UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

PRONÓSTICO DE DEMANDA DE ENERGÍA Y POTENCIA ELÉCTRICA EN EL LARGO PLAZO PARA LA RED DE CHILECTRA S.A. UTILIZANDO TÉCNICAS DE MINERÍA DE DATOS

TESIS PARA OPTAR AL GRADO DE MAGÍSTER EN GESTIÓN DE OPERACIONES

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

ANDRÉS FELIPE PALMA LLEWELLYN

PROFESOR GUÍA:
SR. RICHARD WEBER HAAS

MIEMBROS DE LA COMSIÓN:
SR. LUIS ABURTO LAFOURCADE
SR. ROGER MELLADO ZAPATA

Este trabajo ha sido parcialmente financiado por Conicyt: CONICYT-PCHA/Magíster Nacional Complementario/2013 - 221320517

RESUMEN DE TESIS PARA OPTAR AL TÍTULO DE: Ingeniero Civil Industrial y grado de Magíster en Gestión de Operaciones
FECHA: 2 de diciembre de 2015
Por: Andrés Felipe Palma Llewellyn
Profesor Guía: Richard Weber

PRONÓSTICO DE DEMANDA DE ENERGÍA Y POTENCIA ELÉCTRICA EN EL LARGO PLAZO PARA LA RED DE CHILECTRA S.A. UTILIZANDO TÉCNICAS DE MINERÍA DE DATOS

Las compañías dedicadas a la distribución de energía eléctrica incurren en altas inversiones para mantener una operación continua. Este tipo de condiciones hace aún más necesario el contar con una correcta proyección de demanda que permita precisar las inversiones, tratando de no incurrir en sobre ó sub inversión, lo que puede llevar a la implementación de una red sobre-dimensionada (si se sobreestima) ó a arriesgar la calidad del servicio y la compra de energía a un mayor precio, en el caso de la subestimación.

El presente trabajo de tesis tiene como objetivo construir modelos de pronóstico de demanda de energía y potencia eléctrica con el fin de mejorar la proyección que realiza CHILECTRA S.A. Se realizaron experimentos con redes neuronales artificiales support vector regression y métodos estadísticos de series de tiempo (SARIMA y SARIMAX) para desarrollar cinco modelos predictivos que pronostiquen en un horizonte de 5 años las siguientes series: demanda de energía en el sistema, demanda residencial, demanda comercial, demanda industrial y potencia máxima mensual.

El enfoque utilizado para los modelos corresponde a uno con variables explicativas y rezagos, utilizando tres conjuntos de variables: Climatológicas, Macroeconómicas y de Sistema (red).

Se realizó un pre-procesamiento a las series a ingresar a los modelos, donde destaca una transformación aplicada a la serie de potencia eléctrica, en la cual se elimina el efecto de un patrón intrínseco asociado a las diferencias de demandas diarias dentro de cada semana.

Para la selección de variables se utilizó el método de análisis de regresión para luego ingresar los atributos a los modelos, entregándole de esta forma variables (con rezagos incluidos) que tuvieran una relación lineal y que además cumplían, bajo esta óptica, con la homocedasticidad e independencia temporal de los errores.

Se utilizaron cinco horizontes temporales para evaluar el desempeño los modelos de redes neuronales, support vector regression, SARIMA y SARIMAX, obteniendo las redes neuronales un mejor desempeño, registrando un MAPE de un 2,78% para la demanda de energía, y un 3,74% para la potencia máxima de cada mes, siendo estos valores el promedio de los errores en cada horizonte. Estos valores implican una disminución del error respecto de los modelos previamente utilizados en la empresa de un 3,44% para el pronóstico de energía, y de un 1,44% para el caso de la potencia, logrando de esta forma el objetivo inicial planteado.

Los modelos de demanda sectoriales obtuvieron los siguientes errores en el largo plazo: 3,50% para el que trata la demanda residencial, 4,02% para el de demanda comercial, y 3,81% para el de industrial. Estos valores fueron obtenidos mediante redes neuronales, y nuevamente mostraron un mejor desempeño que los otros métodos.

Los resultados en el largo plazo fueron satisfactorios para todos los modelos finalmente establecidos, considerando en que un 10% de error en este ámbito es evaluado como bueno.

Debido a la alta dependencia de los modelos con la variable IMACEC, demostrada mediante un análisis de sensibilidad, se realizó un experimento de minería de textos para pronosticar los cambios en las expectativas del producto interno bruto, utilizados para pronosticar ésta variable explicativa, en base a las noticias de meses anteriores.

Se entrenó un modelo clasificador de support vector machine que logró tener un máximo de desempeño de 60% de accuracy, lo que no es un buen resultado dado que todavía está abierto a mucha incertidumbre. No obstante aporta evidencia sobre la información que contienen las noticias para predecir posibles cambios futuros en la economía.

Los objetivos de esta tesis fueron cumplidos, al confeccionar e implementar modelos predictivos de las demandas de energía y potencia eléctrica que mejoran la predicción de la demanda a la que se ve enfrentada CHILECTRA. Por otra parte, la implementación de la metodología de minería de textos se transforma en un primer paso para que la empresa pueda adelantarse a cambios macroeconómicos a ocurrir en el país.

A Mamá, Papá y Claudio

AGRADECIMIENTOS

A Nancy, Esteban, Carmen y Karla. Porque con ustedes aprendí que se puede formar familia sin sangre.

A Cristóbal, Miguel y Andrés. Por la amistad, el incondicional apoyo, los buenos ratos, las tareas que hicimos juntos y por estar desde el inicio hasta el final de esta travesía.

A Hugo, Luis, Mijael, Pablo, Carolina, Javier, Naima, Fabián, Joaquín y a todos los entrañables amigos de Coquimbo. Gracias a todos ustedes por los buenos ratos y la amistad que han compartido conmigo.

A Roger Mellado, César Araya y Bernardo Bravo por facilitar la realización de este trabajo, así como también por la grata experiencia que viví con ustedes en la empresa.

A mi profesor guía, Richard Weber, por su constante apoyo y preocupación por la realización de este trabajo. También a Luis Aburto, por los consejos que ayudaron a perfeccionar la tesis.

Tabla de Contenido

Índice de Tablas

Índice de Ilustraciones

Índice de Gráficos

1. Introducción

Una de las tareas de mayor relevancia para una empresa consiste en la planificación. Ésta debe ser la encargada de la confección de un plan que permita un correcto funcionamiento de la empresa además de ser una de las claves para conseguir mejores resultados con una mejor eficiencia.

La confección de la planificación no es algo trivial, y una de las tareas que se destaca dentro de las más importantes es la predicción de la demanda que la empresa recibirá de parte de sus clientes. Esta tarea permitirá llevar a cabo una planificación que visualice las inversiones necesarias a realizar, pero dando la posibilidad de efectuar acciones que minimicen los costos y/ó permitan entregar un mejor servicio.

Es común ver que al realizar una proyección de demanda, se determina realizar inversiones para aumentar la capacidad de producción. A su vez, la importancia de la planificación aumenta a medida que este tipo de inversiones implican un mayor costo y tiempo de desarrollo, como lo es en el caso del sector eléctrico.

Las compañías dedicadas a la generación ó distribución de energía eléctrica incurren en altas inversiones para mantener una operación continua[1]. Este tipo de condiciones hace aún más necesario el contar con una correcta proyección de demanda que permita precisar las inversiones, tratando de no incurrir en sobre ó sub inversión.

Como referencia, la mayor empresa distribuidora del país es Chilectra S.A., la cual presentó inversiones por construcciones en curso por un monto 119 millones de dólares en el año 2014, mientras que realizó compras de energía por 1.250 millones de dólares en el mismo período [49].

Para comprender de mejor forma la dinámica de este tipo de empresas, es necesario explicar el funcionamiento del mercado eléctrico

[1] La ley 20.018 del Ministerio de Economía, Fomento y Reconstrucción obliga a las distribuidoras el suministro permanente de energía a todos sus clientes regulados.

de Chile, dentro del cual existen 5 grupos enfocados en distintas actividades.

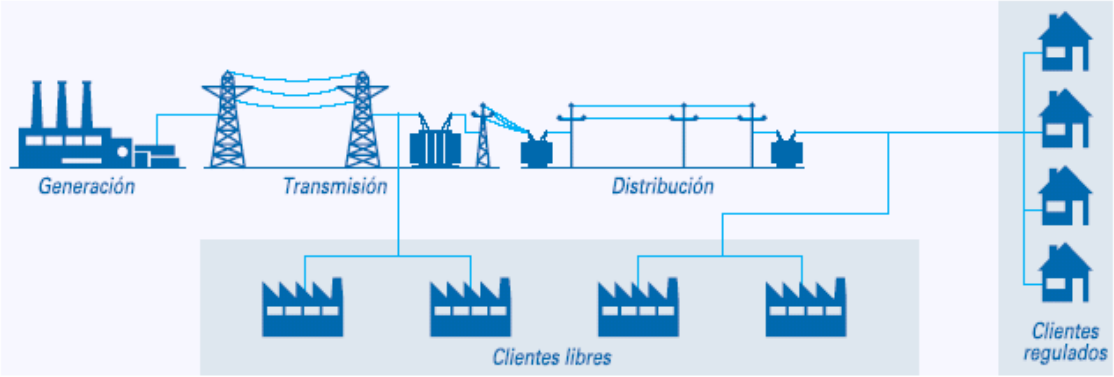


Ilustración 1: Esquema de Mercado Eléctrico / Fuente: www.claseejecutiva.cl

Generación: Encargados de la generación de la energía eléctrica mediante distintos tipos de centrales, como hidroeléctricas, térmicas, eólicas, etc.

Transmisión: Rama encargada de realizar el transporte de energía eléctrica desde las compañías generadoras hasta las distribuidoras a través de líneas de transmisión, subestaciones y otros equipos.

Distribución: Consiste en llevar la electricidad al cliente final, mediante una baja en la tensión eléctrica para que la energía sea apta para el consumo diario de los clientes en su área de concesión.

Consumidores: Son los clientes de las empresas eléctricas que realizan el consumo de la energía, como casas, oficinas, etc.

En base a la magnitud de la demanda de éstos, los consumidores son clasificados entre clientes regulados y clientes libres, a los cuales les aplican distintas tarifas. Cabe destacar que los clientes regulados son clientes de las distribuidoras, mientras que los libres no lo son necesariamente.

Reguladores: La estructura del mercado eléctrico presenta un tendencia hacia un monopolio natural, luego la presencia de las entidades reguladoras tienen como objetivo velar por los consumidores,

de manera que imponen las tarifas a cobrar a éstos (regulados), además de fijar distintas exigencias a las empresas del mercado.

"El principal organismo del Estado que participa en la regulación del sector eléctrico en Chile es la Comisión Nacional de Energía (CNE), quien se encarga de elaborar y coordinar los planes, políticas y normas necesarias para el buen funcionamiento y desarrollo del sector energético nacional, velar por su cumplimiento y asesorar a los organismos de Gobierno en todas aquellas materias relacionadas con la energía" [50].

En este contexto, y para abastecer la demanda de sus clientes, Chilectra S.A. realiza actualmente un plan quinquenal de inversiones, el cual se basa en un modelo de proyección de demanda de energía y potencia eléctrica.

El presente trabajo está dedicado a la proyección de energía y de la demanda de potencia para la red de concesión de Chilectra S.A. en el largo plazo (5 años), con el objetivo de disminuir el error de pronóstico para ayudar a realizar una planificación óptima.

1.1 Definición del Problema

Anterior a la realización de este trabajo, la empresa contaba con los siguientes modelos de proyección:

- Modelo de Demanda Energía Eléctrica en el Sistema
- Modelo de Demanda de Potencia Eléctrica Máxima en el Anillo

Estos modelos eran utilizados para realizar la proyección a 5 años, describiendo cada uno de estos el comportamiento mensual de ambas variables de interés (cada modelo entrega la demanda eléctrica total mensual y la potencia eléctrica máxima mensual, en un horizonte de 60 meses a futuro). En base a los resultados de estos modelos, y en específico, a las tasas de crecimiento anuales respecto de la demanda total de energía en el sistema y la potencia máxima registrada en el año, se define gran parte de la inversión a realizar por la empresa.

Dados estos resultados de proyecciones, una subestimación de la demanda se considera como un caso más crítico que sobreestimación, ya que la empresa distribuidora debe cumplir con la normativa legal vigente que establece la obligatoriedad de las empresas distribuidoras a dar suministro a quien lo solicite dentro de su zona de concesión.

Por lo tanto, la subestimación de la demanda implica el recontratar inversiones haciéndolas más costosas y también arriesgar el suministros eléctrico entregado a los clientes, mientras que la sobreestimación lleva a obtener capacidad ociosa de las instalaciones, además de implicar en algunos casos tener dineros retenidos dentro de la empresa por errores en la confección de presupuestos.

Debido a estos factores, el mejoramiento de las proyecciones a realizar es siempre un objetivo para la empresa, por lo cual se propuso realizar un nuevo modelo que permitiera mejorar el actual desempeño. Además de esto, también surgió la solicitud de generar modelos nuevos para la proyección de demanda de energía para los siguientes sectores de clientes:

- Residenciales
- Comerciales
- Industriales

Para la realización de estas tareas la empresa proveyó de los datos históricos de demanda para todas las series de interés, esto es para la demanda de energía en el sistema, la demanda histórica de potencia máxima, así como también para las demandas de energía sectoriales.

Un punto clave a tratar en este trabajo es la validación estadística de los modelos propuestos, de manera de tener métricas claras de comparación en el desempeño.

1.2 Objetivos Generales y Específicos

1.2.1 Objetivo General

Desarrollar modelos de proyección de Demanda de Energía y Potencia Máxima en el Largo plazo para la red de Chilectra S.A., con el fin de mejorar la proyección que se realiza actualmente.

1.2.2 Objetivos Específicos

Los objetivos específicos que se tienen para este trabajo de tesis son los siguientes:

- Desarrollar modelos de proyección para las demandas de energía eléctrica sectoriales (Residencial, Comercial e Industrial).

- Realizar la validación estadísticas de los modelos de la empresa y también de los nuevos modelos, con el fin de tener métricas claras de comparación.

- Análisis de resultados para entender los factores que explican el comportamiento de la demanda de Energía y Potencia Máxima de la red.

1.3 Relevancia del Tema de Tesis

La proyección de demanda es una tarea crucial para cada empresa, y en Chilectra en específico, ésta permite en el largo plazo determinar:

- Proyección de gastos e ingresos
- Realización de contratos con compañías generadoras para la compra de energía
- Definir el plan de inversión quinquenal

Las decisiones estratégicas de la empresa son derivadas en gran parte de la proyección, sobre todo en lo que respecta a las inversiones a realizar en la red. En consecuencia de esto, el error en el que puede incurrir un pronóstico conlleva a distintitas situaciones en base a si se sobreestima o subestima.

Parte de los errores asociados a una sobreestimación son:

- Sobre utilización de recursos
- Implementación de una red sobre-dimensionada

Mientras que los asociados a una subestimación son:

- Arriesgar calidad de servicio
- Compra de energía a mayor precio

En el último caso, el arriesgar la calidad de servicio puede transformarse en multas por no suministrar energía a los clientes, además que los equipos de la red pueden incurrir en fallas que los deje fuera de funcionamiento, implicando además que se deba gastar en reparación y/ó reemplazo de estos, además de dañar la imagen de la empresa.

Por otra parte, las correctas estimaciones de demandas futuras le permiten a la empresa realizar mejores contratos para la compra de energía.

Tomando esto en cuenta, el mejorar de la proyección de demanda es un primer paso para obtener mejores resultados en el largo plazo para Chilectra S.A.

1.4 Metodología

En esta sección se hará una breve descripción de la metodología aplicada, enmarcando esto en el proceso de extracción de conocimiento a partir de datos, conocido como KDD (Knowledge Discovery in Databases)[4] .

El proceso KDD tiene como objetivo el extraer conocimiento a través de la transformación de datos de bajo nivel (datos muy numerosos que no son de fácil asimilación) en información que pueda ser más compacta, abstracta ó más útil, dependiendo del problema al que se está haciendo frente [4].

Esta metodología tiene un total de 5 etapas, las que son mencionadas a continuación, incluyendo una descripción breve de las tareas realizadas en este trabajo.

1. Selección de los datos a partir de los cuales se extraerá conocimiento.

Los datos utilizados en esta tesis corresponden las demandas mensuales históricas de Energía y Potencia Eléctrica Máxima mensual, en el período correspondiente a Junio 2001 hasta Agosto 2014. También se utilizaron datos climatológicos y del IMACEC en el mismo período para la confección de modelos multivariados.

2. Limpieza y preprocesamiento.

De ser necesario, en esta etapa se remueven los datos erróneos y se define el procedimiento frente a datos faltantes. Estas tareas no se requirieron en los ejercicios desarrollados.

3. Aplicación de transformaciones que faciliten la manipulación, interpretación o modelamiento de los datos.

Los datos históricos fueron todos transformados a la misma escala antes de ser ingresados a los métodos de proyección finales. La escala utilizada fue de [-0,6;0,6] con el objetivo de no saturar la funciones de las redes neuronales.

4. Minería de Datos: aplicación de herramientas analíticas en la búsqueda de patrones e información relevante.

Se aplicaron modelos de Support Vector Regression y Redes Neuronales (multilayer perceptron), probando diversas combinaciones de parámetros de diseño (número de neuronas en la capa de entrada, número de neuronas en la capa oculta, fracción de datos muestrales en los conjunto de entrenamiento y validación). Junto con lo anterior, se desarrollaron modelos SARIMA y SARIMAX para comparar el desempeño respecto de métodos tradicionales de pronóstico.

5. Interpretación y evaluación.

Se evaluó el desempeño predictivo de los modelos construidos en un total de 5 horizontes de pronóstico, de los cuales 3 corresponden a un período de 60 meses, uno a 48 y otro a 36 meses (todos estos horizontes corresponden a distintas particiones del periodo compuesto desde enero de 2007 hasta diciembre de 2013). Los mejores modelos fueron seleccionados de acuerdo al menor MAPE, y que también tratando que cumplieran con la condición tener errores homocedásticos y no autocorrelacionados.

1.5 Resultados Esperados

Se espera poder determinar la metodología para generar una predicción de la demanda de Energía y de Potencia Máxima con la mayor precisión posible respecto de las distintas validaciones que se realizarán. A su vez se espera aportar evidencias respecto al potencial de las técnicas de inteligencia artificial para el pronóstico a largo plazo, ya que en la actualidad estas técnicas son utilizadas mayoritariamente para el pronóstico a corto plazo.

Además de esto es fundamental que esta metodología sea implementada en la empresa, por lo cual no se puede esperar crear un proceso que sea una caja negra, por el contrario, se pretende que la metodología generada se convierta en un sustento principal para la planificación agregada de la cadena productiva de Chilectra S.A.

2. Análisis de la Situación Actual

Con el fin de entender de mejor manera la actualidad de la empresa y la problemática adherente a la proyección de demanda, en éste capítulo se aborda una descripción de la empresa y como su demanda ha ido evolucionando a través de los años, además de realizar una medición del desempeño de la metodología actual que utiliza la empresa para realizar sus proyecciones.

2.1 La Empresa

2.1.1 Historia

Chilectra S.A. es una empresa privada Chilena de distribución eléctrica cuya zona de concesión pertenece a la región Metropolitana de Santiago.

El origen de la compañía se remonta al 1 de agosto de 1921, fecha en que se fusionaron las empresas Chilean Electric Tramway and Light Co. (fundada en 1889) y la Compañía Nacional de Fuerza Eléctrica, que operaba desde 1919 en Santiago, creando así la Compañía Chilena de Electricidad Ltda, o también llamada Chilectra.

La empresa fue estatizada el 14 de agosto de 1970, mediante la Ley n° 17.323 que autorizó a la Corporación de Fomento de la Producción (Corfo) a adquirir todas las acciones y bienes, y desde 1971 se llamó Compañía Chilena de Electricidad S.A.

Trece años después, la compañía inició un proceso de reprivatización que culminó en Agosto de 1987 con la totalidad del capital accionario en el sector privado. Producto de este proceso, en noviembre de 1987, se creó la primera filial de la compañía; Distribuidora Chilectra Metropolitana S.A., que en mayo de 1994 se constituyó como: Chilectra S.A.

Actualmente el 99,1% de la empresa pertenece a Enersis S.A., una de las principales multinacionales eléctricas privadas de Latinoamérica, con participación directa e indirecta en los negocio de generación, transmisión y distribución de la energía eléctrica. Enersis además es controlado por el Grupo Enel.

2.1.2 Zona de Concesión

Chilectra distribuye energía actualmente en 33 comunas de la Región Metropolitana, en una zona de concesión de 2.118 km2. El nivel de venta física de energía, 12.191 GWh total[2], la convierte en la empresa de distribución de energía eléctrica más grande de Chile.

Entre sus clientes figuran industrias, grandes empresas, comercio, entidades fiscales y domicilios particulares. Además la compañía posee dos filiales: Luz Andes Ltda. y Empresa Eléctrica Colina Ltda.

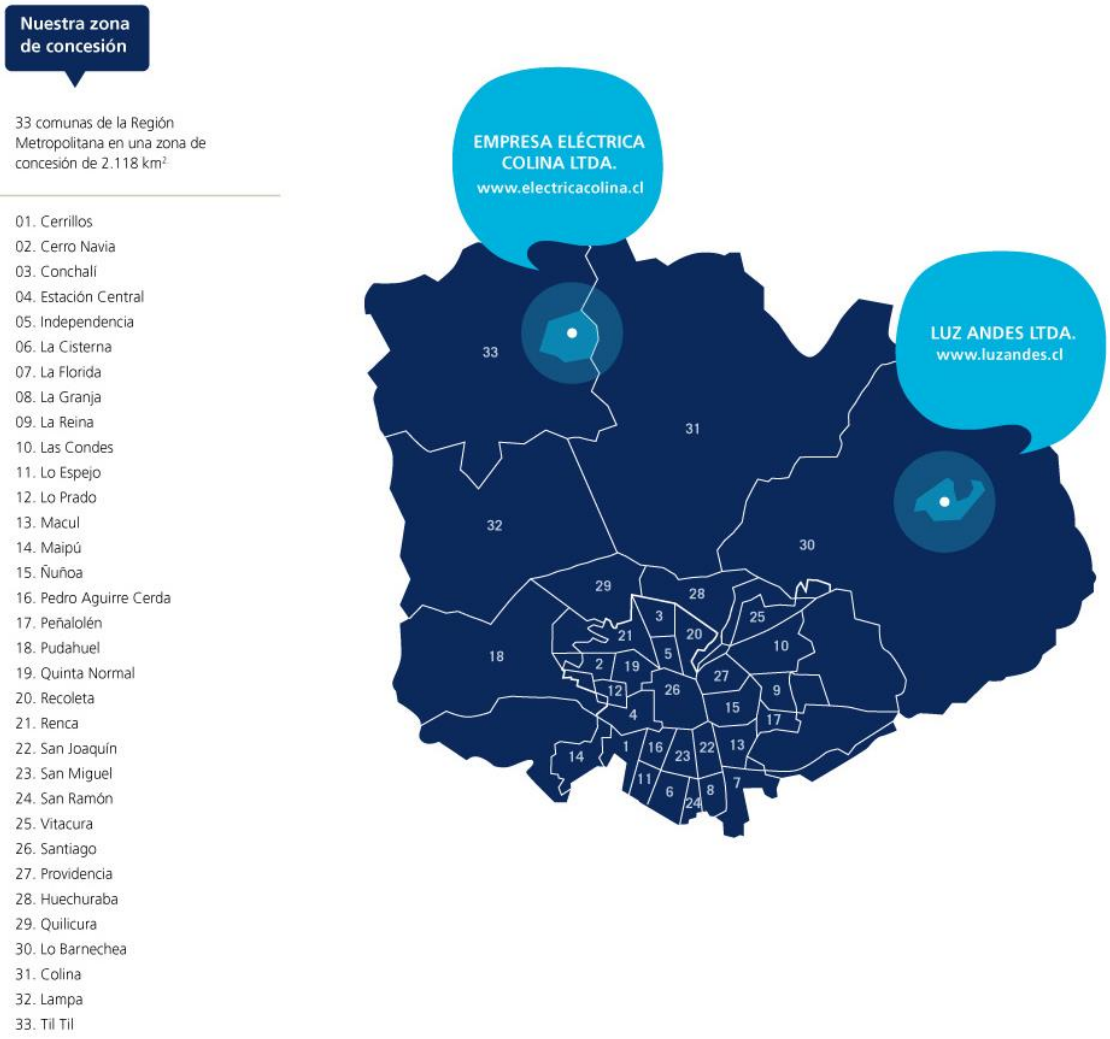


Ilustración 2: Área de Concesión de Chilectra S.A.
Fuente: www.chilectra.cl

[2] Venta correspondiente al año 2013

En esta área, la cual incluye a las filiales, se mide la totalidad de la demanda energética, dado que bajo esta medición la empresa debe de realizar los contratos para la compra de energía con las compañías generadoras. Por otra parte, otra importante métrica para la empresa consiste en la potencia eléctrica que se presenta en sus redes, no obstante para los objetivos de este trabajo se realiza una distinción respecto del área donde esta se mide. En este caso, la potencia se mide en una red que forma un anillo en alta tensión.el cual comprende gran parte del área de concesión.

Se habla entonces del "anillo de subtransmisión", compuesto por una red en AT recibiendo suministro desde seis puntos de interconexión. Las subestaciones de poder abastecidas dan suministro principalmente a la zona urbana de la ciudad de Santiago, concentrando aproximadamente el 94% de la demanda total del sistema.

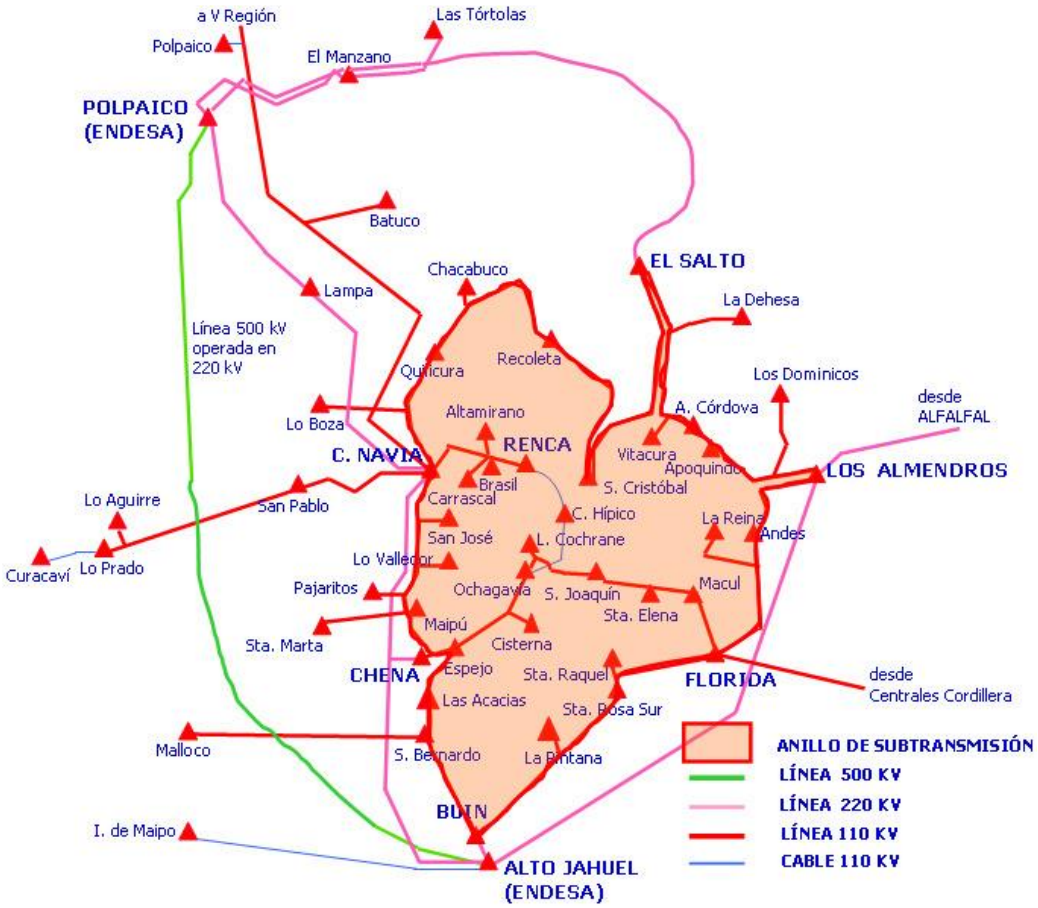


Ilustración 3: Red AT de Chilectra S.A.
Fuente: Chilectra S.A.

2.2 Datos Históricos

Para comprender de mejor forma el comportamiento de la demanda que enfrenta la empresa, a continuación se presentarán los datos respectivos a la evolución de las demandas históricas, no obstante las unidades respectivas están normalizadas con el fin de no revelar información privada de la empresa.

También se agrega como dato relevante a cada serie el coeficiente de variación, que representa el cociente entre la desviación estándar de la serie y su media. Este valor es utilizado como una métrica para ver la estabilidad de la serie a través del tiempo.

2.2.1 Demanda Energía Sistema

La demanda de energía en el sistema (es decir, toda el área de concesión) se presenta en el siguiente gráfico, así como también una tabla donde se muestran los porcentajes de crecimiento en la demanda respecto a cada año (calculada en base a la relación de las demandas totales ocurridas).

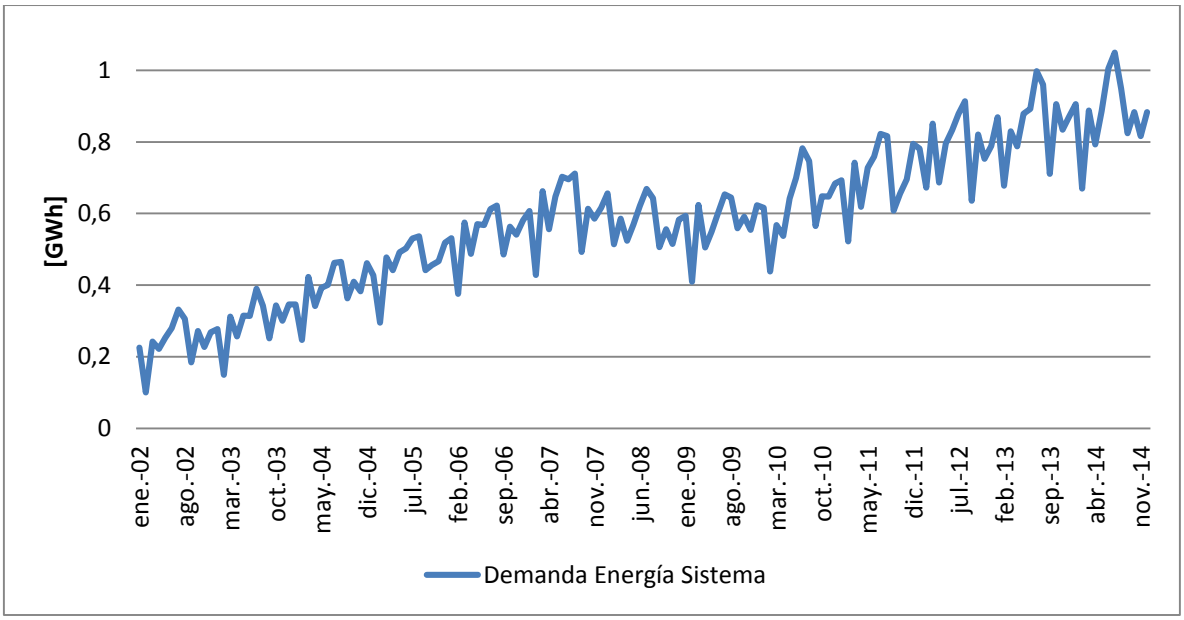


Gráfico 1: Demanda de Energía del Sistema

Coeficiente de Variación: 15,22%

Año	Crecimiento
2003	5,34%
2004	8,07%
2005	6,09%
2006	5,98%
2007	4,90%
2008	-2,19%
2009	-0,21%
2010	3,95%
2011	5,04%
2012	5,19%
2013	4,15%
2014	2,09%

Tabla 1: Crecimientos Anuales de Demanda de Energía Sistema

El crecimiento de la demanda de energía en el sistema presenta a principio de la década del 2000 un crecimiento estables con un peak considerable el año 2004, siendo detectable una tendencia de crecimiento bastante estable a simple vista, así como también una estacionalidad estable a través de los años, sin variaciones muy grandes respecto de la media anual.

Sin embargo el año 2008 y 2009 desaparece la tendencia, debido a la crisis energética que ocurre en el primero de estos años que ocasionó una baja considerable en la demanda (acompañada también de una crisis económica mundial) y además un aumento prominente en los precios de la energía dada la falta de generación en el país.

Se puede ver que posterior a la crisis se vuelve a retomar una tendencia a la alza, aunque con una estacionalidad con mayor varianza, presentándose máximos más marcados en los meses de invierno. Cabe destacar también la baja en el crecimiento que ocurre el último año, que es contemporánea a una baja de crecimiento en la actividad económica del país.[3]

El coeficiente de variación para esta serie es relativamente bajo[4], no obstante ha de compararse con otras series en el largo plazo para

[3] A Enero de 2015, se espera que el crecimiento del PIB del país cierre en un 1,8%
[4] Cuando es menor a 1 se considera que una serie es de baja varianza.

determinar su dificultad de pronóstico, razón por la cual se analizará este ítem más adelante.

Se incluye por último un gráfico de cajas que muestra la dispersión que tiene la estacionalidad de cada mes asociado a la serie. Cabe destacar que estos datos representan únicamente la estacionalidad, dado que la tendencia ha sido extraída mediante una media móvil.

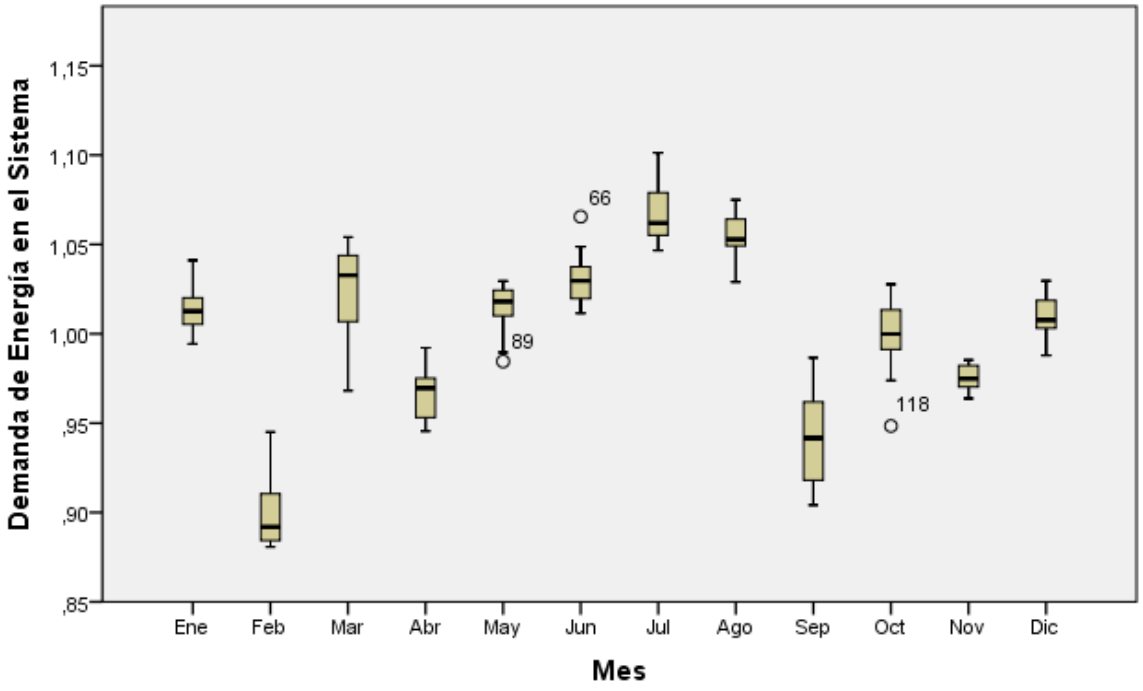


Gráfico 2: Dispersión de estacionalidad de Demanda de Energía en el sistema

2.2.2 Demanda Energía Residencial

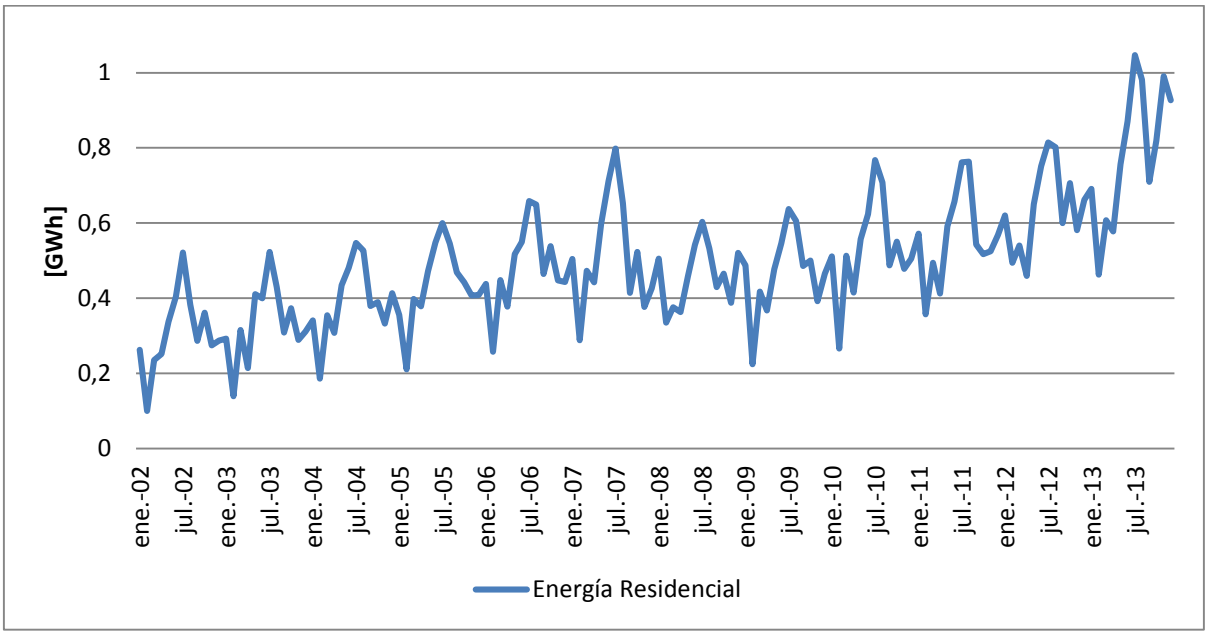


Gráfico 3: Demanda de Energía Residencial

Coeficiente de variación: 14,09%

Año	Crecimiento
2003	2,42%
2004	5,35%
2005	4,00%
2006	3,97%
2007	2,87%
2008	-4,59%
2009	0,60%
2010	5,39%
2011	2,49%
2012	5,92%
2013	10,67%

Tabla 2: Crecimientos Anuales de Demanda de Energía Residencial

La demanda residencial presenta una tendencia muy parecida respecto de la demanda total del sistema, al presentar la misma baja en el año 2008 y una desaceleración el 2009, y una tendencia de crecimiento en el resto de las fechas.

