j,t-1}^{*(i)} \sin\lambda_j^{(i)} + \gamma_1^{(i)} d_t \tag{2.98b}$$

$$s_{j,t}^{*(i)} = -s_{j,t-1}^{(i)} \sin\lambda_j^{(i)} + s_{j,t-1}^{*(i)} \cos\lambda_j^{(i)} + \gamma_2^{(i)} d_t \tag{2.98c}$$

y es llamada modelo *BATS trigonométrico* (TBATS). La ventaja principal del modelo TBATS es que se permite que la estacionalidad fluctúe levemente en el tiempo.

2.4.4. Descomposición estacional por regresión local polinomial (STL)

La descomposición estacional por regresión local polinomial (STL) es un procedimiento de filtrado para descomponer una serie de tiempo en tendencia, estacionalidad y componentes restantes, usando una regresión local ponderada (LOESS) como método para estimar relaciones no-lineales [32]. Asumiendo una descomposición aditiva, la serie de tiempo y_t puede ser descompuesta como $y_t = S_t + T_t + R_t$, donde S_t es el componente estacional, T_t es el componente de ciclo-tendencia, y R_t es el componente restante. Transformaciones multiplicativas pueden ser obtenidas aplicando previamente transformaciones Box-Cox, de ser necesario.

Para propósitos de pronóstico, la serie de tiempo descompuesta puede ser escrita según $y_t = \hat{S}_t + \hat{A}_t$, donde $\hat{A}_t = \hat{T}_t + \hat{R}_t$ es el componente de estacionalidad ajustado [13]. Entonces, las dos series mencionadas son pronosticadas de manera separadas y luego agregadas para construir el pronóstico final. Los pronósticos para los componentes estacionales son producidos usando el método estacional ingenuo (véase sección 2.3.1). Por otro lado, para pronosticar el componente ajustado estacionalmente se consideran dos modelos no estacionales: suavizamiento exponencial de Holt-Winters y ARIMA.

La característica más atractiva de STL, con respecto a otros procedimientos de descomposición radica en su resiliencia a observaciones atípicas en la data, resultando en componentes de sub-serie robustos [33]. La implementación del procedimiento STL está basada en métodos numéricos y no requiere de modelamiento matemático. EL procedimiento se lleva a cabo en un ciclo iterativo de eliminación de tendencia y actualización de componentes estacionales a partir de las sub-series. En cada iteración, los pesos de robustez se forman basados en la estimación del componente irregular, que luego es usado para ponderar observaciones atípicas a través de los cálculos realizados. El ciclo iterativo está formado por dos procedimientos recursivos, una iteración interna que aplica un suavizamiento estacional que actualiza el componente estacional, y luego un suavizamiento que actualiza la componente de tendencia.

2.5. Métodos de Inteligencia Artificial para el pronóstico

2.5.1. Redes neuronales artificiales (ANN)

La evidencia empírica sugiere que las Redes Neuronales Artificiales (ANN) son una herramienta alternativa atractiva para investigadores y practicantes del área del pronóstico; siendo una de las técnicas no-paramétricas que se desempeñan aceptablemente bien en el uso habitual. Caracterizadas como aproximadores universales basados en datos de cualquier función lineal o no-lineal, las ANN se construyen como una alternativa a los métodos de pronóstico estadísticos, así como también para propósitos comparativos. La ventaja principal de las ANN radica en su habilidad para aprender el proceso generador de datos sin requerir supuestos de su comportamiento y forma funcional. Así, las relaciones funcionales que subyacen al proceso son aprendidas y, posteriormente, es posible obtener pronósticos plausibles cuando nuevos datos de entrada están disponibles. Las implementaciones halladas en la literatura usualmente difieren en aspectos como *pre-procesamiento* de los datos, *arquitectura* de la red y procesos de *implementación* y *validación*.

En el contexto del pronóstico de series de tiempo, las ANN son usadas como funciones de aproximación no-lineal debido a su capacidad para mapear el espacio de entrada (variables exógenas y un conjunto de índices de enteros positivos no necesariamente secuenciales que representan valores presentes y rezagados) a un espacio de salida (pronósticos para valores futuros). En cuanto al pronóstico de series de tiempo, es posible limitar el análisis a la estructura de red neuronal *prealimentada* (*feed-forward*) utilizada convencionalmente conocida como *perceptrón multicapa* (MLP), y en particular, a la siguiente forma funcional

$$\hat{y}_t = f(\boldsymbol{x}_t, \boldsymbol{\theta}) = \beta_0 + \sum_{j=1}^{J} \beta_j g\Big(w_{0j} + \sum_{i=1}^{I} w_{ij} x_i\Big) \tag{2.99}$$

donde $\hat{y}_t$ es el pronóstico a un paso calculado utilizando como vectores de entrada $\boldsymbol{x}_t$, para observaciones presentes y rezagadas de las serie de tiempo como ilustra la Fig. (2.3), que también podría incluir variables exógenas. I denota el número de neuronas de entrada x_i de una ANN que forman la *capa de entrada*, y J es el número de unidades de procesamiento o *neuronas* que forman la *capa oculta*. Los valores de entrada son presentados a la ANN como

un conjunto de vectores de entrada aleatorios compuestos de una *ventana móvil* de longitud fija I a través de la serie. Las neuronas transforman las entradas por medio de coeficientes $\boldsymbol{\theta} = (w_{ij}, \beta_j)$, que corresponden a los pesos de la red para las capas ocultas y la capa de salida, respectivamente. Cada capa tiene su propio término de sesgo, que siempre tiene un valor igual a 1. $g(\cdot)$ es una función de transferencia no lineal llamada *función de activación*, que usualmente es acotada, no decreciente y derivable. Usualmente, las funciones de activación son funciones lineales, sigmoides o tangente hiperbólica. Finalmente, dado que el pronóstico de series de tiempo corresponde a un problema de regresión, se utiliza una función lineal en la capa de salida. La salida de la red es comparada con el valor observado para definir un criterio de error a ser minimizado. Las derivadas del error con respecto a los pesos son evaluadas usando el algoritmo de *propagación hacia atrás* (*back-propagation*) [34]. Dicho algoritmo permite que los valores de los pesos w_{ij} y β_j a ser hallados, minimicen algún criterio de ajuste (ej. MSE) a través de todas las N instancias en el set de entrenamiento como en la Ec. (2.100)

$$\min_{\boldsymbol{\theta}} \Big(\sum_{t=1}^{N} (y_t - f(\boldsymbol{x}_t, \boldsymbol{\theta}))^2 + \lambda \sum_{ij} \theta_{ij}^2 \Big), \tag{2.100}$$

donde y_t es el valor observado y f corresponde al modelo ANN.

2.5.1.1. Algoritmo de propagación hacia atrás

Considérese una versión generalizada del problema de minimización planteado en la Ec. (2.100) dada por

$$E(\boldsymbol{w}) = \sum_{p} \|\boldsymbol{t}^p - f(\boldsymbol{x}^p; \boldsymbol{w})\|^2, \tag{2.101}$$

donde $(\boldsymbol{x}^p, \boldsymbol{t}^p)$ corresponde a las observaciones e $y = f(\boldsymbol{x}, \boldsymbol{w})$ es la salida de la red. Notar que $E(\boldsymbol{w})$ es una función diferenciable solo para unidades diferenciables. El grupo de Rumelhart-McClelland [34] propuso una forma de descenso paso a paso para reducir la Ec. (2.101), con la siguiente regla de actualización

$$w_{ij} \leftarrow w_{ij} - \eta \frac{\partial E}{\partial w_{ij}} \tag{2.102}$$

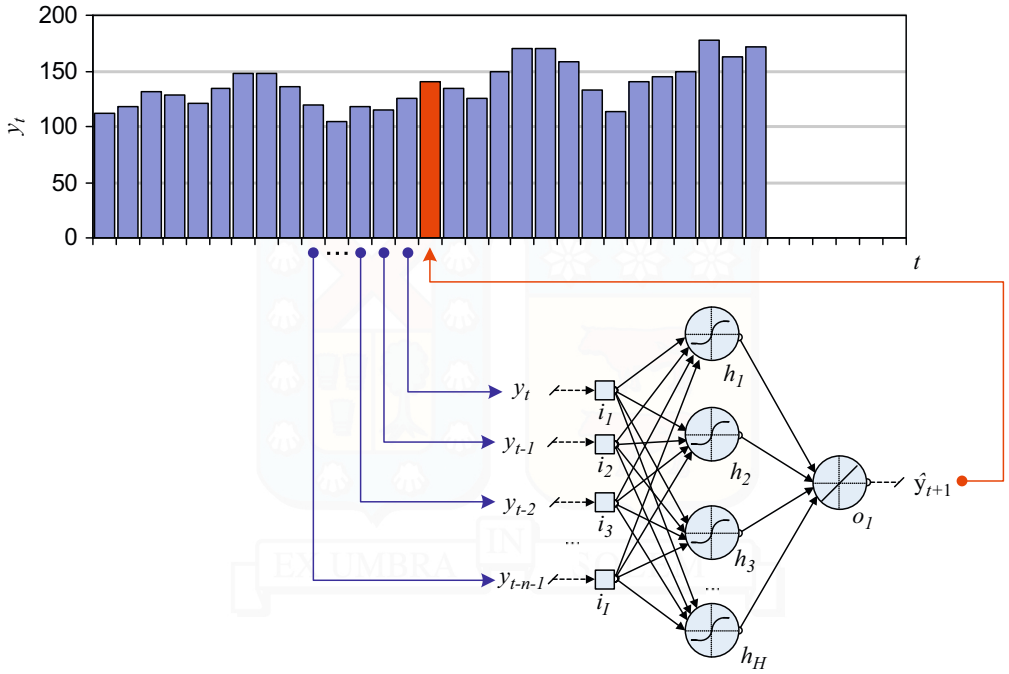


Figura 2.3: MLP autoregresivo para pronóstico

Fuente: [35]

y por ello, la derivada parcial puede ser escrita en la forma

$$\frac{\partial E}{\partial w_{ij}} = \sum_p y_i^p \delta_j^p \tag{2.103}$$

que se conoce como la *regla delta generalizada*. Aquí, el superíndice p se refiere a los cálculos relacionados con el ejemplo p. Más aún, como los valores de δ pueden ser calculados desde la salida hacia la entrada de la red, así el proceso de calcular las derivadas y el algoritmo descendiente son conocidos como *propagación hacia atrás*.

Recordando que cada unidad tiene entrada $x_j = \sum_{i \to j} w_{ij} y_i$ y salida $y_j = f_j(x_j)$. Cada forma del criterio de ajuste E se suma a través de las observaciones, calculándose las derivadas de E^p, que pueden ser sumadas a través de ellas. Por lo anterior, es posible prescindir del superíndice p y tomar calcular las derivadas parciales para E con respecto a los pesos w_{ij} y con respecto a las entradas x_i y salidas y_i de las unidades. Es importante saber qué se mantiene fijo y qué no, en lo que es considerado derivadas *ordenadas*. Cuando se obtiene las derivadas parciales con respecto a los pesos, la función E es considerada

como una función de todos los pesos, por lo que cambios en un peso w_{ij} afectan la entrada y la salida de la unidad j y todas las unidades conectadas a j, incluyendo algunas unidades de salida. Cuando se calculan las derivadas parciales con respecto a la entrada o salida de la red, se permite a todas las otras señales en la red que dependen de la entrada o la salida a seguir su dependencia usual. Así todos los pesos, entradas y salidas, en otras unidades en la misma u otra capa anterior, son mantenidos constantes. Se evalúa $\partial E/\partial x_j$ notando que x_j sólo afecta a las salidas a través de y_j, y esto solo actúa a través de las conexiones con las unidades de salida.

Para las primeras derivadas se tiene

$$\frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial x_j}\frac{\partial x_j}{\partial w_{ij}} = y_i \frac{\partial E}{\partial x_j} = y_i f_j'(x_j)\frac{\partial E}{\partial y_j} = y_i \delta_j \tag{2.104}$$

si se considera $\delta_j = \partial E/\partial x_j$. La primera igualdad proviene de la dependencia de E con los pesos sólo a través de las salidas; la segunda a partir de $x_j = \sum w_{ij} y_i$. Notar que

$$\delta = \frac{\partial E}{\partial x_j} = \frac{\partial E}{\partial y_j} = f_j'(x_j)\frac{\partial E}{\partial y_j} \tag{2.105}$$

Para las unidades de salida $\partial E/\partial y_j$ puede ser calculado directamente a partir de E.

Para unidades de capas anteriores se tiene que

$$\delta_j = f_j'(x_j)\frac{\partial E}{\partial y_j} = f_j'(x_j)\sum_{k:j\to k} w_{jk}\frac{\partial E}{\partial x_k} = f_j'(x_j)\sum_{k:j\to k}\frac{\partial E}{\partial x_k}\frac{\partial x_k}{\partial y_j} \tag{2.106}$$

$$= f_j'(x_j)\sum_{k:j\to k} w_{jk}\delta_k, \tag{2.107}$$

la suma siendo sobre unidades k alimentadas por la unidad j. Esta fórmula ha sido redescubierta en varias ocasiones. Usualmente al proceso de calcular las salidas a partir de las entradas se le conoce como *propagación hacia adelante*, seguido de la antes mencionada *propagación hacia atrás* para calcular δ_i y por lo tanto $\partial E/\partial_{ij}$.

Es posible notar que en la Ec. (2.100), el descenso se aplica a

$$E + \lambda \sum_{ij} w_{ij}^2 = E + \lambda C \tag{2.108}$$

que corresponde de cierta forma a una regularización que reduce la magnitud de los pesos a cada paso con cierto *decaimiento*.

2.5.2. Máquinas de aprendizaje extremo (ELM)

Las máquinas de aprendizaje extremo (ELM) son un algoritmo de aprendizaje relativamente nuevo para entrenar redes neuronales de una capa oculta [36]. Los pesos de entrada y sesgos de capas ocultas son asignados de manera aleatoria en lugar de ser ajustados de manera exhaustiva. Luego, los pesos de salida son calculados a través de una operación inversa en la matriz de salida de la capa oculta. Es así como el aprendizaje neuronal se convierte en un problema de mínimos cuadrados que puede ser resuelto reduciendo el uso de recursos computaciomales. Las ELM se caracterizan por un buen desempeño en cuanto a capacidad de generalización. La gran velocidad de entrenamiento gracias a su mecanismo de aprendizaje libre de iteraciones permite evitar problemas como hallazgo de mínimos locales, criterios de parada y la determinación de tasas de aprendizaje o decaimientos. De manera similar a otros modelos de redes neuronales, las ELM pueden contener cientos de neuronas ocultas y pueden sufrir de sobreajuste. Para aliviar problemas de arquitectura e inicialización, estos modelos usualmente son ensamblados para incrementar la precisión y robustez de sus resultados [37].

Dado un set de entrenamiento con N observcaciones $(\boldsymbol{x}_t, \boldsymbol{y}_t)$, la red neuronal de una sola capa oculta de la Ec. (2.99) puede ser escrita utilizando el producto interno $\boldsymbol{w}_j \cdot \boldsymbol{x}_j$ de la siguiente manera

$$\sum_{j=1}^{J} \beta_j g\left(\boldsymbol{w}_j \cdot \boldsymbol{x}_j + b_j\right) = \boldsymbol{o}_t, \qquad t = 1, \ldots, N \tag{2.109}$$

donde $\boldsymbol{x}_i$ identifica el vector de entrada, $\boldsymbol{y}_t$ es la salida deseada, y $\boldsymbol{o}_t$ es el resultado observado. Si el error de entrenamiento es cero $\sum_{t=1}^{N} \|\boldsymbol{o}_t - \boldsymbol{y}_t\| = 0$, entonces hay pesos de entrada $\boldsymbol{w}_j$, sesgo b_j, y pesos de salida β_j, tales que

$$\sum_{j=1}^{J} \beta_j g\left(\boldsymbol{w}_j \cdot \boldsymbol{x}_j + b_j\right) = \boldsymbol{y}_t, \qquad t = 1, \ldots, N, \tag{2.110}$$

la Ec. (2.110) puede ser reescrita como $\boldsymbol{H\beta} = \boldsymbol{y}$, donde $\boldsymbol{H}$ es la matriz de salida de la capa oculta. Dado que $\boldsymbol{H}$ es una matriz no-cuadrada, ELMs no pueden aproximar el error de entrenamiento cero. Dado que el número de nodos ocultos es usualmente menos que la cantidad de observaciones de entrenamiento, la red se torna un sistema lineal indeterminado y los pesos de salida puede ser determinados mediante el método de mínimos cuadrados. La solución $\boldsymbol{\beta}^* = \boldsymbol{H}^\dagger \boldsymbol{y}$ es dada por las ELM, donde $\boldsymbol{H}^\dagger$ es la matriz inversa generalizada de Moore-Penrose $\boldsymbol{H}$.

2.5.3. Máquinas de vectores de soporte (SVM)

Las máquinas de vectores de soporte para modelos de regresión (SVR) se han transformado en una aproximación poderosa para problemas de predicción. Un mapeo no-lineal es definido para enlazar los datos de entrada (set de entrenamiento) a un espacio dimensional más grande. Teóricamente, en ese nuevo espacio de altas dimensiones, existe una función lineal que permite formular una relación no-lineal entre lo datos de entrada y los datos de salida [38]. Dado un set de entrenamiento donde el vector de entrada $\boldsymbol{x}_i$ está asociado al vector de salida y_i, SVR resuelve el siguiente problema de optimización:

$$
\begin{aligned}
\min_{\boldsymbol{w},b,\xi,\xi^*} \quad & \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^{l}(\xi_i + \xi_i^*) \\
\text{subject to} \quad & (\boldsymbol{w}^T\phi(\boldsymbol{x}_i) + b) - y_i \leq \epsilon + \xi_i, \\
& y_i - (\boldsymbol{w}^T\phi(\boldsymbol{x}_i) + b) \leq \epsilon + \xi_i^*, \\
& \xi_i, \quad \xi_i^* \geq 0, \quad i = 1, \ldots, l.
\end{aligned}
$$

donde ϕ mapea los valores de entrada a un espacio dimensional mayor, ξ_i y ξ_i^* son variables de holgura positivas que representan la distancia entre el valor observado y los valores de acotamiento correspondientes al tubo insensitivo ϵ como en la Fig. (2.4). $C > 0$ es el parámetro de costo que establece un intercambio entre capacidad de generalización y error de entrenamiento [39]. Un problema de programación cuadrática restringido de manera lineal que tiene una solución única y globalmente óptima [40] puede ser resuelto mediante

la obtención de las soluciones del siguiente problema dual

$$\begin{aligned} \min_{\boldsymbol{\alpha},\boldsymbol{\alpha}^*} \quad & \frac{1}{2}(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*)^T\boldsymbol{K}(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*) \\ & + \epsilon\sum_{i=1}^{l}(\alpha_i+\alpha_i^*) + \sum_{i=1}^{l} z_i(\alpha_i-\alpha_i^*) \\ \text{subject to} \quad & \sum_{i=1}^{l}(\alpha_i-\alpha_i^*) = 0, \\ & 0 \le \alpha_i, \quad \alpha_i^* \le C, \quad i = 1,\ldots,l \end{aligned}$$

donde $\boldsymbol{K}(\boldsymbol{x}_i,\boldsymbol{x}_j) = \phi(\boldsymbol{x}_i)^T\phi(\boldsymbol{x}_j)$ se llama la *función de kernel* y corresponde a un producto interno que tiene muchos elementos y puede ser costosa de calcular. Sin embargo, el producto interno en un espacio de dimensiones mayores puede ser calculado de manera eficiente mediante la aplicación del *truco del kernel*. El desempeño de un modelo SVR depende de la elección de una función de kernel que se ajuste al objetivo a aprender, pues es sabido que el rendimiento de SVR depende de la elección de funciones de kernel así como también de los hiperparámetros asociados al modelo. En general, las funciones de kernel más utilizadas son la función lineal, polinomial, tangente hiperbólica y gaussiana.

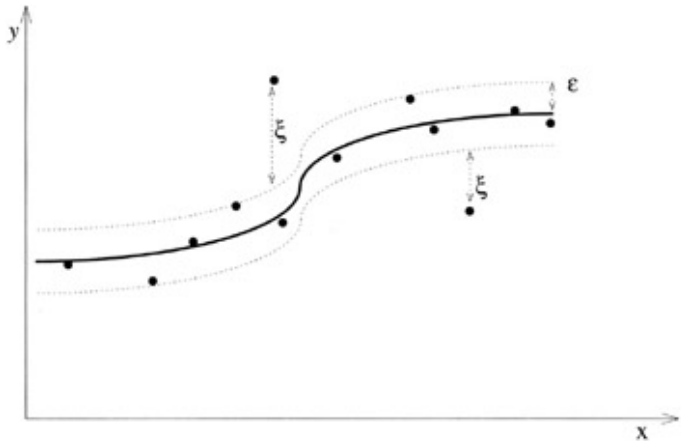


Figura 2.4: Banda insensitiva para regresión no-lineal mediante SVM

Fuente: [41]

2.5.4. Selección de variables

La *ingeniería de características* intenta aumentar la eficacia predictiva de los algoritmos de aprendizaje creando características de los datos sin procesar que facilitan dicho proceso. En este contexto, una *característica* corresponde a aquello que pueda ayudar a resolver un problema.

La *selección de variables* (*feature selection*) corresponde al proceso de seleccionar un subconjunto de variables de entrada relevantes que sean informativos y suficientes para la buena predicción en su posterior uso, la *construcción de modelos*. En general, la selección de variables contempla la construcción, evaluación y transformación de características permitiendo logran un entrenamiento acelerado e incrementando la eficiencia computacional gracias a la reducción de la dimensionalidad de los datos y a un mejor entendimiento del proceso que subyace a los datos generados [35, 42]. La *construcción de variables* corresponde a la creación de nuevas características a partir de las variables de entrada, usualmente haciendo uso del *dominio del problema* (*domain knowledge*). La *evaluación de variables* (*feature evaluation*) se lleva a cabo analizando las propiedades de los datos y reduciendo el espacio de búsqueda de variables a incorporar en los modelos. Finalmente, la *transformación de variables* ayuda a adecuar el preprocesamiento de los datos para facilitar un mejor modelamiento posterior, por ejemplo, vía estandarización de las variables (*coeficiente z*, *escalamiento lineal* o *min-máx*), o la eliminación de componentes de tendencia y estacionalidad en el caso de las series de tiempo vía diferenciación (véase sección 2.4.2.1).

2.5.4.1. Correlación no-lineal (criterio de Información Mutua)

El análisis de las funciones ACF y PACF es comúnmente utilizado para seleccionar variables rezagadas a incorporar en modelos de predicción de serie de tiempo, por ejemplo, con modelos ARIMA y también para implementaciones de modelos no lineales como los de inteligencia artificial. Sin embargo, el análisis de ACF presenta ciertas dificultades cuando es aplicado a series de tiempo de alta frecuencia, dado que los intervalos de confianza serán más angostos debido a su relación con el tamaño muestral, lo que significa que casi todos los rezagos serán significativos. Más aún, el estudio de las relaciones presentes en la serie

se lleva a cabo de manera lineal, lo que limita la identificación de variables para modelos no lineales (ej. modelos IA). Debido a lo anterior, el *criterio de Información Mutua* (MI) es usado como una medida de información teórica para la independencia de dos variables.

Las variables independientes poseen un valor de cero MI, mientras que las variables dependientes tendrán un valor positivo. El criterio MI es adecuado para llevar a cabo tareas de selección de variables ya que captura correlaciones lineales y no-lineales entre las variables rezagadas y las de salida [42, 43].

Sean X e Y dos variables aleatorias continuas con función de densidad de probabilidad conjunta $p(x, y)$ y funciones de densidad marginal $u(x)$ y $v(y)$. El coeficiente MI para X e Y se define como

$$I(X, Y) = \int \int p(x, y) ln \frac{p(x, y)}{u(x)v(y)} dxdy. \tag{2.111}$$

La manera más directa y ampliamente utilizada para estimar el coeficiente MI consisten en particionar los soportes de X e Y en intervalos de tamaño finito para aproximarlo a través de una suma finita [44]. El coeficiente MI toma valores entre 0 y ∞, pero puede ser normalizado considerando $\rho(X, Y) = \sqrt{1 - e^{-2I(X,Y)}}$ como trasformación invertible.

2.5.5. Optimización de hiper-parámetros

Un método por defecto para optimizar parámetros de ajuste en la etapa de entrenamiento es llevar a cabo una *búsqueda exhaustiva* (*grid search*). Esta aproximación es usualmente efectiva, aunque cuando existen muchos parámetros puede ser ineficiente. Una alternativa es usar una combinación de búsqueda exhaustiva con *carreras*. Otra alternativa es la *selección aleatoria* de combinaciones de parámetros de ajuste para cubrir el espacio de parámetros en menor medida, como se muestra en la Fig. (2.5).