La estacionalidad en esta curva está caracterizada por un fuerte peak en los meses de invierno, y por otra parte un valle que ocurre en general en Febrero, asociable al mes típico de vacaciones.

El año 2013 ocurre un crecimiento muy anormal, sobre todo en los meses de invierno, donde se "rompe" la pendiente de crecimiento, pero solo en el sentido en que ésta estaría tomando una razón de cambio aún mayor que la anterior.

2.2.3 Demanda Energía Comercial

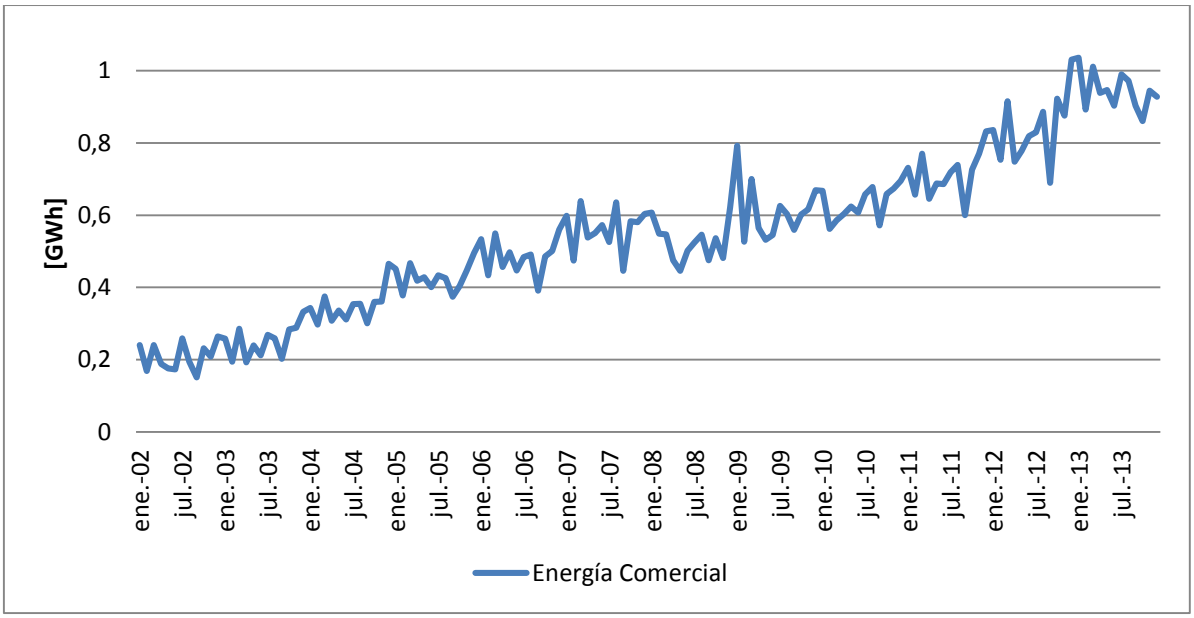


Gráfico 4: Demanda de Energía Comercial

Año	Crecimiento
2003	6,82%
2004	14,07%
2005	10,29%
2006	6,88%
2007	8,36%
2008	-3,67%
2009	8,93%
2010	2,00%
2011	7,70%
2012	11,09%
2013	8,15%

Tabla 3: Crecimientos Anuales de Demanda de Energía Comercial

Coeficiente de Variación: 23,90%

Esta demanda sectorial no presenta una estacionalidad clara, presentando a través de los años crecimientos no tan similares como las dos series tratadas previamente.

Al igual que antes, y demostrando el efecto global y transversal que tuvo la crisis energética, la tendencia de crecimiento se presenta en todos los años menos en el 2008. Aun así, el año 2009 ocurre un peak de demanda en el mes de febrero que dista de las variaciones típicas de otros años, siendo tal efecto que este nivel de demanda mensual llega a ser igualado recién el año 2011.

Dada la irregularidad de la curva de demanda, es esperable que al aplicar modelos de proyección para la serie no se obtengan los mejores resultados residuales, dificultando por ejemplo el objetivo de que el modelo de proyección presente errores homocedásticos.

2.2.4 Demanda Energía Industrial

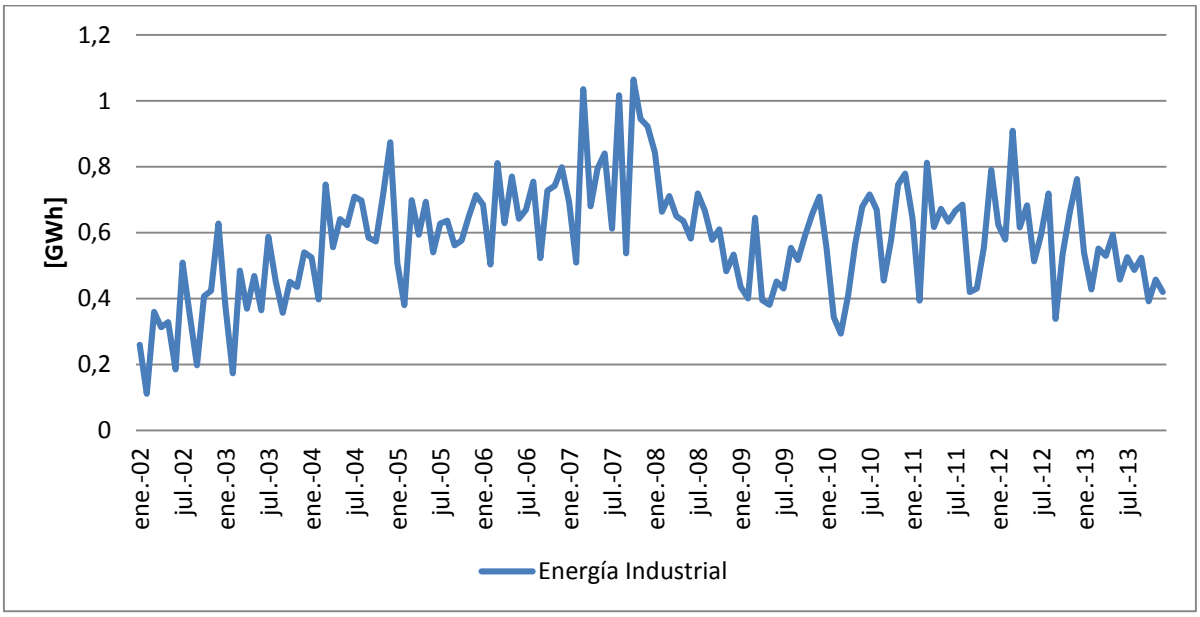


Gráfico 5: Demanda de Energía Industrial

Coeficiente de Variación: 8,04%

Año	Crecimiento
2003	4,15%
2004	10,46%
2005	-1,69%
2006	4,03%
2007	5,00%
2008	-6,75%
2009	-5,52%
2010	2,41%
2011	2,00%
2012	0,81%
2013	-6,01%

Tabla 4: Crecimientos Anuales de Demanda de Energía Industrial

El sector industrial posee un comportamiento muy distinto a las otras demandas sectoriales, presentándose diferencias no solo por el hecho de que no posee una tendencia de crecimiento estable posterior al año 2008, sino que además en años previos a la crisis también se presentan bajas en el consumo (año 2005).

A su vez no se puede apreciar una estacionalidad clara a través de los años, donde el único factor que resalta es que en el mes de marzo se presenta en general un peak en la demanda, aunque no necesariamente es la demanda máxima anual. Este fenómeno presumiblemente se asocia al hecho de que en el mes de Febrero debiese de esperarse una baja en el consumo por la menor actividad económica en la región debido a la "pausa" provocada por las vacaciones, y luego en el mes de marzo se vuelve a tomar el ritmo de trabajo normal.

Al igual que con la demanda comercial, la irregularidad de la curva presumiblemente dificultará la capacidad de pronóstico al aplicar modelo de proyección lineales.

2.2.5 Demanda Potencia Máxima Anillo

A diferencia de las series anteriores, para este caso estamos tratando con una medición física distinta que es la potencia. En la serie que se muestra a continuación, se aprecian los valores de demanda máxima de potencia ocurridos en cada uno de los meses. Específicamente, esta serie contiene las mediciones de la mayor demanda de energía en el "anillo de subtransmisión" ocurrida en un instante.

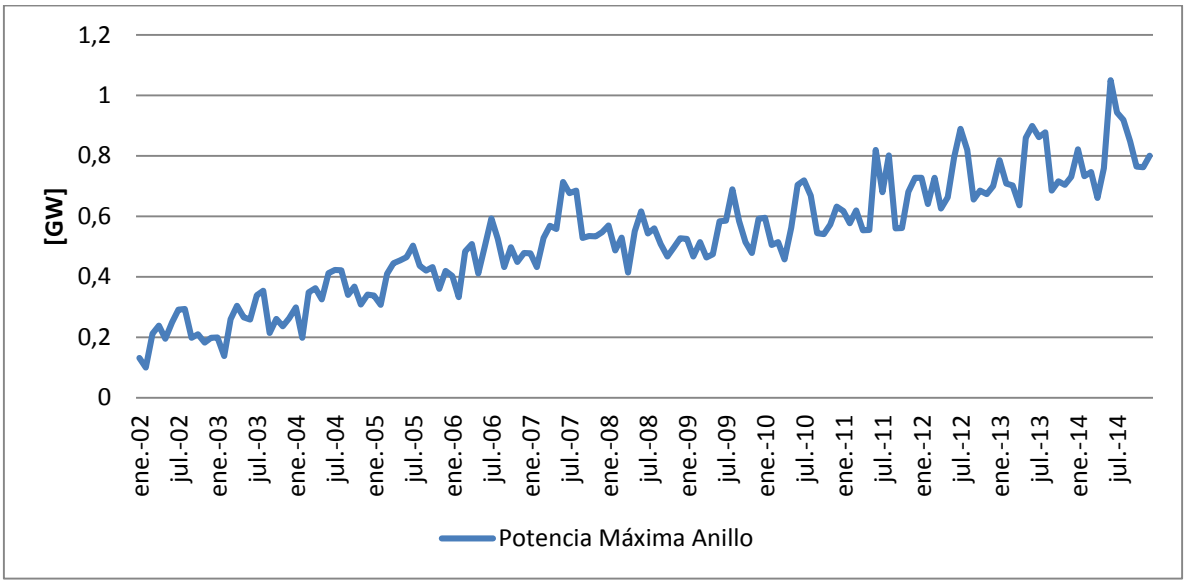


Gráfico 6: Demanda de Potencia Máxima Anillo

Coeficiente de Variación: 14,73%

Año	Crecimiento
2003	4,10%
2004	6,98%
2005	6,31%
2006	6,59%
2007	8,27%
2008	-1,82%
2009	-6,38%
2010	9,19%
2011	6,37%
2012	4,15%
2013	0,52%
2014	8,63%

Tabla 5: Crecimientos Anuales de Demanda de Potencia Máxima en el Anillo

Lo último mencionado hace que esta serie no sea comparable con la medición de las otras no solamente por su unidad física, sino porque ésta presenta lo ocurrido en un instante y no el total (o la suma) de un mes.

Analizando la serie en cuestión, se puede apreciar que al igual que con la demanda de energía del sistema existe una tendencia clara al crecimiento a excepción de los años de crisis energética en el 2008.

Si bien la tendencia es marcada, los crecimientos respectos de la potencia máxima anual año a año no son tan estables, presentándose peaks de crecimiento cercanos al 9%. En términos de crecimientos absolutos, el caso del 2014 es completamente anormal, dado que se desvía fuertemente del crecimiento promedio absoluto[5]. Las causales propuestas para explicar este repentino crecimiento hacen referencia a la entrada de la calefacción eléctrica en los últimos años, que ha ganado terreno sobre la calefacción a parafina y a gas.

Se presentan además claros peaks de demanda en los meses de invierno, siendo siempre este período donde ocurren las demandas máximas anuales. A su vez se presenta otro peak en verano, no obstante este nunca ha superado al de invierno, aunque casi ocurre esto el año 2013, con un peak ocurrido en Noviembre.

El comportamiento de la estacionalidad se ve reflejado en el siguiente gráfico de cajas, donde se aprecia claramente el peak que ocurre en los meses de invierno (Junio, Julio y Agosto).

[5] Por razones de confidencialidad, este valor no será revelado en este trabajo.

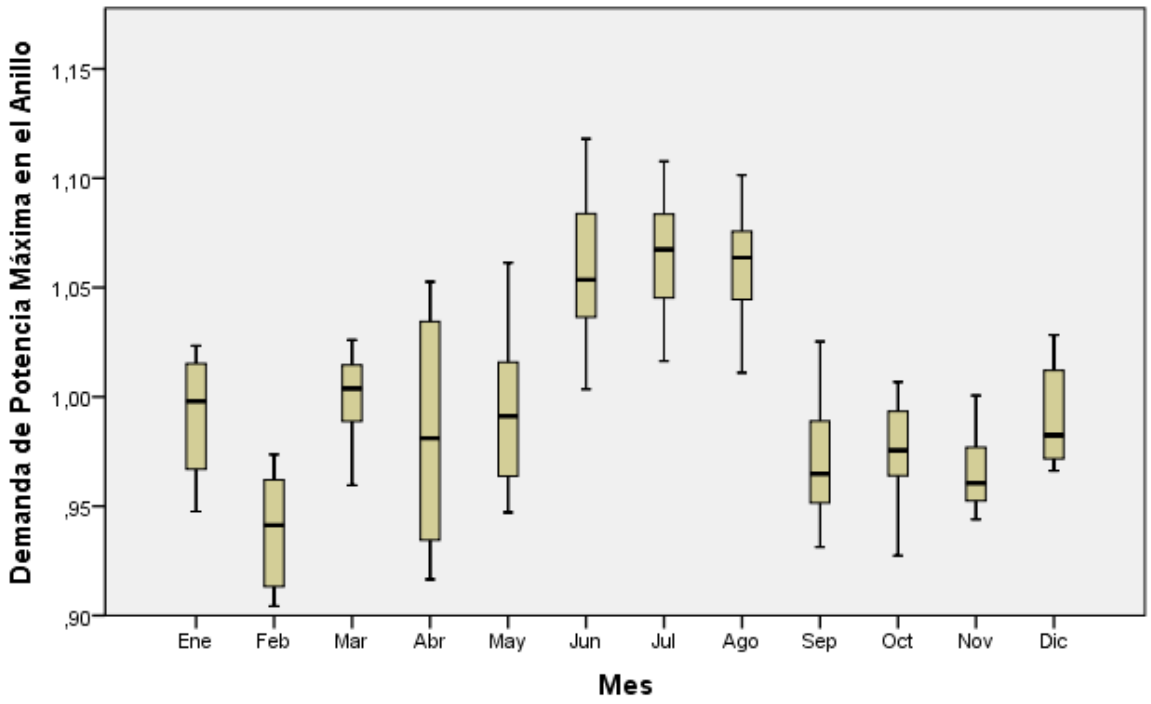


Ilustración 4: Dispersión de estacionalidad de Demanda de Potencia Máxima en el Anillo

2.2.6 Varianza de las Series

En esta sección se compararan las varianzas que tiene cada serie, de manera de visualizar la estabilidad de su comportamiento a través del tiempo. A continuación se mostrarán los coeficientes de variación que presentan las series, y así también los coeficientes de variación que presentan los meses de la serie de estacionalidad de cada serie[6].

Coeficiente de Variación				
Energía	Residencial	Comercial	Industrial	Potencia
15,22%	14,09%	23,90%	8,04%	14,73%

Tabla 6: Coeficiente de variación para cada serie

Coeficiente de Variación de Estacionalidad				
Energía	Residencial	Comercial	Industrial	Potencia
4,59%	8,35%	4,68%	3,78%	3,95%

Tabla 7: Coeficientes de variación para las estacionalidades de cada serie

[6] La serie de estacionalidad de cada serie se obtuvo a través de una descomposición multiplicativa de cada serie.

Coeficiente de Variación					
Meses	**Energía**	**Residencial**	**Comercial**	**Industrial**	**Potencia**
Ene	1,22%	2,95%	1,76%	2,13%	2,65%
Feb	2,41%	2,38%	4,09%	3,28%	2,68%
Mar	2,36%	4,15%	1,95%	6,15%	2,12%
Abr	1,57%	2,30%	3,63%	1,96%	5,42%
May	1,33%	2,59%	2,17%	2,50%	3,24%
Jun	1,73%	2,06%	3,17%	3,23%	3,57%
Jul	1,52%	2,32%	2,02%	4,78%	2,59%
Ago	1,15%	2,77%	2,78%	2,08%	2,36%
Sep	2,56%	2,30%	2,44%	4,28%	2,57%
Oct	2,04%	2,54%	3,00%	4,76%	2,16%
Nov	0,72%	2,49%	2,32%	3,53%	1,70%
Dic	1,14%	1,80%	2,83%	3,47%	2,19%

Tabla 8: Coeficiente de Variación mensual de la estacionalidad para cada serie

De la primera tabla se puede ver que las series son estables al presentar coeficientes similares, y todos menores al 25%. No obstante este coeficiente demuestra no ser tan representativo respecto de la capacidad de pronóstico de cada serie, dado que la serie "industrial" presenta el menor coeficiente, pero a la vez siendo la que presenta un comportamiento más aleatorio y menos similar a la demás.

Los coeficientes de variación de la estacionalidad de cada serie entrega una métrica para la dificultad de pronóstico a enfrentar en los comportamientos repetitivos en que incurren cada serie. La demanda residencial presenta un mayor coeficiente, no obstante, los valores debajo de un 10% muestran que las variaciones entre los meses no son tan grandes, aunque no por esto dejan de ser significativas.

Respecto de la variación mes a mes, se puede ver que la estacionalidad de las series es bastante estable a través del tiempo, existiendo un máximo de varianza de un 6,15%, correspondiente a la demanda industrial en el mes de marzo. Estas bajas varianzas se traducen en que la tendencia intrínseca de cada serie no potencia la varianza de la estacionalidad.

2.3 Modelos Actuales

Para la realización de pronósticos, la compañía actualmente ejecuta dos modelos de regresión lineal múltiple mediante mínimos cuadrados ordinarios (MCO). Las proyecciones solo se realizan para dos series:

- Demanda de Energía de Sistema
- Demanda Máxima de Potencia en el Anillo

Estos modelos son utilizados para realizar una proyección a 5 años del comportamiento de estas series, obteniendo como resultado las demandas respectivas para cada uno de los meses futuros en el horizonte planteado.

Las variables explicativas utilizadas en estos modelos hacen una clara referencia a lo que es mencionado muchas veces en la bibliografía [43], tomando en cuenta la utilización de variables Socioeconómicas, basadas principalmente en el PIB, variables climatológicas como la temperatura, y además se utiliza el precio de la energía. En el caso de la demanda máxima de Potencia en el anillo, se utiliza como variable explicativa la demanda de energía del sistema proyectada.

Cabe destacar que para la utilización de este tipo de modelos se requiere la proyección de las variables explicativas a utilizar.

2.3.1 Evaluación Modelos Actuales

Con el fin de imponer una métrica para la comparación, se realizó la validación de los modelos actuales que implementa la empresa para medir su desempeño.

Para este caso se utilizará como métrica de error la media del error absoluto porcentual (más conocido como MAPE: Mean Absolute percentage error), el cual tiene la siguiente fórmula:

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}|(A_t - F_t)/A_t|$$

Dónde:

- A_t: Valor actual de la serie en el período u observación t
- F_t: Valor Pronosticado para el período u observación t

La evaluación del desempeño se realizó haciendo que el modelo "prediga" el pasado para así medir el error. Dado que se realiza una proyección a 5 años, se ejecutó el modelo para predecir los siguientes horizontes de pronóstico:

- Horizonte 1: Enero 2009 a Diciembre 2013
- Horizonte 2: Enero 2008 a Diciembre 2012
- Horizonte 3: Enero 2007 a Diciembre 2011
- Horizonte 4: Enero 2007 a Diciembre 2010
- Horizonte 5: Enero 2007 a Diciembre 2009

Los datos utilizados para el entrenamiento del modelo corresponden a los siguientes para cada horizonte:

- Horizonte 1: Junio 2001 a Diciembre 2008 (79 datos)
- Horizonte 2: Junio 2001 a Diciembre 2007 (67 datos)
- Horizonte 3: Junio 2001 a Diciembre 2006 (55 datos)
- Horizonte 4: Junio 2001 a Diciembre 2005 (43 datos)
- Horizonte 5: Junio 2001 a Diciembre 2004 (31 datos)

La utilización de estos datos de entrenamiento implica que para los últimos periodos de evaluación no se toma en cuenta el error del primer año para el horizonte 4 y 5, a su vez que no se considera el error del segundo año de pronóstico para el horizonte 5.

En estos horizontes se hará el cálculo del MAPE para finalmente asignar un error promedio al modelo entre los horizontes. Los modelos son ejecutados mediante el software E-views.

Modelo de Demanda de Energía del Sistema

A continuación se muestran los resultados de las ejecuciones del modelo en los distintos horizontes para el modelo de Energía.

Energía Sistema	2009-2013	2008-2012	2007-2011	2007-2010	2007-2009	Promedio
Modelo Chilectra	3,00%	8,71%	7,08%	7,34%	4,95%	**6,22%**

El error promedio del modelo en los distintos horizontes es de 6,22%. Cabe destacar que este error es bueno considerando que en promedio un error de un 10% es aceptable para las proyecciones de largo plazo en general [43].

Modelo de Demanda de Potencia Máxima Anillo

Al igual que antes, en esta sección se detallarán los resultados de las ejecuciones del modelo de Potencia con su desempeño.

Potencia Máxima	2009-2013	2008-2012	2007-2011	2007-2010	2007-2009	Promedio
Modelo Previo	3,68%	7,89%	5,78%	4,33%	4,30%	**5,20%**

El modelo de Potencia muestra sistemáticamente mejores resultados que su par relacionado a la Energía, inclusive tomando en cuenta que el modelo de Potencia máxima utiliza las proyecciones de Energía como factor explicativo.

El error promedio en los horizontes es de 5,2%, 1 punto menos que el modelo de energía.

3. Marco Teórico

En este capítulo se expondrán las metodologías pertinentes para llevar a cabo una proyección de demanda, las cuales fueron utilizadas en este trabajo de tesis.

Los métodos de pronóstico utilizados se pueden dividir en dos conjuntos:

- **Univariados**
- **Multivariados**

Los métodos univariados son aquellos que asumen que el comportamiento futuro de la variable a pronosticar sólo depende de su comportamiento en el pasado, mientras que los métodos multivariados consideran que la variable a pronosticar puede ser explicada mediante su relación con otras variables relacionadas, de las cuales además podemos conocer su comportamiento en el futuro.

En las siguientes secciones se detallarán las técnicas de métodos univariados y multivariados utilizados en este trabajo.

3.1 Métodos Univariados

Este enfoque busca explicar el comportamiento futuro de una serie cronológica de datos en base a su pasado, es decir, dejando de lado variables exógenas para obtener una noción limpia de cómo debería comportarse de forma intrínseca (endógena).

3.1.1 Promedios Móviles

La técnica de promedio móvil consiste en utilizar como pronóstico para el próximo período el promedio de una cantidad fija de datos [7]. El término "móvil" se acuña dado que para el siguiente periodo a pronosticar, entrará un dato nuevo y saldrá el dato más antiguo que fue utilizado en la proyección anterior, luego para realizar los siguientes pronósticos el promedio se "mueve" a través de los datos.

De esta manera, se utiliza como pronóstico para el siguiente periodo el promedio de los n valores pertenecientes a los datos más recientes de la serie de tiempo. Utilizando una expresión matemática, se obtiene:

$$Y_{t+1} = \frac{1}{n}\sum_{i=0}^{n-1} Y_{t-i}$$

Dónde:

- Y_t: Valor de la serie en el periodo t
- n: Número de observaciones a considerar

En la Ecuación, el término n indica que, conforme se tiene una nueva observación de la serie de tiempo, se reemplaza la más antigua de la ecuación y se calcula un nuevo promedio. En este caso se supone que todas las observaciones de la serie de tiempo son igualmente importantes para la estimación del parámetro a pronosticar, no obstante existen variaciones de éste método.

El método mostrado anteriormente se denomina **promedio móvil simple**, pero existe otro método llamado **promedio móvil ponderado** [7], donde no todas las observaciones utilizadas tienen la misma importancia. La ecuación que describe a este último método está dada por:

$$Y_{t+1} = \sum_{i=0}^{n-1} Y_{t-i} * w_i \qquad con \sum_{i=0}^{n-1} w_i = 1$$

Al agregar los ponderadores w_i, se puede dar más preponderancia a observaciones que se consideran más importantes para predecir el siguiente periodo. En general, siempre se da un mayor al valor al ponderador correspondiente al periodo más reciente a la proyección [7].

El método de promedios móviles es muy útil cuando se tiene información no desagregada, y cuando no se conoce otro método más sofisticado que permita predecir con mayor confianza.

3.1.2 Modelos ARIMA[7]

Los modelos ARIMA son la integración de dos tipos de modelos, los auto regresivos y los de media móvil (ARIMA es la sigla en inglés "Auto Regressive Integrated Moving Average), desarrollado por Box-Jenkins, los cuales se basan en el tratamiento de la correlación de la serie.

Cabe destacar que para la utilización de este tipo de modelos, se requiere que la serie de tiempo sea estacionaria (que no tenga una tendencia de crecimiento, por ejemplo) y que satisfaga homocedasticidad (varianza relativamente constante en el tiempo). Si bien estas condiciones parecen sesgar bastante al universo de series de tiempo ante el cual se pueden aplicar estos métodos, este hecho se resuelve aplicando el modelo ARIMA a una transformación de la serie, por ejemplo la diferencia entre periodos, ó simplemente una transformación logarítmica.

Como ya se mencionó, un modelo ARIMA utiliza elementos auto-regresivos (AR) y de medias móviles (MA), así como también órdenes de integración (I).

Cada término AR corresponde al uso de valores rezagados de observaciones pasadas en el modelo de regresión de la observación actual:

$$AR(p): \quad y_t = a_1 y_{t-1} + a_2 y_{t-2} + \cdots + a_p y_{t-p} + \epsilon_t$$

En este caso, los modelos AR tratan de describir el valor futuro de la serie en base a sus observaciones anteriores más un error.

Cada término MA corresponde al uso de valores rezagados de errores pasados en el modelo de regresión de la observación actual, asumiendo este tipo de modelos que el valor futuro de la serie puede ser predicho en base a los errores anteriores.

$$MA(q): \quad y_t = \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q}$$

[7] Basado en Hyndman, Rob J; Athanasopoulos, George. "8.9 Seasonal ARIMA models". Forecasting: principles and practice.

Cada orden de integración I corresponde a una diferenciación de la serie. Si la diferenciación es estacional (trimestres, semestres, años) entonces el modelo se denomina Seasonal-ARIMA o SARIMA, que a su vez puede contener elementos estacionales de auto-regresión (SAR) o medias móviles (SMA):

$$Diferenciación\ ordinaria\colon\ y_t - y_{t-1} = \epsilon_t$$

$$Diferenciación\ estacional\colon\ y_t - y_{t-T} = \epsilon_t$$

La caracterización general de un modelo SARIMA es:

$$SARIMA(p,d,q)x(P,D,Q)_T$$

Donde p es la cantidad de términos auto-regresivos, q de términos de media móvil, P de términos auto-regresivos estacionales, Q de términos de media móvil estacionales, d es el orden de integración ordinaria, D es el orden de integración estacional y T la referencia para la integración estacional. La elección de estos parámetros determinará la precisión del pronóstico, y este proceso se realiza mediante la metodología desarrollada por Box-Jenkins en base al tratamiento de la correlación de la serie.

3.2 Métodos Multivariados

Estas metodologías de pronóstico suponen que la variable de interés es dependiente de otras variables, por ende para pronosticar su valor futuro se debe conocer el tipo de relación que tiene con estas variables que explican su comportamiento, para luego proyectar su valor en el futuro.

Una características de estos métodos, como se mencionó anteriormente, es que dado que proyectan los valores futuros de la variable de interés en base a los valores de las variables explicativas, es necesario conocer el valor futuro de estas variables, y en caso de no conocerlo, se requiere de realizar una proyección de éstas.

3.2.1 Regresión

La regresión es un método en el cual el valor de una variable de interés es expresado mediante una ecuación que involucra a otras variables exógenas. En otras palabras, el valor de la variable de interés

(ó variable dependiente) es igual a una función en base a otras variables exógenas. Esto es:

$$Y = f(X)$$

Dónde:

- Y: Variable de Interés
- X: Variable(s) Explicativa(s)
- f: Función de Regresión

El punto principal de éste método, y causante de la diversificación del mismo, está en la función de regresión, dado que ésta es la que describe la relación entre las variables explicativas y la dependiente.

Una vez que se tiene la función de regresión, y los valores de las variables explicativas a futuro, el pronóstico está listo para realizarse. Sin embargo, la función de regresión "perfecta" no existe (o es muy difícil de encontrar) por lo cual la utilización de esta metodología se basa en proponer una función de regresión que además está acompañada de un error, llegando de esta forma a la siguiente relación:

$$Y = f(X) + \varepsilon$$

Dónde el nuevo término "ε" es un error que acompaña a la función de regresión y que flexibiliza el hecho de que la función exprese la totalidad de la varianza de los valores de la variable de interés.

Tomando esto en cuenta, el método se enfoca en proponer una función de regresión que minimice el error a través de los datos que se tienen disponibles, tanto de la variable objetivo como de las variables explicativas.

La forma más clásica de regresión es la **regresión lineal**, donde la relación entre la variable dependiente y las explicativas queda descrita de la siguiente forma:

$$Y = f(X) + \varepsilon \quad => \quad Y_t = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_n x_n + \varepsilon_t$$

Dónde:

- y es la variable dependiente
- $x_1, x_2, \ldots, x_n$ son las variables explicativas (o regresores)

- $b_0, b_1, \ldots, b_n$ son los coeficientes de regresión (b_0 es llamado el intercepto, o término libre, y corresponde al valor que toma y cuando todas las explicativas valen cero)
- ε_t es el error aleatorio que no puede ser predicho

Cuando este tipo de ecuación es propuesta, los coeficientes de regresión son estimados mediante la metodología de Mínimos Cuadrados Ordinarios (MCO) y pasan a ser el factor principal a estimar para realizar el pronóstico.

La diversificación de éste método se da cuando la ecuación propuesta no es lineal, ante lo cual se utilizan distintas transformaciones para expresar la relación entre las variables explicativas y la dependiente. Algunas de las otras formas de las ecuaciones de regresión proponer son:

- Logarítmica: $Y = b_0 + b_1 \ln(X)$
- Cuadrática: $Y = b_0 + b_1 X^2$
- Cúbica: $Y = b_0 + b_1 X^3$
- Potencia: $Y = b_0 + X^{b_1}$ ó $\ln(Y) = \ln(b_0) + b_1 \ln(X)$
- Exponencial: $Y = b_0 exp^{b_1 X}$ ó $\ln(Y) = \ln(b_0) + b_1 X$

3.2.2 ARIMAX y SARIMAX

Los modelos ARIMA y SARIMAX tienen la particularidad de explicar la variable dependiente en base a su comportamiento pasado, así como también a los errores previos de estimación. Aun así, puede existir una variable externa que explique el comportamiento de la serie a predecir, y el incluirla en un modelo ARIMA podría mejorar el rendimiento de este modelo.

Los modelos ARIMAX y SARIMAX hacen referencia a la integración de variables externas a los modelos propuestos por Box-Jenkins [16]. De esta forma, los modelos generales quedan definidos mediante la siguiente notación:

$$SARIMAX(p,d,q)x(P,D,Q)_T Y$$

Donde Y representa las variables externas del proceso.

Para ilustrar, suponiendo que se tiene un modelo de la forma $ARIMAX(p,d,q)X$, donde X es la única variable explicativa, la ecuación de transferencia estaría representada por:

$$\Delta y_t = c + x_t + \varepsilon_t$$

3.3 Técnicas de Inteligencia Artificial

Hasta el momento solo se han mencionado técnicas de pronóstico que se basan en proponer una estructura de relación entre la variable a pronosticar con su pasado ó con otras variables. Los métodos de inteligencia artificial son métodos que permiten encontrar una relación de manera automática, sin necesidad de proponer una estructura previamente.

Las metodologías de inteligencia artificial para el pronóstico reciben su nombre dado que se basan en paradigmas de aprendizaje supervisado, en el cuál se utilizan datos de "entrenamiento" para deducir una función.

Los métodos de inteligencia artificial más utilizados son las Redes Neuronales Artificiales y el Support Vector Machines para regresión, técnicas que serán descritas dadas su importancia en este estudio.

3.3.1 Redes Neuronales Artificiales

Las Redes Neuronales Artificiales (RNA) son un modelo matemático que emula el funcionamiento del cerebro humano. En términos generales, una red consiste en un gran número de unidades simples de proceso, denominas neuronas, que actúan en paralelo y están conectadas mediante vínculos ponderados [18].

El funcionamiento del sistema nervioso se da por la interconexión existente entre neuronas, que al recibir una de estas un estímulo, se produce una reacción química que "procesa" el impulso para luego pasarlo a la siguiente neurona, siempre y cuando este estímulo supere el umbral de propagación. En caso de pasar el umbral, el estímulo llega

a una neurona final que produce una reacción final (un movimiento en el cuerpo, por ejemplo).

Éste funcionamiento es el que imitan las RNA, donde una serie de unidades simples de proceso, denominadas neuronas artificiales, están interconectadas entre sí. Llevado al ámbito aplicado, una (o varias) neurona recibe un dato de entrada (estímulo), que es procesado mediante una función para luego traspasar este valor a las neuronas con las cuales tiene conexión, pero por cada conexión existe un ponderador que multiplica al valor que será procesado por cada neurona. El procesamiento del estímulo se repite hasta que se llega a una(s) neurona final que produce un resultado final, que en el ámbito de este estudio sería el valor futuro de la variable a pronosticar.

El punto clave de las RNA se encuentra en los ponderadores de la red, dado que con una correcta modificación de estos mediante un proceso de aprendizaje, la red mejora su desempeño respecto del objetivo para la cual fue diseñada, como por ejemplo, disminuir el error de pronóstico.

Las RNA pueden ser descritas como un algoritmo que trata de estimar una función desconocida, donde sólo se conoce una cantidad finita de elementos que pertenecen al dominio, y otra cantidad de elementos finita que pertenecen al conjunto imagen.

Bajo esta premisa, las RNA permiten la estimación de funciones complejas, e inclusive una red lo suficientemente grande, con una estructura y ponderadores adecuados, es capaz de aproximar cualquier función con el nivel de precisión que se desee [38].

3.3.1.1 Caracterización de una RNA

La aplicación de una red neuronal se puede dar en muchos ámbitos, y por lo que incumbe a este estudio, se tratará el caso en que se aplique un *aprendizaje supervisado*, que se da cuando se requiere estimar una función en base a datos de entrenamiento, los que muestran los resultados deseados que el algoritmo debe replicar.

En términos generales, la estructura de funcionamiento de una red neuronal sería la siguiente:

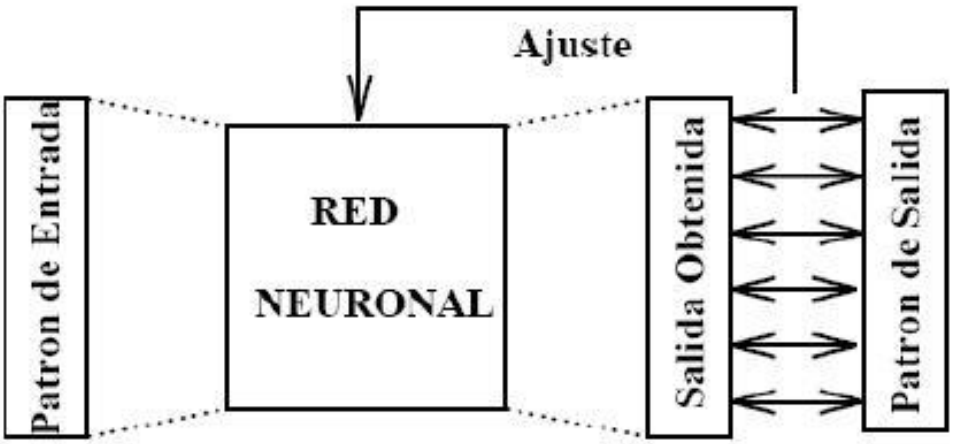


Ilustración 5: Funcionamiento de una RNA con aprendizaje supervisado

Existe un patrón de entrada, que para llevarlo al caso de pronóstico pueden ser los valores pasados de la variable de interés y/ó los valores de las variables explicativas, los cuales son ingresados a la red neuronal. Ésta arroja un resultado y se compara con el resultado deseado, que sería en este caso un valor conocido de la variable a pronosticar, y en base al error obtenido se realiza un ajuste a la red.