Existe una gran cantidad de modelos donde los procesos anteriormente descritos permitan hallar valores razonables para los hiper-parámetros en relativamente poco tiempo. Sin embargo, existen algunos modelos donde la eficiencia en un espacio de búsqueda pequeño pueden cancelar otras optimizaciones. Por ejemplo, algunos modelos pueden usar *submodelos* donde M combinaciones de parámetros de ajuste son evaluadas, potencialmente menos de M modelos ajustados serán requeridos. Esta aproximación es considerada mejor

cuando una búsqueda exhaustiva es usada.

(a) Búsqueda exhaustiva **(b)** Búsqueda aleatoria

Figura 2.5: Optimización de hiper-parámetros

Fuente: [45]

2.6. Demanda de energía eléctrica

Durante las últimas décadas, varios países han decidido optar por las vías de la liberalización de mercados. Pese a las diferencias entre ellos, la motivación para la liberación de los sectores de energía eléctrica a nivel mundial mezcla un común ideológico y razones políticas. En particular, existe la creencia de que el éxito de la liberación del mercado en otras industrias puede ser duplicado en el sector energético y la "necesidad" de separar o desagregar estructuras de monopolios integrados verticalmente que tradicionalmente han administrado la generación, el transporte y la distribución de la energía eléctrica. La competencia ha sido justificada por los beneficios percibidos de introducir fuerzas de mercado en una industria previamente vista como un monopolio natural con grandes economías verticales. El distanciamiento con el carácter de monopolio natural ha sido posible, a su vez, debido a cambios en las tecnologías de generación y mejoras en la transmisión. Así, la motivación detrás de la liberalización de la electricidad es su forma final, promover el aumento de la eficiencia, estimular la innovación técnica y conducir a una inversión eficiente [46].

La liberación de los mercados de potencia fue liderada por Chile. La reforma que comenzó en 1982, basada en la idea de separar las compañías de generación y distribución donde la energía eléctrica era pagada de acorde a una fórmula basada en el costo, un sistema de despacho con costos marginales de tarifado y un sistema de intercambio energético para cumplir con contratos con clientes. La privatización a gran escala comenzó en 1986 y condujo a la desintegración vertical parcial del sector y la formación de un mecanismo de intercambio energético masivo. Las reformas chilenas fueron seguidas por países como Inglaterra, Escocia, Noruega, Suecia, Finlandia, Dinamarca, Australia, Estados Unidos y Canadá. En general, el número de mercados de electricidad liberalizados está creciendo constantemente en todo el mundo, pero la tendencia es más visible en Europa. Algunos de estos mercados han operado de manera satisfactoria durante décadas, sin embargo otros han tenido que someterse a varios cambios para mejorar su rendimiento. Los beneficios usualmente incluyen una tendencia clara en los precios de la electricidad y un uso eficiente de los activos en el sector eléctrico. Sin embargo, pese a que los precios netos de electricidad

en general han disminuido, los nuevos impuestos han sido aplicados a los precios han revertido estos efectos en varios casos. En particular, la tendencia a la baja de los precios no es aparente - de existir -, para pequeños y medianos clientes industriales y especialmente para el consumidor doméstico.

Otra controversia corresponde a la capacidad de los mercados de energía liberalizados para proporcionar incentivos suficientes para la inversión en capacidad de nueva generación (o transmisión). En el nuevo entorno, las decisiones de inversión ya no son planificadas de manera centralizada, sino que son el resultado de fuerzas competitivas. En consecuencia, generalmente el uso de tecnologías intensivas en capital con largos tiempos de construcción es evitado, incluso si sus costos marginales son bajos. En su lugar, se prefieren las plantas de generación que se pueden construir a corto plazo (como las plantas alimentadas por gas). Pero incluso entonces, la expectativa de precios más bajos puede hacer que los inversionistas privados pospongan los gastos en la capacidad de nueva generación o la ampliación de la red de transmisión. Esto pone a los responsables políticos bajo presión para intervenir. En consecuencia, hay un debate en curso sobre si establecer *pagos de capacidad* (como en algunos países de América Latina y España), organizar mercados de capacidad (como en el noreste de Estados Unidos) o tener mercados de "solo energía"(como en Australia y Nueva Zelanda).

La idea básica de los pagos de capacidad (introducida originalmente en Chile en 1982) es otorgar a cada generador un pago diario que es una medida de la contribución del generador a la confiabilidad del sistema de energía, es decir, su disponibilidad. La evidencia internacional sugiere, sin embargo, que los pagos de capacidad crean incentivos deficientes para aliviar el problema de la capacidad e incluso pueden empeorarlo. Por ejemplo, los generadores pueden intentar aumentar los pagos de capacidad al hacer disponibles menos recursos de capacidad, lo que aumenta, en lugar de disminuir, la probabilidad de escasez.

Los sistemas de *pago por capacidad basados en la cantidad* (a diferencia de los pagos por capacidad basados en el precio discutidos anteriormente) generalmente han tomado la forma de *mercados de capacidad instalada* (ICAP). El objetivo principal de la introducción de estos mercados ha sido garantizar que se asegure que la capacidad adecuada sea entregada diaria o estacionalmente para cumplir con los requisitos de carga y reserva

del sistema. Los distribuidores que venden electricidad a los consumidores finales deben cumplir con sus obligaciones de capacidad, que equivalen a sus cargas mensuales máximas esperadas más un margen de reserva. Pueden lograr esto, ya sea mediante transacciones internas o bilaterales, o mediante el mercado de capacidad en el que los generadores venden un derecho de retiro que permite al operador del sistema recuperarlos en caso de escasez. A medida que los mercados maduraban, los coordinadores del mercado se dieron cuenta de la necesidad de fomentar la confiabilidad de generación y eliminar una fuente potencial de poder de mercado. En consecuencia, se desarrollaron *créditos de capacidad no forzada* (UCAP), que se calculan tomando el ICAP y ajustándolo sobre la base de la confiabilidad del generador.

En los mercados de solamente "de energía", el precio mayorista de la electricidad proporciona una compensación por los costos fijos y variables. El "precio" a pagar por esto son los *peaks* de precios, es decir, cambios abruptos y generalmente no anticipados en el *precio spot* que en casos extremos pueden llevar a quiebras de compañías de energía que no están preparadas para asumir tales riesgos. Los *peaks* de precios deberían enviar señales a los inversores de que se necesita capacidad de nueva generación. Sin embargo, si las alzas son raras y no muy extremas, pueden no proporcionar suficiente motivación. En tal caso, pueden ser necesarios incentivos regulatorios (por ejemplo, pagos de capacidad) para impulsar inversiones oportunas y adecuadas. Un problema social relacionado es si los consumidores están dispuestos a aceptar *peaks* de precios. De lo contrario, se necesitan límites de precios protectores, que nuevamente requieren incentivos regulatorios para la inversión en nueva capacidad.

2.6.1. Pronóstico de demanda eléctrica

El pronóstico de demanda eléctrica ha incrementado su importancia desde el desarrollo de los mercados competitivos eléctricos. Los costos de sobre o sub-contratar y luego vender o comprar energía en el mercado de balance a tiempo real ha incrementado tanto que pueden conducir a grandes pérdidas financieras. La minimización del volumen de riesgo, especialmente a corto plazo, jamás ha tenido tanta importancia como lo tiene para las compañías energéticas como en hoy en día. Los métodos revisados en las secciones (2.4)

y (2.5) constituyen un conjunto enriquecido de herramientas que pueden ser aplicadas al pronóstico a corto plazo. Éstos difieren en complejidad y rendimiento de pronóstico, pero todos sirven al mismo propósito. Desafortunadamente, no existen un único mejor modelo. Cada proceso de demanda eléctrica debe ser abordado de manera individual y la aproximación óptima puede ser seleccionada sólo después de un estudio comparativo del comportamiento del modelo. Las técnicas de preprocesamiento pueden ser útiles en el proceso de preselección de modelos e identificación de parámetros, sobre todo cuando la disponibilidad de datos de entrada y su calidad puede limitar no sólo el rango de modelos a considerar, sino también el rendimiento del pronóstico.

2.6.2. Caracterización de la curva de demanda eléctrica

Antes de llevar a cabo el proceso de modelamiento y predicción, es importante mencionar ciertos asuntos con respecto al pronóstico de demanda eléctrica. Se debe tener en cuenta que la precisión de pronóstico no sólo depende de la eficiencia numérica del algoritmo empleado, sino también de la calidad de los datos analizados y la habilidad de incorporar importantes factores exógenos en los modelos. Para el pronóstico a corto plazo, un gran número de variables pueden ser consideradas, tales como factores temporales, datos de clima, precios de la electricidad, eventos sociales e incluso segmentaciones por tipo de cliente.

2.6.2.1. Observaciones vacías y atípicas

Si los datos de entrada del modelo de pronóstico son deficientes, será una tarea difícil o imposible obtener un buen pronóstico, sin importar qué tan bueno es un modelo. Los datos obtenidos por ejemplo, minuto a minuto, usualmente son irregulares y están llenos de observaciones faltantes (NA). Un problema relacionado es la manipulación de condiciones de demanda observadas pero anómalas. Si el comportamiento de la demanda es anormal en cierto día, esta desviación de las condiciones normales puede ser reflejada en los pronósticos futuros. Una posible solución para este problema es tratar las observaciones anormales como observaciones atípicas y usar procesos de filtrado correctivo para preprocesar los datos y producir observaciones de calidad que puedan servir como argumento de entrada para los

modelos de pronóstico. Desafortunadamente, los algoritmos correctivos automatizados a veces no funcionan satisfactoriamente y conocimiento de humanos expertos es requerido para supervisar el proceso.

2.6.2.2. Factores temporales

Los factores temporales o *de calendario* que influencian a los sistemas de carga incluyen la época del año, el día de la semana y la hora del día. Además existen diferencias en los perfiles de demanda entre estaciones y entre días de semana y días de fin de semana. La demanda en diferentes días de semana también puede comportarse de manera diferente: los días lunes y viernes pueden tener diferentes estructuras que los días entre ellos. Finalmente, los perfiles de carga durante festivos y sus días adyacentes pueden desviarse del comportamiento típico. Los días festivos también son más difíciles de pronosticar debido a la poca frecuencia de sus ocurrencias.

2.6.2.3. Condiciones climáticas

Además de los factores temporales, las condiciones climáticas son las variables exógenas más influyentes, especialmente para el pronóstico a corto plazo. Ciertas variables climáticas pueden ser consideradas, pero la temperatura, humedad y nubosidad son los predictores más utilizados. El enfoque habitual de STLF utiliza el escenario meteorológico pronosticado como entrada. Sin embargo, uno de los desarrollos recientes en el pronóstico climático es el llamado *enfoque ensamblado*. Dicho enfoque, consiste en calcular múltiples pronósticos con ponderaciones de probabilidad asignadas. En lugar de utilizar pronósticos puntuales, utiliza múltiples escenarios para el valor futuro de una variable meteorológica. A su vez, estas entradas generan múltiples pronósticos de carga, que naturalmente contienen mucha más información que solo la carga esperada. Además de predicciones horarias más precisas, la descripción probabilística de la demanda futura también se puede utilizar como entrada para los sistemas de apoyo al a toma de decisiones. Desafortunadamente, la mayoría de los servicios meteorológicos no proporcionan descripciones probabilísticas de las variables meteorológicas, sino solo pronósticos puntuales.

Capítulo 3

Caso de estudio: Pronóstico de demanda eléctrica a corto plazo en Francia

3.1. RTE

Réseau de Transport d'Électricité (RTE) es el operador de sistemas de transmisión de electricidad de Francia, responsable de la operación, mantención y desarrollo del sistema de transmisión de alto voltaje francés, siendo el más grande de Europa. Su misión es proveer de acceso a una fuente de electricidad económica, segura y limpia a todos sus consumidores. Asimismo, en aspectos operativos, RTE se asegura de que a cada momento exista un balance entre la oferta y la demanda de electricidad en Francia. Posee más de 105.000 km de líneas entre 63.000 y 400.000 voltios y 500 líneas de redes fronterizas que conectan a Francia con redes de 33 países europeos, ofreciendo así oportunidades de intercambio esenciales para la optimización del sistema eléctrico a nivel de intercambio económico.

3.2. Análisis de datos exploratorio

Antes de modelar la demanda eléctrica en el dominio temporal, es importante entender su comportamiento en el tiempo. Para ello, una de las aproximaciones frecuentemente adoptada por los científicos de datos es el Análisis de Datos Exploratorio (EDA), que enfatiza

las representaciones gráficas de los datos. El análisis corresponde a un proceso cíclico de extracción e interpretación de patrones, cuyo objetivo es complementar la construcción de modelos basándose en los hallazgos del EDA. Asimismo, dicho análisis también tiene como objetivo la búsqueda de patrones inesperados y el desarrollo de descripciones enriquecidas de la información disponible.