Ante esto es posible describir que el fin de la red neuronal es estimar ó "aprender" una función que permita transformar los datos de entrada en la salida deseada.

Descrito este funcionamiento general, se procede a caracterizar los elementos básicos de un RNA.

Neurona Artificial

La neurona de una RNA, basada en las neuronas del ser humano, está diseñada para recibir un estímulo, procesarlo y generar una respuesta. Se puede caracterizar una neurona mediante el siguiente esquema:

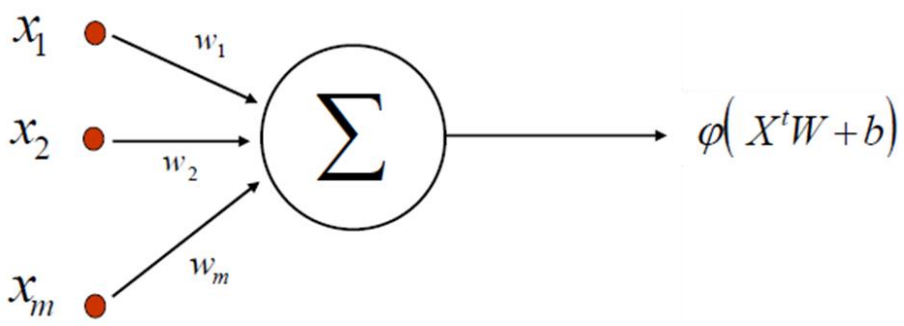


Ilustración 6: Esquema de una Neurona Artficial

Dónde:

- $x_1, \ldots, x_m$: Input que recibe la neurona
- $w_1, \ldots, w_m$: Peso sináptico de cada input
- Σ: Función de agregación
- φ: Función de Propagación

El input que recibe la neurona correspondería a los valores de las variables explicativas por ejemplo. Estas son "ingresadas" multiplicadas por su peso sináptico, que es un valor que permite otorgarle una ponderación a los estímulos ó input que recibe la neurona, dado que con el aprendizaje que realizará la neurona (o la red neuronal) los pesos darán más importancia a unos inputs que ha otros.

La entrada de datos con los pesos son luego "agregados" mediante una función, que generalmente es la función de suma (es decir, se suman los valores que entran a la neurona), no obstante también existen otro tipo de funciones de agregación como la multiplicatoria. Por ejemplo, la entrada de datos con sus pesos y la función de agregación se podrían expresar de la siguiente forma:

$$\sum_{i=1}^{m} x_i w_i \quad \text{ó} \quad \prod_{i=1}^{m} x_i w_i$$

Formas de Agregación de input de la neurona

Finalmente para el proceso de una neurona, esta debe de traspasar el estímulo que recibió. La analogía es que el input que recibe la neurona es "procesado" mediante una función de propagación (ó activación), la cual recibe como entrada la agregación de los datos de entrada con los pesos sinápticos y determinará si éstos fueron

suficientes para activar a la neurona y así "propagar" el impulso, ó en caso contrario simplemente no propagar información a otra neurona, al igual que como ocurre en el cerebro humano.

Dado que las funciones de activación tienen una finalidad ideológicamente binaria (activar o no), las funciones matemáticas utilizadas tienen un recorrido que va entre 0 y 1. Algunas de las funciones de activación más comunes son:

- Función Identidad: $f(x) = x$
- Función Tangente Hiperbólica: $f(x) = \tanh(x)$
- Función Logística: $f(x) = \frac{1}{1+\exp(-ax)}$
- Función Escalón: $f(x) = \begin{cases} 1 & si\ x \geq 0 \\ 0 & si\ x < 0 \end{cases}$
- Función Lineal a tramos: $f(x) = \begin{cases} 1 & si \quad x \geq c \\ x & si - c < x < c \\ -1 & si \ \ x \leq -c \end{cases}$

Estas funciones son las que finalmente procesan el input que han recibido para así pasarlo a otra neurona en el caso que estemos ante una red neuronal. Luego el producto final que genera una neurona, suponiendo una función de agregación como suma, y a la tangente hiperbólica como la función de propagación, sería el siguiente:

$$\varphi\left(\sum_{i=1}^{m} x_i w_i\right) = \tanh(\sum_{i=1}^{m} x_i w_i)$$

Por otra parte, también se puede destacar que si estamos resolviendo un problema con una sola neurona, y estamos utilizando la función identidad para propagar, estamos entonces ante una regresión lineal dado que los pesos sinápticos cumplirían la misma función que los factores de regresión.

La red neuronal

Las neuronas descritas anteriormente tienden a asociarse de distinta forma para resolver algún problema. A este tipo de asociaciones se les llaman *arquitectura* ó *topología* de la red y describen la forma en la que se asocian las neuronas de la red, por ejemplo, a qué neurona (ó a que neuronas) una pasa su "señal" de propagación.

La topología más clásica, y que es utilizada para los métodos de pronóstico, es la *feedforward*, que se caracteriza por utilizar capas de neuronas, en las cuales existen neuronas en paralelo que no transfieren información entre sí, pero si lo hacen con las neuronas de la siguiente capa, hasta llegar a la capa de salida donde la(s) producen el output final. En la siguiente ilustración se muestra una topología simple de una topología *feedforward*.

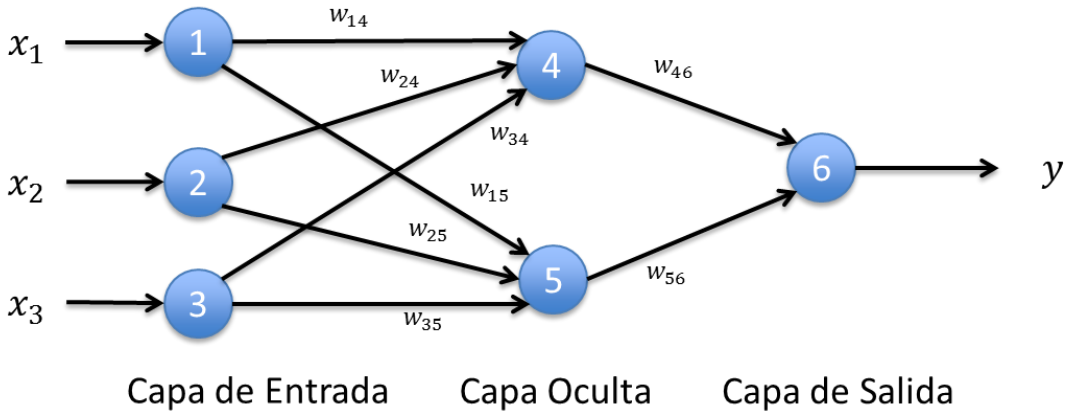


Ilustración 7: Ejemplo de Red Feedforward

En este caso la primera capa recibe los datos de 3 variables, donde cada estimulo se lleva a una sola neurona de la capa de entrada, luego esta señal es procesada en cada neurona para que genere un estímulo a la siguiente capa, la cual se llama capa oculta dado que esta entremedio de la capa de entrada y salida.

Cabe destacar que en este tipo de arquitecturas es muy común que una neurona de una capa pase su estímulo (o valor de la función de propagación) a todas las neuronas de la siguiente capa. Luego, los estímulos que reciben las neurona de la siguiente capa corresponden a lo producido por las neuronas de la capa previa, pero además ponderada por un peso sináptico correspondiente, dado que por cada conexión que hay en la arquitectura hay un peso sináptico involucrado. Este proceso

sigue hasta que se llega a la neurona final de la capa de salida, que es la que produce el valor que se espera de la función a estimar.

La estimación de la función se da a partir de la modificación de los ponderadores de las conexiones entre neuronas, ya que es en estos pesos donde se guarda el conocimiento en la red neuronal. Los ponderadores se van cambiando en una serie de iteraciones donde a la red se la entrena con una serie de registros que se tienen de la función a estimar, es decir, se poseen varias observaciones de los valores que tomaría la función y los argumentos que acompañan a esa función (en el caso del pronóstico, los argumentos serían los valores de las variables explicativas en un momento, y el valor de la variable de interés en ese mismo instante sería el valor objetivo de la función a estimar).

En este entrenamiento a la red se ingresan los datos como el argumento de la función a estimar, y cuando estos son procesados por la red hasta obtener el valor final, se compara con el valor deseado y se calcula un error. Luego para ir cambiando los ponderadores existen varios algoritmos, y uno de los más empleados es el *"Backpropagation"*[8], donde los ponderadores de la red se actualizan siguiendo la dirección negativa del gradiente del error (método de descenso del gradiente) [26].

El proceso de aprendizaje de la red tiene un tiempo de duración variable dependiendo de una serie de factores, como la cantidad de observaciones involucradas, la arquitectura de la red, entre otros. Además de esto, existe un riesgo de que la red "memorice" el conjunto de entrenamiento (overfitting), ante lo cual en muchos casos se utilizan condiciones de término para el aprendizaje. Las condiciones más utilizadas son:

- Error Máximo Permitido: Se propone una meta de un indicador de error (MAPE, Error Cuadrático Medio, etc.) en el conjunto de entrenamiento, que en caso de ser alcanzado se termina el entrenamiento.
- Número máximo de épocas: Una época es el nombre que se da a la presentación de todos los ejemplos del conjunto de

[8] Ver Anexos

entrenamiento a la red (en general, el entrenamiento de una red es realizado con varias épocas).

- Desempeño de la red en un conjunto de Validación: Una parte de las observaciones disponibles se dejan fuera del proceso de entrenamiento para formar un conjunto de validación. En el proceso de entrenamiento, a intervalos regulares, se evalúa el desempeño de la red en el conjunto de validación para verificar desempeño, poniendo como objetivo una meta en el error por ejemplo. Una vez que se alcanza el valor meta se detiene el proceso de entrenamiento.

La gran cantidad de posibilidades que existen para el diseño de una red hace que esta tarea pueda ser muy compleja. Considerando que no es práctica probar todas las posibilidades, han surgido diferentes heurísticas y reglas basadas en la experiencia. No obstante, no existe una heurística capaz de diseñar una red que tenga un buen desempeño en cualquier conjunto de datos [11].

Temas como la determinación automática del número de capas o neuronas ocultas están actualmente bajo investigación, lo que hace que, en la práctica, el método más común para el diseño de Redes Neuronales sea el de "prueba y error" [26], cuya duración podría ser prolongada dado que no se debe descartar una red mientras ésta no haya completado su aprendizaje.

3.3.2 Support Vector Regression

El algoritmo de Support Vector Regression (SVR) proviene de una clase de algoritmos llamados Support Vector Machines (SVM), que en términos generales son ampliamente usados para la clasificación de un conjunto de datos, acción que se realiza a través de un hiperplano que separa el espacio que conforman los datos cumpliendo dos objetivos:

- Minimizar la cantidad de registros mal clasificados
- Que el hiperplano tenga el mayor margen de separación de los conjuntos de datos

En este caso, el algoritmo de SVR es básicamente el mismo, pero en vez de tratar de clasificar un conjunto de datos, busca los márgenes óptimos para una regresión. Esta búsqueda se hace a través de un problema de optimización donde se utiliza una función de pérdida, cuyo objetivo es aplicar un error a las observaciones que quedan fuera del margen, pero las que quedan adentro no poseen error.

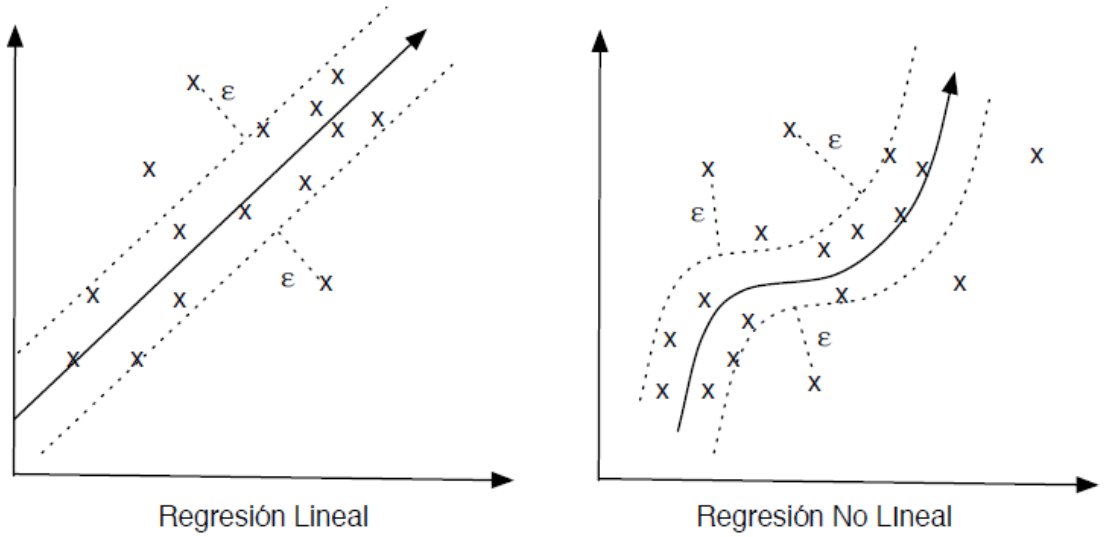


Ilustración 8: SVR lineal y no lineal

Los ejemplos anteriores muestran que para el caso lineal es intuitiva la creación de un hiperplano que acapare con sus márgenes la mayoría de las observaciones, no obstante en el caso no lineal no se puede que el ocupar un hiperplano como tal (una recta en el caso bidimensional) dado que no sería lo óptimo para crear un margen que permita hacer una correcta clasificación.

Para tratar esta problématica, el método de SVR (y SVM en general) hacen la utilización de las funciones de Kernel, las cuales son aplicadas al conjunto de datos para que estos sean linealmente separables.

Funcionamiento del Algoritmo[9]

Se tiene un conjunto de datos dado por $\{(x_1, y_1), \dots, (x_l, y_l)\} \in \mathbb{R}^d \times \mathbb{R}$, donde $\mathbb{R}^{\boldsymbol{d}}$ es el espacio que denota la dimensionalidad de $\boldsymbol{x_i}$. El objetivo principal de SVR es encontrar una función $\boldsymbol{f}(\boldsymbol{x})$ tal que tiene a lo más una desviación de "$\boldsymbol{\varepsilon}$" respecto de los valores $\boldsymbol{y_i}$ para todo el conjunto de entrenamiento, y además cumpla con la condición ser lo más plana posible. En otras palabras, no importan los errores si es que son menores que $\boldsymbol{\varepsilon}$, pero no se aceptará una desviación más grande que ésta.

Para graficar este método, supondremos que $\boldsymbol{f}$ es lineal, tomando la siguiente forma:

$$f(x) = \langle w, x \rangle + b \qquad donde\ w \in \mathbb{R}^d, b \in \mathbb{R}$$

El objetivo de que la función sea lo más plana posible se plasma sobre la búsqueda de un menor $\boldsymbol{w}$. Una forma de lograr esto es mediante la minimización de la norma euclidiana, i.e. $\|\boldsymbol{w}\|^{\boldsymbol{2}}$. Este problema puede ser plasmado como uno de optimización de la siguiente forma:

$$\begin{aligned} &\min && \|w\|^2 \\ s.a. \quad & && y_i - \langle w, x_i \rangle - b \leq \varepsilon \\ & && \langle w, x_i \rangle + b - y_i \leq \varepsilon \end{aligned}$$

En este caso existe una suposición de que la función $\boldsymbol{f}$ actualmente puede aproximar todos los pares (x_i, y_i) con precisión $\boldsymbol{\varepsilon}$, ó en otras palabras que el problema es factible. Dado que esto no siempre es el caso, se agregan variables de holgura $\boldsymbol{\xi_i, \xi'_i}$ que hacen el problema factible, pero además se interpretan como el error que se permite. Luego la formulación es la siguiente:

$$\begin{aligned} &\min && \|w\|^2 + C \sum_{i=1}^{l} (\xi_i + \xi'_i) \\ s.a. \quad & && y_i - \langle w, x_i \rangle - b \leq \varepsilon + \xi_i \\ & && \langle w, x_i \rangle + b - y_i \leq \varepsilon + \xi'_i \end{aligned}$$

[9] Basado en "A Tutorial on Support Vector Regression" (Smola y Schölkopf, 1998) y "Support Vector Machines Explained" (Fletcher, 2009)

$$\xi_i, \xi'_i \geq 0$$

La constante $C > 0$ determina una compensación entre lo plano que es la función y la cantidad de desviaciones mayores que ε que serán toleradas.

En la siguiente figura se grafica como solamente los puntos que quedan afuera del margen permitido aportan al costo de la función objetivo, siendo esta penalización lineal dependiendo de la distancia al margen permitido.

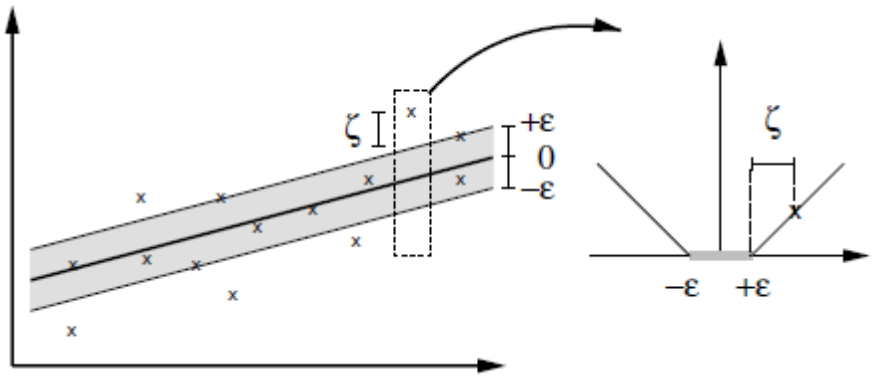


Ilustración 9: Asignación del error a los puntos que quedan fuera de la banda

El vector w es el que terminará por describir la función que se busca para luego realizar el pronóstico. No obstante, en este caso se ha supuesto una función lineal, pero como muchas veces no es el caso, y la relación entre las variables explicativas y la dependiente es no lineal, se utiliza un *preprocesamiento* del conjunto de entrenamiento x mediante la utilización de una función de Kernel, las cuales transforman el espacio, esto es por ejemplo:

$$\phi: \mathbb{R}^2 \rightarrow \mathbb{R}^3$$

El objetivo de utilizar esta función es que se permite tratar el problema de optimización de la misma forma antes descrita, quedando éste formulado de la siguiente forma:

$$\min \quad \|w\|^2 + C\sum_{i=1}^{l}(\xi_i + \xi'_i)$$

$$s.a. \quad y_i - \langle w, \phi(x_i)\rangle - b \leq \varepsilon + \xi_i$$
$$\langle w, \phi(x_i)\rangle + b - y_i \leq \varepsilon + \xi'_i$$

$$\xi_i, \xi'_i \geq 0$$

Las funciones de Kernel más utilizadas para SVR son:

- Lineal: $\phi(x) = x_i x_j$
- Polinomial: $\phi(x) = \left(\gamma x_i x_j + coef\right)^a$
- Radial Basis Function: $\phi(x) = exp(-\gamma|x_i - x_j|^2)$
- Sigmoide: $\phi(x) = \tan(\gamma x_i x_j + coef)$

3.4 Minería de Datos

"La minería de datos es la integración de un conjunto de áreas que tienen como propósito la identificación de un conocimiento obtenido a partir de las bases de datos que aporten un sesgo hacia la toma de decisión" (Molina y otros, 2001).

La minería de datos es una rama del conocimiento científico que afronta el problema de transformar datos de bajo nivel (datos que son muy numerosos para ser interpretados con facilidad) en información que puede cumplir el objetivo de verificar una hipótesis del usuario, ó el descubrimiento de nuevos patrones [21]. Este objetivo se logra a través de la utilización de distintas técnicas y algoritmos que en conjunto con la interpretación del usuario pueden generar conocimiento valioso en base a un gran conjunto de datos.

El término ha sido acuñado en varias metodologías, por ejemplo en el proceso de descubrimiento de conocimiento en bases de datos (*Knowledge Discovery in Database*, KDD), que cuenta de varias etapas definidas para lograr el objetivo (ver ilustración), pero además existen otros procesos que también estandarizan los pasos a seguir para lograr el objetivo de la minería de datos, como lo son los procesos CRISP-DM (Cross Industry Standard Process for Data Mining por sus siglas en inglés, ó Proceso Estandarizado a través de las industrias para la minería de datos) ó SEMMA (Sample, Explore, Modify, Model and

Assess, ó Muestra, Exploración, Modificación, Modelización y Valoración), y si bien tienen distintos enfoques, existen muchas fases o tareas a realizar que son homologables [2].

Tomando esto en cuenta, la minería de datos puede ser interpretada como una serie de metodologías interdisciplinarias que permiten encontrar correlaciones significativas, ajuste de modelos, ó relaciones entre variables a partir de los datos disponibles. Esta amplia gama de objetivos incluye también la tarea realizar una predicción [21], lo que hace que sea pertinente considerar sus metodologías para llevar a cabo el objetivo de este trabajo.

No obstante, su beneficio para la predicción no sólo recae en las técnicas antes mencionadas, sino que también con una metodología más global que hace referencia a todo el proceso sobre la construcción del modelo de predicción.

Los proyectos de minería de datos, en general, siguen una metodología como las planteadas anteriormente (KDD, SEMMA, CRISP-DM), no obstante, existen pasos de estas metodologías que son homologables con las otras, y por ende es factible describir un proyecto de minería de datos cumple, independiente de la metodología, con las siguientes etapas [2]:

1. **Filtrado de Datos**: En general los datos con los que se cuentan para realizar un proyecto no es necesariamente el óptimo, y esto no solo se debe a un concepto de la cantidad de datos, sino que también de su organización y estado, lo que impide que estos datos sean ingresados luego a algún algoritmo. Para enfrentar esta problemática se utiliza un preprocesamiento de los datos, evaluando la calidad de estos, realizar una limpieza (eliminar datos que no sirven), aplicación de filtros o transformaciones según los requerimientos, de esta forma quitando datos incorrectos y/ó inválidos del análisis, tratando de minimizar la cantidad de "basura" que va a ser analizada posteriormente [21].

2. Selección de Variables: Es muy común que la cantidad de datos que se poseen para un proyecto sea demasiado grande, inclusive si todos los datos que se poseen están "limpios". Luego, para enfocar el análisis, se seleccionan solamente algunas categorías de estos datos, lo

que puede ser interpretado como seleccionar las variables que se estiman tendrán una mayor influencia en la detección de un patrón.

3. **Extracción de Conocimiento**: Una vez que se tienen los datos filtrados y seleccionados, la utilización de una técnica de minería de datos, ad-hoc al objetivo del proyecto, permite confeccionar un modelo que encarna los patrones observados en los valores de las variables del problema, ó las relaciones de asociación entre ellas.

4. **Interpretación y Evaluación**: La confección de los modelos lleva a la posterior evaluación de su desempeño, y efectivamente validar si el modelo replica los comportamientos observados. Este proceso de validación (y selección en caso de tener varios modelos) permite confirmar una hipótesis y/ó descubrir un nuevo patrón a través de la interpretación de los resultados.

De este listado de tareas, que son mencionados ó incluidos de manera tácita en las metodologías de data mining (KDD, CRISP-DM, SEMMA), las técnicas de predicción pertenecerían a la extracción del conocimiento, pero las otras 3 tareas también son de importancia para la construcción de un modelo de pronóstico, sobre todo las tareas de selección de variables e interpretación y evaluación.

En las siguientes secciones se detallarán enfoques utilizados para la selección de variables así también como los indicadores a utilizar para la validación de los modelos (etapa de interpretación y evaluación).

3.4.1 Selección de Variables[10]

Dado el problema de modelación que se trata en esta tesis, existen 3 enfoques para realizar la selección de variables, los cuales se diferencian entre sí por cómo interactúan con las otras etapas de un proyecto de minería de datos. Los enfoques son:

- Filtros
- Metodos Wrapper
- Métodos Embebidos

[10]Basado en Guyon,I. y Elisseeff, A. "An Introduction to Variable and Feature Selection", Journal of Machine Learning Research 3 (2003) 1157-1182

3.4.1.1 Métodos de Filtro

Estas metodologías seleccionan variables sin importar el método de extracción de conocimiento a utilizar. En general se basan en realizar un ranking de utilidad[11] de las potenciales variables explicativas, y en base a un criterio de corte dejar una porción afuera del análisis mientras que las otras pasan a la siguiente etapa.

Unos de los métodos más utilizados para esto es realizar un ranking de correlación de las variables disponibles con la variable de interés a pronosticar. Luego, un criterio de corte básico sería el dejar de lado las variables que no están significativamente correlacionadas con la variable de interés.

Los métodos de filtro son rápidos de aplicar, no obstante tienden a entregar información redundante al modelo a aplicar a posteriori, dado que no consideran la relación entre las variables explicativas. Dado esto, los métodos de filtros a veces son aplicados sólo como una etapa previa a la selección de variables.

La siguiente figura resume el funcionamiento de este tipo de selección de variables y su relación con las otras etapas de un proyecto de minería de datos.

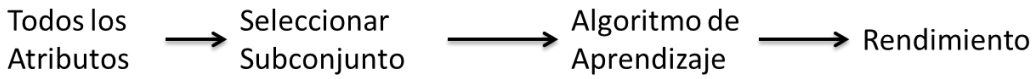


Ilustración 10: Método de Filtro

3.4.1.2 Métodos Wrapper

Los métodos wrapper (ó envolventes) utilizan subconjuntos de atributos que son ingresados al algoritmo de aprendizaje a utilizar en el proyecto, y en base a un criterio interno de selección[12], se selecciona el mejor subconjunto.

[11] La utilidad hace referencia a que tanta relación directa tienen con la variable de interés a pronosticar por ejemplo.

[12] Un criterio como la ponderación que tiene cada variable dentro del algoritmo sería un ejemplo. Un criterio externo al algoritmo sería la performance que éste alcanza.

Dado que explorar todas las combinaciones de subconjuntos crece de manera exponencial con cada atributo, y además es un proceso computacionalmente muy demandante probar tantas combinaciones de variables en el algoritmo de aprendizaje, se utilizan heurísticas de búsqueda, donde las más utilizadas son la selección hacia adelante (*Foward Selection*) y la eliminación hacia atrás (*Backward elimination*). En la primera se parte con un modelo sin variables, de manera que en cada iteración por venir se ingresa la variable más relevante. Por otra parte, en *Backward elimination* el modelo inicial considera todos los atributos y en cada iteración se elimina la variable menos relevante.

Estos algoritmos, en general, paran cuando ya no quedan variables relevantes fuera del modelo (*Foward Selection*) ó todas las variables que están en el modelo son lo suficientemente relevantes (*Backward elimination)*.

Esta metodología se distingue por incluir en el proceso de selección al algoritmo a utilizar en el proyecto de minería de datos.

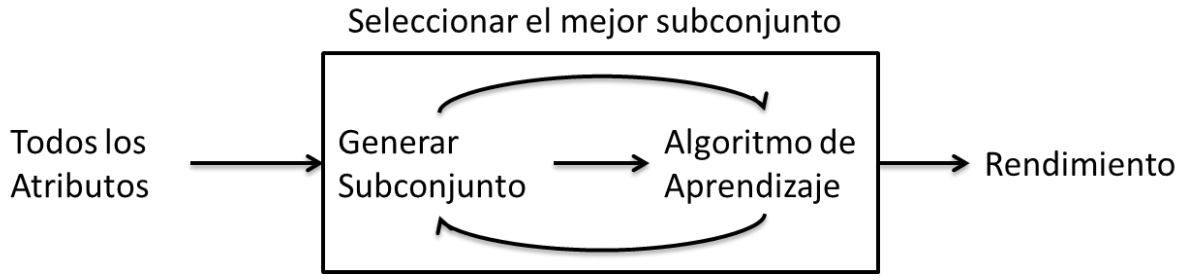


Ilustración 11: Método Wrapper

Una metodología wrapper que llama la atención en el ámbito de la predicción es al análisis de regresión, que pasa a ser descrita a continuación.

Análisis de Regresión

El análisis de Regresión es una técnica estadística utilizada para estudiar la relación entre las variables explicativas con la dependiente. El proceso consta de la utilización de modelos de regresión lineal múltiple (de allí su nombre), y además de un proceso de selección de variables en base a los coeficientes estimados.

La selección de atributos de éste método se realiza a través pruebas estadísticas sobre los coeficientes estimados en una regresión lineal múltiple. En particular, mediante la utilización de la prueba de T de Student [13], se evalúa la probabilidad de que un coeficiente en particular sea igual a cero, utilizando el *p-valor*, que representa la probabilidad de que un coeficiente sea cero, y por ende comprobar si puede ser omitido del modelo. El nivel de significancia impuesto para esta prueba es de un 10%.[14]

Tomando en cuenta esta prueba estadística, se utiliza un método de regresión llamado *Stepwise Backward*, el cual es descrito en los siguientes pasos:

1. Se comienza estimando un modelo que incluye todas las variables explicativas propuestas
2. Para cada variable en el modelo, se calcula el *p-valor* de los coeficientes asociado a la prueba estadística
3. Se selecciona el *p-valor* más grande entre los calculados
    a. Si el *p-valor* es mayor a 0.1, la variable asociada al coeficiente es sacada del modelo
    b. Si ningún *p-valor* es mayor a 0.1, se procede al paso 4
4. Se estima el nuevo modelo, y se calculan los nuevos coeficientes y sus *p-valor*, así como también los *p-valor* de las variables que no están en el modelo, correspondiendo este al valor que tendría en caso de que aquella variable fuese la siguiente en integrarse al modelo.
5. Se selecciona el *p-valor* menor de las variables que no están en el modelo
    a. Si el *p-valor* es menor a 0.1, la variable asociada al coeficiente entra al modelo
    b. Si ningún *p-valor* es menor a 0.1, se procede al paso 6
6. Se repiten los pasos 2 al 5 hasta que el mayor *p-valor* de los coeficientes dentro del modelo sea menor a 0.1, y que el menor *p-valor* de las variables afuera del modelo sea mayor a 0.1

[13] https://www.encyclopediaofmath.org/index.php/Student_test

[14] Esto implica que si un *p-valor* es mayor a 0.1 no se puede rechazar la hipótesis de que el coeficiente es cero, luego se omite la variable asociada al coeficiente.

Este algoritmo asegura que todas las variables que se incluyen en el modelo final sean significativas, así como también los rezagos de éstas incluidos.

Esta metodología de selección de variables tiene variaciones, que están basadas en los métodos *Foward Selection* y *Backward Elimination*, donde el primero es un algoritmo que empieza sin variables explicativas y que solamente va agregando variables a medida que su *p-valor* es menor a 0,1. El segundo método parte con todas las variables y sólo elimina a medida que el *p-valor* es mayor a 0,1. Las metodologías *Stepwise* son una combinación de las dos mencionadas anteriormente, dado que permiten tanto la eliminación como la integración de variables.

3.4.1.3 Métodos Embebidos

La metodología final de selección de variables es una profundización de los métodos wrapper, dado que se basan en encontrar el subconjunto óptimo de atributos que cumple con un objetivo de desempeño del algoritmo a utilizar en el proyecto.

Estos métodos son computacionalmente más demandantes que los wrappers, y casi siempre son confeccionados de manera específica para cada caso. En general, los métodos embebidos buscan dos objetivos a optimizar:

- Minimizar el error del Algoritmo (maximizar el desempeño)
- Minimizar el número de atributos incluidos

El primer objetivo es algo que siempre se debe de buscar, mientras que el segundo responde a un objetivo más cercano a la generalidad del patrón a encontrar, además de tratar de evitar la sobre-complejidad del modelo (que además podría afectar a la interpretación de los resultados).

A diferencia de los métodos *Wrapper*, acá hay una evaluación del rendimiento final que tuvo el algoritmo de aprendizaje en el objetivo final (por ejemplo, se evalúa el error de pronóstico).

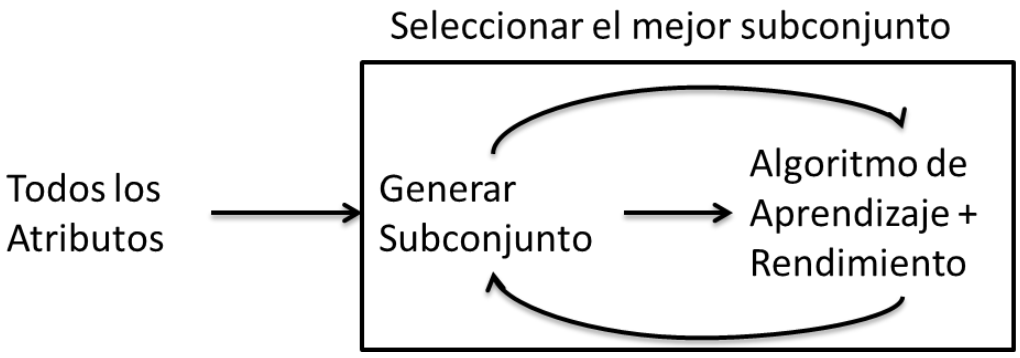


Ilustración 12: Método Embebido

3.4.2 Validación

Esta etapa es crucial para cualquier proyecto de minería de datos, y es la que permite evaluar si el modelo que estamos construyendo cumple ó no con el objetivo que buscamos.

Para llevar a cabo este procedimiento, se dividen los datos disponibles para generar dos conjuntos que tienen distintas funciones. Estos conjuntos son llamados:

- Conjunto de Entrenamiento
- Conjunto de Validación

Sobre el primero de estos conjuntos se aplica el algoritmo de aprendizaje con el objetivo de estimar una función (de clasificación, regresión, etc.). Una vez que se ha estimado esta función con el conjunto de entrenamiento, se pasa a utilizar el conjunto de validación.

En el caso de este conjunto se aplica la función estimada, utilizando como argumento de la función los datos de las variables explicativas que se poseen. El resultado de esto es que se obtiene para cada observación del conjunto de validación un valor estimado de la variable de interés. Luego, como sabemos cuál es valor verdadero al cual debería haber llegado la función, calculamos un error el cual permite evaluar el rendimiento de la función en un caso de la vida real.

En otras palabras, el objetivo del conjunto de validación es poner a prueba a la función estimada en el de entrenamiento.

En general, el conjunto de entrenamiento posee más observaciones que el conjunto de validación, siendo alrededor del 70% de los datos totales disponibles, mientras que el 30% correspondería al de validación.

Dependiendo del tipo de objetivo que se busca con el proyecto de minería de datos, se utilizan distintas medidas de error que miden el rendimiento. Dado que en este trabajo se busca realizar un pronóstico, se expondrán algunas de las medidas clásicas de error utilizadas en este caso.

Suponiendo que f_i represente el valor pronosticado para una observación, y_i el valor real a obtener en esa observación, y n el número total de observaciones a evaluar, los errores más utilizados son:

- MAE (Mean Absolute Error ó error medio absoluto)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i|$$

- MAPE (Mean Absolute Percentage Error ó error porcentual medio absoluto)

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{f_i - y_i}{n}\right|$$

- MSE (Mean Squared Error ó Error Cuadrático Medio)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(f_i - y_i)^2$$

- RMSE (Root-Mean Squared Error ó Raíz del Error Cuadrático Medio)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f_i - y_i)^2}$$

- NRMSE (Normalized Root-Mean Squared Error ó Raíz del Error Cuadrático Medio normalizado)

$$NRMSE = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(f_i - y_i)^2}}{y_{max} - y_{min}}$$

4. Desarrollo y Aplicación de la Metodología

En este capítulo se abordará la descripción y aplicación de la metodología, la cual abarca la construcción de los modelos de proyección de demanda con sus respectivas variables explicativas, la configuración de "rezagos" a utilizar[15], y la aplicación de las distintas metodologías de extrapolación.

La metodología seleccionada se desenvuelve en el marco de un trabajo desarrollado para la empresa y que considera la utilización del software "Intelligent Forecaster", el cual permite trabajar con distintos métodos de pronóstico, además de realizar distintas configuraciones entre la variable dependiente a pronosticar y las explicativas.

4.1 Definición de Metodología a utilizar

Para el presente proyecto se ha optado por la realización de un modelo multivariado de pronóstico, esto considerando que el modelo actual que posee la empresa utiliza esta metodología. No obstante, para este proyecto también se consideró la utilización de rezagos como variables de entrada, utilizando así un enfoque mixto en el cual se explicarán los comportamientos futuros de las series a pronosticar en base a los valores de otras variables en ese momento, pero también en considerando los valores pasados de esas mismas variables explicativas, además del comportamiento pasado de las series a pronosticar.

Los métodos de pronóstico univariados puros fueron descartados dado que no permiten realizar un análisis de sensibilidad respecto de variables externas (factor solicitado por la empresa), lo que también limita al momento de hacer proyecciones en base a escenarios.

Las metodologías multivariadas implican la investigación de posibles variables explicativas a ser incluidas en los modelos, además de sus respectivas proyecciones en caso de ser necesarias. Estas

[15] Configuración de Rezagos: Referencia a los datos pasados que tienen influencia en el comportamiento futuro de la serie.

proyecciones serán realizadas con metodologías univariadas, las cuales serán detalladas en el siguiente capítulo.

Una de las razones por las cuales se decidió considerar este enfoque mixto recae en el hecho de que al agregar rezagos como variables de entrada, se remueve en parte la dependencia sobre la exactitud del pronóstico de las variables explicativas [1].