Con respecto al estudio de la demanda eléctrica a corto plazo, diversos hallazgos empíricos han sido sistemáticamente reportados en la literatura, los que son compartidos por la mayoría de los sistemas de operación en el mundo. En esta sección, dichos hechos "estilizados" [46] son ilustrados utilizando la base de datos de demanda eléctrica en Francia que contiene observaciones de la carga del sistema en intervalos de media hora desde enero de 2015 a diciembre de 2016.

3.2.1. Visualización de la serie de tiempo

Como se menciona en la sección (1.1), los datos de EED fueron obtenidos a partir de la base de datos pública ofrecida por RTE, que provee de datos históricos de demanda eléctrica hasta 1996, sin observaciones vacías [12]. La serie de carga para el período 2015-2016 se muestra en la Fig. (3.1), que ilustra claramente la estacionalidad anual que reina en la serie. La estacionalidad es considerablemente diferente para los períodos de invierno con respecto a los de verano. Niveles altos de demanda son hallados en invierno y verano comparados con temporadas de otoño y primavera debido al uso de calefacción eléctrica y aire acondicionado, respectivamente. El perfil de carga y la variación durante ambos años se ilustra en la Fig. (3.2), donde patrones constantes de distribución diaria y anual son observados en las variaciones de demanda. El histograma de la demanda en la Fig. (3.3) revela dos *peaks* a 50,000 MW y 54,000 MW, presentando sesgo hacia la derecha, es decir; la mayoría de los datos se encuentran bajo la demanda media. La estadística descriptiva para el año 2015 se presenta en la Tabla (3.1).

La Fig. (3.4) muestra los desplazamientos producidos en el ciclo medio intradiario condicionado al mes del año. El nivel de demanda más bajo es observable entre 4:00 y 5:00 a través del año. Una alza entre 6:00 y 8:00 es notable, que está asociada al comienzo de las actividades humanas rutinarias, horas laborales y períodos de alta demanda de las industrias,

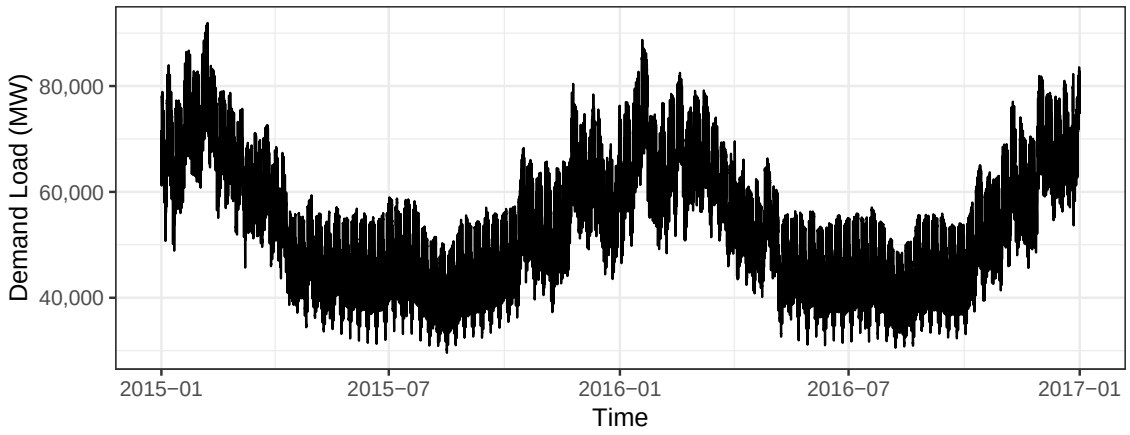


Figura 3.1: EED cada media hora en Francia, obtenida desde RTE para el período 2015-2016

servicios y uso doméstico. La EED alcanza un máximo a las 13:00, a excepción de las horas de la tarde en los meses de invierno, cuando el máximo global se halla aproximadamente a las 19:00, cuando la carga por iluminación está al máximo. Un mínimo local aparece usualmente luego de las 16:00, ocurriendo antes en los meses fríos, y moviéndose hacia las 21:00 en los meses cálidos. Además, los días de verano son menos curvos e irregulares que aquellos pertenecientes a otras estaciones del año.

La Fig. (3.5) muestra el perfil diario promedio para cada día de la semana. Existen tres patrones diferentes que pueden ser observados. De lunes a viernes, el comportamiento de la demanda eléctrica es bastante similar, aunque el día lunes comienza con bajos niveles de demanda asociados al fin de semana. A nivel diario, las curvas son similares para los días laborales, de lunes a viernes. Además, los viernes poseen una baja en los niveles de demanda desde las 13:00, asociado con el fin de los días laborales. Los días de fin de semana muestra un decrecimiento sistemático en la demanda eléctrica comparado con el resto de la semana asociado al bajo consumo industrial y comercial. Además, los niveles de carga en los días sábado son mayores y diferentes al día domingo en las horas de actividad humana.

La Fig. (3.6) presenta la demanda media por día de semana condicionada en el mes del año. El nivel de EED se incrementa y disminuye dependiendo de la etsación del año a la cual pertenece el mes. Los días martes, miércoles y jueves poseen mayor demanda comparado con el resto de la semana. Un patrón constante es observado a través de los

meses, como en la Fig. (3.4), donde el sábado y domingo corresponden a los períodos de menor demanda asociada a niveles bajos de actividad económica.

TABLA 3.1**:** Estadística descriptiva para la demanda eléctrica(MW) (in-sample)

Media	Mediana	Máx.	Min.	Desv. Estándar	Simetría	Curtosis
54222	52868	91934	29590	11611.98	0.45	-0.36

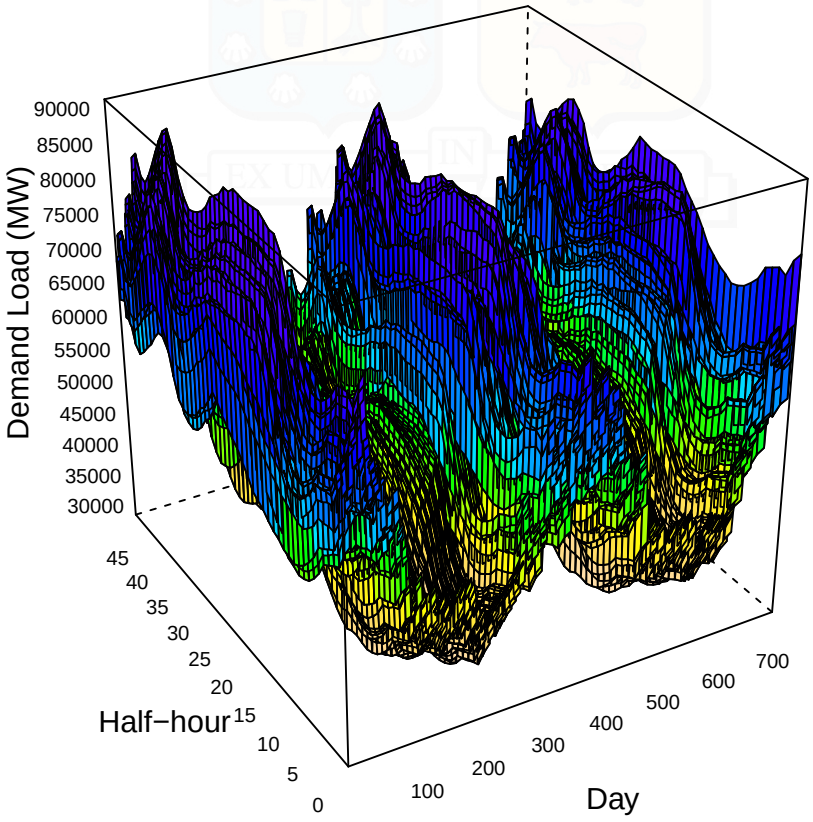


Figura 3.2: Representación 3D de la curva de carga cada media hora para el período 2015-2016

El análisis de la ACF y PACF ayuda al reconocimiento de estructuras de dependencia ya patrones regulares en la curva de carga [5, 47]. Para poder entender la estacionalidad y las estructuras de correlación, los primeros 336 rezagos se muestran en la Fig. (3.7), correspondiendo a una semana de observaciones.

La Fig. (3.7) muestra que la serie es no-estacionaria y fuertemente autocorrelacionada, pues el nivel de autocorrelación está por sobre 0.55 durante toda la semana de rezagos, sin disminución en su nivel. Como era de esperarse, un fuerte patrón estacional emerge cada

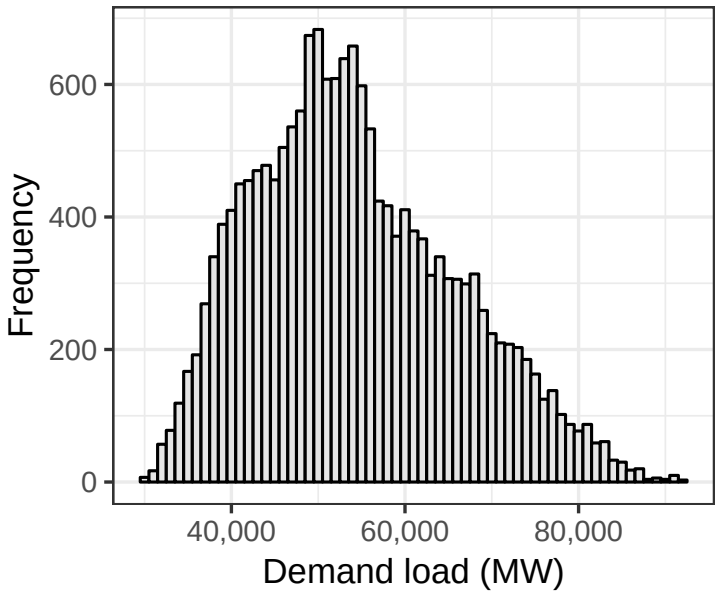


Figura 3.3: Histograma de observaciones semi-horarias para EED

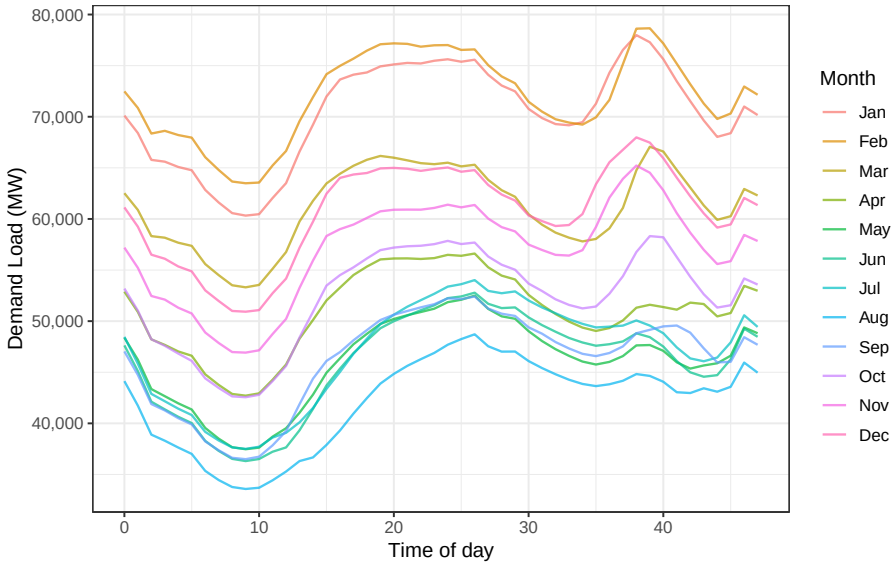


Figura 3.4: Ciclo medio intra-diario para cada mes del año (in-sample)

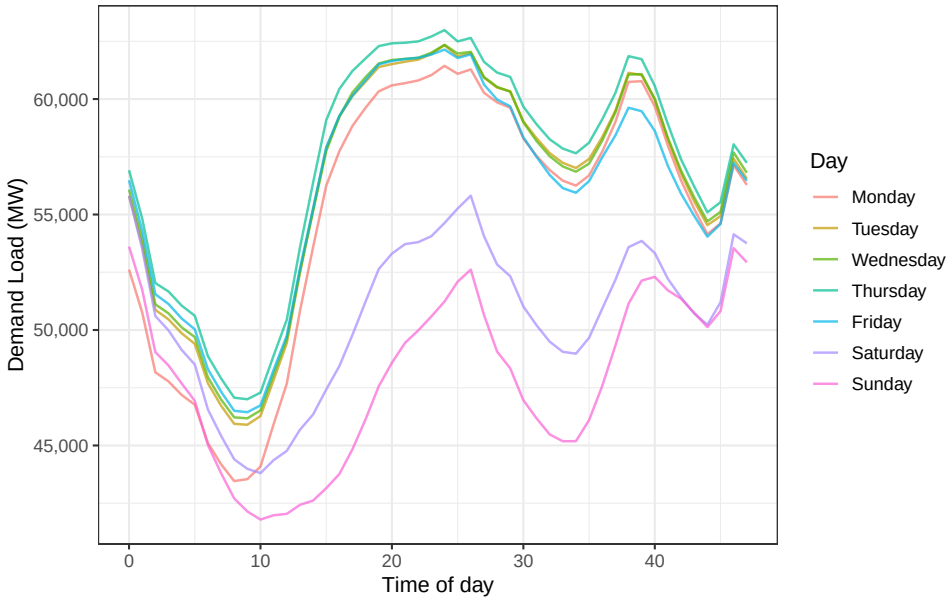


Figura 3.5: Ciclo medio intra-diario para cada día de la semana(in-sample)

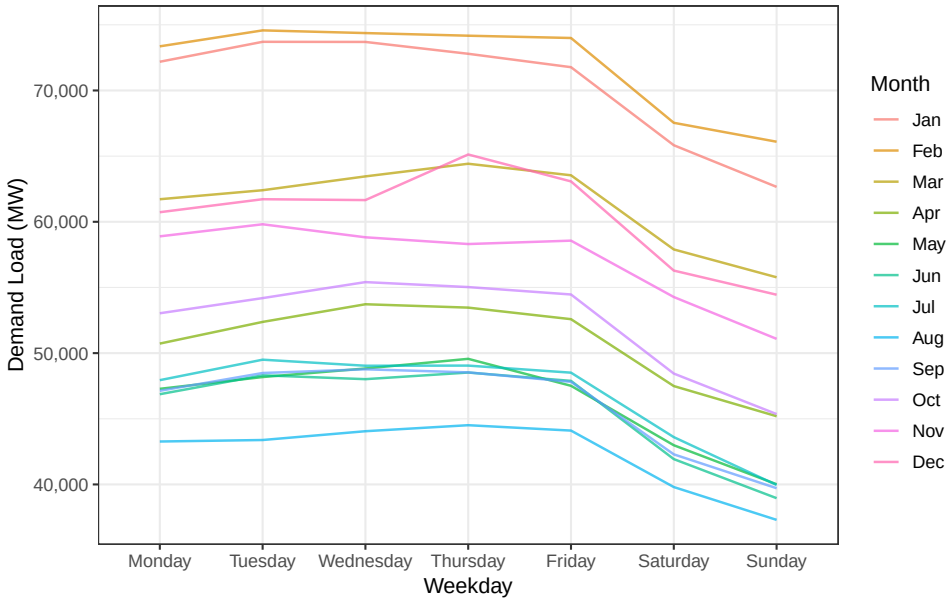


Figura 3.6: Carga media por día de semana para cada mes del año (in-sample)

48 observaciones, lo que corresponde al número de observaciones en un día. La presencia de estacionalidad se confirma por alzas periódicas en la PACF. La dependencia más fuerte puede ser observada en los rezagos 1 y 336, correspondiendo a la observación previa y la observación hace una semana atrás, respectivamente. Los otros peaks de autocorrelación se encuentran, en orden de importancia decreciente, en los días 1, 6, 2, 3, 4 y 5.

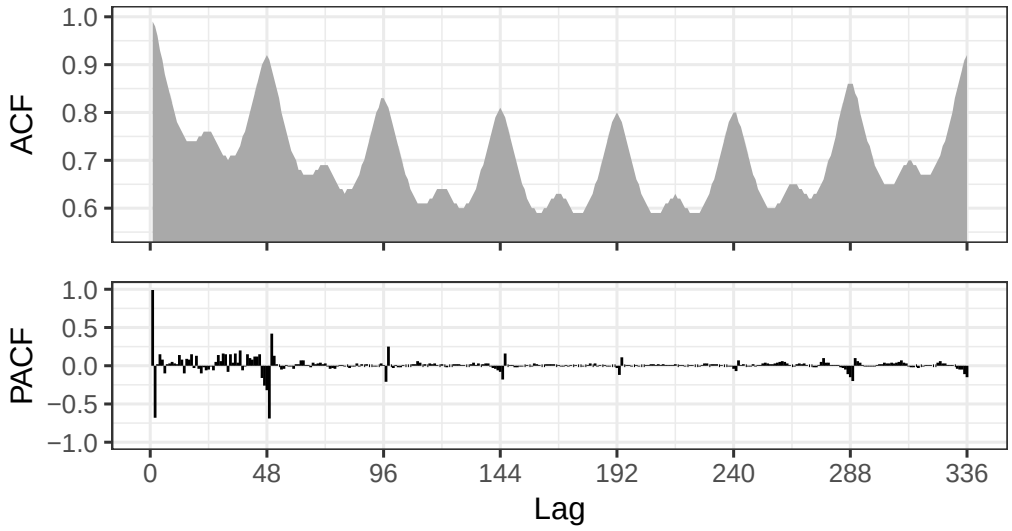


Figura 3.7: Funciones ACF (superior) y PACF (inferior) para la demanda eléctrica semi-horaria (in-sample)

La Descomposición estacional por regresión local polinomial (STL) permite complementar de manera visual el análisis de datos exploratorio, y también forma parte de los modelos estadísticos a evaluar para el pronóstico de demanda eléctrica. Como se revisó anteriormente, la demanda semi-horaria en Francia exhibe dos períodos estacionales. Es posible obtener una descomposición utilizando como estacionalidad periódica $m_1 = 48$ y $m_2 = 336$, para el ciclo diario y semanal, respectivamente. Los tres componentes obtenidos al aplicar la descomposición STL se muestran en la Fig. (3.8). Como era de esperar, los patrones diarios y semanales son observados. El componente de tendencia muestra el movimiento general de la serie, ignorando la estacionalidad y las pequeñas fluctuaciones aleatorias. Las escalas verticales permiten comparar el rango de los componentes, mostrando que la mayor fluctuación ocurre en el ciclo intradiario.

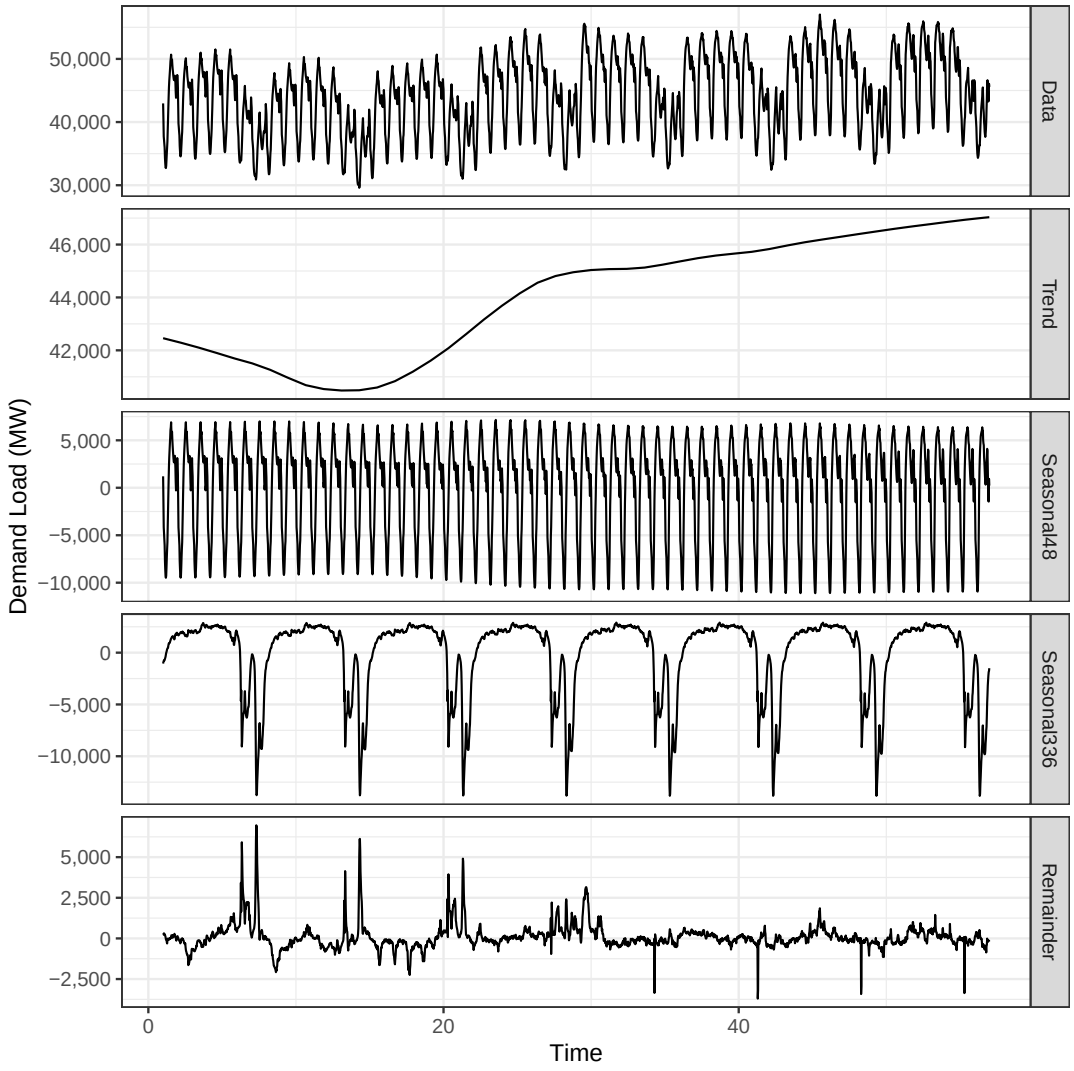


Figura 3.8: Descomposición estacional por Loess (STL) de la demanda semi-horaria para el período comprendido entre Agosto a Septiembre de 2015

3.3. Configuración experimental

3.3.1. Data

Como se introdujo en la sección (3.2.1), el experimento estudia el set de datos de demanda eléctrica en Francia obtenida desde RTE para el período 2015-2016. El número total de observaciones es 35,088 (dos años de 365 y 366 días × 24 horas × 2 observaciones por hora).

Los días especiales (observaciones inusuales) tienen un gran impacto en la demanda

eléctrica. Algunos estudios proponen algoritmos para reducir el error de pronóstico asociado a los días especiales [48, 49, 50] o usan variables dummy para diferenciar dichos períodos de las observaciones usuales [51]. Por otro parte, otros autores eligen dejar las periodicidades naturales de la serie intactas mediante la interpolación de la demanda en períodos correspondientes a las dos semanas adyacentes [29, 52, 53, 54]. Dado que el presente estudio corresponde a capacidad predictiva en línea, antes de ajustar y evaluar los modelos, se elige utilizar esta última aproximación y suavizar feriados y observaciones inusuales halladas en el set de datos.

Los datos son divididos en tres subconjuntos que no se traslapan: el set de entrenamiento (D_{train}), que corresponde al primer 75 % de los datos de 2015, el set de validación (D_{valid}),que contiene el 25 % restante de los datos de 2015, y el set de evaluación (D_{test}), que contiene todos los datos de 2016. Se utiliza D_{train} para llevar a cabo procedimientos de selección de variables descritos en la próxima sección. Los modelos basados en AI utilizan D_{train} y D_{valid} para ajuste de hiperparámetros. D_{test} es utilizado para evaluar la capacidad predictiva para ambos modelos estadísticos y de inteligencia artificial.

3.3.2. Selección de variables

Pese a que la inclusión de variables exógenas como el clima y las condiciones del sistema en la especificación de modelos podrían mejorar sustancialmente las predicciones de demanda [2, 46], los errores de pronóstico en la práctica serán mayores debido a la incertidumbre asociada al pronóstico climático necesario para generar predicciones [55]. Además, cuando las áreas cubiertas por la compañía eléctrica son pequeñas o poseen alta variabilidad en sus sub-regiones, su uso no ofrecerá una ventaja significativa [56]. Más aún, tener acceso a variables climáticas confiables puede ser difícil e incluso costoso, ya que muchos operadores de sistema ofrecen datos de demanda histórica de manera gratuita, pero reservan los datos climáticos asociados a ésta. Debido a lo anterior, la incorporación de variables climáticas en los modelos de pronóstico ha sido considerada impráctica por algunos autores [29, 52, 57]. Varios estudios consideran las metodologías univariadas como suficientes para el corto plazo, argumentando que las variables climáticas cambian en una manera suave y que tal comportamiento debiese ser capturado por la serie de demanda

misma [53, 56, 58, 59]. Basado en lo anterior, los modelos propuestos y revisados en este trabajo se basan en la información contenida por la serie de tiempo y explotan sus características univariadas.