Una de las decisiones más importantes de todo esto proyecto reside en seleccionar qué tipo de metodología de pronóstico utilizar. Dado el objetivo de este trabajo de tesis y las metodologías actualmente utilizadas en la empresa para realizar el pronóstico, se optó utilización técnicas de inteligencia artificial, con el fin de aprovechar la capacidad que tienen éstas al no asumir el tipo de relación que existe entre las variables de entrada.

Si bien en la bibliografía estas técnicas han sido utilizadas mayoritariamente para la proyección a corto plazo [23] [39], para este estudio presentan un buen potencial al poder captar un tipo de relación entre las variables explicativas y la variable dependiente, además que actualmente la empresa utiliza un modelo de regresión, por lo cual la utilización de modelos de inteligencia artificial también puede aportar más información a algún tipo de relación no captada en el modelo previo.

No obstante lo anterior, también se desarrollaron modelos SARIMA y SARIMAX con el fin de comparar los resultados de los métodos de inteligencia artificial con métodos estadísticos clásicos de pronóstico.

Dicho esto, y considerando lo planteado en el presente capítulo, se procedió mediante una metodología de minería de datos para la construcción de los modelos de pronóstico. Las implicancias de esto consideran la realización de 4 grandes etapas:

- Filtrado (pre-procesamiento) de datos
- Selección de Atributos
- Aplicación y validación de una metodología de Pronóstico
- Análisis e interpretación de Resultados

La primera etapa presenta metodologías estándar que son más bien decididas (en algunas ocasiones) en base a los métodos que se

aplicarán en etapas posteriores[16]. No obstante se consideró pertinente la realización un estudio sobre las mismas series para tratar de detectar algún tipo de anomalía o transformación aplicable que represente de mejor manera el fenómeno a pronosticar.

Para la selección de variables, se ha optó por la utilización del análisis de regresión con *Stepwise Backward*, un método del tipo *wrapper*, el cual se basa en realizar una estimación de una función lineal con todas variables explicativas incluidas, e ir eliminando cada una de estas en base a un criterio estadístico. El objetivo de ocupar este método es encontrar las relaciones lineales que puedan tener las variables a pronosticar con las explicativas, de forma que esta combinación de variables seleccionadas sean ingresadas a los métodos de pronóstico, teniendo como base que cumplen con una relación lineal y dejando así a los métodos de inteligencia artificial la tarea de encontrar otro tipo de relación no detectada.

Independiente de lo anterior, el enfoque que se quiere dar a la selección es basado en resultados, por lo cual los resultados del análisis de regresión serán utilizados sólo como una base, ya que el objetivo final de esta tesis es desarrollar un mejor modelo que el actual que tiene la empresa.

Dicho esto, se agregaran variables al resultado del análisis de regresión, adiciones que estarán basadas en presunciones aportadas por expertos en distribución eléctrica. Además, también se desea que con esto se ponga a prueba a los métodos de inteligencia artificial para estimar relaciones no lineales.

Para la validación de los modelos, se ha decidido utilizar el MAPE como métrica de error, dado que es una métrica muy intuitiva al estar basada en porcentajes, además que por solicitud de la empresa se ha solicitado no publicar las magnitudes de la demanda (solamente se ha permitido publicar las formas de las curvas).

Además del MAPE, se buscan otros dos objetivos para asegurar la calidad del pronóstico:

[16] Por ejemplo, se realizan escalamiento a algunas series para que luego no saturen los valores de las funciones de las redes neuronales.

- Homocedasticidad de los errores
- Independencia Inter-temporal de los errores

Estos objetivos están basados en los supuestos de la regresión lineal, y en conjunto, en caso de cumplirse, implican que el modelo tiene un error sistemático homogéneo, y que con el paso del tiempo este error se mantiene.

A pesar de esto, considerando la metodología a utilizar (se consideran rezagos), la independencia de los errores se ha dejado como un objetivo secundario dado que el utilizar rezagos rompe con esta condición [47].

Cabe mencionar que tanto para la homocedasticidad e independencia temporal de los errores, el conjunto de errores a utilizar para evaluar estas condiciones corresponde al total de la muestra, incluyendo de esta forma tanto los errores de entrenamiento como los de validación.

Para probar la condición de homocedasticidad, se utilizó la prueba de Cox-Stuart [10], la cual al no ser rechazada indica que no hay suficiente evidencia para asumir que la muestra tiene un cambio monotónico de su dispersión, ó en otras palabras, que es heterocedástica.

La métrica utilizada para establecer que hay ó no independencia de los errores, es la autocorrelación y la autocorrelación parcial, primando como criterio la segunda. Se consideró que de no existir rezagos con correlaciones altamente significativas, se podría establecer la independencia de los errores.

A su vez, considerando que los datos que se disponen no son muchos (son registros mensuales), la independencia de los errores se establecerá cuando no hayan autocorrelaciones altamente significativas en un período de un año. Esto debido que al aumentar la cardinalidad de los rezagos, menor es la cantidad de datos bajo los cuales se calculan las autocorrelaciones.

Todas las pruebas de autocorrelaciones y homocedasticidad se efectúan sobre los errores absolutos.

4.1.1 Particiones de Datos Utilizadas

Se consideró un primer conjunto de entrenamiento bajo el cual se realizará la selección de atributos (mediante análisis de regresión). Este conjunto comprende el siguiente periodo:

- Junio de 2001 hasta Diciembre de 2006 (para Demanda de Energía de Sistema y Demanda de Potencia máxima en el anillo) lo que corresponde a 67 registros
- Enero de 2001 hasta Diciembre de 2006 (para demandas de energías sectoriales) lo que corresponde a 72 registros.

En estos mismos conjuntos se realizará también el entrenamiento de los métodos de inteligencia artificial, así como también se construirán los modelos SARIMA y SARIMAX.

El conjunto de validación se queda corresponde al horizonte desde Enero de 2008 hasta Diciembre de 2013 (72 registros), el cual es dividido en los siguientes 5 segmentos:

- Enero de 2007 hasta Diciembre 2011 (60 registros)
- Enero de 2008 hasta Diciembre 2012 (60 registros)
- Enero de 2009 hasta Diciembre 2013 (60 registros)
- Enero de 2007 hasta Diciembre 2010 (48 registros)
- Enero de 2007 hasta Diciembre 2009 (36 registros)

Los primeros 3 horizontes son utilizados para medir los errores de una predicción que tiene como último registro de datos conocidos como el mes anterior (i.e. para el primer horizonte sería Diciembre de 2006), no obstante, para los últimos 2 esto cambia, quedando la siguiente configuración:

- Último registro conocido:
  - Diciembre de 2005 para horizonte desde Enero de 2007 hasta Diciembre 2010
  - Diciembre de 2004 para horizonte desde Enero de 2007 hasta Diciembre 2009

La implicancia del último dato conocido marca el punto de partida del horizonte de pronóstico, por lo cual para el primero de éstos se

evalúa el error en base a lo proyectado para el segundo, tercer, cuarto y quinto año, mientras que para el último sólo se evalúa los últimos tres años. La utilización de este tipo de horizontes permite ver el desempeño enfocado en los años posteriores del pronóstico, además de entregar más etapas de prueba para seleccionar el modelo.

El no evaluar los primeros años de estos horizontes radica en el hecho de que el error de estos periodos estaría sesgado al estar estos incluidos en parte en los horizontes utilizados para realizar la selección de variables, y como se especifica en la bibliografía [20], el conjunto de pruebas no debe ser utilizado para realizar la selección de atributos.

4.1.2 Especificación de Software y Hardware

Para la realización de esta tesis se utilizó el software "Intelligent Forecaster" en su versión 2.3.21.0, desarrollado por la empresa BIS-lab, el cual cuenta con funcionalidades que permiten el desarrollo de modelos de pronóstico mediante Redes Neuronales Artificiales y Support Vector Regression, entre otros métodos.

Los modelos SARIMA y SARIMAX fueron ejecutados en el software "Gretl" (Gnu Regression, Econometrics and Time-Series Library) en su versión 1.10.1.

Todos los experimentos y procesamientos de este trabajo de tesis fueron ejecutados en una laptop Samsung R-580, con las siguientes especificaciones:

- Procesador Intel Core i5 M430 (2 núcleos físicos; 2,27 Ghz)
- 3 GB de Memoria RAM
- Windows 7 Professional – 64 Bits

4.1 Preprocesamiento de Series a Pronosticar

Según los pasos en común que se incluyen en los procesos estandarizados de minería de datos descritos en el capítulo anterior, se realizó un procesamiento a las series antes de ser utilizadas para la proyección.

La primera parte del preprocesamiento corresponde a la limpieza de datos, eliminando *"outliers"*[17] y completando las series en caso de que existan valores perdidos. Estos procedimientos no fueron necesarios dado que no existen *outliers* ni valores perdidos.

Considerando los métodos de pronóstico a utilizar, se realizó un escalamiento lineal de las series a un intervalo entre -0,6 y 0,6 con el objetivo no saturar las funciones de activación de las de que los valores que se introduzcan en las funciones de propagación (en el caso de la redes neuronales). Esta transformación fue aplicada tanto para las series a pronosticar como para las series explicativas.

La fórmula de la función que transforma linealmente los datos, considerando que se tiene un rango original dado por $[A, B]$ y se requiere llevarlo al intervalo $[C, D]$, es la siguiente:

$$f(x) = C\left(1 - \frac{x - A}{B - A}\right) + D\left(\frac{x - A}{B - A}\right)$$

Dónde x es un valor que pertenece al rango $[A, B]$.

Con esta transformación los procesos de entrenamiento de los métodos de pronóstico a utilizar realizan la comparación, respecto a los valores obtenidos versus el valor deseado, utilizando los valores escalados por esta función. El resultado final que arroja el programa es llevado a las unidades reales mediante la transformación inversa de esta función, que corresponde a la misma fórmula de arriba pero invirtiendo el uso de los valores de los rangos.

[17] Corresponden a valores atípicos de una muestra, los cuales son numéricamente muy distintos.

Realizando un análisis de las series a pronosticar, se consideró pertinente recordar que representa cada valor mensual que toman las series a pronosticar. Esto quedaría descrito de la siguiente forma:

- Demanda de Energía Eléctrica en el Sistema: Suma de las demandas totales de energía eléctrica ocurrida en cada día en el sistema
- Demanda de Energía Sectoriales (Residencial, Comercial e Industrial): Ídem, pero acotada sólo al sector correspondiente
- Demanda de Potencia Máxima en el Anillo: Demanda Máxima de Energía Eléctrica ocurrida en una hora del mes[18], considerando sólo el anillo.

Al analizar estas definiciones, se detectan dos problemáticas. La primera de éstas hace referencia a las series de Demanda de Energía, las cuales al representar una suma de eventos ocurridos cada día (la demanda diaria de energía), están afectadas por un efecto calendario.

Esto se traduce en que si comparamos los valores de demanda de un mes en un año respecto al valor del mismo mes, pero en el próximo, la diferencia entre ambos no estará explicada solamente por un factor de tendencia, sino que también a que existen diferencias entre la cantidad de días laborales.

Este efecto hace necesario incluirlo en la proyección, dado que para factores de planificación si es requerido tomarlo en cuenta en el futuro. Ante esto, se decidió que en vez de aplicar un tipo de transformación a las series de energía, se utilizaría una variable explicativa que trate el efecto calendario a través de una relación que incluya las diferencias existentes en la demanda de cada día de la semana.

La segunda problemática detectada tiene relación con la serie de potencia eléctrica. Los valores registrados corresponden a un evento que ocurrió en un instante del mes, que a diferencia de la energía, no representa una suma de un continuo de eventos.

[18] Recordar que la Potencia eléctrica se define como la energía eléctrica entregada o absorbida por un elemento en un tiempo determinado.

Esta condición genera una falta de "historia" que no permite explicar cómo se ha llegado a este nivel de demanda máxima (faltaría el continuo de los datos).

Se ideó sobrellevar esta problemática mediante el uso de otras variables de entrada para el pronóstico de la potencia, las cuales harían referencia a otros registros de potencia a considerar en el mismo mes (la segunda mayor potencia, la potencia media, etc.). No obstante se descartó esta idea por la necesidad de pronosticar estas mismas variables explicativas, que al final de cuentas requerirían de la confección de un modelo de pronóstico como el de potencia máxima.

Al considerar que la serie de potencia máxima carece de más información que permita explicar su comportamiento, donde ni siquiera se tienen datos en qué día u hora de la semana ocurrió el evento, se planteó realizar una transformación a la serie.

En este caso no hay un efecto que sea causa de la diferencia de cantidad de días en el mes, pero si existe un patrón intrínseco que vendría a ser el comportamiento típico que tienen los días, haciendo alusión a las relaciones de la demanda máxima de potencia diaria que existe dentro de una semana.

Si bien no se espera que este efecto sea ampliamente determinante para fijar el valor de la demanda de Potencia Máxima, si se mantiene la sospecha de que es lo suficientemente grande como para cambiar en, al menos, un 1% el valor de la potencia máxima del mes, y si además esto pasará para un mes de máximo anual, cambiaría por ende el valor del crecimiento anual de esta serie.

La hipótesis que nace de esto se traduce en que existe la creencia de que la demanda máxima de potencia en un mes esta explicada por dos grupos de factores:

- Factores Tendenciales y Estacionales: Asociados al movimiento propio de la serie, explicando la mayor parte de esta.
- Factor día: Una pequeña parte del valor se explica por el día de la semana en que ocurrió la demanda máxima.

La suma de estos factores resultaría en la demanda registrada en un mes, la que gráficamente podría explicarse en la siguiente manera:

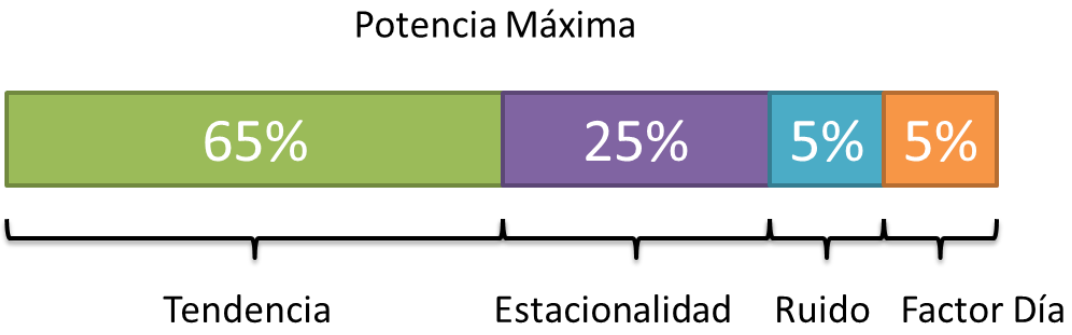


Ilustración 13: Factores que explican la Potencia Máxima (valores hipotéticos)

Con la transformación que se quiere aplicar, se eliminaría este factor día de la serie, permitiendo comparar los valores solamente asociados al factor mayoritario, qué es el que se considera pertinente para realizar la proyección de la serie.

Para ejemplificar un poco mejor lo que se quiere lograr con la transformación a esta serie, se expone el siguiente ejemplo:

"*Si se determina que el Lunes es el día de potencia máxima dentro de la semana, y el día Jueves corresponde a un 95% de ese valor (mediante promedios históricos), entonces una potencia máxima de un mes que ocurrió en un día jueves se vería aumentada en un 5,26%*[19] *al aplicar esta transformación".*

El aumento que aplica la transformación[20] busca utilizar como escenario que todas las demandas ocurrieron en el mismo día de la semana, específicamente en el día de mayor potencia máxima dentro de la semana. De esta forma se pretenden trabajar con la serie que representaría los potenciales valores de la potencia máxima en caso de ocurrir en el día de mayor demanda en el mismo día.

[19] Corresponde al valor dado por [(1 / 0,95) – 1]*100% = 5,26%

[20] Notar que la transformación aumentaría ó mantendría el valor de las potencias máximas, pero nunca lo disminuiría.

La transformación se basa en calcular la siguiente relación para todos los días de todas las semanas de las cuales se tiene registros de potencia[21]:

$$r_{i,w} = \frac{PMAX_{i,w}}{PMAX_{S,w}}$$

Dónde:

- $r_{i,w}$: Relación entre la potencia máxima del día "i" de la semana "w" respecto del día seleccionado para calcular la relación
- $PMAX_{i,w}$: Potencia Máxima del día "i" de la semana de la semana "w"
- $PMAX_{S,w}$: Potencia Máxima del día seleccionado para calcular la relación, en la semana "w"

Hecho esto, se pasa a calcular el promedio de estas relaciones respecto de todas las semanas, que sería representado a través de la siguiente ecuación.

$$R_i = \frac{\sum_w r_{i,w}}{|w|}$$

Dónde:

- R_i: Relación histórica del día de la semana "i" respecto al día seleccionado para calcular la relación
- $|w|$: Cardinalidad del conjunto de semanas

Una vez calculado los siete valores R_i, la serie de potencia máxima mensual es transformada en base al día en que ocurrió cada demanda. La transformación final correspondería a la siguiente ecuación.

$$\widehat{PMAX}_{t,\iota} = \frac{PMAX_{t,i}}{R_i}$$

Dónde:

[21] Se tienen los datos de Potencia diaria desde el 4 de Junio de 2001 hasta el 31 de Diciembre de 2013

- $\widehat{PMAX}_{t,i}$: Potencia Máxima Transformada del período "t" que ocurrió en un día "i"
- $PMAX_{t,i}$: Potencia Máxima del período "t" que ocurrió en un día "i"

Cabe destacar que, como se mencionó antes, esta transformación debe aumentar o mantener el valor de la serie en cada registro, por lo cual, para la estimación de los valores R_i, donde se calcula una relación respecto a un día de la semana, se busca que este día sea el que tenga mayor demanda histórica, de tal forma que cuando se calculen los R_i, el valor del máximo sea igual a 1.

Si bien este ejercicio se podría ejercer con otro día, eliminando esta restricción de los valores de R_i (aunque igual suprimiendo el efecto del factor día en la potencia), la causa por la cual la transformación actual sólo aumenta ó mantiene el valor de la serie tiene directa relación con el hecho de que es más costoso subestimar la demanda que sobreestimarla. Luego, con la serie transformada se estará entrenando a los algoritmos con valores más altos de potencia, disminuyendo de esta forma la posibilidad de subestimar la demanda al realizar una proyección.

Un último detalle metodológico para realizar esta transformación tiene relación con los datos a considerar para el cálculo de R_i, dado que no se incluyeron en el cálculo aquellas semanas que contuvieran días festivos, ya que estas tienen una distribución de la demanda atípica entre sus días (los días festivos presentan mucha menos demanda de potencia y energía respecto de los días laborales).

Descrito el procedimiento metodológico, se muestra a continuación los resultados obtenidos.

Valores de R_i

El día que cumplió con la condición de tener la mayor demanda histórica de potencia fue el día martes, lo que llevó a los siguientes valores de R_i:

Día	Valor Ri	Cambio que implica
Lunes	0,996	0,40%
Martes	1	0,00%
Miércoles	0,998	0,20%
Jueves	0,997	0,30%
Viernes	0,982	1,83%
Sábado	0,844	18,48%
Domingo	0,814	22,85%

Tabla 9: Valores de Ri

Se puede ver que la demanda de potencia en los días Sábado y Domingo son considerablemente menores respecto del resto de la semana, mientras que la diferencia más significativa entre los días laborales ocurre respecto al Viernes. En la parte derecha de la tabla se muestra el cambio porcentual que implica el valor de R_i al aplicar la transformación.

En base a estos valores, se pasó a transformar la serie de potencia, obteniendo los siguientes resultados.

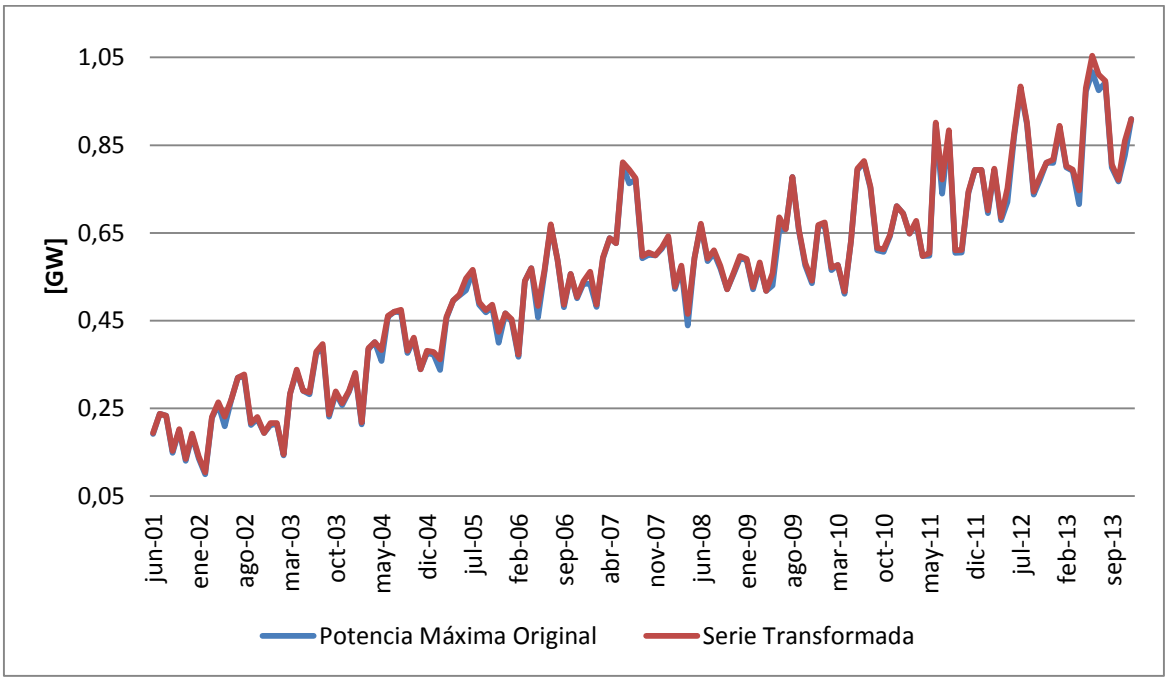


Gráfico 7: Serie de Potencia Original vs Transformada

Crecimientos de Máximos Anuales		
Año	**Serie Original**	**Serie Transformada**
2002	6,82%	7,23%
2003	5,18%	5,18%
2004	5,69%	5,60%
2005	6,31%	6,17%
2006	6,59%	6,59%
2007	8,27%	8,45%
2008	-7,55%	-7,71%
2009	6,54%	6,35%
2010	1,89%	2,06%
2011	4,79%	4,79%
2012	4,15%	4,33%
2013	2,03%	3,52%

Tabla 10: Comparación de Crecimientos de Máximas Demandas Anuales

El gráfico muestra que la diferencia entre las series es casi nula, no obstante puede tener cambios significativos a la hora de comparar los crecimientos de las máximas anuales, indicador de gran importancia para la empresa.

Este hecho se ejemplifica mejor para el caso del año 2013, donde la demanda máxima de potencia ocurrió un día viernes, lo que ocasionó que la demanda aumentará al transformar la serie.

4.2 Estudio de Variables Relevantes para la Metodología

Dada la metodología seleccionada para desarrollar la proyección de demanda (modelo causal), en esta sección se estudiaran y validaran las variables más relevantes para describir el comportamiento de la demanda eléctrica que presenta la red Chilectra S.A., tanto para su demanda de energía eléctrica en el sistema (incluidas las demandas sectoriales[22]), como para la potencia máxima de cada mes en el anillo.

En la bibliografía es recurrente encontrar que algunas de las variables a utilizar en este tipo de modelos a largo plazo son [43]:

- Producto Interno Bruto (PIB)
- Tamaño de la Población
- PIB per Cápita
- Temperatura Media
- Pérdidas del sistema
- Precio de la Energía
- Velocidad del Viento
- Humedad Relativa
- Cantidad de Horas de Sol

Estas variables pueden ser aglomeradas en dos conjuntos, unas pertenecientes al segmento de *Macroeconómicas* (PIB, PIB per Cápita, Tamaño de la Población), otras describen *Características Propias del Sistema* (Precio de la Energía, Factor de Carga y Pérdidas del Sistema) y las restantes en se agrupan en *Climatológicas*.

[22] Demandas Sectoriales: Referencias a las demandas del sector Residencial, Industrial y Comercial.

La bibliografía apunta a que estos grupos de variables permiten describir de buena manera la demanda eléctrica en el largo plazo, y es más, en el modelo implementado en la empresa se utilizan variables pertenecientes a estos 3 segmentos. No obstante, un modelo de proyección de demanda no debe incluir todas estas variables por distintos motivos, como lo son los siguientes:

- *Redundancia de Información*: En diversas ocasiones muchos de los factores que se incluyen en un modelo presentan correlaciones muy marcadas, lo que implicaría, en caso de incluirse 2 variables con correlaciones lineales fuertes, que se está entregando información redundante al modelo, hecho que puede causar que se distorsione la relevancia de una variable en función de otra cuando se ejecute el método de estimación[23]**.**

- *Complejidad del modelo*: En un caso hipotético, el modelo de proyección con mejor performance (menor error de pronóstico por ejemplo) puede incluir un gran número de variables y al mismo tiempo rezagos de éstas, no obstante la ejecución y actualización del mismo puede volverse un proceso muy engorroso al considerar muchas series de datos. Para evitar que esto ocurra se debe incluir un numero razonable de variables que permita explicar de buena forma la variable dependiente, procurando a la vez que la ejecución y actualización del modelo no sea muy engorrosa.

Tomando en cuenta lo anterior, no existe una combinación única y óptima de variables a utilizar en un modelo. Luego, en base a criterios como el error y la usabilidad del modelo, se debe realizar una selección de variables que permita obtener un conjunto de factores explicativos que no sea redundante y pueda explicar de buena forma el comportamiento de la serie a pronosticar.

[23] Método de Estimación: Metodología como Redes Neuronales Artificiales ó Regresión mediante MCO que permite captar la relación entre las variables explicativas y la dependiente y así realizar la proyección.

Para cumplir estos objetivos, la selección de variables se realizará en base a la relación que posee cada factor con la serie a explicar, y también considerando la posibilidad de acceder a datos de calidad del factor en cuestión.

4.2.1 Variables a Considerar y Recopilación de Datos

En base a lo planteado anteriormente, en este estudio se considerarán variables macroeconómicas, climatológicas y aquellas que describen características propias del sistema.

Una condición que limita las posibilidades de variables a considerar es que la periodicidad de los datos de Energía y Potencia es mensual, por ende para formar un modelo con variables explicativas se debe contar con registros mensuales de los factores a considerar. Este hecho afecta al momento de incluir variables como el PIB o la Población dado que su periodicidad no es mensual, no obstante una forma de utilizarlos es repetir el mismo valor para cada año, ejerciendo de esta forma un efecto para marcar la tendencia de crecimiento más que para describir la estacionalidad presentada dentro del año.

En las siguientes secciones se detallará las variables que fueron consideradas en el estudio para cada uno de los efectos de

4.2.1.1 Variables Macroeconómicas

Para los factores de ésta índole, que pueden ser subdivididos en Económicos y demográficos, se considera de vital importancia la utilización del PIB debido a su gran influencia al representar la actividad económica del país, lo que tiene influencia directa en la demanda de energía y Potencia para la red.

Dicho esto, y con el fin de mitigar el efecto de la periodicidad de éste dato, se decidió utilizar el Índice Mensual de Actividad Económica[24] (IMACEC) que publica el Banco Central de Chile, índice representativo de la actividad económica de Chile, cuyo propósito es medir la evolución de la actividad económica a precios constantes, basándose en los bienes y servicios que componen el PIB del país y emulando por lo tanto parte de su comportamiento.

Si bien el IMACEC abarca la actividad económica de todo el país, y en este caso sólo interesa la actividad económica de la región metropolitana, no existe un índice o publicación que contenga esta información de manera mensual (el BBCC solo publica el PIB anual de cada región, sin descomposición mensual).

Respecto de las variables demográficas, estadísticas como la población de la región metropolitana están sujetas al censo realizado cada 10 años, y si bien existen estudios que realizan proyecciones sobre la población en cada región, se decantó por utilizar el IMACEC dado que una mayor población debiese de también reflejar una mayor actividad económica.

La obtención del IMACEC se realizó desde la página del Banco Central de Chile, utilizando la serie empalmada con base 2008, obteniéndose así datos desde enero de 2003 hasta Agosto de 2014.

En las series de frecuencia mensual debe tenerse en consideración el denominado "efecto calendario", por el cual no pueden considerarse similares los meses con diferente número de días laborales o festivos. Este efecto se hace muy notable, por ejemplo, cuando el período de semana santa cambia de mes de un año a otro, lo que tergiversa cualquier comparación con el comportamiento de esos meses. Sin embargo, el efecto está siempre presente y es necesario medirlo si se quieren hacer comparaciones más consistentes entre los meses de un mismo año o a través de los años.

[24] http://www.bcentral.cl/estadisticas-economicas/metodologias-estadisticas/pdf/Imacec.pdf

Como la disparidad entre los meses radica en la diferencia entre sus días (feriados, laborales, sábados, domingos) lo conveniente es incluir una variable que recoja y explique esta disparidad, a la que se denominó *"Laboralidad"*, y cuya estimación depende de los distintos niveles de demanda de energía para los distintos días de la semana en un mes. Se busca con esta variable representar el número de días laborales equivalentes que posee un mes, en relación a la demanda de energía que necesita el sistema. Luego el cálculo de ésta se resume como la suma de cocientes entre el consumo de un día respecto del consumo del día miércoles de esa misma semana, realizado esto para todos los días de cada mes.

4.2.1.2 Variables Climatológicas

Con el objetivo de estudiar y confirmar el efecto climatológico en la demanda energética, se recopilaron una serie de datos históricos de datos desde el sitio web de la Dirección Meteorológica de Chile.

Los datos recopilados corresponden a los registrados por la estación meteorológica de Quinta Normal dado que su ubicación central es considerada como la que mejor representa la "realidad" climática bajo la cual está sometida el área de concesión de la empresa.

En total se recopilaron 50 variables correspondientes a distintos registros meteorológicos, dentro de los que destacan:

- Temperatura (Media, Máxima, Mínima, Extremas)
- Humedad Relativa
- Velocidad y Dirección del Viento
- Horas de Sol
- Precipitaciones
- Presión Atmosférica
- Nubosidad

La importancia de las variables climatológicas recae sobre la descripción estacionalidad de la demanda, permitiendo explicar en parte los distintos comportamientos de la serie en los meses del año.

Todas las variables recopiladas fueron utilizadas para realizar el análisis de correlación como un primer filtro de variables, obteniéndose registros mensuales desde el año 1996 hasta el 2014.

4.2.1.2 Variables de Sistema

Debido a que para realizar un pronóstico con un modelo causal es necesario hacer proyecciones de las variables explicativas, solamente se utilizó el Precio de la Energía para esta sección dado que la misma empresa realiza proyecciones sobre este. Este factor también presenta una arista importante en el estudio dado que permitiría plasmar la elasticidad del consumo ante los cambios en el precio, afectando este tanto a clientes regulados como no regulados, siendo éstos últimos posiblemente más sensibles al tener la capacidad de negociar una tarifa.

Las pérdidas del sistema fueron descartadas debido a que no aportan mucha información al tener valores estables alrededor de su promedio (5,7%), confirmado esto por su baja desviación estándar (0,09%).

4.2.2 Selección de Atributos mediante Análisis de Regresión

En esta sección se muestran los resultados de la aplicación del método de Análisis de Regresión con *Stepwise Backward*, donde se busca la mejor combinación de variables explicativas que modelen linealmente, y con buen desempeño, las series de demanda energética y de potencia eléctrica.

Este método, funciona en base a las variables a las cuales esté expuesto en un inicio, luego se aplicaron distintas combinaciones de variables explicativas y sus respectivos rezagos para iniciar cada análisis de regresión.

En concordancia con la bibliografía [43], las combinaciones de variables propuestas para los modelos consisten de la siguiente estructura:

- IMACEC (Variable Socioeconómica)
- Laboralidad (Variable Socioeconómica)
- Precio de la Energía (Variable Característica del Sistema)

- Temperatura[25]
- Otra Variable Climatológica
- Variables Binarias Dummies (o de apoyo) para meses

No se realizaron pruebas que incluyeran a todas las variables climatológicas dado que, por su estacionalidad, es altamente probable que presenten una correlación elevada entre sí, entregando información redundante al modelo. Además, en caso de probar con todas las variables y rezagos, se generan experimentos donde la cantidad de coeficientes de regresión es mayor al número de observaciones.

Las variables Binarias *Dummies* fueron utilizadas con el objetivo de ayudar de manera previa al modelo a describir la estacionalidad presente. Son 12 variables que indican el mes de la observación, mediante la asignación del valor 1 si es que se está en un mes, y 0 para las demás. De esta forma entre las 12 variables para una observación siempre habrá una que tendrá el valor 1, mientras que las demás serán iguales a Cero.

Si bien la selección *Stepwise* arroja como resultado un modelo compuesto solo de variables relevantes, no necesariamente cumple con las condiciones de MCO. Con fines de construir un buen modelo, el procedimiento de selección mediante análisis de regresión se da por terminado cuando una combinación de variables entrega un modelo con errores *homocedásticos* e independientes entre sí.

La homocedasticidad se comprobó a través de la prueba de Cox-Stuart [10][26], donde la hipótesis nula corresponde a que la serie no presenta un cambio en su dispersión. La independencia de los errores se calculará solo en base a las funciones de autocorrelación (normal y parcial) aplicadas a los errores cuadráticos.

A su vez, antes de realizar la estimación mediante MCO, las variables en cuestión (incluida la dependiente) son escaladas al intervalo [0,1;1]. El objetivo de esto consiste en eliminar los efectos por diferencias de magnitudes entre las variables, además que al hacer esto los coeficientes de regresión serán comparables entre sí al tener

[25] Una o más variables haciendo referencia a la temperatura, como la temperatura media, máxima y/ó mínima de un mes. Las variables utilizadas harán referencia a lo necesario para la utilización de sus respectivas causas.

[26] Ver Anexos: Prueba de Cox Stuart

magnitudes acordes, deprendiéndose fácilmente del análisis la incidencia que tiene cada variable en la regresión.

Cabe destacar que en los experimentos realizados, las series de demandas energéticas y potencia (que son las variables dependientes de la regresión), no sufren ningún tipo de descomposición, es decir, las series solamente son escaladas, pero la estacionalidad y tendencia están presente en éstas al momento de estimar la regresión.

La aplicación de éste método fue realizada en los siguientes horizontes de tiempo:

- Junio de 2001 hasta Diciembre de 2006 - para Demanda de Energía de Sistema y Demanda de Potencia máxima en el anillo
- Enero de 2001 hasta Diciembre de 2006 - para demandas de energías sectoriales

En todos estos horizontes se utilizaron todos los datos para entrenar a la regresión lineal, dado que en este caso no importa el pronóstico al estar realizando una selección de atributos.

A continuación se muestran los resultados del análisis de Regresión para las distintas series a pronosticar, así como también la descripción de otros procesos realizados para solucionar problemas.

4.2.2.1 Demanda de Energía en el Sistema

La estimación de modelos de regresión para esta serie fue realizada en base a 20 combinaciones distintas de variables explicativas, siguiendo la estructura descrita anteriormente, mientras que las variaciones de rezagos propuestas incluyen la utilización de registros hasta un año atrás respecto al periodo a pronosticar, utilizando, en el caso de las variables explicativas, el mismo periodo a pronosticar ($t+1$), el periodo actual (t), un semestre atrás respecto del periodo a pronosticar ($t-5$) y un año atrás ($t-11$), períodos que son considerados representativos dado que pretenden mostrar la relación entre un valor de la demanda respecto de la demanda del mes anterior, de un semestre atrás, y de un año atrás. Fueron incluidos también

variaciones de estos rezagos con una diferencia de 1 mes ($t-10$ y $t-12$ para el caso de $t-11$).

Se añadieron además rezagos de la misma variable dependiente, utilizando los mismos que en el caso de las variables explicativas, exceptuando el caso del periodo $t+1$, qué es el objetivo. La estructura de la función de regresión propuesta es plasmada en la siguiente ecuación.

$$y_{t+1} = b_0 + b_1 y_t + b_2 y_{t-1} + b_3 y_{t-5} + b_4 y_{t-11} + b_5 x_{t+1} + b_6 x_t + b_7 x_{t-1} + b_8 x_{t-5} + b_9 x_{t-11} + e_t$$

Dentro de las pruebas realizadas, un resultado constante fue la eliminación de la variable correspondiente a la temperatura. Dada las características entre las relaciones a proponer para el método, era de esperarse éste resultado al no tener esta variable, al igual que otras climatológicas, una relación lineal con el consumo de la energía. Este comportamiento puede ser plasmado por el hecho de los peaks de consumo que ocurren en invierno y verano, dado que cuando se presentan temperatura baja en invierno, se presenta un peak en el consumo, pero a la vez cuando sube en verano también se presenta otra alza en la demanda. Luego, para que la variable de Temperatura Media pudiese ser incluida en el modelo, se le realizó una transformación correspondiente al valor absoluto de la diferencia que presenta cada mes respecto del promedio histórico total de la temperatura media. Según los cálculos realizados, la temperatura media histórica considerando datos desde 2003 a 2014, es de 14,81°C. Luego, la nueva variable puede ser expresada de la siguiente forma:

$$\widehat{T_t} = |T_t - 14{,}81|$$

Dónde:

- T_t: Temperatura Media en el período t
- $\widehat{T_t}$: Desviación Respecto de la Temperatura media histórica en el período t

A esta nueva variable se le apodó como *"T° positiva"* debido a que para toda desviación respecto de la media la serie toma valores mayores a cero, impidiendo además que exista un valor negativo. En el siguiente cuadro se puede ver la comparación entre ambas series de temperatura.

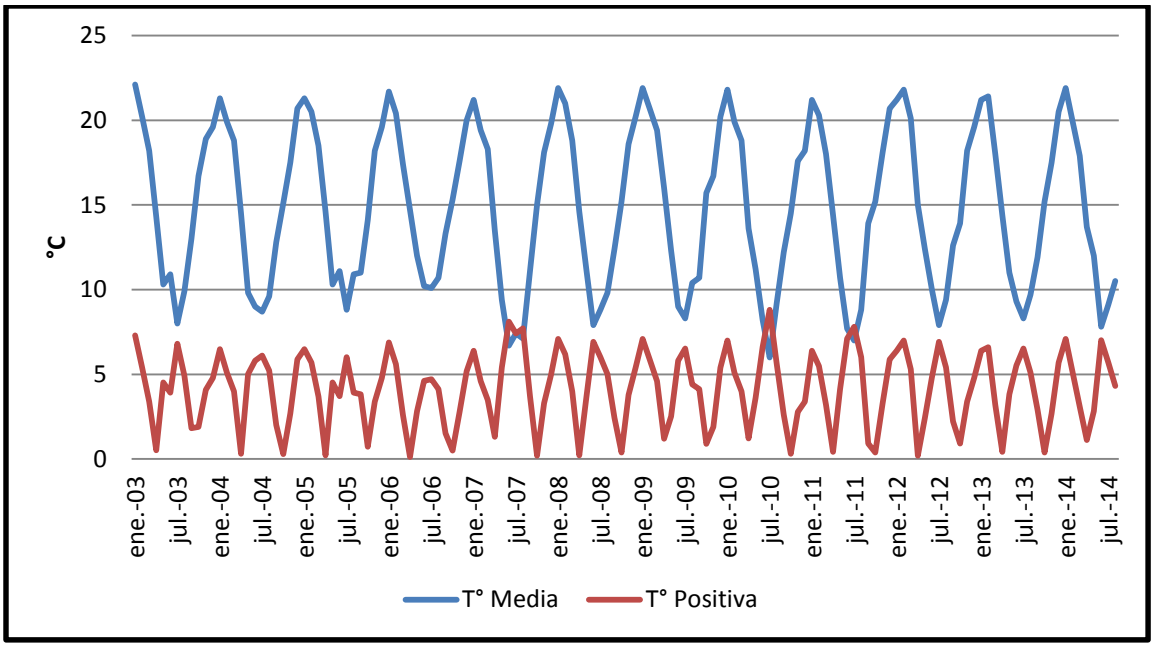


Gráfico 8: Comparación entre la Temperatura Media y su Transformada

La aplicación de esta transformación permitió encausar una relación lineal entre la temperatura media y la demanda de energía eléctrica en el sistema de Chilectra S.A., dado que en esta ocasión, tanto para verano como para invierno la *T° positiva* crece, al igual que lo hace la demanda energética.

Realizando pruebas con ésta variable se llegó al siguiente modelo de Regresión Lineal Múltiple qué cumple con los criterios mencionados antes.

$$ES_{t+1} = b_0 + b_1 ES_t + b_2 ES_{t-3} + b_3 ES_{t-10} + b_4 ES_{t-11} + b_5 L_{t+1} + b_6 \hat{T}_{t+1} + b_7 IM_{t+1} + b_8 IM_{t-6} + b_9 HR_{t+1} + b_{10} HS_{t+1} + b_{11} P_{t+1} + e_t$$

Donde:

- ES_t: Demanda de Energía Eléctrica en el Sistema en el Período t
- L_t: Laboralidad del Periodo t
- $\hat{T}_t$: T° Positiva en el período t
- IM_t: IMACEC del período t
- HR_t: Humedad Relativa Promedio a las 2pm en el período t
- HS_t: Horas de Sol en el período t
- P_t: Precio de la Energía en el mes t

El detalle de los resultados se muestra en el siguiente cuadro:

Serie de Tiempo	Rezago	Coeficiente de Regresión	Error Estándar	Estadístico T	P-valor
Demanda de Energía Eléctrica en el Sistema	0	0,2327	0,04	5,8167	0
	-3	-0,1286	0,0483	-2,6621	0,0097
	-10	-0,1795	0,0573	-3,1304	0,0026
	-11	0,1006	0,0568	1,771	0,0811
Laboralidad	1	0,186	0,0271	6,8604	0
T° Positiva	1	0,113	0,0145	7,8138	0
IMACEC	1	0,6077	0,0652	9,3269	0
	-6	0,3122	0,0676	4,6176	0
Horas de Sol	1	-0,1248	0,0267	-4,6692	0
Humedad Relativa Promedio 2pm	1	-0,0741	0,0352	-2,1039	0,0391
Precio de la Energía	1	-0,1264	0,0178	-7,1118	0

Tabla 11: Resultados de Análisis de Regresión para Demanda de Energía en el Sistema

Se aprecia que, a excepción del término constante, todas las variables presentan un *p-valor* menor a 0.1, implicando así que todas las variables son significativas para el modelo.

A su vez se puede apreciar la gran influencia que ejerce el IMACEC sobre la proyección, dado que sus dos coeficientes asociados son los más altos para el modelo de regresión. Se destaca también la influencia positiva de la variable *T° Positiva* mediante su coeficiente, comprobando la hipótesis planteada al momento de generar la variable.

El desempeño del modelo es además bastante bueno al tener un valor de R^2 superior a 0,95.

Estadísticas	Valor
Observaciones Incluidas	67
Coeficientes Incluidos	12
R-Cuadrado	0,9627
R-Cuadrado Ajustado	0,9565

Tabla 12: Desempeño de Análisis de Regresión para la Demanda de Energía en el Sistema

La homocedasticidad queda demostrada por los resultados del test de Cox-Stuart:

Cambio en la Dispersión (Hipótesis Alternativa)	p-valor
Cox-Stuart (Dispersión)	0,2962

Tabla 13: Prueba de Homocedasticidad para los errores del Modelo de Energía del Sistema

El *p-valor* no ayuda a rechazar la hipótesis nula, por lo cual se asume que los errores presentan una varianza constante.

Por otro lado, las gráficas de autocorrelación no permitieron obtener resultados ideales, pero sí lo suficientemente buenos como para terminar el análisis de regresión.

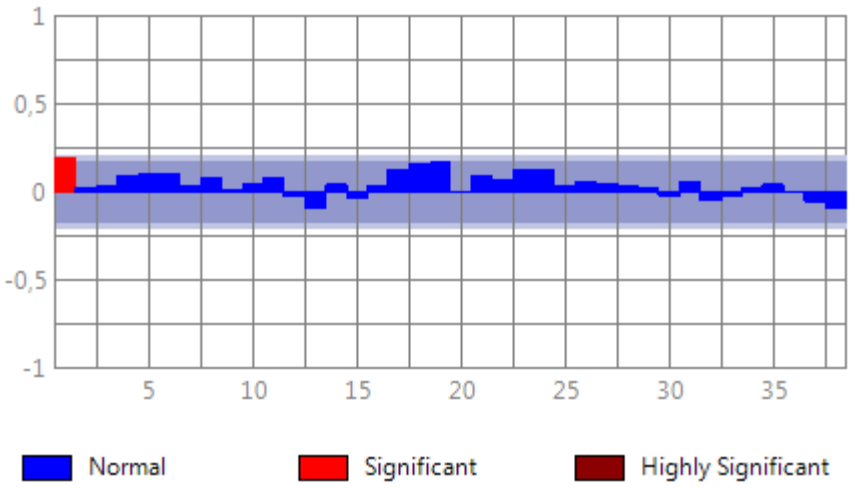


Gráfico 9: Autocorrelaciones del Error para Modelo lineal de Demanda Energética en el Sistema

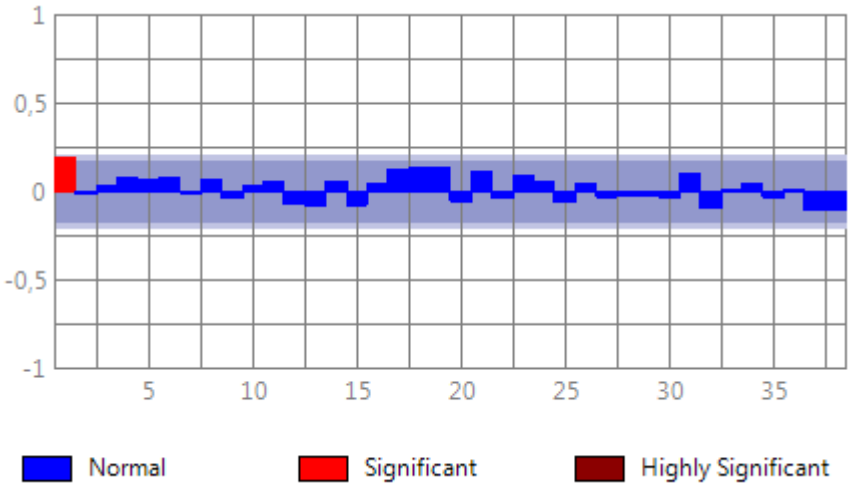


Gráfico 10: Autocorrelaciones Parciales del Error para Modelo lineal de Demanda Energética en el Sistema

En estas se puede apreciar que existe una autocorrelación significativa entre un error y el error anterior. Este resultado es de esperarse dada la cantidad de rezagos utilizados, no obstante dado que no existe una alta significancia del registro, se considera como una performance lo suficientemente buena para llevarla a cabo en los modelos de inteligencia artificial.

4.2.2.2 Demanda de Energía Residencial

Los experimentos para esta serie incluyeron la misma estructura de variables, pero mayores variaciones en los rezagos respecto del caso anterior debido a la dificultad de encontrar un modelo que cumpliera con los criterios.

La ecuación lineal que responde a los resultados finales es la siguiente:

$$RE_{t+1} = b_1 RE_t + b_2 RE_{t-7} + b_3 RE_{t-12} + b_4 D2_{t+1} + b_5 D6_{t+1} + b_6 L_{t+1} + b_7 IM_{t-1} + b_8 IM_{t-3} + b_9 HS_{t+1} + b_{10} P_{t+1} + b_{11} \hat{T}_{t+1} + e_t$$

Donde:

- RE_t: Demanda de Energía Eléctrica en el sector Residencial en el periodo t
- L_t: Laboralidad del Periodo t
- $\hat{T}_t$: T° Positiva en el período t
- IM_t: IMACEC del período t
- HS_t: Horas de Sol en el período t
- P_t: Precio de la Energía en el mes t
- $D2_t$: Variable Dummie Binaria para el mes de Febrero en el periodo t
- $D6_t$: Variable Dummie Binaria para el mes de Junio en el periodo t

Serie de Tiempo	Rezago	Coeficiente de Regresión	Error Estándar	Estadístico T	P-valor
Demanda Residencial	0	0,3855	0,0723	5,3331	0
	-7	-0,2379	0,0577	-4,1223	0,0001
	-12	-0,2502	0,0767	-3,261	0,0018
V. Dummie (Febrero)	1	-0,1321	0,0239	-5,5381	0
V. Dummie (Junio)	1	0,0835	0,0227	3,6753	0,0005
Laboralidad	1	0,2826	0,0391	7,2329	0
Precio Energía	1	-0,1371	0,0292	-4,6913	0
Horas de Sol	1	-0,1592	0,0226	-7,0543	0
IMACEC	-1	0,2305	0,0772	2,9857	0,0041
	-3	0,4781	0,109	4,3871	0
T° Positiva	1	0,1497	0,0262	5,7186	0

Tabla 14: Resultados de Análisis de Regresión para la Demanda Residencial

Los coeficientes de regresión arrojan un resultado parecido al de la Demanda del Sistema al tener el IMACEC el coeficiente de mayor importancia (aunque con rezago en este caso). Cabe destacar también en el análisis la presencia de variables dummies como relevantes, pero solamente la de dos meses. La correspondiente a Julio pareciera responder a que en éste mes se presenta el peak de demanda anual, mientras que el de Febrero lo haría para el peak que ocurre en Verano.

Los resultados de la regresión también son buenos al tener un buen coeficiente de determinación cercano a la unidad.

Estadísticas	Valor
Observaciones Incluidas	72
Coeficientes Incluidos	12
R-Cuadrado	0,9278
R-Cuadrado Ajustado	0,9143

Tabla 15: Desempeño de Análisis de Regresión para la Demanda Residencial

La homocedasticidad de los errores también fue cumplida, así como también la independencia de los errores en el tiempo, no presentándose ninguna correlación significativa, tanto para las autocorrelaciones normales como para las parciales, lo que se muestra a continuación.

Cambio en la Dispersión (Hipótesis Alternativa)	p-valor
Cox-Stuart (Dispersión)	0,2649

Tabla 16: Prueba de Homocedasticidad para los errores del Modelo de Energía Residencial

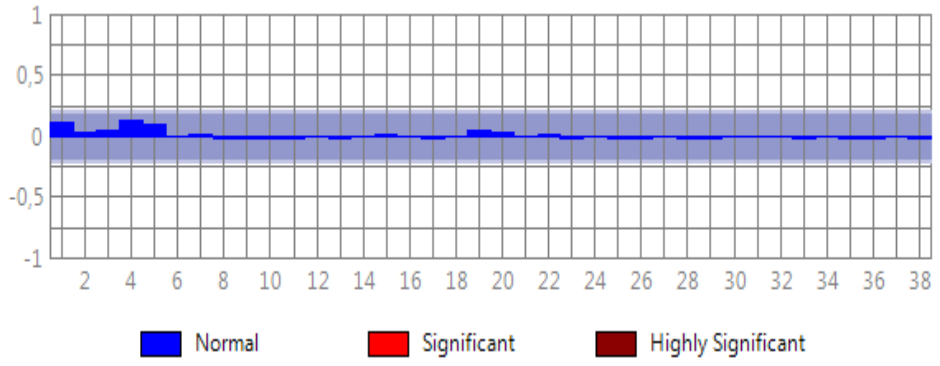


Gráfico 11: Autocorrelaciones del Error para Modelo lineal de Demanda Energética Residencial

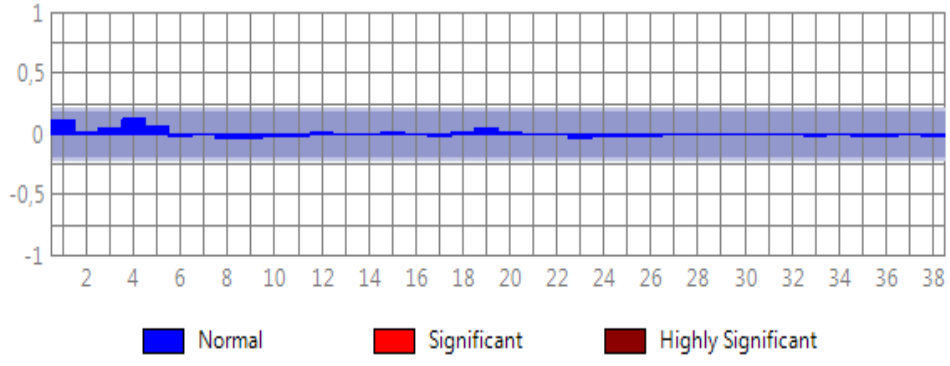


Gráfico 12: Autocorrelaciones Parciales del Error para Modelo lineal de Demanda Energética Residencial

4.2.2.3 Demanda de Energía Comercial

Los resultados para esta serie no fueron óptimos al probar más de 20 combinaciones de variables, con sus respectivas variaciones en rezagos, y no poder cumplir con los criterios de Homocedasticidad e Independencia de los errores. No obstante se espera que con la aplicación de los algoritmos de inteligencia artificial se puedan cumplir.

La combinación de variables y rezagos seleccionada es la que alcanzó un mayor valor del coeficiente de determinación.

La ecuación obtenida es la siguiente:

$$CO_{t+1} = b_0 + b_1 CO_t + b_2 CO_{t-1} + b_3 CO_{t-3} + b_4 CO_{t-7} + b_5 CO_{t-9} + b_6 L_{t+1} + b_7 \hat{T}_{t+1} + b_8 IM_{t+1} + b_9 P_{t+1} + b_{10} P_t + e_t$$

Donde:

- CO_t: Demanda de Energía Eléctrica en el sector Residencial en el periodo t
- L_t: Laboralidad del Periodo t
- $\hat{T}_t$: T° Positiva en el período t
- IM_t: IMACEC del período t
- P_t: Precio de la Energía en el mes t

Serie de Tiempo	Rezago	Coeficiente de Regresión	Error Estándar	Estadístico T	P-valor
Término Constante		-0,0627	0,0147	-4,2698	0,0001
Demanda Comercial	0	0,5207	0,09	5,786	0
	-1	0,1571	0,0739	2,1255	0,0375
	-3	-0,1888	0,0802	-2,3533	0,0218
	-7	-0,2866	0,0815	-3,515	0,0008
	-9	0,1621	0,0803	2,019	0,0478
Laboralidad	1	0,2058	0,0388	5,2987	0
T° Positiva	1	0,0892	0,0229	3,901	0,0002
IMACEC	1	0,5243	0,1013	5,1777	0
Precio Energía	1	-0,3757	0,1115	-3,3689	0,0013
	0	0,4029	0,1118	3,6042	0,0006

Tabla 17: Resultados de Análisis de Regresión para la Demanda Comercial

Nuevamente se presenta la preponderancia del IMACEC al ser el que tiene el coeficiente de mayor valor, pero también destacan en importancia la demanda ocurrida en el periodo anterior así como también el precio de la energía, tanto en el rezago a pronosticar como el actual. Este resultado permite especular sobre la importancia del precio de la energía para el comercio y como éste afecta a la demanda.

Los resultados globales del análisis de regresión, así como los de la prueba de homocedasticidad y la independencia de los errores se muestran a continuación.

Estadísticas	Valor
Observaciones Incluidas	72
Coeficientes Incluidos	11
R-Cuadrado	0,8946
R-Cuadrado Ajustado	0,8776

Tabla 18: Desempeño de Análisis de Regresión para la Demanda Comercial

Cambio en la Dispersión (Hipótesis Alternativa)	p-valor
Cox-Stuart (Dispersión)	0,0003

Tabla 19: Prueba de Homocedasticidad para los errores del Modelo de Energía Comercial

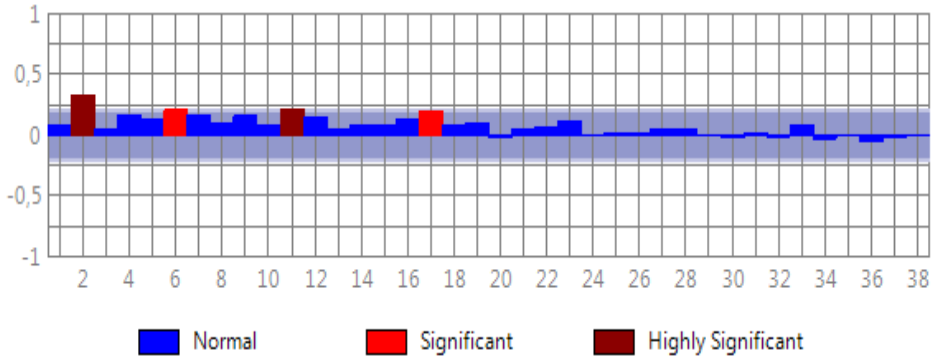


Gráfico 13: Autocorrelaciones del Error para Modelo lineal de Demanda Energética Comercial

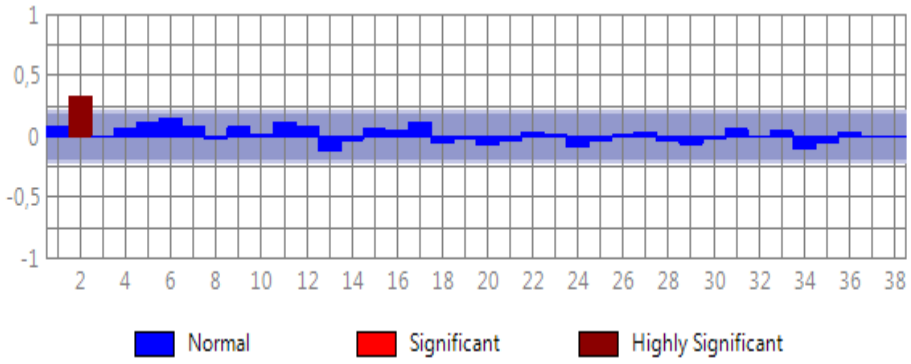


Gráfico 14: Autocorrelaciones Parciales del Error para Modelo lineal de Demanda Energética Comercial

4.2.2.4 Demanda de Energía Industrial

Para el consumo energético industrial no pudo lograrse una combinación de variables que permitiera cumplir con los criterios de estimación impuestos. En este caso se violó la independencia de errores, no obstante se pasó la prueba de homocedasticidad.

La ecuación que representa la relación lineal para la demanda de energía industrial es la siguiente:

$$IN_{t+1} = b_0 + b_1 IN_t + b_2 IN_{t-1} + b_3 IN_{t-4} + b_4 D3_{t+1} + b_5 D5_{t+1} + b_6 D6_{t+1} + b_7 D8_{t+1} + b_8 D10_{t+1} + b_9 D12_{t+1} + b_{10} L_{t+1} + b_{11} IM_{t+1} + b_{12} IM_t + b_{13} IM_{t-5} + b_{14} IM_{t-6} + e_t$$

Donde:

- IN_t: Demanda de Energía Eléctrica en el sector Industrial en el periodo t
- L_t: Laboralidad del Periodo t
- IM_t: IMACEC del período t
- $D3_t$: Variable Dummie Binaria para el mes de Marzo en el periodo t
- $D5_t$: Variable Dummie Binaria para el mes de Mayo en el periodo t
- $D6_t$: Variable Dummie Binaria para el mes de Junio en el periodo t

- $D8_t$: Variable Dummie Binaria para el mes de Agosto en el periodo t
- $D10_t$: Variable Dummie Binaria para el mes de Octubre en el periodo t
- $D12_t$: Variable Dummie Binaria para el mes de Diciembre en el periodo t

Serie de Tiempo	Rezago	Coeficiente de Regresión	Error Estándar	Estadístico T	P-valor
Demanda Industrial	0	0,3163	0,1034	3,0582	0,0034
	-1	0,3073	0,0819	3,7546	0,0004
	-4	0,2101	0,0879	2,3889	0,0202
V. Dummie (Octubre)	1	-0,256	0,0731	-3,4999	0,0009
V. Dummie (Diciembre)	1	-0,2187	0,0728	-3,0045	0,0039
V. Dummie (Marzo)	1	-0,2792	0,0943	-2,9591	0,0045
V. Dummie (Mayo)	1	-0,1967	0,0578	-3,4006	0,0012
V. Dummie (Julio)	1	-0,2585	0,0826	-3,1317	0,0027
V. Dummie (Agosto)	1	-0,2199	0,0659	-3,3385	0,0015
Laboralidad	1	0,4905	0,1042	4,7078	0
IMACEC	1	1,9472	0,3699	5,2641	0
	0	-1,5328	0,3848	-3,9838	0,0002
	-5	-1,0901	0,3029	-3,5992	0,0007
	-6	0,6222	0,3382	1,8396	0,071

Tabla 20: Resultados de Análisis de Regresión para la Demanda Industrial

En este caso la preponderancia del IMACEC es nuevamente muy grande al analizar los coeficientes de regresión. No obstante, y como se verá a continuación, la performance del modelo no es muy buena dado que no se alcanza un coeficiente de determinación mayor a 0,8 lo que se refleja en un hecho poco esperado: La gran relevancia que toman las variables dummies, dado que se ocupan en general como una ayuda para la descripción de la estacionalidad de las series.

Estadísticas	Valor
Observaciones Incluidas	72
Coeficientes Incluidos	15
R-Cuadrado	0,7706
R-Cuadrado Ajustado	0,7143

Tabla 21: Desempeño de Análisis de Regresión para la Demanda Industrial

También es destacable el hecho de que existen coeficientes con un valor mayor a 1 en el caso del IMACEC, no obstante hay una "compensación" dado que en sus mismos rezagos anteriores tiene coeficientes mayores a la unidad pero con el signo contrario.

A continuación se muestran los resultados de la prueba de homocedasticidad y los gráficos de autocorrelación de los errores.

Cambio en la Dispersión (Hipótesis Alternativa)	p-valor
Cox-Stuart (Dispersión)	0,5847

Tabla 22: Prueba de Homocedasticidad para los errores del Modelo de Energía Industrial

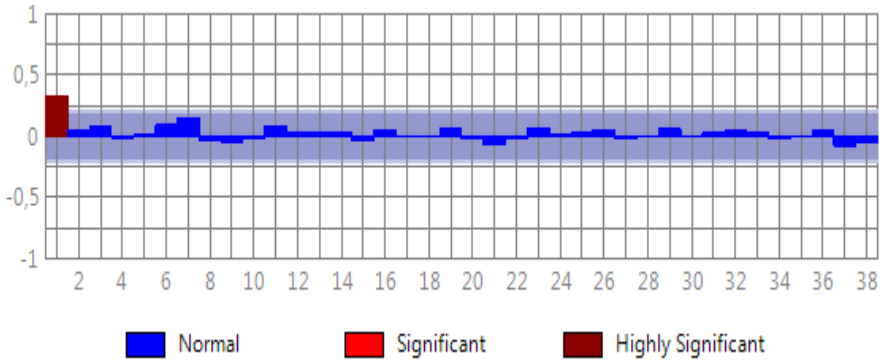


Gráfico 15: Autocorrelaciones del Error para Modelo lineal de Demanda Energética Industrial

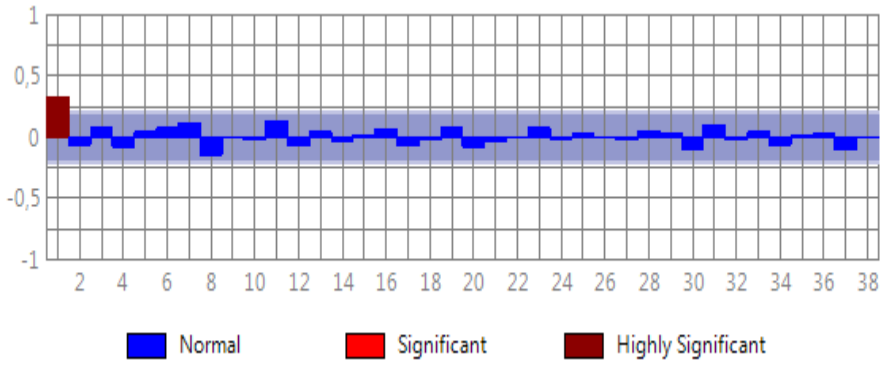


Gráfico 16: Autocorrelaciones Parciales del Error para Modelo lineal de Demanda Energética Industrial

La dependencia del error respecto del periodo anterior estaría realizando una clara referencia al supuesto que se viola al introducir rezagos en los modelos de regresión lineal, no obstante se espera que esto pueda ser enmendado al ocupar otros métodos que pueden captar distintas relaciones entre las variables.

4.2.2.5 Demanda Máxima de Potencia en el Anillo

Para la construcción de estructuras variables explicativas para el caso de la potencia máxima en el anillo, se realizan dos cambios notorios respectos de los modelos de energía.

El primer cambio consiste en la utilización de variables respectivas a la temperatura como la máxima y mínima promedio de cada mes, así como también las extremas. La lógica detrás de esto concuerda con la definición de la serie a explicar. La potencia máxima es un registro que ocurre en un momento determinado del mes, donde la demanda de energía eléctrica en un momento alcanzó su peak[27], luego es de esperar que las causas de este hecho estén asociadas a alguna condición crítica dentro del mes.

El otro cambio notable consiste en que el modelo de Potencia Máxima utiliza la energía demandada como variable explicativa, lo cual resulta bastante lógico dada la definición de la Potencia (que básicamente puede ser expresada como el cociente entre Energía y una unidad de Tiempo).

En consecuencia de lo anterior, y tomando en cuenta la preponderancia que tiene el IMACEC en los modelos de energía, el IMACEC no formará parte de las variables explicativas a proponer.

Los experimentos para esta serie no lograron reproducir los criterios establecidos, violándose tanto la independencia de los errores como la homocedasticidad. Sin embargo, se logra un buen coeficiente de determinación.

La ecuación lineal obtenida en este caso es la siguiente:

$$P_{t+1} = b_0 + b_1 P_t + b_2 L_{t+1} + b_3 Tmin_{t+1} + b_4 ES_{t+1} + e_t$$

[27] Recordar que la Potencia Eléctrica es la cantidad de energía entregada o absorbida por un elemento en un tiempo predeterminado.

Donde:

- P_t: Demanda de Máxima de Potencia en el Anillo en el periodo t
- L_t: Laboralidad del Periodo t
- ES_t: Demanda de Energía Eléctrica en el Sistema en el período t
- $Tmin_t$: Temperatura Mínima Promedio en el período t

Serie de Tiempo	Rezago	Coeficiente de Regresión	Error Estándar	Estadístico T	P-valor
Término Constante		0,1417	0,0299	4,7466	0
Demanda Máxima de Potencia en el Anillo	0	0,102	0,0594	1,7158	0,0904
Laboralidad	1	-0,2192	0,0326	-6,7259	0
T° Mínima	1	-0,0586	0,0145	-4,0365	0,0001
Demanda de Energía en el Sistema	1	0,9421	0,0668	14,0987	0

Tabla 23: Resultados de Análisis de Regresión para la Demanda Máxima de Potencia en el Anillo

Estadísticas	Valor
Observaciones Incluidas	67
Coeficientes Incluidos	5
R-Cuadrado	0,9146
R-Cuadrado Ajustado	0,91

Tabla 24: Desempeño de Análisis de Regresión para la Demanda Máxima de Potencia en el Anillo

Destaca para esta ocasión la simplicidad del modelo obtenido y, al mismo tiempo, la buena performance alcanzada. La gran preponderancia que tiene la demanda de energía responde claramente a la relación física intrínseca que tiene con la Potencia. También se destaca que variables como la T° Máxima ó otras variables climatológicas fueron descartadas, y al mismo tiempo la incluida presenta la más baja preponderancia en el modelo. Se desprende de esto que las relaciones de las variables climáticas con la potencia no son lineales, y considerando esto al momento de implementar algoritmos de inteligencia artificial se agregarán otras variables climatológicas.

A continuación se muestran los resultados de la prueba de homocedasticidad y los gráficos de autocorrelaciones.

Cambio en la Dispersión (Hipótesis Alternativa)	p-valor
Cox-Stuart (Dispersión)	0,0351

Tabla 25: Prueba de Homocedasticidad para los errores del Modelo de Potencia Máxima

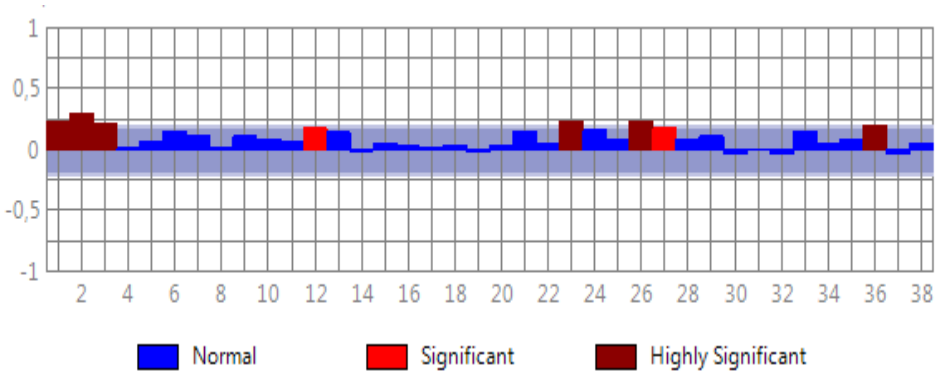


Gráfico 17: Autocorrelaciones del Error para Modelo lineal de Demanda Máxima de Potencia en el Anillo

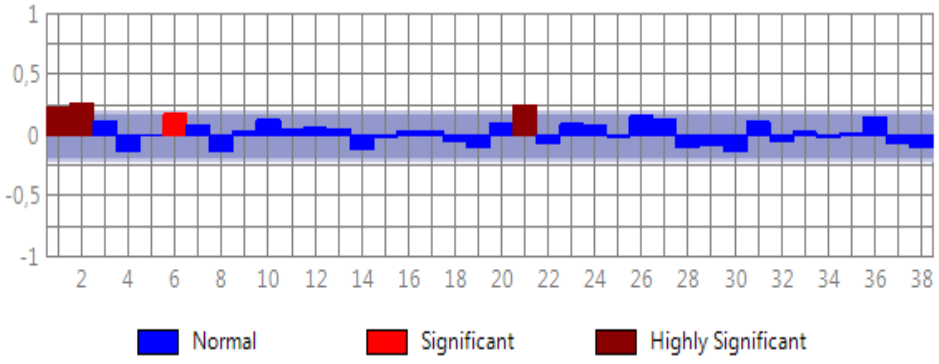


Gráfico 18: Autocorrelaciones Parciales del Error para Modelo lineal de Demanda Máxima de Potencia en el Anillo

Los gráficos de autocorrelación arrojan una cantidad considerables de rezagos que muestran dependencia significativa con el error actual. Por otra parte, la prueba de homocedasticidad es rechazada, aunque dado el *p-valor*, no alcanza a serlo a un nivel de significancia del 99%[28].

[28] Para que esto ocurriese el p-valor debiese ser menor a 0,01

Las variables seleccionadas según el análisis muestran la importancia de tener un pronóstico certero de la energía demanda en el sistema para reproducir los mismos resultados para la potencia. Esta necesidad se repetirá también para las otras variables explicativas utilizadas que requieren de tener una proyección certera para poder obtener una buena performance.

En la siguiente sección se tratarán los métodos utilizados para proyectar las variables explicativas relevantes.

4.2.3 Variables Seleccionadas

Cómo se mencionó al final del capítulo anterior, el análisis de regresión es una base para la selección de variables finales a utilizar. Las variables y los respectivos utilizados se muestran a continuación, donde los rezagos aparecen entre paréntesis, y marcados en negrita están los cambios agregados. Por último, se agrega un asterisco a la derecha de la variable que resultó ser la más relevante en el análisis de regresión.

Modelo de Demanda de Energía en el Sistema

- Demanda de Energía en el Sistema (0, -3, -10, -11)
- Laboralidad (1, **0**)
- Temperatura Media (transformada) (1)
- IMACEC (1,-6) *****
- Horas de Sol (1)
- Humedad Relativa a las (2pm) (1)
- Precio de la Energía (1)

Modelo de Demanda de Energía Residencial

- Demanda de Energía Residencial (0, -6, -7, **-11**, -12)
- Laboralidad (1)
- Precio de la Energía (1)
- Horas de Sol (1)
- IMACEC (**1**,-1,-3,-11) *****
- Temperatura Positiva (1)

Modelo de Demanda de Energía Comercial

- Demanda de Energía Comercial (0, -1, -3, -7, -9)
- Laboralidad (1)
- Temperatura Positiva (1)
- IMACEC (1, **-1, -3, -7, -9**) *****
- Precio de la Energía (1, 0)

Modelo de Demanda de Energía Industrial

- Demanda de Energía Industrial (0, -1, -4, **-11**)
- Laboralidad (1)
- IMACEC (1, 0, -5, -6, **-11**) *****
- **Precio de la Energía (1)**

Modelo de Demanda de Potencia Máxima en el Anillo

- Dda. Potencia Máxima en el Anillo (0, **-2, -7, -9, -11**)
- Variables Dummies para cada mes (1)
- Laboralidad (1)
- **Temperatura Máxima promedio mensual (1)**
- Demanda de Energía en el sistema (1) *****
- **Humedad Relativa (1)**
- Temperatura Mínima promedio mensual (1)

Se agregaron variables climatológicas como la temperatura máxima para que, en conjunto con la temperatura mínima, se pudiera entregar a los algoritmos una descripción de la varianza que tiene la temperatura en cada mes, específicamente para el caso de la potencia máxima.

Además, si bien con la temperatura mínima se pudo lograr una buena relación en el caso de la potencia, se teme que no entregue suficiente información sobre las condiciones en que ocurrió la potencia máxima del mes[29].

[29] La potencia máxima de cada mes ocurre en horarios laborales, mientras que la temperatura mínima de cada día ocurre en un intervalo entre las 6am y las 8am.

Por otro lado, el precio de la energía entra en todos los modelos de energía, dado que se quiso probar la importancia que esta serie tiene en la demanda.

Respecto a los rezagos agregados, responden a la iniciativa de querer agregar más información para que los métodos de inteligencia artificial capten la relación existente.

4.2.3.1 Endogeneidad de IMACEC

Un punto no mencionado anteriormente tiene que ver con la endogeneidad del IMACEC respecto de las series de Energía. Según [19], "Una variable es endógena cuando sus valores están determinados dentro del modelo y es predeterminada o exógena cuando sus valores se determinan fuera del modelo". En este caso, es lógico pensar que el IMACEC presentaría endogeneidad respecto de las demandas de energía, sobre todo la demanda del sistema, dada que una mayor demanda energética representaría una mayor actividad económica, y a su vez una mayor actividad económica una mayor demanda energética.

La problemática de esto se traduce en que, en los modelos estimados mediante MCO, implica un sesgo en el estimador de esta variable explicativa [19], al estar correlacionados los errores de estimación de la variable dependiente con la variable independiente.

Independiente del problema de endogeneidad, este sesgo antes mencionado es esperable en los modelos planteados para el análisis de regresión dado que se utilizan variables con rezagos, lo que genera de por sí una correlación intertemporal entre los errores de estimación [47]. No obstante, se realizaron pruebas y se revisó bibliografía pertinente para confirmar la utilización de esta variable.

Según Urrutia y Sanchéz [45], entre el IMACEC y el consumo de energía a nivel nacional no existe una relación de causalidad, resultado obtenido a través de la prueba de causalidad de Granger [51]. Este hecho presenta un primer antecedente para negar la endogeneidad entre las variables a pronosticar en este trabajo y el IMACEC, no obstante, existe una diferencia crucial entre las series utilizadas en la prueba, dado que la demanda de energía a nivel nacional no presenta el mismo comportamiento que la demanda que percibe de energía que percibe la red de Chilectra.