En el contexto propuesto, la selección de variables corresponde al proceso de construir, evaluar y transformar la información contenida en la serie de tiempo de demanda eléctrica para generar pronósticos buenos y precisos.

En cuanto a la construcción de variables, la serie de demanda puede ser codificada de manera determinística para los modelos AI, como en el caso del modelo TBATS (véase sección 2.4.3.8), es posible hacer uso de pares de seno-coseno para cada estacionalidad identificada dentro de la serie como

$$x_{m_i,1} = \text{sen}\left(\frac{2\pi t}{m_i}\right), \qquad x_{m_i,2} = \cos\left(\frac{2\pi t}{m_i}\right) \tag{3.1}$$

con $m_1 = 48$ y $m_2 = 336$ para la estacionalidad diaria y semanal, respectivamente.

Como muestra la Fig. (3.7), la curva de demanda exhibe estructuras de correlación adicionales de longitud no-estacional que pueden ser incorporadas al vector de entrada junto a las variables dummy que codifican la estacionalidad de manera determinística. Para integrar realizaciones rezagadas de la serie de demanda en la forma de términos no-lineales autoregresivos, se selecciona un conjunto de variables rezagadas relevantes e informativas a partir de una ventana de datos de una semana en el tiempo, llevando a cabo selección de variables basando la decisión en la ACF y el coeficiente MI.

La Fig. (3.9) muestra el coeficiente MI normalizado y ordenado de manera jerárquica para los rezagos a una semana. Es posible notar que la curva del coeficiente MI se aplana gradualmente, sin mejora significativa aproximadamente luego de la variable en el ranking 50, que es establecido como punto de corte. Por ello, se considera la selección de las 50 variables mejor evaluadas por el análisis del criterio MI, como principales variables para los modelos basados en IA.

La transformación de variables en series de tiempo busca un adecuado preprocesamiento de las características para facilitar el modelamiento. Como se describe en [43], la evolución en el tiempo de la curva de carga es no-estacionaria, con la presencia de autocorre-

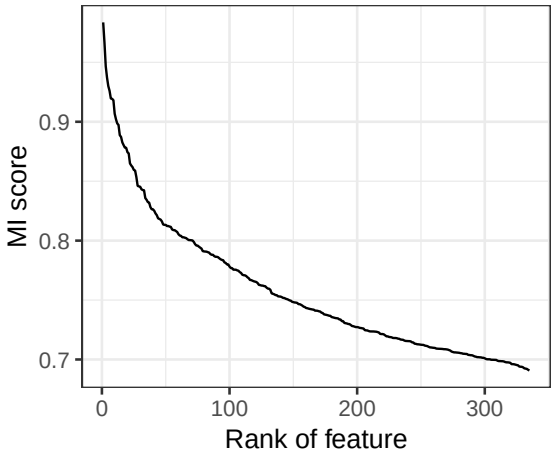


Figura 3.9: Coeficiente MI para cada variable en orden jerárquico

laciones fuertes y con decaimiento lento, como se muestra en la Fig. 3.7. Así, previo a ajustar los modelos, una dirección de experimentación interesante es filtrar la no-estacionariedad de la serie mediante la aplicación de los operadores de rezagos $(1 - L^{48})(1 - L^{336})$ o $(1 - L)(1 - L^{48})(1 - L^{336})$ a la serie de demanda, y_t, como etapa de preprocesamiento análoga al procedimiento para modeos ARIMA [43, 52]. Las Figs. (3.10) y (3.11) muestran la manera en que la EED decae cuando esta codificación estocástica específica es utilizada. Al momento de considerar modelos AI, las variables de entrada también son estandarizadas mediante el coeficiente z, escalamiento lineal y el método min-máx. Cada operación de preprocesamiento es revertida luego para evaluar el pronóstico producido por cada modelo IA. Las conjuntos de variable de entrada para cada modelo están listadas en la Tabla (3.2).

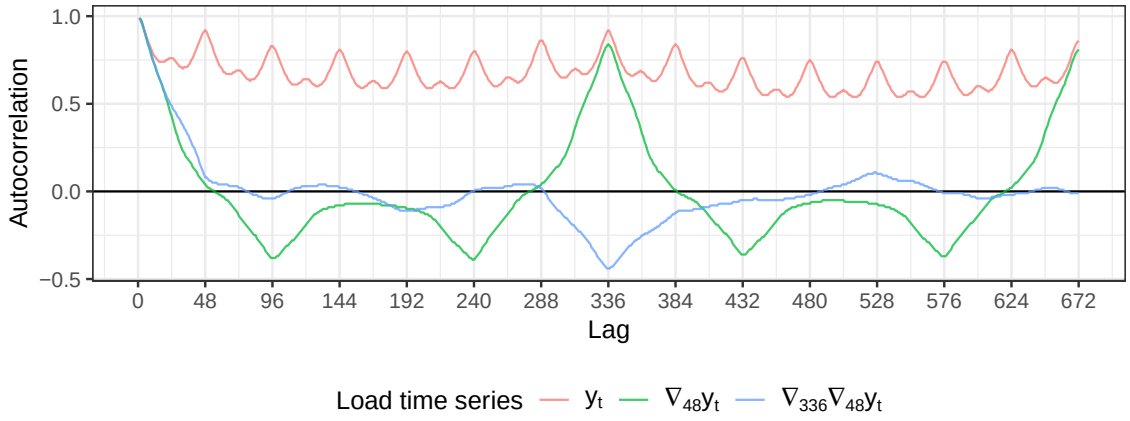


Figura 3.10: Funciones de autocorrelación para la serie EED diferenciada dos veces

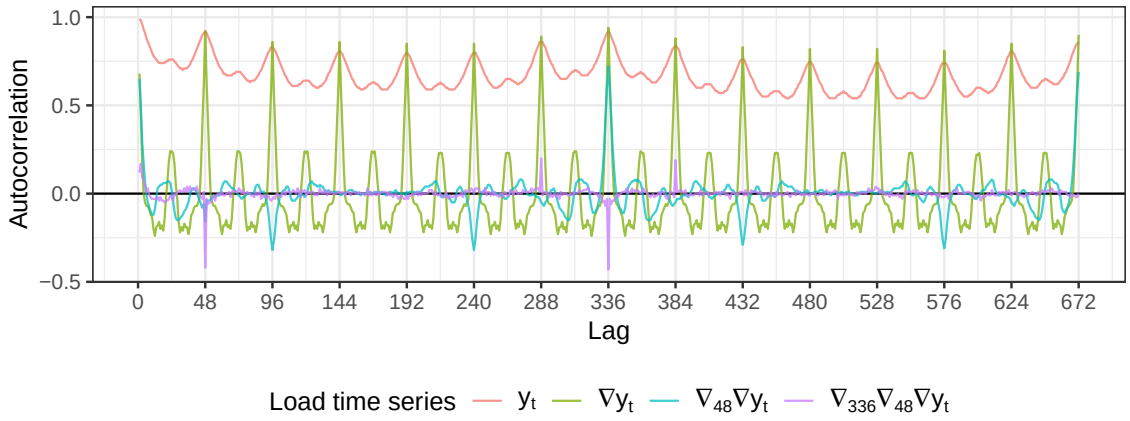


Figura 3.11: Funciones de autocorrelación para la serie EED diferenciada tres veces

TABLA 3.2**:** Variables seleccionadas por criterio MI

Conjunto de Variables	Variables rezagadas utilizadas para pronóstico
MI	1 - 13, 43 - 55, 95 - 99, 145, 241 - 242, 285 - 293, 330 - 336
$MI_{\nabla_{48}\nabla_{336}}$	1 - 39, 326 - 336
$MI_{\nabla_1\nabla_{48}\nabla_{336}}$	1 - 4, 49, 90 - 92, 95, 97 - 99, 101, 106, 110, 117, 124, 125, 128, 131, 132, 136, 137, 142, 144, 145, 151, 156, 159, 163, 165, 166, 170, 175, 179, 192 - 194, 206, 210 - 212, 227, 229, 240, 241, 289, 331, 335, 336

$\nabla_{48}\nabla_{336} = (1 - L^{48})(1 - L^{336})$
$\nabla_1\nabla_{48}\nabla_{336} = (1 - L)(1 - L^{48})(1 - L^{336})$

3.4. IMPLEMENTACIÓN EN R

Los siguientes paquetes y consideraciones han sido contempladas para implementar los modelos revisados en la sección (2.4), utilizando los paquetes de computación estadística disponibles en el lenguaje R [11].

Preprocesamiento de los datos y EDA

- Manipulación de datos: `data.table, dplyr, tidyverse, tsibble`
- Fechas: `lubridate`
- Gráficos: `ggplot2, plotly, lattice, gridExtra`

- Métricas de rendimiento: `Metrics`
- Entrenamiento de modelos: `caret`
- Computación paralela de modelos: `parallel, doMC`

STL

- Paquete `R`: `forecast`
- Función: `mstl`
- Ventana estacional: Periódica
- Componentes estacionales: 48, 336
- Box-Cox: Automático
- Residuos: SES, ARIMA

DSHW

- Paquete `R`: `forecast`
- Función: `bats`
- Componentes estacionales: 48, 336
- Corrección AR(1): Verdadero
- Box-Cox: Automático
- Tendencia: Verdadero

TBATS

- Paquete `R`: `forecast`
- Función: `bats`

- Componentes estacionales: 48, 336
- Corrección AR(1): Verdadero
- Box-Cox: Automático
- Tendencia: Verdadero

ANN

- Paquete R: `nnet`
- Función: `nnet`
- Algoritmo de Optimización: BFGS
- Hiper-parámetros: `size`, `decay`
- Función de activación: sigmoide
- Cantidad de iniciaciones: 20
- Combinación de pronóstico: media
- Tolerancia al error: 10^{-4}
- Iteraciones máximas: 1000
- Validación: Validación cruzada para series de tiempo ($75 - 25\,\%$)

ELM

- Paquete R: `nnfor`
- Función: `elm`
- Algoritmo de Optimización: Mínimos cuadrados
- Hiper-parámetros: `size`

- Función de activación: sigmoide
- Combinación de pronóstico: media
- Cantidad de iniciaciones: 20
- Validación: Validación cruzada automática para parámetro `size`
- Estimación para capa de salida: `lasso` con validación cruzada

SVM

- Paquete `R`: `kernlab`
- Función: `ksvm`
- Algoritmo de Optimización: SMO
- Hiper-parámetros: `cost`, `sigma`
- Kernel: Radial Basis Function (RBF)

$$\boldsymbol{K}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\Big(\frac{-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2\sigma^2}\Big) \tag{3.2}$$

- Tolerancia al error: 10^{-3}
- Epsilon en tubo insensitivo: 0.1
- Validación: Validación cruzada para series de tiempo ($75 - 25\,\%$)

Cada método genera pronósticos a un paso en adelante de manera iterativa, es decir; los pronósticos generados son utilizados como entradas para que los modelos puedan generar un pronóstico recursivo de varios pasos hacia adelante.

3.5. Análisis de resultados

La utilización de la métrica MAPE (véase sección 2.3.4.2) para evaluar el rendimiento de pronóstico se ha establecido como un estándar en la industria en cuanto a la medición de la precisión del pronóstico. Esto debido al hecho de que es capaz de capturar la proporcionalidad entre el error de pronóstico y la carga observada [55]. Rendimientos similares a los del MAPE fueron obtenidos con las métricas escala-dependientes MAE y RMSE. La Tabla (3.3) contiene el rendimiento en el set de evaluación para el horizonte de pronóstico de un día en adelante ($h = 48$), desde enero 2016 a diciembre 2016. Para el período de evaluación, se llevan a cabo pronósticos multi-paso considerando el esquema de ventana rodante (véase sección 2.3.4.3), re-estimando el modelo para cada ventana usando los hiper-parámetros obtenidos en el proceso de validación. Las mismas variables seleccionadas fueron utilizadas para cada modelo de la familia AI.