Luego, el argumento que evitaría la endogeneidad entre el IMACEC y la demanda de energía que recibe Chilectra se basa en que la variable macroeconómica representa la actividad de todo Chile, mientras que la zona de concesión de la empresa sólo se enfoca en la región metropolitana.

Con el fin de comprobar esto, se realizó la prueba de Granger sobre las series desestacionalizadas de energía del sistema con la serie del IMACEC. En este caso las pruebas demostraron que no se obtienen los mismos resultados que obtuvieron Urrutia y Sánchez, dado que la prueba arrojó que existe una relación bidireccional, es decir, el IMACEC causa la demanda de energía, y la demanda de energía causa el IMACEC, lo que sería un primer indicio de la endogeneidad existente entre ambas variables[30].

Estos resultados llevaron a la realización de prueba de Hausman, utilizada en [97] para contrarrestar el problema de endogeneidad. Una variación de esta prueba permite determinar si existe efectivamente endogeneidad entre la variable dependiente y la independiente. Los resultados de esta prueba arrojaron que efectivamente existe endogeneidad entre las variables[31].

Algo a destacar sobre estos resultados tiene relación con el modelo de proyección de IMACEC que proponen Urrutia y Sánchez [45], donde utilizan consumo total de energía del país como variable explicativa[32]. Con las pruebas realizadas se han entregado indicios de que el modelo expuesto podría ser sesgado, factor que no es menor dado que es uno de los modelos utilizados por el Banco Central de Chile [6], y que además presenta buenos resultados.

A pesar del sesgo que produce la utilización de estas variables, la utilización de variables macroeconómicas derivadas del PIB son introducidas en varios modelos de pronóstico de energía [12] [43], lo que induce que este sesgo a generarse es un costo a asumir para obtener proyecciones a largo plazo.

---

[30] Detalles en Anexos 8.F Prueba de Granger

[31] Detalles de resultados en Anexos 8.G Prueba de Hausman

[32] En específico utilizan los despachos de energía eléctrica, y además se utiliza el supuesto de que los despachos son iguales a la demanda al considerar que las pérdidas que ocurren en distribución son parte de la generación.

4.3 Pronóstico de Variables Relevantes

La metodología seleccionada implica la proyección de las variables objetivos con el fin de cumplir la relación planteada en el análisis de regresión en el futuro.

Los modelos causales de proyección tienen la desventaja que su performance no sólo depende de la calidad ó veracidad de la relación entre las variables explicativas y la dependiente propuesta, sino que también de la calidad del pronóstico realizado sobre las variables explicativas.

Considerando las validaciones realizadas al modelo anterior implementado en la empresa, donde se obtuvo un MAPE menor a un 5% en promedio entre varios horizontes de 5 años, se propuso utilizar la misma metodología de pronóstico de variables explicativas en primera instancia para posteriormente evaluarlas e implementar cambios en caso de que sea necesario.

En las siguientes sub-secciones se abordarán las metodologías utilizadas para la proyección de las variables explicativas, separadas según su respectiva clasificación, a utilizar en los modelos de Energía y Potencia Máxima en el Anillo, incluyendo a su vez los respectivos resultados

4.3.1 Pronóstico de Variables Climatológicas

Los modelos de Energía y Potencia utilizan, en su conjunto, las siguientes variables climatológicas para su proyección:

- Temperatura Media Mensual
- Temperatura Máxima[33] y Mínima Promedio del mes
- Humedad Relativa Promedio a las 2pm en el mes
- Horas de Sol en el Mes

La proyección se realiza mediante un proceso en el cual se calculan los valores promedios de cada mes a través de los años y una vez obtenido una curva con, por ejemplo, las horas de sol promedio

[33] Es una variable que es considerada importante al momento de describir los peaks de demanda de energía y potencia que ocurren en Verano.

para cada mes a través de los años, se utilizan estos valores como la proyección, repitiendo los valores a través de los años del horizonte de pronóstico.

La proyección se podría expresar de la siguiente forma:

$$Tmax_m = \sum_{2001}^{2014} Tmax_{m(i)} /14$$

Dónde:

- $Tmax_m$: Temperatura máxima para el mes "m" (cualquiera del año)
- $Tmax_{m(i)}$: Temperatura máxima registrada en el mes "m" en el año i

Este procedimiento se compara a realizar un pronóstico de mediante promedio móvil simple para un horizonte de un período, donde luego este pronóstico se repite para los 4 períodos posteriores (completando así la proyección a 5 años que se requiere para el modelo final).

De esta forma, el procedimiento de proyección se repite para todas las variables explicativas, donde se realizaron pruebas para comparar cual es la cantidad de años anteriores a considerar que permite realizar el mejor pronóstico, ó en otras palabras, determinar el número de observaciones óptimas para el promedio móvil que minimice el error.

En específico, se realizaron pruebas de pronósticos considerando **3, 5, 7 y 10** observaciones de años anteriores para los métodos de promedios móviles de cada mes.

Una vez realizadas los pronósticos con distintos números de observaciones a considerar para la media móvil, y en distintos horizontes de pronóstico, se eligió el modelo que presenta el menor error en promedio a través de los distintos horizontes de pronóstico.

Para el análisis de los resultados de la proyección, se utilizaron 5 horizontes de proyección al igual que los utilizados para la validación del modelo original que existía en la empresa, cada uno correspondiente a

un período de 5 años. Los horizontes de proyección utilizados para la validación fueron:

- 2009-2013
- 2008-2012
- 2007-2011
- 2006-2010
- 2005-2009

4.3.1.1 Resultados para T° Media

Con el fin de comparar los desempeños de cada uno de los modelos de Promedio Móvil, en la siguiente tabla se resume los errores obtenidos a través de los horizontes y los distintos métodos.

Método	MAPE					
	2009-2013	2008-2012	2007-2011	2006-2010	2005-2009	**Promedio**
P.M.(3)	5,73%	6,42%	8,89%	7,77%	5,46%	**6,86%**
P.M.(5)	5,79%	6,36%	7,94%	7,01%	5,71%	**6,56%**
P.M.(7)	5,64%	6,05%	7,53%	6,90%	5,65%	**6,36%**
P.M.(10)	5,52%	5,94%	7,38%	6,77%	5,58%	**6,24%**

Tabla 26: Desempeño Final de Pronóstico para T° Media

Como se aprecia en la tabla, el método con mejor desempeño global es el método de Promedio Móvil considerando 10 años anteriores para realizar el pronóstico. A partir de esto, este es el método seleccionado para realizar el pronóstico de la variable explicativa *"T° Media"* en la validación de los modelos de Energía y Potencia.

4.3.1.2 Resultados para T° Mínima

Método	MAPE					
	2009-2013	2008-2012	2007-2011	2006-2010	2005-2009	**Promedio**
P.M.(3)	16,45%	18,39%	28,10%	22,20%	15,86%	**20,20%**
P.M.(5)	17,70%	17,95%	24,83%	20,51%	15,49%	**19,30%**
P.M.(7)	17,96%	17,65%	24,01%	19,50%	14,80%	**18,78%**
P.M.(10)	16,97%	16,40%	23,19%	18,92%	15,06%	**18,11%**

Tabla 27: Desempeño Final de Pronóstico para T° Mínima

Nuevamente, la media móvil que muestra mejor desempeño es la que utiliza 10 rezagos. Si bien en este el MAPE con el cual se trabaja es mayor, no representa un error tan grande al ver el MAE, dado que los

errores bordean 1°C de magnitud, similar al caso de la Temperatura media.

4.3.1.3 Resultados para T° Máxima

Método	MAPE					
	2009-2013	2008-2012	2007-2011	2006-2010	2005-2009	**Promedio**
P.M.(3)	4,15%	4,87%	4,43%	3,90%	4,32%	**4,334%**
P.M.(5)	3,99%	4,11%	4,24%	4,00%	4,33%	**4,133%**
P.M.(7)	3,56%	4,42%	4,39%	4,22%	4,04%	**4,126%**
P.M.(10)	3,93%	4,62%	4,40%	4,05%	4,42%	**4,284%**

Tabla 28: Desempeño Final de Pronóstico para T° Máxima

En esta ocasión el utilizar 7 rezagos para la media móvil obtiene el menor error, por lo cual se utilizan los pronósticos de esta metodología para la construcción de los modelos de energía y el de potencia.

Se aprecia que los métodos tienen casi el mismo error, el que además es bastante bajo considerando el MAPE.

4.3.1.4 Resultados para Humedad Relativa a las 2pm

Método	MAPE					
	2009-2013	2008-2012	2007-2011	2006-2010	2005-2009	**Promedio**
P.M.(3)	16,19%	14,97%	18,66%	18,04%	14,37%	**16,45%**
P.M.(5)	18,12%	18,66%	18,81%	22,79%	15,01%	**18,68%**
P.M.(7)	18,33%	20,99%	21,09%	23,02%	10,81%	**18,85%**
P.M.(10)	17,74%	21,87%	21,21%	22,98%	14,67%	**19,69%**

Tabla 29: Desempeño Final de Pronóstico para humedad relativa a las 2pm

Para el pronóstico de la Humedad Relativa Promedio en el mes a las 2pm, la metodología que considera una media móvil con 3 rezagos obtiene los mejores resultados.

4.3.1.5 Resultados para Horas de Sol

Método	MAPE					
	2009-2013	2008-2012	2007-2011	2006-2010	2005-2009	**Promedio**
P.M.(3)	10,16%	11,97%	11,57%	10,51%	13,49%	**11,54%**
P.M.(5)	9,86%	10,87%	11,27%	10,33%	12,44%	**10,96%**
P.M.(7)	9,63%	10,71%	11,51%	11,16%	13,36%	**11,27%**
P.M.(10)	9,89%	10,71%	12,19%	11,25%	13,19%	**11,45%**

Tabla 30: Desempeño final de pronóstico para horas de sol

Para la última variable climatológica a utilizar en los modelos de proyección de Energía y el de Potencia, el promedio móvil considerando 5 rezagos cumple con un mayor desempeño.

4.3.2 Pronóstico de IMACEC

La proyección de esta variable se basó en 2 aspectos principales:

- Estacionalidad del IMACEC
- Expectativas de crecimiento del PIB

La primera premisa hace referencia a un aspecto clave de cualquier serie, que proviene por el comportamiento "clásico" de una serie dentro de un período ó estación. En este caso el IMACEC presenta una clara estacionalidad asociada a la actividad económica típica que tiene cada mes del año.

La segunda premisa tiene que ver con la relación existente entre el IMACEC y el PIB. Dado que el IMACEC representa el 90% bienes y servicios que componen el PIB [52], es lógico asumir que el IMACEC debiese de representar una imagen mensual del indicador anual antes mencionado, luego, comparando el IMACEC total de un año respecto de uno adyacente, se debiesen de tener crecimientos similares a los que se registran en el PIB [13].

En base a esto último, se esperaría que las expectativas de crecimiento del PIB que publica el Banco Central de Chile (de manera mensual) fuesen igualmente similares a las tasas de crecimiento anuales del IMACEC en el futuro.

Esta situación permite realizar una proyección de la variable explicativa utilizando como base las expectativas de crecimiento del PIB

que publica el BB.CC., lo que finalmente se traduce en realizar un pronóstico que cuadre con estas expectativas, pero que además tenga la estacionalidad correspondiente.

De esta forma, para la proyección del PIB se desarrolló una metodología donde se estima la estacionalidad típica del IMACEC en base a la relación (cociente) que existe entre el valor del mes de Enero de un año con los otros meses. La relación se calcula en base al IMACEC de 5 años previos al horizonte de pronóstico, para luego calcular el promedio de las relaciones existentes entre el valor de la serie en el mes de Enero respecto de los otros meses.

Una vez calculada esta relación promedio, se utilizaron las expectativas del PIB (que incluyen el crecimiento esperado del PIB del año actual y la de dos años sucesores) junto con la función de Excel "*Goal Seek*"[34], de manera de calcular los valores que debiese seguir el IMACEC para lograr los crecimientos estipulados, además de cumplir con la estacionalidad promedio calculada.

El detalle de la metodología y su aplicación se encuentra en Anexos.

4.3.2.2 Validación de Metodología

La metodología propuesta fue validada en los mismos horizontes de pronóstico que en las otras variables (que son los mismos que se utilizaron para evaluar el desempeño del modelo que existía en la empresa). En cada uno de los horizontes se utilizaron las expectativas de crecimiento del PIB de Diciembre del año anterior al comienzo del horizonte, es decir, si el horizonte de pronóstico es desde 2009 a 2014, se utilizaron las expectativas de Diciembre de 2008.

A continuación se mostrarán los resultados generales obtenidos en cada horizonte.

[34] Este trabajo fue realizado con una versión en Inglés del Software.

Desempeño Promedio

Métrica	Horizontes de Pronóstico					Promedio
	2009-2013	2008-2012	2007-2011	2006-2010	2005-2009	
MAPE	2,85%	2,16%	2,90%	3,37%	2,28%	**2,71%**
MAE	3,255	2,263	2,958	3,372	2,228	**2,815**

Tabla 31: Errores de Pronóstico de metodología de IMACEC

El desempeño del pronóstico del IMACEC es bastante bueno, considerando que un error de un 2,71% en el largo plazo es considerado bajo [43].

4.4 Aplicación de Técnicas de Inteligencia Artificial para el Pronóstico de Demandas

Cuando ya se ha terminado con el pronóstico de las variables explicativas a utilizar en el modelo, se procede a continuación con la aplicación de los métodos finales de pronóstico, correspondientes a Support Vector Regression y Redes Neuronales Artificiales.

La realización de estos experimentos se llevó a cabo utilizando los mismos horizontes de pronóstico que fueron usados anteriormente, los cuáles son:

- Horizonte 1: Enero 2009 a Diciembre 2013
- Horizonte 2: Enero 2008 a Diciembre 2012
- Horizonte 3: Enero 2007 a Diciembre 2011
- Horizonte 4: Enero 2007 a Diciembre 2010
- Horizonte 5: Enero 2007 a Diciembre 2009

Estas metodologías de pronóstico fueron aplicadas al conjunto de variables explicativas seleccionadas, incluyendo los rezagos que se consideraron relevantes durante el análisis de regresión.

Los datos utilizados para el entrenamiento en cada horizonte son:

- Horizonte 1: Junio de 2001 hasta Diciembre 2008 (91 registros)
- Horizonte 2: Junio de 2001 hasta Diciembre 2007 (79 registros)
- Horizonte 3: Junio de 2001 hasta Diciembre 2006 (67 registros)
- Horizonte 4: Junio de 2001 hasta Diciembre 2005 (55 registros)
- Horizonte 5: Junio de 2001 hasta Diciembre 2004 (43 registros)

En las siguientes secciones se muestran los mejores resultados obtenidos de la aplicación de cada uno de éstos métodos a las series a pronosticar y en cada uno de los horizontes correspondientes.

Al igual que en el caso del análisis de regresión, se propuso como condición de término, en caso de ser alcanzable, elegir el modelo con menor error de pronóstico pero que además cumpliera con las condiciones de Homocedasticidad e independencia temporal de los errores.

4.4.1 Aplicación de Support Vector Regression

En base a las variables seleccionadas en el análisis de regresión y sus respectivos rezagos, se realizaron pruebas de SVR con tres kernel distintos: "*Radial Basis Function*" (RBF de ahora en adelante), polinomial, y lineal.

En las distintas pruebas realizadas se probaron con distintas configuraciones, donde se incluyeron las siguientes variaciones:

- Variación del parámetro "γ" (Gamma) de RBF
  - Valores utilizados: 0.1, 3.1, 5.1, 7.1 y 9.1
- Kernel Polynomial de Grado 3
- Kernel Lineal
- Valor de Epsilon: 0.1, 0.2, 0.3, …, 3.1
- Costo: 0.5, 0.7, 0.8, 1, 1.3, 1.5, 1.8, 2

Dado a que las primeras pruebas se obtuvieron buenos resultados con unas configuraciones (para la Demanda de Energía del Sistema), para los experimentos posteriores se acotaron las configuraciones posibles.

Cabe destacar que para hacer el entrenamiento de los modelos se ocupó la mayor cantidad de datos disponibles, significando esto, por ejemplo, que en el caso de la demanda de energía del sistema, donde sólo se contaban con datos desde Junio de 2001, que a medida que el horizonte de pronóstico se iba haciendo más lejano a la fecha antes mencionadas, más muestras eran ingresadas al modelo.

Los detalles de los pronósticos realizados se incluyen en Anexos en las secciones 8.D.a y 8.D.c.

4.4.1.1 Demanda de Energía del Sistema

El mejor modelo corresponde a la siguiente configuración de SVR:

- Gamma (de RBF): 0.1
- Costo: 0.3
- Epsilon: 0.1

A continuación se muestran la tabla con los resultados obtenidos de esta configuración en todos los horizontes de pronóstico.

Métricas	Horizontes					
	2009-2013	2008-2012	2007-2011	2007-2010	2007-2009	Promedio
MAPE Entrenamiento	2,98%	2,64%	2,61%	2,87%	2,68%	**2,75%**
MAPE Validación	5,33%	4,42%	3,96%	7,70%	10,21%	**6,32%**
Cox-Stuart (P-valor)	0,009	**0,1496**	**0,0872**	0,0002	0,0009	
Errores Independientes	No	No	No	No	No	

Tabla 32: Desempeño de SVR en Pronóstico de Energía Sistema

El método tuvo resultados aceptables respecto a los errores, no obstante, el promedio del desempeño no mejora respecto del modelo de la empresa (MAPE promedio de 6,21%), además que en algunos casos el pronóstico es sistemáticamente menor que el valor esperado.

Cabe también destacar que en todos los horizontes no se cumplió la condición de independencia de los errores, visualizando esto a través de la cantidad de rezagos que tienen tanto una autocorrelación como autocorrelación parcial significativa.

Respecto de la homocedasticidad de los errores mediante la prueba de Cox-Stuart, en tres de cinco casos no se cumplió a un nivel de significancia del 99% (se remarcan los resultados en que no hay suficiente evidencia para descartar homocedasticidad).

4.4.1.2 Demanda de Energía Residencial

La configuración utilizada en el modelo SVR que obtuvo mejores resultados es la siguiente:

- Gamma (de RBF): 0.1
- Costo: 0.7
- Epsilon: 0.1

Los resultados obtenidos se muestran en la siguiente tabla.

Métricas	Horizontes					
	2009-2013	2008-2012	2007-2011	2007-2010	2007-2009	Promedio
MAPE Entrenamiento	3,04%	2,75%	2,90%	2,82%	2,97%	**2,89%**
MAPE Validación	6,39%	5,03%	5,46%	5,06%	7,31%	**5,85%**
Cox-Stuart (P-valor)	**0,1214**	**0,281**	0,0125	**0,2295**	**0,1338**	
Errores Independientes	No	Si	No	Si	No	

Tabla 33: Desempeño de SVR en Pronóstico de Energía Residencial

Este es el primer resultado del cual no se tiene una métrica de comparación previa, por lo cual se considera que en cuanto a nivel de error, el método de SVR se desempeña de buena manera según los parámetros señalados en la bibliografía [43].

La homocedasticidad de los errores se cumple en cuatro de 5 horizontes, lo que es considerado como un buen resultado, no obstante el modelo tiene problemas para explicar los peak de demanda que ocurren en los meses de invierno.

En cuanto a la independencia de los errores, solamente se logró en uno de los 4 horizontes, en el cuál no se presentan rezagos significativos (respecto de la autocorrelación parcial) hasta después del primer año.

4.4.1.3 Demanda de Energía Comercial

La configuración utilizada en el modelo SVR que obtuvo mejores resultados es la siguiente:

- Gamma (de RBF): 0.1
- Costo: 3
- Epsilon: 0.1

Métricas	Horizontes					
	2009-2013	2008-2012	2007-2011	2007-2010	2007-2009	Promedio
MAPE Entrenamiento	4,30%	4,17%	4,01%	3,84%	4,10%	**4,08%**
MAPE Validación	15,09%	7,33%	7,61%	6,82%	5,37%	**8,44%**
Cox-Stuart (P-valor)	0,0001	**0,0501**	0,0081	0,0094	**0,21**	
Errores Independientes	No	No	No	No	No*	

Tabla 34: Desempeño de SVR en Pronóstico de Energía Comercial

De la tabla se puede ver que en el primer horizonte el algoritmo no entrega buenos resultados, no obstante los otros 4 tienen un desempeño, al menos en nivel de errores, similar.

Los pronósticos no pudieron captar la forma general de la demanda de energía comercial, repitiéndose en cuatro ocasiones que la curva pronosticada es un acercamiento a la tendencia de la serie original, pero falla al momento de captar la estacionalidad.

Considerando este desempeño, en dos casos se estableció homocedasticidad, aunque se considera que es un resultado fortuito y no representa la capacidad del modelo para captar el comportamiento intrínseco de la curva.

Este hecho se refuerza viendo que en todos los horizontes no hay independencia de los errores. Se marca el último horizonte con un asterisco dado que solamente existe un rezago que muestra una autocorrelación parcial significativa, factor que en algunos casos puede ser efecto de la utilización de variables rezagadas en el modelo, lo que daría una ventana a que fuese igual considerado como un buen resultado en el marco de la independencia de los errores.

4.4.1.4 Demanda de Energía Industrial

La configuración utilizada en el modelo SVR que obtuvo mejores resultados es la siguiente:

- Gamma (de RBF): 0.1
- Costo: 3
- Epsilon: 0.1

Métricas	Horizontes					
	2009-2013	2008-2012	2007-2011	2007-2010	2007-2009	Promedio
MAPE Entrenamiento	3,30%	2,42%	2,77%	3,36%	3,91%	**3,15%**
MAPE Validación	5,20%	6,57%	4,73%	5,42%	5,34%	**5,45%**
Cox-Stuart (P-valor)	**0,2295**	0,0002	0,0037	**0,2295**	**0,5235**	
Errores Independientes	Si	No	No*	No*	No	

Tabla 35: Desempeño de SVR en Pronóstico de Energía Industrial

El desempeño en cuanto al nivel de errores puede ser considerado como bueno, no solamente por el valor promedio del error, sino que también por el hecho de que no existe mucha variación entre los distintos horizontes.

No obstante, al igual que en el caso anterior, el pronóstico no capta de buena forma el comportamiento de la serie, donde a veces ni siquiera se capta la tendencia, y en la mayoría de los casos es una curva con poca varianza y sin crecimiento que trata de minimizar el error[35].

La homocedasticidad se alcanza en tres casos, aunque dado lo mencionado antes no toma mucha importancia este análisis, así como tampoco el analizar la independencia de los errores.

4.4.1.5 Demanda de Potencia Eléctrica en el Anillo

La configuración utilizada en el modelo SVR que obtuvo mejores resultados es la siguiente:

- Gamma (de RBF): 0.1
- Costo: 2
- Epsilon: 0.1

Métricas	Horizontes					
	2009-2013	2008-2012	2007-2011	2007-2010	2007-2009	Promedio
MAPE Entrenamiento	3,05%	2,92%	3,01%	3,07%	2,96%	**3,00%**
MAPE Validación	7,58%	6,52%	8,63%	9,85%	10,71%	**8,66%**
Cox-Stuart (P-valor)	0,0002	0	0,0009	0,0009	**0,0525**	
Errores Independientes	No	No	No	No	No	

Tabla 36: Desempeño de SVR en Pronóstico de Potencia Máxima en el Anillo

El desempeño del algoritmo no logró mejorar el error respecto del modelo anterior, obteniendo en promedio un error mayor en un 3%.

En general se puede ver que el modelo lograr captar la tendencia de crecimiento, exceptuando el primer horizonte donde no sólo se logra en el primer año de pronóstico.

A su vez, el modelo no logra homocedasticidad en cuatro de los cinco horizontes, mientras que en ninguno se logra la independencia de los errores.

[35] Ver Anexos sección 8.D.a

4.4.2 Aplicación de Redes Neuronales Artificiales

La utilización de las redes neuronales incluyó considerable cantidad de configuraciones, tomando en cuenta la cantidad de factores que pueden influir en la calidad del modelo.

Considerando que al utilizar registros mensuales no son muchos los datos disponibles, y por ende tampoco pueden ser muchos los datos de entrada a la red, las pruebas realizadas consideraron una arquitectura de red pequeña, solamente utilizando una capa oculta.

Las funciones de activación utilizadas fueron la logística y la tangente hiperbólica. Por otra parte, los pesos iniciales eran generados de manera aleatoria.

Los detalles de los resultados se pueden ver en la sección 8.D.b de Anexos.

4.4.2.1 Demanda de Energía del Sistema

La Configuración de red utilizada con los mejores resultados fue la siguiente:

- 8 neuronas en la capa oculta
- Función de Activación logística ($\alpha = 1$)
- Tasa de Aprendizaje: 0,85
- Tasa de Decrecimiento (por Época): 0,96
- Épocas de Entrenamiento: 1.000

Métricas	Horizontes					
	2009-2013	2008-2012	2007-2011	2007-2010	2007-2009	Promedio
MAPE Entrenamiento	3,40%	2,88%	2,08%	3,06%	1,05%	**2,49%**
MAPE Validación	2,21%	3,10%	2,51%	4,01%	2,08%	**2,78%**
Cox-Stuart (P-valor)	**0,3915**	**0,1496**	0,0125	**0,4244**	**0,2863**	
Errores Independientes	No*	No*	Si	No	No*	

Tabla 37: Desempeño de RNA en Pronóstico de Energía Sistema

El nivel de errores que entrega el algoritmo es muy bajo, presentándose solamente una alza mayor en el horizonte 2006-2010, aunque en términos de pronóstico al largo plazo sigue siendo un

desempeño muy bueno, siendo esto además reforzado por el hecho de que el modelo no solamente capta de muy buena manera la tendencia de la serie, sino que también la estacionalidad que esta presenta[36].

En cuatro de los horizontes no se pudo descartar la homocedasticidad de los errores, no obstante de utilizar un nivel de confianza del 99% por ejemplo, no se podría descartar, llegando a cumplir la condición en todos los horizontes.

Respecto a la independencia de los errores, solamente se logró en un caso bajo los criterios establecidos, no obstante, en otros tres horizontes (marcados con un asterisco) solamente aparece un rezago con correlación altamente significativa, y como se mencionó anteriormente, en teoría esto se debe a la utilización de variables rezagadas en el pronóstico.

4.4.2.2 Demanda de Energía Residencial

La configuración utilizada es la siguiente:

- 8 neuronas en la capa oculta
- Función de Activación: Tangente Hiperbólica
- Tasa de Aprendizaje: 0,85
- Tasa de Decrecimiento (por Época): 0,99
- Épocas de Entrenamiento: 1.000

Métricas	Horizontes					
	2009-2013	2008-2012	2007-2011	2007-2010	2007-2009	Promedio
MAPE Entrenamiento	2,16%	2,02%	2,04%	1,77%	1,62%	**1,92%**
MAPE Validación	2,96%	3,79%	3,41%	3,65%	3,68%	**3,50%**
Cox-Stuart (P-valor)	**0,1214**	**0,0708**	**0,1849**	**0,2295**	**0,5235**	
Errores Independientes	Si	No	Si	No	No*	

Tabla 38: Desempeño de RNA en Pronóstico de Energía Residencial

El nivel de errores, al igual que en el caso anterior, es bastante bueno y no presenta grandes variaciones entre los horizontes. Si bien se lograron buenos resultados, existieron dos puntos que no pudieron ser bien pronosticados, siendo estos representados por peaks de demanda que ocurrieron en invierno.

[36] Ver Anexos sección 8.D.b

Además se destaca que en todas las pruebas se obtuvo la homocedasticidad de los errores, lo que representó la primera prueba en lograr esto.

Por último, la independencia de los errores se alcanza en dos casos, mientras que un tercero solamente presenta un rezago significativo.

Dada el bajo nivel y la homocedasticidad de los errores, la independencia temporal de estos pasaría a un segundo plano dadas las configuraciones de las variables de entrada utilizadas.

4.4.2.3 Demanda de Energía Comercial

Configuración de la red:

- 8 neuronas en la capa oculta
- Función de Activación logística ($\alpha = 1$)
- Tasa de Aprendizaje: 0,5
- Tasa de Decrecimiento (por Época): 0,99
- Épocas de Entrenamiento: 1.000

Métricas	Horizontes					
	2009-2013	2008-2012	2007-2011	2007-2010	2007-2009	Promedio
MAPE Entrenamiento	3,74%	3,16%	2,59%	2,60%	2,24%	**2,87%**
MAPE Validación	4,71%	4,22%	3,73%	3,80%	3,64%	**4,02%**
Cox-Stuart (P-valor)	**0,0895**	**0,1102**	**0,0614**	**0,029**	**0,4049**	
Errores Independientes	Si	Si	Si	Si	No	

Tabla 39: Desempeño de RNA en Pronóstico de Energía Comercial

Nuevamente se obtuvo buenos resultados respecto del nivel de error, donde el mayor error de pronóstico se ve reflejado en los peaks de demanda ocurridos en los primeros de enero de 2009. A pesar de esto, la tendencia de la serie fue captada de buena manera en todos los modelos.

La homocedasticidad, al igual que la independencia temporal de los errores se consiguió en cuatro de los cinco horizontes, reflejando de esta forma que el desempeño en los peaks de demanda mencionados anteriormente no tuvieron una fuerte influencia respecto de la dispersión de los errores.

4.4.2.4 Demanda de Energía Industrial

Configuración de la red utilizada:

- 6 neuronas en la capa oculta
- Función de Activación logística ($\alpha = 1$)
- Tasa de Aprendizaje: 0,85
- Tasa de Decrecimiento (por Época): 0,99
- Épocas de Entrenamiento: 1.000

Métricas	Horizontes					
	2009-2013	2008-2012	2007-2011	2007-2010	2007-2009	Promedio
MAPE Entrenamiento	3,20%	3,75%	2,48%	2,70%	3,10%	**3,05%**
MAPE Validación	4,25%	3,78%	3,96%	3,62%	3,45%	**3,81%**
Cox-Stuart (P-valor)	**0,1214**	**0,1496**	0,0009	**0,1078**	**0,1338**	
Errores Independientes	No*	Si	Si	Si	No*	

Tabla 40: Desempeño de RNA en Pronóstico de Energía Industrial

El desempeño general del algoritmo es bastante bueno, siendo la primera evidencia de esto el MAPE que se presentó a través de los horizontes. A diferencia que con la aplicación de SVR, el bajo error en este caso si está representado por una mejora sustancial respecto de cómo el pronóstico capta la tendencia de la serie en el conjunto de validación, y si bien existen peaks en la demanda que no son explicados correctamente, si se entrega un buen pronóstico que permite ver la tendencia general de la demanda.

La homocedasticidad de los errores solo fue rechazada en un horizonte de los total de cinco, mientras que los errores mostraron independencia en tres casos bajo los criterios establecidos, mientras que en los restantes solamente aparece un rezago altamente significativo dentro de un año de horizonte de análisis.

4.4.2.5 Demanda de Potencia Máxima en el Anillo

Configuración de red utilizada:

- 6 neuronas en la capa oculta
- Función de Activación logística ($\alpha = 1$)
- Tasa de Aprendizaje: 0,85
- Tasa de Decrecimiento (por Época): 0,99
- Épocas de Entrenamiento: 1.000

Métricas	Horizontes					
	2009-2013	2008-2012	2007-2011	2007-2010	2007-2009	Promedio
MAPE Entrenamiento	2,77%	2,81%	2,43%	2,29%	1,53%	**2,39%**
MAPE Validación	3,78%	4,40%	3,98%	3,31%	3,24%	**3,74%**
Cox-Stuart (P-valor)	**0,1338**	0,0146	0,0125	0	0	
Errores Independientes	No	No	No*	No	No	

Tabla 41: Desempeño de RNA en Pronóstico de Potencia Máxima en el Anillo

El nivel de errores para la potencia máxima vuelve a ser bajo, donde el modelo captó de buena forma la tendencia de la serie, no obstante existieron peaks de demanda, correspondientes a meses de invierno, que no pudieron ser explicados de buena forma en algunos horizontes, en especial el ocurrido el año 2007. No obstante, resulta interesante ver que algunos de estos peaks si son explicados de mejor forma cuando pertenecen a distintos horizontes.

La homocedasticidad de los residuos no fue descartada en uno de los horizontes, resultado que se explica debido al comportamiento más aleatorio de la serie respecto de las de energía.

Respecto de la independencia de los errores, en ningún caso se logró, aun así en un horizonte de pronóstico se dio el caso de que se presentaba un rezago con correlación altamente significativa.

4.4.3 Aplicación de SARIMA y SARIMAX

La aplicación de los métodos SARIMA y SARIMAX fueron realizados mediante la metodología de Box-Jenkins [16], tomando en cuenta que al estimar los modelos las series fueron integradas hasta ser estacionarias, y la agregación de más términos fue detenida al momento de presentar una autocorrelograma sin correlaciones significativas, incluyendo esto también a las parciales.

4.4.3.1 Demanda de Energía del Sistema

El modelo SARIMA estimado para esta serie fue:

$$SARIMA(1,1,2)x(1,1,0)_{12}$$

SARIMA	2009-2013	2008-2012	2007-2011	2007-2010	2007-2009	Promedio
MAPE Entrenamiento	2,32%	2,61%	2,33%	2,16%	2,01%	**2,29%**
MAPE Validación	12,56%	6,79%	10,10%	7,57%	9,24%	**9,25%**

Tabla 42: Desempeño de SARIMA en Pronóstico de Demanda de Energía en el Sistema

El modelo SARIMAX corresponde al mismo modelo SARIMA pero agregando las variables IMACEC y Temperatura media mensual (transformada). Ambas series son diferenciadas para entrar al modelo. Los resultados obtenidos fueron:

SARIMAX	2009-2013	2008-2012	2007-2011	2007-2010	2007-2009	Promedio
MAPE Entrenamiento	3,39%	2,45%	3,87%	2,13%	2,05%	**2,78%**
MAPE Validación	7,09%	6,44%	10,86%	7,21%	7,48%	**7,82%**

Tabla 43: Desempeño de SARIMAX en Pronóstico de Demanda de Energía en el Sistema

4.4.3.2 Demanda de Energía Residencial

El modelo SARIMA estimado para la demanda de energía residencial fue:

$$SARIMA(6,1,6)x(0,1,1)_{12}$$

Los resultados obtenidos fueron:

SARIMA	2009-2013	2008-2012	2007-2011	2007-2010	2007-2009	Promedio
MAPE Entrenamiento	3,58%	5,74%	5,43%	4,58%	4,01%	**4,67%**
MAPE Validación	4,46%	10,37%	9,50%	7,42%	5,26%	**7,40%**

Tabla 44: Desempeño de SARIMA en Pronóstico de Demanda de Energía Residencial

La variación SARIMAX de este modelo incluyó nuevamente a las variables IMACEC y a la misma variable de temperatura media. Los resultados fueron:

SARIMAX	2009-2013	2008-2012	2007-2011	2007-2010	2007-2009	Promedio
MAPE Entrenamiento	2,29%	4,22%	3,62%	4,51%	3,57%	**3,64%**
MAPE Validación	4,56%	9,27%	10,65%	7,40%	6,37%	**7,65%**

Tabla 45: Desempeño de SARIMAX en Pronóstico de Demanda de Energía Residencial

4.4.3.3 Demanda de Energía Comercial

Para la energía comercial se estimó el siguiente modelo SARIMA:

$$SARIMA(1,1,2)x(1,1,1)_{12}$$

Se obtuvieron los siguientes resultados:

SARIMA	2009-2013	2008-2012	2007-2011	2007-2010	2007-2009	Promedio
MAPE Entrenamiento	3,34%	5,01%	4,20%	4,67%	3,81%	**4,21%**
MAPE Validación	10,31%	8,11%	8,79%	6,77%	10,44%	**8,88%**

Tabla 46: Desempeño de SARIMA en Pronóstico de Demanda de Energía Comercial

La variación SARIMAX de este modelo incluyó nuevamente a la variable IMACEC con la temperatura media (trasformada), llegando al siguiente desempeño:

SARIMAX	2009-2013	2008-2012	2007-2011	2007-2010	2007-2009	Promedio
MAPE Entrenamiento	3,57%	5,17%	4,45%	3,89%	4,22%	**4,26%**
MAPE Validación	10,96%	6,18%	7,70%	6,53%	6,32%	**7,54%**

Tabla 47: Desempeño de SARIMAX en Pronóstico de Demanda de Energía Comercial

4.4.3.4 Demanda de Energía Industrial

El modelo SARIMA estimado fue:

$$SARIMA(3,1,2)x(1,1,1)_{12}$$

SARIMA	2009-2013	2008-2012	2007-2011	2007-2010	2007-2009	Promedio
MAPE Entrenamiento	4,99%	5,40%	4,90%	4,68%	4,34%	**4,86%**
MAPE Validación	14,21%	27,12%	13,18%	10,98%	21,08%	**17,31%**

Tabla 48: Desempeño de SARIMAX en Pronóstico de Demanda de Energía Industrial

En este caso, el modelo SARIMAX, que incluye el mismo proceso SARIMA, añade solamente la variable IMACEC.

SARIMAX	2009-2013	2008-2012	2007-2011	2007-2010	2007-2009	Promedio
MAPE Entrenamiento	4,87%	5,53%	5,06%	4,61%	10,71%	**6,16%**
MAPE Validación	10,27%	27,33%	10,56%	9,93%	20,64%	**15,75%**

Tabla 49: Desempeño de SARIMAX en Pronóstico de Demanda de Energía Industrial

4.4.3.5 Demanda de Potencia Eléctrica en el Anillo

El proceso SARIMA estimado corresponde a:

$$SARIMA(1,1,8)x(1,1,1)_{12}$$

SARIMA	2009-2013	2008-2012	2007-2011	2007-2010	2007-2009	Promedio
MAPE Entrenamiento	3,02%	3,39%	4,66%	3,73%	2,83%	**3,53%**
MAPE Validación	5,32%	10,69%	8,31%	6,07%	4,63%	**7,00%**

Tabla 50: Desempeño de SARIMAX en Pronóstico de Demanda de Potencia Máxima en el Anillo

Para el caso particular de la Potencia, se utilizó como variable explicativa la Demanda de Energía del sistema pronosticada, pero por el método SARIMAX detallado anteriormente, de manera de hacer una analogía a lo desarrollado por los métodos anteriores.

En este caso, las variables agregadas en el modelo SARIMAX corresponde a la energía pronosticada y a la temperatura máxima mensual promedio.

SARIMAX	2009-2013	2008-2012	2007-2011	2007-2010	2007-2009	Promedio
MAPE Entrenamiento	3,04%	3,49%	5,25%	3,97%	2,87%	**3,72%**
MAPE Validación	8,35%	11,52%	8,20%	6,17%	4,35%	**7,72%**

Tabla 51: Desempeño de SARIMAX en Pronóstico de Demanda de Potencia Máxima en el Anillo

4.5 Análisis Comparativo de Resultados

A continuación se muestran tablas que comparan los resultados obtenidos por cada método, además de comparar con los resultados del modelo previo de la empresa en los casos correspondientes.

Demanda de Energía en el Sistema

Energía Sistema	2009-2013	2008-2012	2007-2011	2007-2010	2007-2009	Promedio
SVR	5,33%	4,42%	3,96%	7,70%	10,21%	**6,32%**
RNA	2,21%	3,10%	2,51%	4,01%	2,08%	**2,78%**
SARIMA	12,56%	6,79%	10,10%	7,57%	9,24%	**9,25%**
SARIMAX	7,09%	6,44%	10,86%	7,21%	7,48%	**7,82%**
Modelo Chilectra	3,00%	8,71%	7,08%	7,34%	4,95%	**6,22%**

Tabla 52: Comparación de resultados para demanda de energía en el sistema

De los cinco modelos, el mejor resultado fue obtenido por las redes neuronales artificiales, que bajaron el error respecto del modelo de Chilectra en un 3,44%.

Este resultado es considerado como una mejora notable, tomando en cuenta que el nivel del modelo previo ya era considerado como bajo. Además, el modelo de RNA consistentemente obtiene mejores resultados que el modelo previo, lo que demuestra su superioridad.

Respecto a los resultados de SVR, si bien no lograron mejorar los errores en general, en dos horizontes si logran una disminución. Este hecho abre la posibilidad a seguir explorando opciones, dado que existen otras funciones de kernel a probar para tratar de convertir el espacio a uno linealmente separable.

Por otra parte, los modelos SARIMA y SARIMAX no pudieron obtener buenos resultados en general, aunque en el horizonte 2008-2012 pudieron ambos mejorar en desempeño al modelo de Chilectra.

Demanda de Energía Residencial

Energía Residencial	2009-2013	2008-2012	2007-2011	2007-2010	2007-2009	Promedio
SVR	6,39%	5,03%	5,46%	5,06%	7,31%	**5,85%**
RNA	2,96%	3,79%	3,41%	3,65%	3,68%	**3,50%**
SARIMA	4,46%	10,37%	9,50%	7,42%	5,26%	**7,40%**
SARIMAX	4,56%	9,27%	10,65%	7,40%	6,37%	**7,65%**

Tabla 53: Comparación de resultados para demanda de energía residencial

La comparación de resultados muestra que el modelo de RNA vuelve a tener un desempeño mejor respecto de SVR, así como también respecto de SARIMA y SARIMAX. Sistemáticamente se muestra que se obtienen menores errores en todos los horizontes.

Otro resultado que favorece la elección de RNA sobre SVR es el hecho de que el modelo de RNA obtuvo errores homocedásticos en todos los horizontes de evaluación, lo que demuestra no solo el bajo nivel de errores, sino que también la estabilidad de estos.

Demanda de Energía Comercial

Energía Comercial	2009-2013	2008-2012	2007-2011	2007-2010	2007-2009	Promedio
SVR	15,09%	7,33%	7,61%	6,82%	5,37%	**8,44%**
RNA	4,71%	4,22%	3,73%	3,80%	3,64%	**4,02%**
SARIMA	10,31%	8,11%	8,79%	6,77%	10,44%	**8,88%**
SARIMAX	10,96%	6,18%	7,70%	6,53%	6,32%	**7,54%**

Tabla 54: Comparación de resultados para demanda de energía comercial

Al igual que en el caso anterior, se repite el resultado de que el modelo de RNA disminuye el error de manera sistemática respecto de los otros modelos.

Además también se da el hecho de que los residuos de los modelos de RNA muestran homocedasticidad en más horizontes respecto de los errores con SVR.

Para el pronóstico de esta serie se dio que por primera vez, la metodología de SARIMAX sobrepasa en desempeño a la metodología de SVR.

Demanda de Energía Industrial

Energía Industrial	2009-2013	2008-2012	2007-2011	2007-2010	2007-2009	Promedio
SVR	5,20%	6,57%	4,73%	5,42%	5,34%	**5,45%**
RNA	4,25%	3,78%	3,96%	3,62%	3,45%	**3,81%**
SARIMA	14,21%	27,12%	13,18%	10,98%	21,08%	**17,31%**
SARIMAX	10,27%	27,33%	10,56%	9,93%	20,64%	**15,75%**

Tabla 55: Comparación de resultados para demanda de energía industrial

Las redes neuronales vuelven a tener un mejor desempeño, hecho que se acompaña con la mayor estabilidad de los residuos también.

Es destacable que dentro de la dificultad existente para el pronóstico de esta serie, que tiene una tendencia cambiante a través de los registros obtenidos, además de no presentar una estacionalidad clara, se lograron obtener buenos resultados de pronóstico, al menos respecto al nivel de errores, tanto para RNA como para SVR.

No obstante, como se comentó en la sección anterior, se pudo ver que solamente el modelo de RNA captaba de buena manera la cambiante tendencia de la serie, mientras que el resultado para SVR difiere. Resultados de este tipo remarcan que un bajo error promedio no siempre será bueno.

Demanda de Potencia Máxima en el Anillo

Potencia Máxima	2009-2013	2008-2012	2007-2011	2007-2010	2007-2009	Promedio
SVR	7,58%	6,52%	8,63%	9,85%	10,71%	**8,66%**
RNA	3,78%	4,40%	3,98%	3,31%	3,24%	**3,74%**
SARIMA	5,32%	10,69%	8,31%	6,07%	4,63%	**7,00%**
SARIMAX	8,35%	11,52%	8,20%	6,17%	4,35%	**7,72%**
Modelo Chilectra	3,68%	7,89%	5,78%	4,33%	4,30%	**5,20%**

Tabla 56: Comparación de resultados para demanda de potencia máxima

Al comprobar los errores promedio a través de los horizontes de evaluación, se aprecia que el modelo que tiene un menor error es el de redes neuronales, bajando el error promedio en un 1,46%, lo que al igual que con el modelo de energía en el sistema, se considera como una buena mejora dado el bajo nivel del modelo previo.

A diferencia de casos anteriores, acá el modelo de RNA no lograr superar en todos los horizontes el desempeño del modelo de Chilectra, ocurriendo esto en un horizonte específicamente. No obstante, la

diferencia no es grande, donde el MAPE del modelo previo es menor en un 0,1% respecto del modelo de RNA.

A su vez, el modelo de SVR solamente obtuvo mejores resultados en un horizonte respecto del modelo previo. Aunque el resultado destacable en este caso en específico, es que los métodos estadísticos clásicos obtuvieron mejores resultados que SVR.

4.5.1 Modelos Finales

Las tablas de resultados muestran que el algoritmo de redes neuronales artificiales tiene un desempeño constantemente mejor que las demás metodologías comparadas en este estudio.

Para graficar de mejor forma el desempeño de estos modelos de redes neuronales, que fueron los que obtuvieron mejor desempeño, se dispone a continuación una tabla con los errores promedio que tiene cada modelo al pronosticar en los distintos años del horizonte:

Modelo	1er año	2do año	3er año	4to año	5to año
Energía	2,23%	2,37%	2,70%	3,37%	3,24%
Energía Ch.[37]	3,38%	4,38%	5,22%	8,09%	10,12%
Potencia	3,32%	3,59%	3,63%	4,18%	3,98%
Potencia Ch.	3,31%	3,66%	5,08%	5,41%	7,19%
Residencial	3,36%	3,59%	3,74%	3,14%	3,67%
Comercial	4,53%	3,53%	3,91%	4,32%	3,80%
Industrial	3,21%	3,74%	4,23%	3,95%	3,93%

Tabla 57: Errores promedios de los modelos en los distintos años de pronóstico

El modelo de energía presenta mejor desempeño respecto del anterior a lo largo de los 5 años de pronóstico, lo que demuestra que el modelo nuevo tiene consistentemente un mejor desempeño.

Por otro lado, el desempeño en potencia del modelo de RNA solamente tiene un desempeño peor en el primer año, aunque la diferencia es mínima. En los años restantes muestra un mejor desempeño, lo que permite establecer que tiene una mejor capacidad para predecir en el largo plazo.

[37] La "Ch." hace referencia a los resultados del modelo de Chilectra.

A continuación presentan gráficos que comparan los errores de los modelos finales de redes neuronales versus los de Chilectra a través de un periodo de 5 años. Para esto, se promediaron los errores que cometía cada modelo a través de cada periodo posterior al último dato conocido, tomando en cuenta el error promedio porcentual (MPE) y el MAPE.

Demanda de Energía en el Sistema[38]

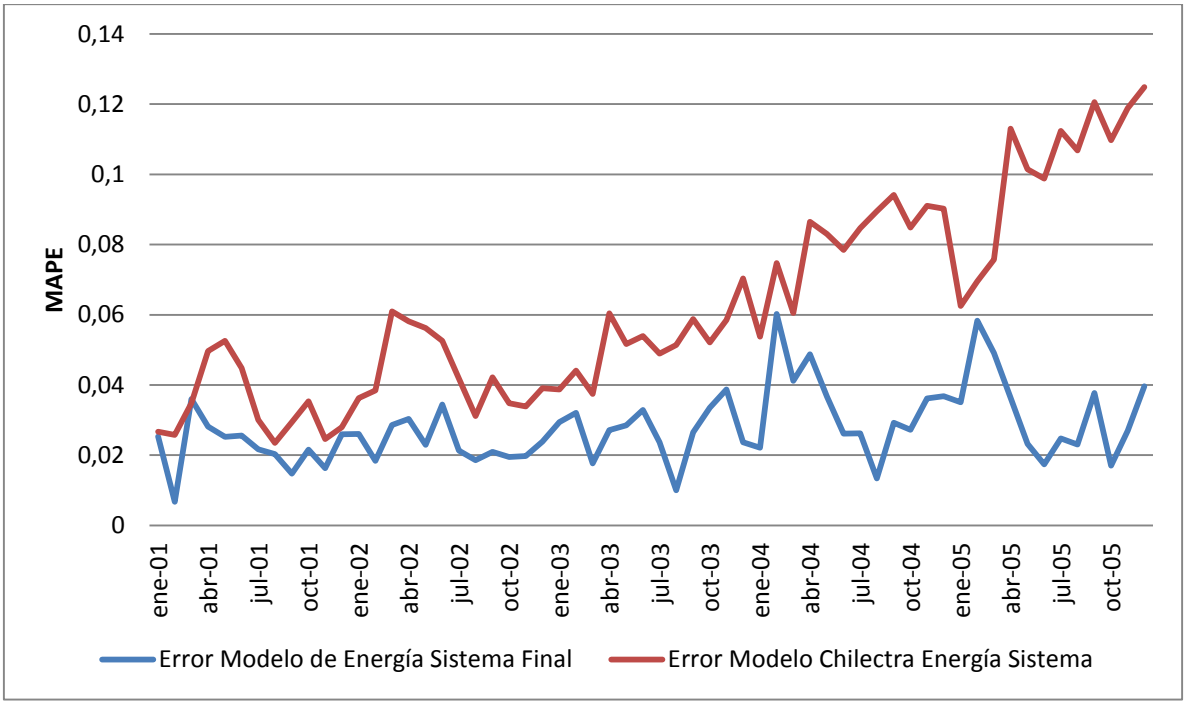


Gráfico 19: Evolución de MAPE para modelos de Energía del Sistema

[38] Las fechas que se muestran en los gráficos son hipotéticas

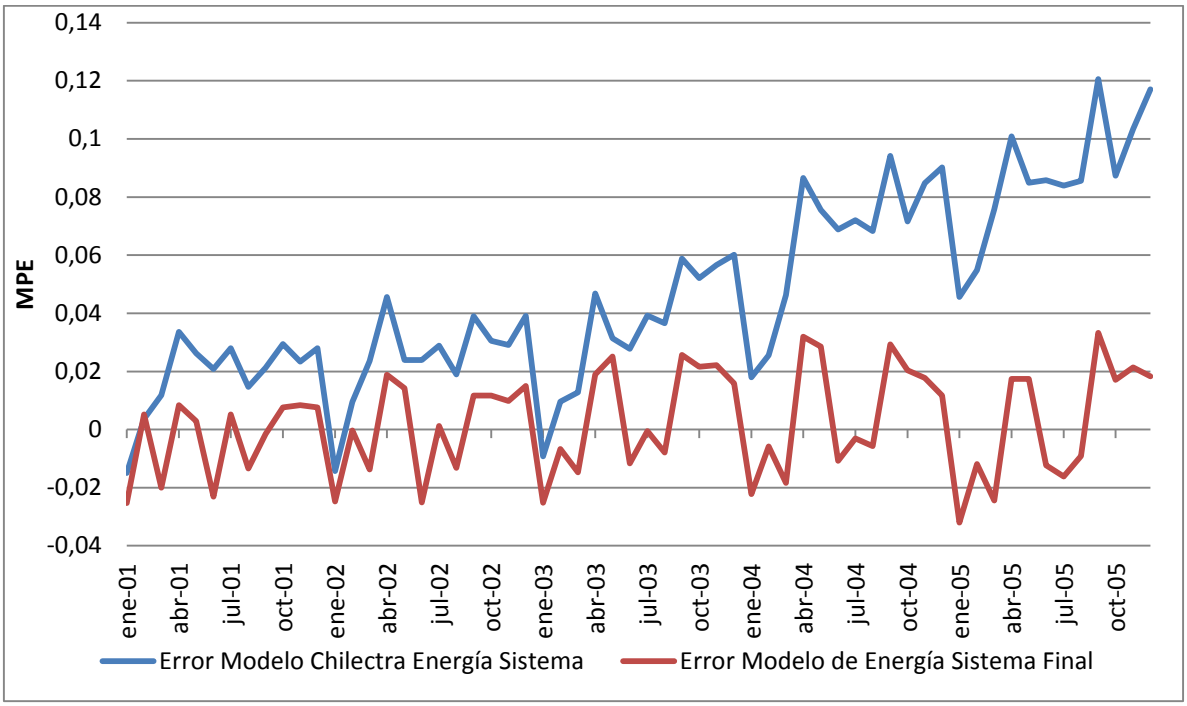


Gráfico 20: Evolución de MPE para modelos de Energía del Sistema

Del primer gráfico se desprende que el desempeño del modelo construido supera en todo ámbito al modelo de Chilectra, dado que tiene menores errores en el corto como en el largo plazo. Mientras que los residuos del modelo de Chilectra presentan una clara tendencia al alza, el modelo de Redes Neuronales presenta un proceso casi estacionario respecto a sus residuos, a pesar de los peaks que presentan desde el cuarto año de pronóstico.

Por otra parte, el segundo gráfico muestra que el modelo Chilectra tiende a sobreestimar la demanda de energía eléctrica en el sistema, mientras que el modelo creado tiene errores que varían entre la sobre estimación y la subestimación, no presentándose una tendencia clara si es más hacia un suceso u otro.

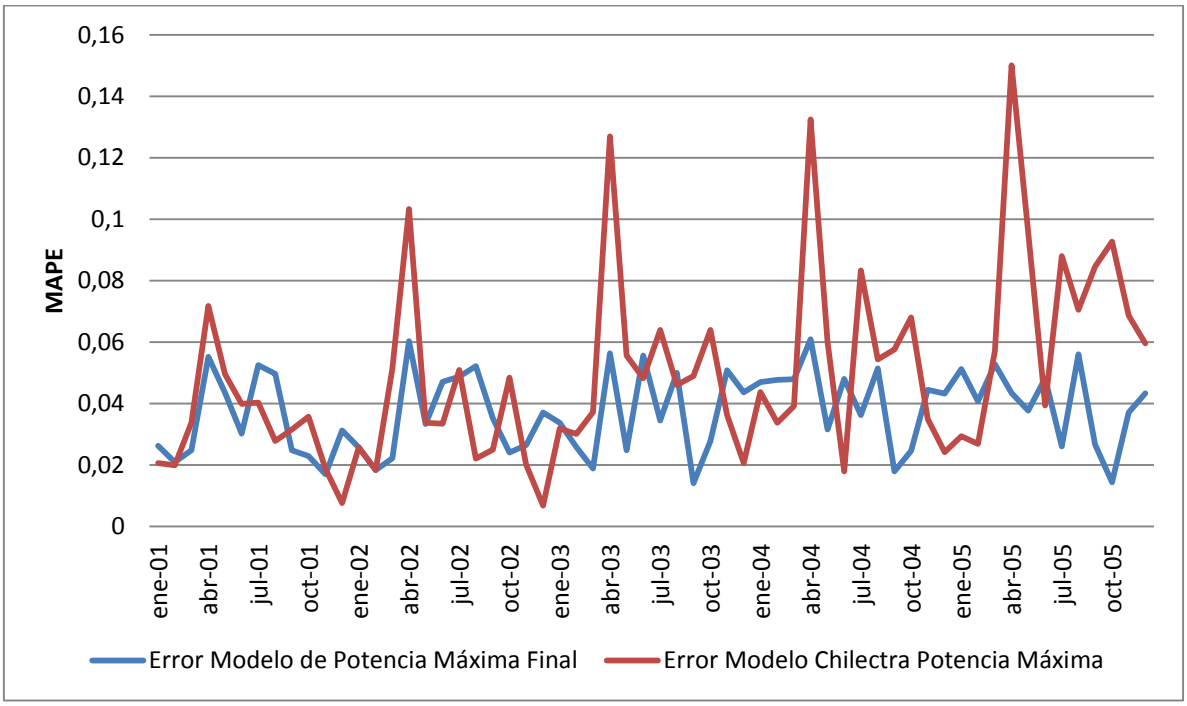


Gráfico 21: Evolución de MAPE para modelos de Potencia Máxima

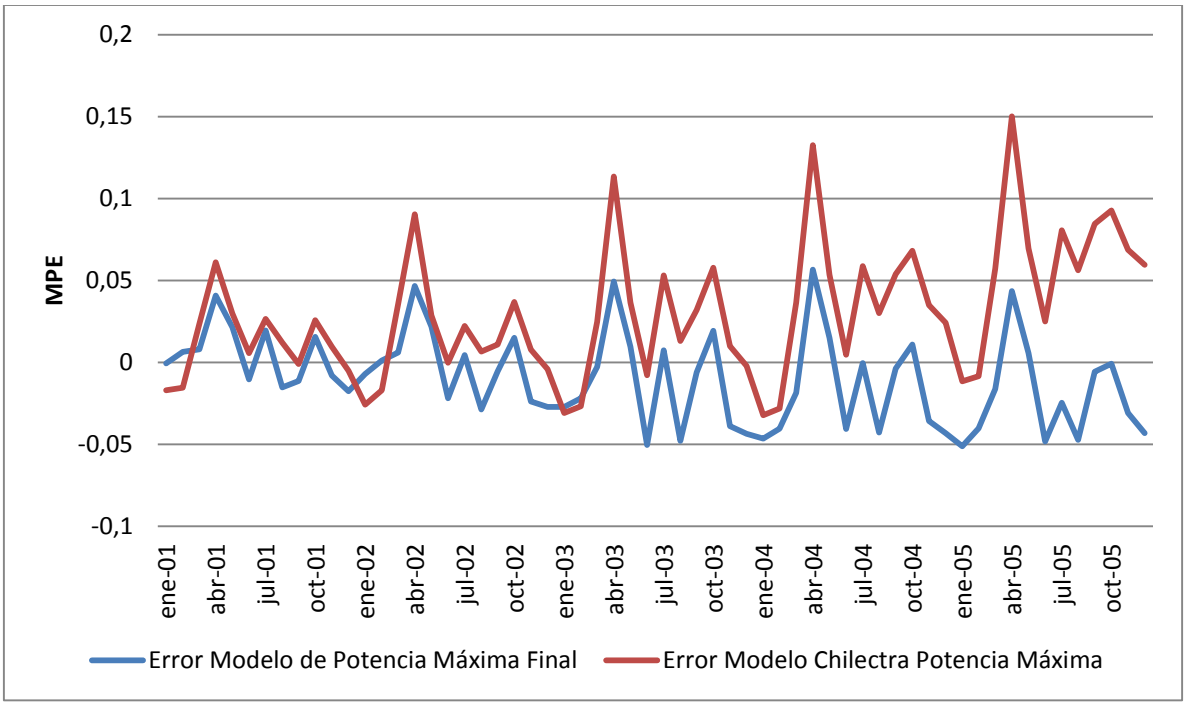


Gráfico 22: Evolución de MPE para modelos de Potencia Máxima

Los pronósticos de potencia muestran que el modelo de Redes Neuronales es menos propenso a la sobreestimación, además que se apegan en mayor medida al comportamiento de la serie a través del horizonte de pronóstico.

A su vez se ve que el modelo de Chilectra tiende a tener una sobre estimación considerable en los meses de Abril. Aun así, los modelos presentan un comportamiento muy similar en el corto plazo (1 año), y a medida que avanza el horizonte de pronóstico se comienzan a diferenciar.

Modelos Sectoriales

En el caso de los modelos restantes de energía, si bien no existe un modelo para comparar, se puede ver que los errores a través de los años no presentan una tendencia clara al alza, lo que si bien no implica directamente la existencia de homocedasticidad, rechaza el hecho de que exista una tendencia de crecimiento en los errores a través del tiempo.

Una métrica que se consideró importante a la hora de analizar los errores de los modelos es la raíz del error cuadrático medio normalizado, ó NRMSE por sus siglas en inglés, el cual es utilizado como una medida de la varianza explicada por un modelo, factor relevante a la hora de evaluar la capacidad explicativa de éste. Su interpretación se encuentra en que mientras menor sea su valor, mayor es la cantidad de varianza explicada.

Para comparar los resultados en términos de explicación de la varianza, la siguiente tabla muestra el NRMSE, promediado entre todos los horizontes, para todos los modelos en cuestión:

Modelo	NRMSE
Energía	11,38%
Energía Ch.	24,09%
Potencia	16,39%
Potencia Ch.	21,44%
Residencial	9,32%
Comercial	12,70%
Industrial	16,41%

Tabla 58: NRMSE de los modelos

La diferencia entre los NRMSE de los modelos de energía es de un 12,7%, mientras que para el caso de los modelos de potencia es de un 5,04%. Estas diferencias en las métricas son reflejadas como la cantidad de varianza extra que explican los nuevos modelos.

Los modelos de demanda sectorial Comercial y Residencial presentan NRMSE cercanos a los obtenidos por el modelo nuevo de energía del sistema, mientras que el correspondiente a la demanda Industrial presenta un valor muy cerca al de potencia, resultado que es considerado como bueno tomando en cuenta que esta serie presenta irregularidades tanto en su tendencia cambiante como también al no tener una estacionalidad clara.

En base a los resultados mostrados previamente, para cada serie a pronosticar, el modelo final implementado es el de Redes Neuronales Artificiales, con las configuraciones especificadas en la sección anterior, utilizando también las variables y sus respectivos rezagos determinados en este capítulo.

5. Análisis de Modelos Finales

En el presente capítulo se analizan los modelos finales establecidos anteriormente, con el objetivo de ver los pronósticos reales que estos realizarán al momento de ser utilizados, así como también ver cuál es la composición interna de estos respecto de la importancia que tiene cada variable para la proyección.

La primera parte de este capítulo consta de un análisis de sensibilidad de los modelos, basado este procedimiento en la realización de una proyección verdadera[39].

En base a estos resultados se espera poder divisar no sólo la relación existente entre las variables exógenas y la dependiente, sino que también visualizar como estas relaciones si estas reflejan un comportamiento de los clientes en la realidad.

Por último, en base a la importancia que tendría la variable más importante se propone una metodología para mejorar su pronóstico.

5.1 Análisis de Sensibilidad

Las redes neuronales artificiales han sido consideradas como un modelo tipo "caja negra", término que hace referencia a la dificultad para interpretar la relación entre las variables que captan los modelos, esto debido a la gran cantidad de interconexiones que existen en la red.

No obstante, es posible determinar el efecto que tiene cada variable realizando un análisis de sensibilidad a una red ya entrenada.

De ésta forma, en el presente capítulo se detallan los resultados obtenidos en base a un análisis de sensibilidad de que se realizó utilizando los modelos finales seleccionados en el capítulo anterior.

[39] Qué no tuviese como objetivo validar un modelo, sino que permita hacer una proyección.

La metodología utilizada consta de la realización de un pronóstico para cada uno de los modelos, considerando a este pronóstico como el caso base.

Este pronóstico se realizó entrando a las configuraciones de redes neuronales antes descritas en un escenario que comprometía datos desde Enero de 2003 hasta Noviembre de 2014 para las series de demanda de energía en el sistema y potencia máxima, mientras que para las series de demanda sectoriales se incluyeron datos desde Enero de 2003 hasta Diciembre de 2013.

Con el fin de generar un buen pronóstico, para cada pronóstico base se realizó una partición de los datos, utilizando tres conjuntos:

- Entrenamiento (65% de los datos)
- Validación (25% de los datos)
- Generalización (10% de los datos)

Como es presumible, el primer conjunto de datos es el que se muestra a la red para entrenarla, mientras que el segundo y tercer conjunto cumplen una función prácticamente igual, que consiste en validar el desempeño del pronóstico que se está realizando. La diferencia entre estos conjuntos está en los datos que comprenden: El conjunto de validación corresponde datos que están justo después de los de entrenamiento, mientras que el de generalización es el posterior al de validación.

El objetivo de esta partición era el de tener distintas métricas temporales del desempeño, donde además se utilizó un error ponderado sobre los tres conjuntos para dar mayor énfasis a los periodos más actuales. El error ponderado fue utilizado como métrica para parar antes el entrenamiento de la red (mientras se entrenaba el modelo de manera simultánea se visualizaban los errores en los otros conjuntos).

Para el cálculo del error, el conjunto de entrenamiento tenía una ponderación de un 20%, el de validación un 40% y el de generalización un 40% también. De esta forma, al último 35% de los datos se les da una ponderación mayor respecto de los de entrenamiento, siendo los datos del conjunto de generalización los que tienen mayor importancia.

Esta configuración de entrenamiento llevo a que se obtuvieran, para todos los modelos de pronóstico, errores homocedásticos e independientes temporalmente.

Además del pronóstico, se calculó un intervalo de confianza al 95% para éste en base a los errores obtenidos en los procesos de validación de cada modelo (i.e. los errores obtenidos en el capítulo 4). Esto se realizó asumiendo la normalidad de los errores, y luego calculando la desviación estándar de éstos.

Al implicar una normalidad de los errores, el intervalo de confianza calculado es simétrico respecto del pronóstico base (igual probabilidad de subestimar que sobreestimar).

Una vez calculado este pronóstico base, se pasó a replicar estos pronósticos pero modificando los valores de las variables exógenas para el horizonte de pronóstico. Por ejemplo, el primer dato modificado para el modelo de demanda de energía del sistema corresponde a diciembre de 2014[40].

Las modificaciones se realizaron aumentando y disminuyendo los valores futuros de las variables exógenas en tres magnitudes distintas para cada caso. Por ejemplo en el caso del IMACEC se aumentó (y disminuyó) en un 1%, 2% y 3% todos los valores pronosticados, mientras que para las variables de temperatura se realizaron variaciones en 1°C, 2° y 3°.

El objetivo final de estas variaciones consistió en calcular una sensibilidad promedio que tiene cada modelo respecto de una variable exógena. El cálculo final de la sensibilidad se realizó según la siguiente ecuación:

$$S_{m,x} = \frac{1}{3} * \sum_{i=1}^{3} \frac{\left(\Delta m_{x+}^{i} - \Delta m_{x-}^{i}\right)}{2i}$$

[40] Para todos los datos anteriores de las variables exógenas se utilizaron los valores reales y no los pronosticados.

Donde:

- $S_{m,x}$: Sensibilidad del modelos "m" respecto de variable exógena "x"
- Δm_{x+}^{i}: Variación porcentual de modelo "m" cuando variable exógena "x" aumenta en un "i" porciento.
- Δm_{x-}^{i}: Variación porcentual de modelo "m" cuando variable exógena "x" disminuye en un "i" porciento.

En la formula anterior se puede ver se utiliza la sustracción en el numerador de la división entre las variaciones, esto debido a que se espera que un cambio en otro sentido para la variable tenga un efecto contrario al de su par cuando aumenta el valor del factor exógeno.

Al ser esta sensibilidad un promedio de los cambios existentes, solamente se espera ver la importancia que tiene una variable para determinar los valores futuros de la serie a pronosticar, y no tiene como objetivo determinar si la relación la forma de la relación entre la serie de interés y las variables exógenas utilizadas en el modelo.

A continuación se muestran los resultados de los pronósticos bases y las sensibilidades promedio calculadas para todos los modelos.

5.1.1 Modelo de Energía Sistema

Pronóstico Base

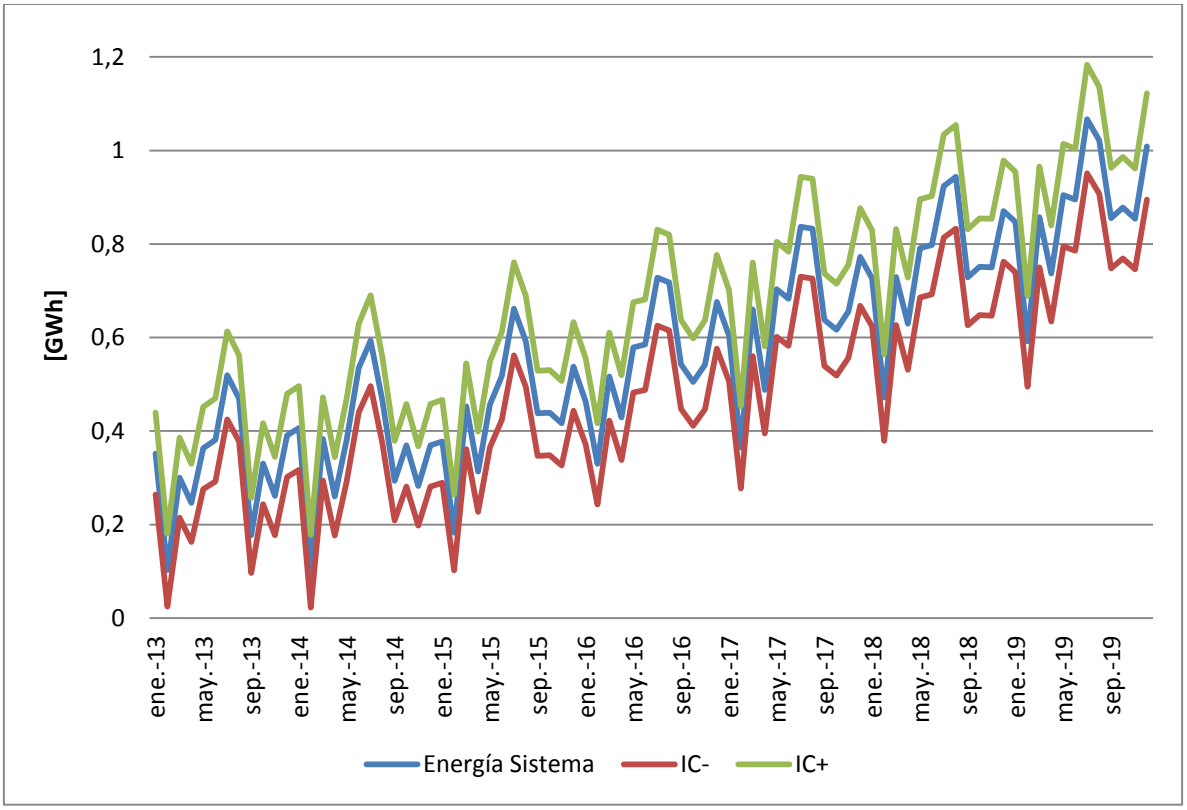


Gráfico 23: Pronóstico de Demanda de Energía en el Sistema

Intervalo de Confianza al 95%: [-3,96% ; 3,96%]

Crecimientos	
2014	2,09%[41]
2015	3,55%
2016	4,44%
2017	4,30%
2018	4,18%
2019	4,47%

Tabla 59: Crecimientos anuales pronosticados para Demanda de Energía en el Sistema

[41] Crecimiento real, no es un valor pronosticado

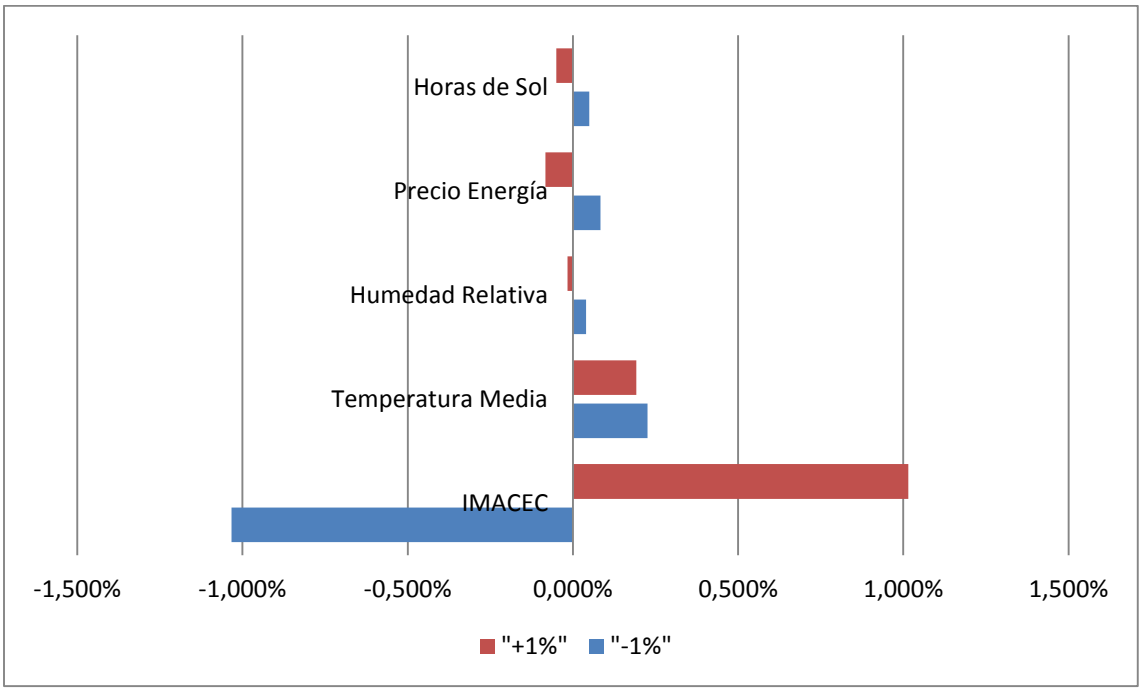


Gráfico 24: Sensibilidades de la demanda de energía en el sistema

Sensibilidades

IMACEC: 1,024%

Humedad Relativa a las 2pm: 0,028%

Temperatura Media: 0,296%

Horas de Sol: -0,050%

Precio de la Energía: -0,050%

De los resultados anteriores se puede ver que el modelo de energía, además de cumplir con las condiciones antes planteadas, logra pronosticar una curva que representa de buena forma la estacionalidad propia del consumo a través del año. Respecto de la tendencia, en específico sobre los crecimientos pronosticados, se puede ver que existe una relación muy directa con las expectativas del PIB utilizadas para hacer el pronóstico, las cuales son las siguientes:

Expectativas de Crecimiento del PIB:

- 2014: 1,09%
- 2015: 3,1%
- 2016: 4%

Considerando que la última expectativa se repite para los siguientes años del horizonte de pronóstico, los crecimientos proyectados son muy cercanos a estos valores, donde la mayor diferencia se produce en el último horizonte, pronosticándose un 4,47%.

Esta condición se avala también con la alta sensibilidad que muestra el modelo respecto del IMACEC, donde una variación de un 1% de esta variable genera un cambio promedio en el pronóstico de un 1,024%.

La segunda variable que tiene más incidencia es la temperatura, lo que refleja el efecto que generan los cambios meteorológicos en el consumo, marcando así la estacionalidad de la predicción. Algo a destacar sobre la relación con esta variable, es que cualquier variación hace aumentar el consumo, lo que lleva a intuir que el movimiento de la temperatura media hacia cualquier dirección activa en casi igual medida un peak de demanda en invierno ó en verano.

Se recalca por último que la sensibilidad del precio es similar a la planteada en [9], donde se estable que la elasticidad de la demanda de energía eléctrica en el país respecto del precio es de -0,063, lo que valida aún más los resultados.