A partir de la Tabla (3.3) es posible verificar que todos los modelos considerados son capaces de superar el modelo estacional ingenuo utilizado como línea base de comparación. Como era de esperarse a partir de la revisión bibliográfica, el modelo de Holt-Winters con estacionalidad doble supera a los otros modelos en todos los casos con un MAPE dentro del set de evaluación de 1.66 %. Por lo tanto, DSHW es establecido como punto de comparación estadístico para comparar los modelos restantes. A partir de los modelos de Descomposición Estacional por Regresión Local Polinomial (STL), el uso de modelos ARIMA para pronosticar el residual de la descomposición STL es preferible por sobre el suavizamiento exponencial, con una diferencia en MAPE de 0.25 %. El método TBATS se desempeña peor que todos los otros modelos estadísticos, probando que la descomposición aplicada no es adecuado para este set de datos de demanda en particular.

De acuerdo al MAPE, el mejor modelo AI es la Red Neuronal Artificial que utiliza el operador de diferencias doble como etapa de preprocesamiento. Rendimientos similares son obtenidos para la Máquina de Aprendizaje Extremo, con el mismo preprocesamiento de los datos.

El mejor modelo de Máquinas de Vectores de Soporte es el único modelo que se desempeña mejor utilizando el operador de diferencias triple. En general, el rendimiento

TABLA 3.3: Métricas de evaluación para métodos entrenadas con una ventana rodante fija

Método de predicción	Parámetros del modelo	Métrica de rendimiento		
		RMSE (MW)	MAE (MW)	MAPE(%)
Modelo estacional ingenuo	-	6194	4593	8.29
ANN	Neuronas ocultas: 20 Peso de decaimiento: 0.01	3195	2332	4.12
$ANN_{\nabla_{48}\nabla_{336}}$	Neuronas ocultas: 6 Peso de decaimiento: 0.1	**2106**	1354	**2.41**
$ANN_{\nabla_1\nabla_{48}\nabla_{336}}$	Neuronas ocultas: 6 Peso de decaimiento: 0.1	2260	1454	2.64
ELM	Neuronas ocultas: Auto Pesos de salida: Lasso	3372	2497	4.55
$ELM_{\nabla_{48}\nabla_{336}}$	Neuronas ocultas: Auto Pesos de salida: Lasso	**2140**	1398	**2.50**
$ELM_{\nabla_1\nabla_{48}\nabla_{336}}$	Neuronas ocultas: Auto Pesos de salida: Lasso	2264	1414	2.55
SVM	Costo: 4 Sigma: 0.0254	4298	3215	5.85
$SVM_{\nabla_{48}\nabla_{336}}$	Costo: 1 Sigma: 0.0453	2534	1666	2.96
$SVM_{\nabla_1\nabla_{48}\nabla_{336}}$	Costo: 0.25 Sigma: 0.0165	2332	1396	**2.55**
DSHW	Parámetros de suavizamiento $\alpha, \beta, \delta, \omega, \phi$	**1301***	922	**1.66***
STL + EXP	Parámetros de suavizamiento α	1749	1179	2.15
STL + ARIMA	parámetros ARIMA (p,d,q)	1487	1054	1.90
TBATS	Parámetros de suavizamientos ω, ϕ, (p,q)	3014	2195	4.04

Entradas en negrita destacan el mejor modelo condicionado a las variables de entrada por técnica, por RMSE y MAPE.

obtenido a través del uso de una codificación de la estacionalidad determinística no es superior al uso de una codificación estocástica, con la incorporación de rezagos relevantes. Así, es esperable que modelos complejos y con más variables relevantes logren superar los resultados obtenidos con modelos univariados.

La Fig. (3.12) muestra la precisión de pronóstico en el set de evaluación para los cuatro mejores métodos considerando un horizonte de evaluación de un día para cada técnica. El perfil edl error es similar entre los modelos, con un incremente en el nivel de MAPE observable para todos los modelos. Dos alzas de error son halladas en la mañana; entre 06:00 y 10:00, y en la tarde, a partir de 13:00 a 17:00. Tales características pueden ser asociadas a la variabilidad en el comportamiento de la demanda eléctrica a partir del sistema en esas horas en específico. Más aún, tales horas son más difíciles de predecir para los modelos propuestos, sin importar la metodología utilizada para obtener el pronóstico. Por otro lado, el crecimiento sostenido del error puede ser atribuido al uso de una estrategia

de pronóstico recursiva. En general, la varianza del pronóstico se incrementará con el horizonte de pronóstico, lo que implica que si se promedia los valores de error absoluto o cuadrado en el set de evaluación, se están combinando resultados con diferentes varianzas.

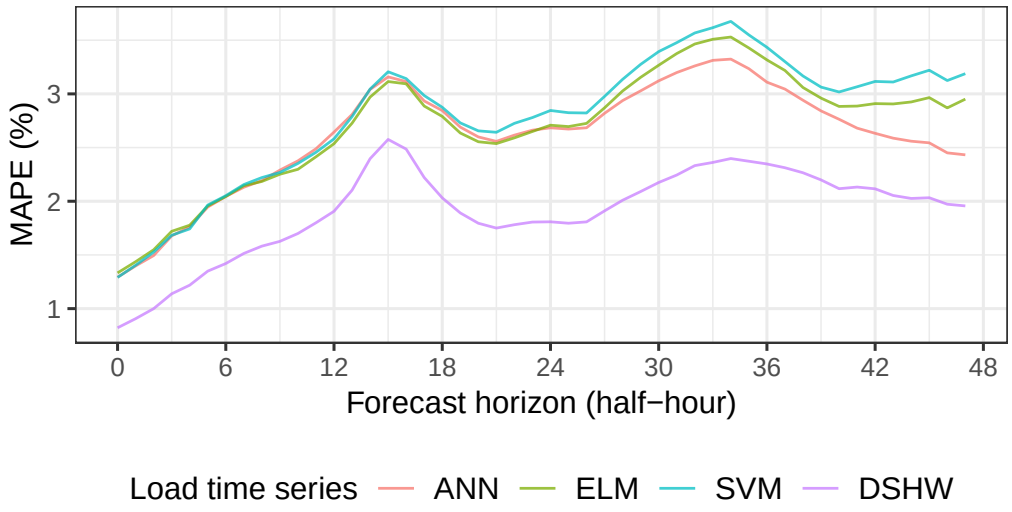


Figura 3.12: Resultados de MAPE con respecto al horizonte de pronóstico (testing set)

En la Fig. (3.13), el desempeño de los mejores modelos por cada técnica condicionados por el día de la semana es presentado. Dada la metodología de pronóstico propuesta, en general los días de semana tienen un menor MAPE que los días de fin de semana. Más aún, los modelos AI tienen mayor dificultad para pronosticar los días viernes, lo que puede ser atribuido a las observaciones cercanas al fin de semana. Por otro lado, es aparente que la similaridad entre los días martes, miércoles y jueves los hace más fáciles de pronosticar para todos los modelos. En la Fig. (3.14) se muestra el error de los mejores modelos para cada mes del año. Es posible observar un incremento en el error de pronóstico independiente de la técnica utilizada para algunos meses. En particular, la existencia de meses más dificíles de pronosticar como Mayo, Julio y Noviembre, puede estar asociado con la irregularidad de las semanas con feriados suavizados y su efecto en el resto de la semana.

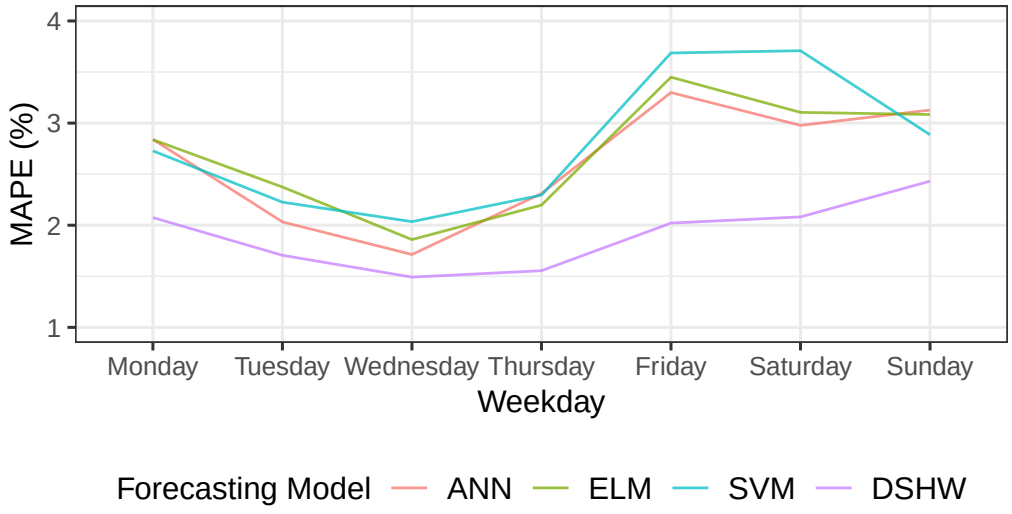


Figura 3.13: MAPE por día de semana

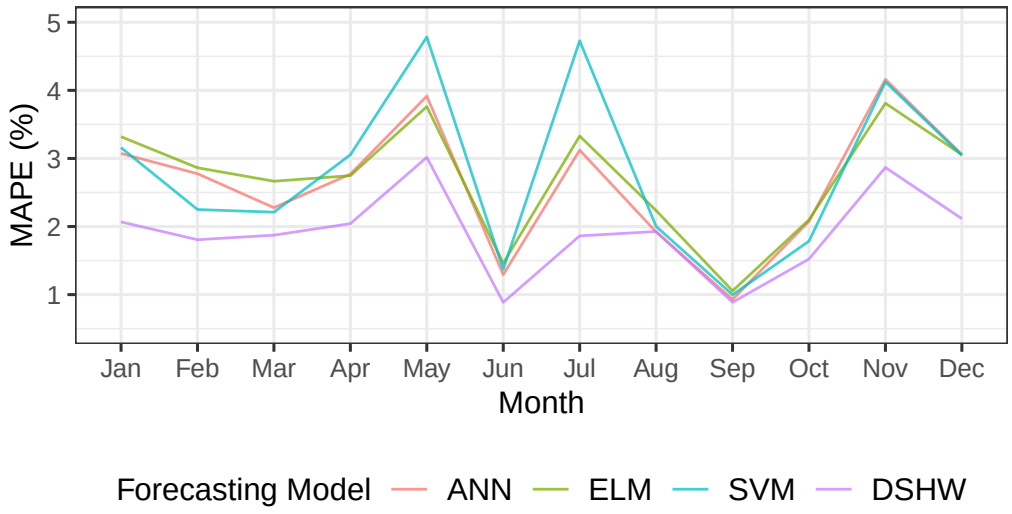


Figura 3.14: MAPE mensual

Capítulo 4

Conclusiones

En este documento, se evalúa el rendimiento de pronóstico de modelos univariados para el pronóstico de demanda semi-horaria a corto plazo a un día en adelante. Los modelos univariados para el pronóstico de carga no se basan en variables explicativas, como variables meteorológicas. Por lo tanto, los estos moedlos pueden ser utilizados cuando los datos meteorológicos no están disponibles o no son confiables.

La base de datos pública del operador del sistema de transmisión Francés se utilizó desde Enero de 2015 hasta Diciembre de 2016. El Análisis de Datos Exploratorio realizado respalda la característica de doble estacionalidad conocida que está presente en la serie temporal. Las visualizaciones permiten realizar un análisis de la duración de las temporadas conocidas en la serie de carga. La relación establecida entre el valor actual y los valores pasados se analizó a través de la función de autocorrelación lineal y el criterio de información mutua no lineal. Los datos de demanda eléctrica son sometidos a selección de variables, en la que se generan variables ficticias que codifican la estacionalidad de la serie de una manera determinista. Dicha alternativa se compara con el enfoque que utiliza operadores de diferencias estacionales para codificar la estacionalidad de una manera estocástica. Los conjuntos de variables obtenidos se comparan por medio de diferentes modelos basados en inteligencia artificial. Por otro lado, los modelos estadísticos conocidos por su buen desempeño en el pronóstico de series temporales estacionales se han considerado como modelos de referencia.

El mejor modelo estadístico es capaz de predecir la demanda de electricidad para un

día en adelante antes con un MAPE de 1,66 % durante un año dentro del set de evaluación. Ningún modelo basado en AI pudo superar rendimiento del mejor modelo estadístico (DSHW). Los mejores resultados se obtuvieron con un modelo de Red Neuronal Artificial con seis neuronas y doble diferenciación como procedimiento de preprocesamiento de datos. La Máquina de Aprendizaje Extremo también fue evaluada con resultados similares obtenidos. En particular, el modelo de Máquinas de Vector de Soporte se desempeñó mejor cuando utilizó la triple diferenciación en la etapa de preprocesamiento. Después de la diferenciación estacional, el patrón estacional principal y la tendencia se eliminan y se mejora el rendimiento. Los resultados muestran que, independientemente del modelo considerado, hay horas, días y meses que son más difíciles de predecir que otros. Dos *peaks* se encuentran en el perfil de error diario. Además, los días de fin de semana son más difíciles de predecir en comparación con los días de la semana. El mismo fenómeno se halla en los meses de Mayo, Julio y Noviembre.

Los modelos revisados en este trabajo están disponibles y se pueden implementar fácilmente utilizando paquetes dentro del lenguaje de estadística estadístico R [11] por cualquier practicante. Por lo tanto, modelos más complejos, y técnicas que incorporen variables explicativas deberían superar las metodologías propuestas.

4.1. Panorama y direcciones futuras

El estudio presentado, sienta las bases para el desarrollo de modelos de pronóstico de demanda eléctrica más complejos. El artículo que subyace este documento sirve como línea base y guía para científicos de datos e investigadores que busquen una introducción gentil al pronóstico de demanda eléctrica, su problemática, y las aproximaciones para su solución. La literatura actual ofrece un sinnúmero de modelos, metodologías e hibridaciones que prometen minimizar el error de pronóstico. Sin embargo, la mayoría de estos estudios carece de un carácter integrado de análisis. Usualmente, la literatura no contempla un análisis de datos exploratorio como estudio previo para el modelamiento o el pronóstico de demanda eléctrica. Más aún, muchos artículos no consideran incluso una metodología como línea base a mejorar (ej. un modelo estacional ingenuo). Pese a que todos los pronósticos

están equivocados, los investigadores llevan años persiguiendo un pronóstico precisos. Como la mayoría de las técnicas originales han sido utilizadas, los autores han comenzado a "combinarlas" para proponer un "nuevo modelo híbrido", algunos aportando valor a la resolución del problema, pero la mayoría realiza una mínima contribución a la literatura. La precisión reportada de dichos estudios es usualmente impresionante, incluso muchas veces demasiado buena para ser verdad. Dichas prácticas conllevan consecuencias negativas para el área de investigación, mostrando siempre que el método propuesto en el artículo supera a todas las otras técnicas para datos específicos, haciendo de las conclusiones difíciles de generalizar. Asimismo, rara vez los estudios propuestos pueden ser reproducidos por cualquier practicante, haciendo del proceso de replicación una tarea críptica y tediosa, lo que limita el progreso de la investigación y el desarrollo.

Es muy importante para los investigadores y practicantes entender que *una técnica universalmente superior simplemente no existe*. La naturaleza de los datos y su jurisdicción determinan qué técnica debe ser utilizada. El enfoque siempre debe ser entendiendo las necesidades del negocio primero, analizando los datos, y luego de un proceso de prueba y error, obtener cuál es la mejor técnica para dicho set de datos en específico. Notar que el error de pronóstico además podría diferir significativamente para diferentes compañías, para diferentes zonas de ella y para diferentes períodos de tiempo.

Las direcciones futuras deben incluir la novedad dentro del esquema de investigación, resolviendo nuevos problemas, proponiendo nuevas metodologías y técnicas a nuevos set de datos; presentando así nuevos hallazgos. Algunas áreas que quedan abiertas a investigación contemplan la variabilidad climática, el impacto del uso de vehículos eléctricos, la generación eólica y solar de energía, la eficiencia energética y la respuesta a la demanda.

En particular, a partir de este trabajo es posible elaborar modelos que integren la temperatura, el clima y variables geográficas en el pronóstico. Trabajos futuros podrían considerar un modelo para cada observación del día, la evaluación de aproximaciones multi-etapa con modelos ensamblados y la incorporación de variables exógenas atigentes a los métodos que admitan pronóstico multivariado.

Bibliografía

[1] Tao Hong et al. Energy forecasting: Past, present, and future. *Foresight: The International Journal of Applied Forecasting*, (32):43–48, 2014.

[2] Tao Hong and Shu Fan. Probabilistic electric load forecasting: A tutorial review. *International Journal of Forecasting*, 32(3):914–938, 2016.

[3] Petra Vrablecová, Anna Bou Ezzeddine, Viera Rozinajová, Slavomír Šárik, and Arun Kumar Sangaiah. Smart grid load forecasting using online support vector regression. *Computers & Electrical Engineering*, 65:102–117, 2018.

[4] Rob J Hyndman and Shu Fan. Density forecasting for long-term peak electricity demand. *IEEE Transactions on Power Systems*, 25(2):1142–1153, 2010.

[5] Hristos Tyralis, Georgios Karakatsanis, Katerina Tzouka, and Nikos Mamassis. Exploratory data analysis of the electrical energy demand in the time domain in greece. *Energy*, 2017.

[6] Tao Hong. *Short Term Electric Load Forecasting*. PhD thesis, North Carolina State University, 2010.

[7] Ramu Ramanathan, Robert Engle, Clive WJ Granger, Farshid Vahid-Araghi, and Casey Brace. Short-run forecasts of electricity loads and peaks. *International journal of forecasting*, 13(2):161–174, 1997.

[8] Tao Hong. Crystal ball lessons in predictive analytics. *EnergyBiz Mag*, pages 35–37, 2015.

[9] AK Srivastava, Ajay Shekhar Pandey, and Devender Singh. Short-term load forecasting methods: A review. In *Emerging Trends in Electrical Electronics & Sustainable Energy Systems (ICETEESES), International Conference on*, pages 130–138. IEEE, 2016.

[10] Réseau de transport d'électricité. RTE datascience.net challenge, 2018.

[11] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018.

[12] French transmission system operator RTE. Réseau de transport d' électricité website, 2018.

[13] Rob J Hyndman and George Athanasopoulos. *Forecasting: principles and practice*. OTexts, 2014.

[14] George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[15] Robert Shumway. *Time Series Analysis and Its Applications : With R Examples*. Springer, Cham, Switzerland, 2017.

[16] Gilbert Strang. *Introduction to linear algebra*. Cambridge Press, Wellesley, MA, 2016.

[17] Carlos M Jarque and Anil K Bera. Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics letters*, 6(3):255–259, 1980.

[18] Rob J Hyndman and Anne B Koehler. Another look at measures of forecast accuracy. *International journal of forecasting*, 22(4):679–688, 2006.

[19] Max Kuhn. The caret package. Online: consultado en 20/01/2018. http://topepo.github.io/caret/data-splitting.html#data-splitting-for-time-series.

[20] Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

[21] Gerda Claeskens. *Model selection and model averaging*. Cambridge University Press, Cambridge New York, 2008.

[22] Gareth James. *An introduction to statistical learning : with applications in R*. Springer, New York, NY, 2013.

[23] Rob J Hyndman, Anne B Koehler, Ralph D Snyder, and Simone Grose. A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of forecasting*, 18(3):439–454, 2002.

[24] Rob Hyndman, Anne B Koehler, J Keith Ord, and Ralph D Snyder. *Forecasting with exponential smoothing: the state space approach*. Springer Science & Business Media, 2008.

[25] Robert Goodell Brown. *Statistical forecasting for inventory control*. McGraw/Hill, 1959.

[26] Charles C Holt. Forecasting seasonals and trends by exponentially weighted moving averages. *International journal of forecasting*, 20(1):5–10, 2004.

[27] Everette S Gardner Jr and ED McKenzie. Forecasting trends in time series. *Management Science*, 31(10):1237–1246, 1985.

[28] Peter R Winters. Forecasting sales by exponentially weighted moving averages. *Management science*, 6(3):324–342, 1960.

[29] James W Taylor. Short-term electricity demand forecasting using double seasonal exponential smoothing. *Journal of the Operational Research Society*, 54(8):799–805, 2003.

[30] Chris Chatfield. The holt-winters forecasting procedure. *Applied Statistics*, pages 264–279, 1978.

[31] Alysha M De Livera, Rob J Hyndman, and Ralph D Snyder. Forecasting time series with complex seasonal patterns using exponential smoothing. *Journal of the American Statistical Association*, 106(496):1513–1527, 2011.

[32] Robert B Cleveland, William S Cleveland, and Irma Terpenning. Stl: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics*, 6(1):3, 1990.

[33] Marina Theodosiou. Forecasting monthly and quarterly time series using stl decomposition. *International Journal of Forecasting*, 27(4):1178–1195, 2011.

[34] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533, 1986.

[35] Sven F Crone and Nikolaos Kourentzes. Feature selection for time series prediction–a combined filter and wrapper approach for neural networks. *Neurocomputing*, 73(10-12):1923–1936, 2010.

[36] Guang-Bin Huang, Qin-Yu Zhu, and Chee-Kheong Siew. Extreme learning machine: theory and applications. *Neurocomputing*, 70(1-3):489–501, 2006.

[37] Song Li, Lalit Goel, and Peng Wang. An ensemble approach for short-term load forecasting by extreme learning machine. *Applied Energy*, 170:22–29, 2016.

[38] Wei-Chiang Hong. Electric load forecasting by seasonal recurrent svr (support vector regression) with chaotic artificial bee colony algorithm. *Energy*, 36(9):5568–5578, 2011.

[39] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):27, 2011.

[40] Chia-Nan Ko and Cheng-Ming Lee. Short-term load forecasting using svr (support vector regression)-based radial basis function neural network with dual extended kalman filter. *Energy*, 49:413–422, 2013.

[41] Nello Cristianini. *An introduction to support vector machines : and other kernel-based learning methods*. Cambridge University Press, Cambridge New York, 2000.

[42] Irena Koprinska, Mashud Rana, and Vassilios G Agelidis. Correlation and instance based feature selection for electricity load forecasting. *Knowledge-Based Systems*, 82:29–40, 2015.

[43] Georges A Darbellay and Marek Slama. Forecasting the short-term demand for electricity: Do neural networks stand a better chance? *International Journal of Forecasting*, 16(1):71–83, 2000.

[44] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical review E*, 69(6):066138, 2004.

[45] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(Feb):281–305, 2012.

[46] Rafal Weron. *Modeling and forecasting electricity loads and prices: A statistical approach*, volume 403. John Wiley & Sons, 2007.

[47] David C Hamilton and Donald G Watts. Interpreting partial autocorrelation functions of seasonal time series models. *Biometrika*, 65(1):135–140, 1978.

[48] O Hyde and PF Hodnett. Rule-based procedures in short-term electricity load forecasting. *IMA Journal of Management Mathematics*, 5(1):131–141, 1993.

[49] Dipti Srinivasan, CS Chang, and AC Liew. Demand forecasting using fuzzy neural computation, with special emphasis on weekend and public holiday forecasting. *IEEE Transactions on Power Systems*, 10(4):1897–1903, 1995.

[50] Kyung-Bin Song, Young-Sik Baek, Dug Hun Hong, and Gilsoo Jang. Short-term load forecasting for the holidays using fuzzy linear regression method. *IEEE transactions on power systems*, 20(1):96–101, 2005.

[51] Agostino Tarsitano and Ilaria L Amerise. Short-term load forecasting using a two-stage sarimax model. *Energy*, 133:108–114, 2017.

[52] James W Taylor, Lilian M De Menezes, and Patrick E McSharry. A comparison of univariate methods for forecasting electricity demand up to a day ahead. *International Journal of Forecasting*, 22(1):1–16, 2006.

[53] James W Taylor. Triple seasonal methods for short-term electricity demand forecasting. *European Journal of Operational Research*, 204(1):139–152, 2010.

[54] SR Brubacher and G Tunnicliffe Wilson. Interpolating time series with application to the estimation of holiday effects on electricity demand. *Applied Statistics*, pages 107–116, 1976.

[55] Henrique Steinherz Hippert, Carlos Eduardo Pedreira, and Reinaldo Castro Souza. Neural networks for short-term load forecasting: A review and evaluation. *IEEE Transactions on power systems*, 16(1):44–55, 2001.

[56] Lacir J Soares and Marcelo C Medeiros. Modeling and forecasting short-term electricity load: A comparison of methods with an application to brazilian data. *International Journal of Forecasting*, 24(4):630–644, 2008.

[57] Bo-Juen Chen, Ming-Wei Chang, et al. Load forecasting using support vector machines: A study on eunite competition 2001. *IEEE transactions on power systems*, 19(4):1821–1830, 2004.

[58] Mashud Rana and Irena Koprinska. Forecasting electricity load with advanced wavelet neural networks. *Neurocomputing*, 182:118–132, 2016.

[59] Ergun Yukseltan, Ahmet Yucekaya, and Ayse Humeyra Bilge. Forecasting electricity demand for turkey: Modeling periodic variations and demand segregation. *Applied Energy*, 193:287–296, 2017.