5.1.2 Modelo de Energía Residencial

Pronóstico Base

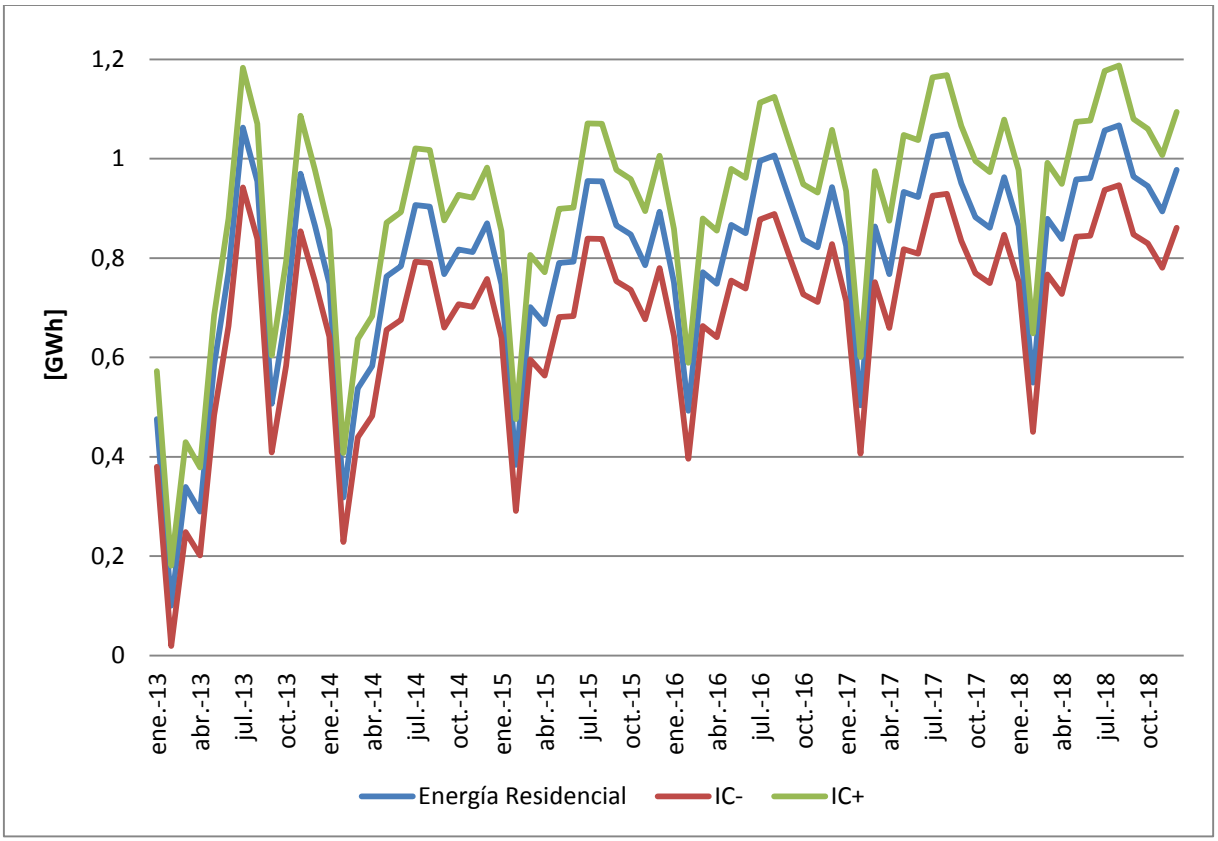


Gráfico 25: Pronóstico de demanda de energía residencial

Intervalo de Confianza al 95%: [-4,09% ; 4,09%]

Crecimientos	
2014	4,00%
2015	1,82%
2016	1,95%
2017	1,72%
2018	1,18%

Tabla 60: Crecimientos anuales pronosticados para demanda de energía residencial

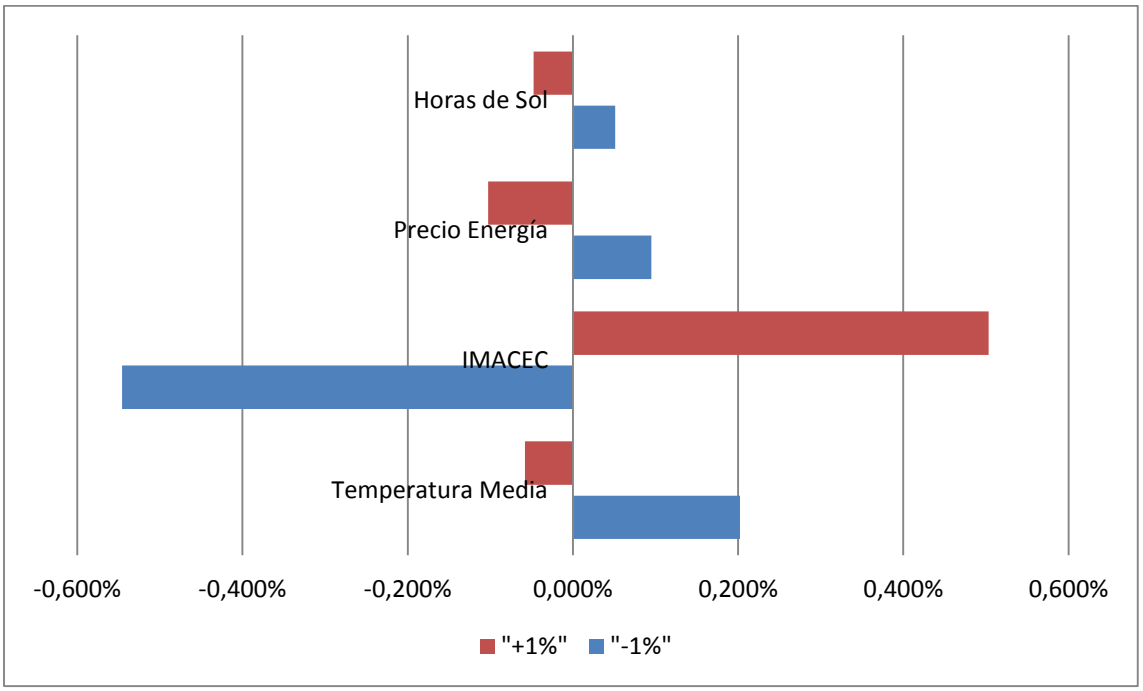


Gráfico 26: Sensibilidades de la demanda de energía residencial

Sensibilidades

T° Media: 0,120%

IMACEC: 0,525%

Precio de la Energía: -0,099%

Horas de Sol: -0,050%

Del gráfico se puede ver que el peak de demanda que ocurre al final del 2013 no permite tener un mes de mayor ó igual demanda en el largo plazo, a pesar de que el modelo cumple con las condiciones de homocedasticidad de los errores.

Al igual que antes, el IMACEC vuelva a ser la variable más relevante aunque con menor preponderancia respecto del modelo de energía del sistema. Esto hace que la temperatura media tenga una preponderacia más comparable a la actividad económica, así como también lo presenta el precio de la energía.

A diferencia del modelo anterior, los crecimientos van disminuyendo posterior al año 2014, factor que refleja la menor preponderancia de la actividad económica en este modelo.

5.1.3 Modelo de Energía Comercial

Pronóstico Base

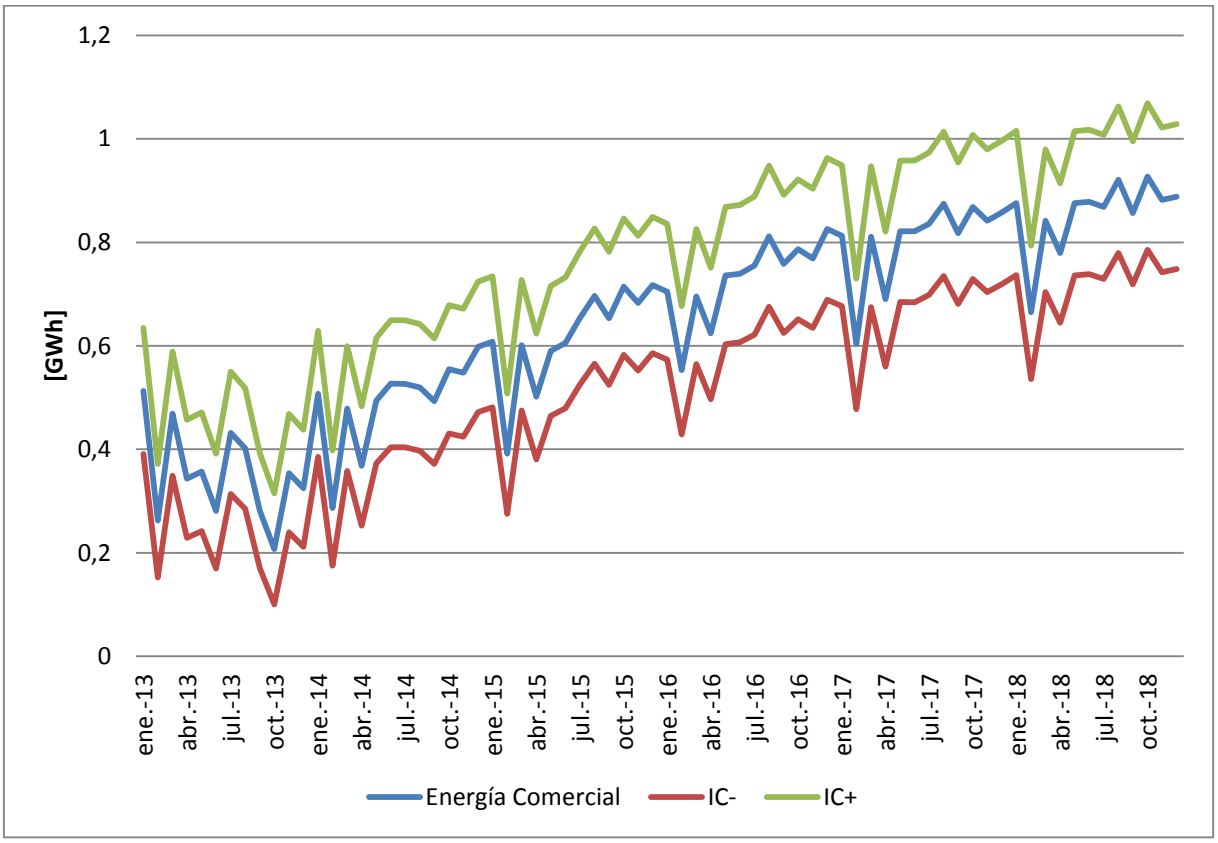


Gráfico 27: Pronóstico de Demanda de Energía Comercial

Intervalo de Confianza al 95%: [-4,78% ; 4,78%]

Crecimientos	
2014	5,84%
2015	4,98%
2016	4,21%
2017	2,70%
2018	1,77%

Tabla 61: Crecimientos anuales pronosticados para Demanda de Energía Comercial

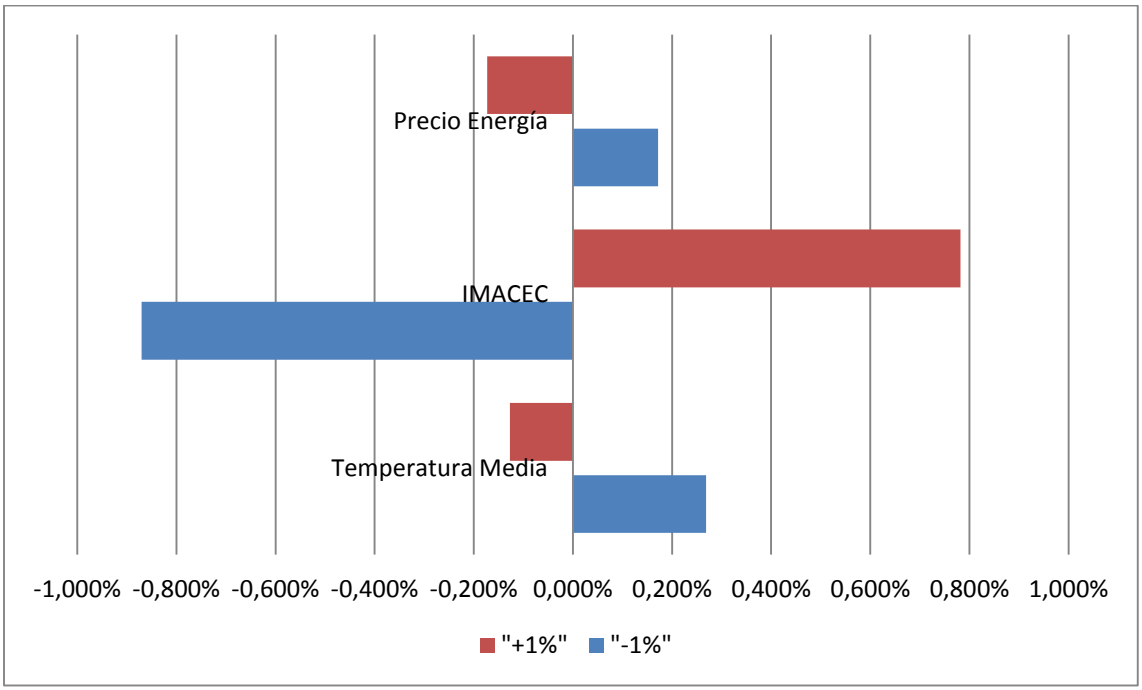


Gráfico 28: Sensibilidades de demanda de energía comercial

Sensibilidades

Temperatura Media: 0,217%

IMACEC: 0,823%

Precio de la Energía: -0,173%

Como pasa con el pronóstico de energía residencial, los crecimientos de la demanda son decrecientes, no obstante se muestra una fuerte relación con el IMACEC, aunque así también con el precio de la energía.

Este modelo es el que presenta una mayor sensibilidad ante el cambio del precio de la energía, aunque sigue estando esta variable en el último escalafón de importancia, dado que la temperatura media tiene mayor preponderancia.

5.1.4 Modelo de Energía Industrial

Pronóstico Base

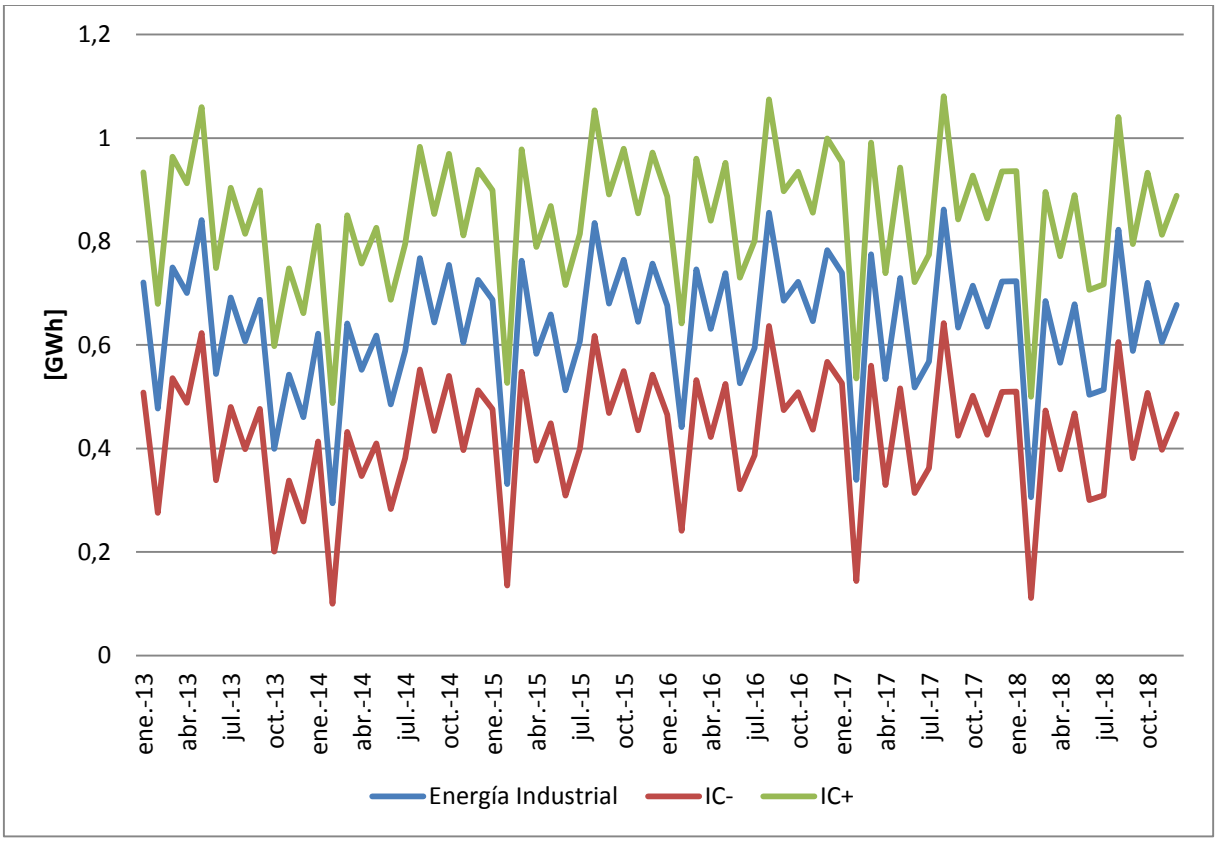


Gráfico 29: Pronóstico de Demanda de Energía Industrial

Intervalo de Confianza al 95%: [-4,48% ; 4,48%]

Crecimientos	
2014	-0,23%
2015	0,95%
2016	0,39%
2017	-0,49%
2018	-0,68%

Tabla 62: Crecimientos anuales pronosticados para Demanda de Energía Industrial

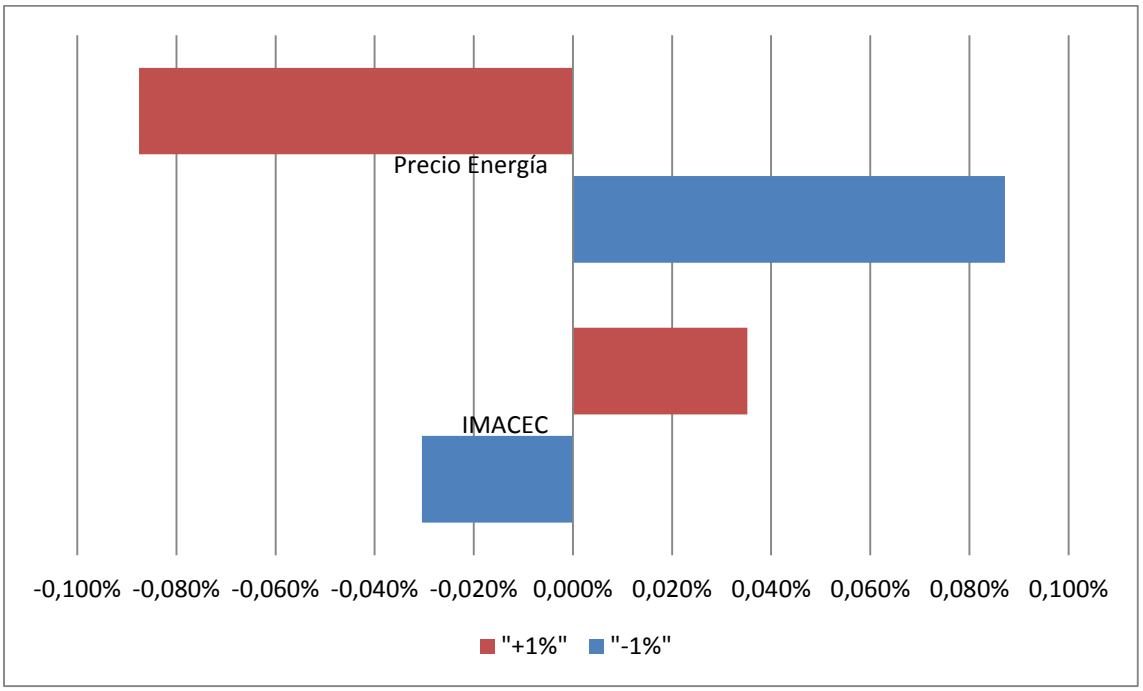


Gráfico 30: Sensibilidades de demanda de energía industrial

Sensibilidades

IMACEC: 0,033%

Precio de la Energía: -0,087%

Esta probablemente fue la serie más difícil de pronosticar dado que no se pudo encontrar una variable exógena que fuera preponderante. Los crecimientos pronosticados son comparables con los que han ocurrido en el pasado, aunque mantienen una varianza mucho menor, esto debido probablemente a que el IMACEC pronosticado tiene una tendencia que varía menos que en la realidad.

Es destacable notar que este es el único modelo de energía donde el IMACEC no presenta la mayor preponderancia, mostrando así la influencia que tiene el precio de la energía en el sector industrial.

5.1.5 Modelo de Potencia Máxima en el Anillo

Pronóstico Base

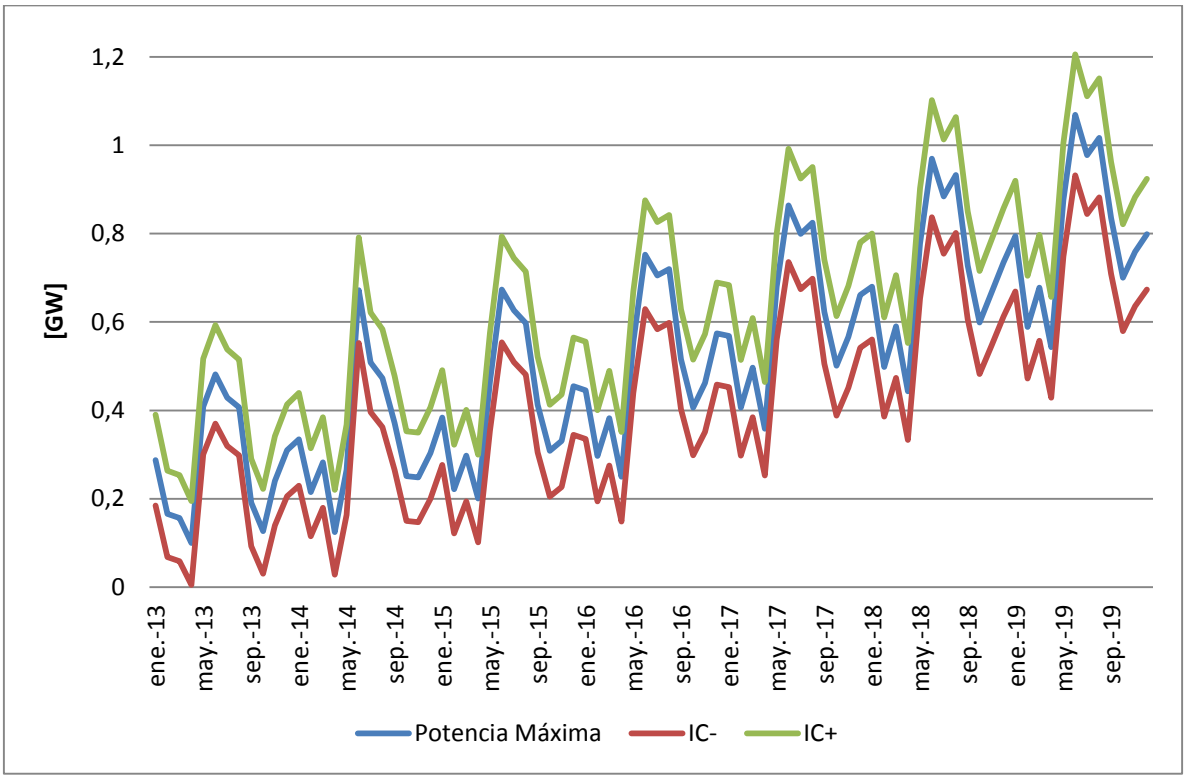


Gráfico 31: Pronóstico de Demanda de Potencia Máxima en el Anillo

Intervalo de Confianza al 95%: [-4,36% ; 4,36%]

Crecimientos	
2014	7,46%[42]
2015	0,06%
2016	2,85%
2017	3,95%
2018	3,61%
2019	3,26%

Tabla 63: Crecimientos pronosticados de máximas anuales para Demanda de Potencia Máxima en el Anillo

[42] Crecimiento real, no es un valor pronosticado.

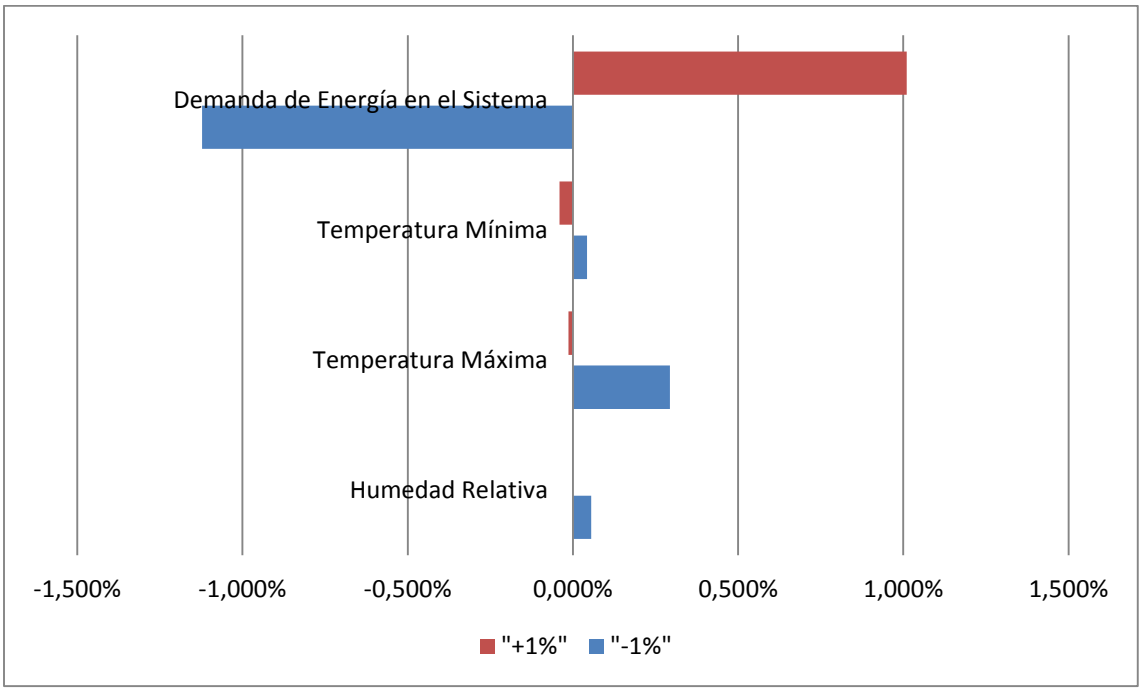


Gráfico 32: Sensibilidades de Demanda de Potencia Máxima en el Anillo

Sensibilidades

Demanda de Energía en el Sistema: 1,10%

Humedad Relativa a las 2pm: -0,03%

Temperatura Máxima: -0,16%

Temperatura Mínima: -0,04%

La curva proyectada muestra que se logró replicar de buena forma una estacionalidad que representa la realidad de la demanda de potencia, además de presentar una curva de crecimiento acorde a los datos anteriores.

Cabe destacar el efecto que tuvo la transformación de los datos para entrenar el modelo, dado que el primer crecimiento de la máxima pronosticado muestra la diferencia que existió entre los datos que verdaderamente ocurrieron versus los transformados, ya que en estos últimos este crecimiento de un 7,46% es en realidad menor, por lo cual

el crecimiento pronosticado, comparando solamente las máximas de los datos transformados, es mayor a un 0,06%.

Esto refleja también que el modelo tiene la capacidad no dejarse influenciar por outlayers, dado que estos podrían haber hecho que el crecimiento pronosticado hubiese sido más alto.

Respecto de las variables más relevantes, la demanda de energía es la que tiene más preponderancia, haciendo eco de su relación física.

Luego destaca la temperatura máxima, y es importante que esta variable tenga más importancia que la temperatura mínima, dado que esta última variable representa un promedio de temperaturas que ocurren en horarios donde no existe una alta demanda[43], mientras que la temperatura máxima ocurre, en general, en un horario cercano al medio día, lo que representa un horario en el cual se demanda más energía.

5.1.6 Variable más relevante

Los modelos mostraron una fuerte relación con el IMACEC, siendo la variable que genera más cambios en los pronósticos de energía, lo que a su vez también se traduce en un cambio en los crecimientos de potencia máxima anual pronosticada.

Este hecho era esperable considerando la fuerte relación que existe entre la demanda energética y la actividad económica de un país.

Si se toma en cuenta que el gasto que realiza Chilectra en compra de energía anualmente (1.250 millones de dólares en 2014), la variación de un 1% en el IMACEC puede generar una diferencia muy relevante en le presupuestación, lo que significaría una diferencia de 128 MM [USD] siguiendo los valores de sensibilidad obtenidos, y tomando como referencia el gasto del año 2014.

El problema que esto conlleva con la realización de pronósticos, es que el IMACEC se proyecta en base a un dato externo de expectativa de crecimiento del PIB, el cual proviene del Banco Central de Chile, y este

[43] Estas temperaturas ocurren en horas de la madrugada

dato es publicado con una frecuencia mensual, presentando cambios a medida que hay más datos disponibles.

Al tomar en cuenta la metodología planteada para el pronóstico del IMACEC, para los primeros tres años de proyección, el cambio en un punto para el crecimiento del PIB genera el mismo cambio en el IMACEC dado que la metodología obliga a que se cumpla esta relación. No obstante, para los años posteriores del horizonte de pronóstico, donde se repite la última expectativa de crecimiento del PIB, el cambio en esta variable puede implicar grandes cambios en el largo plazo.

Se calculó la sensibilidad del IMACEC al cambio en un 1% en la última expectativa de crecimiento del PIB, lo que puede llegar a cambiar la suma total de los índices de IMACEC en un 12,10%.

Si se toma en cuenta que la sensibilidad de la demanda de energía del sistema al IMACEC es de un 1,024%, y que la demanda potencia máxima aumenta en un 1,1% cuando aumenta en un 1% la demanda de energía, entonces, se puede llegar a la siguiente conclusión:

- El aumento en un 1% de la expectativa de crecimiento del PIB para dos años más adelante produciría:
  - Un aumento de un 12,4% en el pronóstico de la cantidad total de Energía a demandarse en el Sistema
  - Un aumento de un de un 13,6% en la demanda de potencia máxima

Bajo esta situación además, hay que tomar en cuenta que es un cambio que puede ocurrir de un mes a otro, y de llevarse a cabo, las magnitudes en las que varían las proyecciones pueden ocasionar otra toma de decisiones, ya sea en inversiones o en la realizaciones de contratos de energía.

Las consecuencias de lo que se acaba de señalar lleva que todo intento por mejorar el pronóstico del IMACEC, y en consecuencia de las expectativas de crecimiento del PIB, tendría un impacto directo en el desempeño de los modelos, sobre todo porque finalmente las metodologías propuestas terminan siendo un algoritmo que transforma expectativas del PIB en pronósticos de energía y potencia.

Es por esto que a continuación se propone una metodología para lograr mejorar la situación que hacen frente los modelos, al estar supeditados a las publicaciones del BB.CC.

5.2 Metodología de Minería de Textos

Como se mencionó anteriormente, el IMACEC pasa a ser una variable determinante al momento de estimar los pronósticos de Energía y Potencia, por lo cual todo esfuerzo por mejorar su pronóstico equivale a mejorar los pronósticos relativos a la empresa.

La preponderancia de esta variable explicativa, sumado al hecho de que presenta endogeneidad con las demandas energéticas, permite explicar que finalmente el modelo implementado es una máquina que traduce expectativas de crecimiento del PIB[44] en la demanda futura de energía eléctrica mes a mes por 5 años.

Ante esto, y considerando que la empresa dispone de una gran base de datos de noticias, se trató de comprobar la siguiente hipótesis:

"El conjunto de noticias publicadas en un período contiene información sobre cómo cambiará el estado de la economía".

Esta hipótesis se basa en el hecho de que el conjunto de noticias describe el estado actual de la economía. Luego, la hipótesis puede ser comparable a las cadenas de markov, donde lo que se plantea es que al estar en un estado se sabe cuáles serán los posibles siguientes estados dentro de la cadena.

En este aspecto, la hipótesis propuesta sería análoga a decir que las noticias describen el estado de la cadena de markov en el que se encuentra actualmente la economía, y por ende sería posible saber cuál será (con una probabilidad) el siguiente estado.

Dicho esto, la metodología que se propone para mejorar el pronóstico del IMACEC se basa en crear un modelo de predicción que

[44] Recordar que las expectativas de crecimiento del PIB son utilizadas para realizar el pronóstico del IMACEC.

utilice el conjunto de noticias como input para conseguir estimar el cambio en las expectativas del PIB que publica el Banco Central.

En un contexto técnico, se pretende realizar una metodología de clasificación mediante minería de textos, que es una rama de la minería de datos dedicada al análisis de textos para tareas como la clasificación de las temáticas de estos, pero también puede responder a los otros objetivos mencionados en el marco teórico.

Bajo esta clasificación, se pretende que en base a las noticias disponibles de un mes se determine si en el próximo mes las expectativas de crecimiento del PIB crecerán o bajarán.

La idea de utilizar textos para clasificarlos es una de las bases de la minería de textos, de donde se desprende claramente el Sentiment Analysis, rama enfocada en descifrar si un texto presenta una opinión positiva o negativa respecto a un tema [44].

No obstante, la clasificación puede ser utilizada como una forma de predecir. Una situación que puede ser tomada como ejemplo, es cuando un banco utiliza una base de clientes para confeccionar un modelo que aprenda a clasificar entre buenos y malos clientes. Luego, cuando un cliente nuevo entra al banco, al aplicar el modelo (y por ende clasificarlo) se estaría realizando un pronóstico sobre si el cliente será bueno ó malo, dado que a priori no se puede considerar que el modelo es perfecto.

Bajo esta misma óptica se aplica la idea planteada, donde se utiliza como factor discriminador el cambio en las expectativas del PIB que publica el Banco Central, dejando como opciones binarias si crecen ó no las expectativas del PIB.

Tener el conocimiento sobre el cambio futuro de las expectativas ayudaría al pronóstico realizado en un instante al dar un primer indicio sobre cómo serán los errores de pronóstico, dado que si se sabe que las expectativas publicadas en el próximo mes bajarán, entonces se podría esperar que en el largo plazo existirá una sobreestimación, y viceversa.

5.2.1 Revisión Bibliográfica

Antes de revisar otras metodologías aplicadas de minería de textos, se consideró pertinente mencionar la particularidad que presenta estos procesos por tener pasos comunes entre todos sus proyectos, los cuales hacen relación al tratamiento previo que se debe realizar a los documentos para poder procesarlos posteriormente. Estas etapas son comparables a la etapa de pre-procesamiento de KDD y hacen referencia al proceso que transforma uno o más documentos de textos a un formato en el cual pueden ser aplicadas las otras etapas de minería de datos, como la selección de atributos y la utilización de algoritmos para la clasificación ó agrupamiento.

Dada la necesidad de utilizar estas técnicas, se pasan a describir a continuación.

Del texto a la minería de datos

Como se mencionó anteriormente, la particularidad de la minería de textos se encuentra en el preprocesamiento que permite convertir uno o varios documentos de textos en registros de datos con atributos a los cuales se les puede aplicar los algoritmos que se deseen. Bajo esta óptica, el objetivo principal de este preprocesamiento se basa en la representación que se le da a un documento de texto, donde el elemento básico son las palabras.

Esta representación de los documentos se basa en un modelo conocido como Vector Space Model, donde un texto es convertido en un conjunto de palabras que aparecen al menos una vez. Este conjunto es el que se representa los atributos que tiene cada documento, para luego ser procesados.

La segmentación permite realizar una representación donde cada documento está representado por un vector que contiene N valores $(x_{i1}, x_{i2}, \dots, x_{iN})$, representando x_{ij} alguna métrica de relevancia de la palabra "j" en el documento "i" [4]. Luego, el conjunto de todos estos vectores (que representan cada uno a un documento) es llamado un Vector Space Model [4].

Cabe destacar también que los N valores que posee cada vector representan el total de palabras utilizadas entre todos los documentos que pasaron un filtro previo, es decir, sin tomar en cuenta preprosiciones, pronombres, etc.

La representación final del conjunto está sujeta a la utilización de una métrica de importancia que tiene cada palabra dentro de un conjunto de documentos. Las métricas más comunes son [4] [24]:

- Frecuencia
- Binario
- TF-IDF (Term frequency – Inverse document frequency)

La primera simplemente cuenta la ocurrencia de un término dentro de cada documento. Por otro lado, la segunda solamente hace diferencia entre si una palabra aparece o no en documento (valor 1 si aparece, 0 en caso contrario). Por último, la metodología TF-IDF [46] es una forma más compleja para determinar la importancia de cada término en un documento, dado que se combina la frecuencia de ocurrencia de un término en un documento con la frecuencia de ocurrencia del mismo pero en el conjunto total de documentos. De esta forma, se puede determinar que una palabra tiene mucha importancia en un documento si es que esta aparece mucho en éste, pero poco dentro del conjunto total de documentos.

Existen varias formas de representar la importancia ponderada de un término dentro de un documento: frecuencia, binaria y tf-idf .Una de las formas más comunes de calcular la importancia ponderada es contar el número de ocurrencia para cada término en un documento dado. Una representación binaria consiste en que el valor 1 indica que el término está presente en el documento, de lo contrario, el valor 0 indica la ausencia del término. Una forma más compleja de calcular la importancia ponderada es llamada tf-idf (term frequency inverse document frequency), la cual combina la frecuencia de un término con una medida de rareza del término en el conjunto completo de documentos [46].

Para llevar a cabo la construcción del Vector Space Model de manera eficiente, se utilizan en general las siguientes etapas:

- Tokenización
- Filtrado de Stopwords
- Stemming

Estas etapas serán descritas a continuación en más detalle dado que son parte de la metodología aplicada en este experimento.

Tokenización

Como ya se mencionó previamente, la primera etapa para tratar los documentos consiste en desmenuzarlos en las palabras que los componen, ó en otras palabras, transformar cada documento en una serie de "*Tokens".* El proceso de tokenización depende del lenguaje del texto, considerando que cada lenguaje tiene distintas reglas gramáticas, la tarea de división resulta fácil para alguien que maneja el idioma, pero no así para un software [46].

El punto clave para distinguir entre las distintas palabras está en las delimitaciones entre tokens, que si bien la base sería un espacio en blanco, esto no siempre es así. Por ejemplo en el lenguaje español se utilizan palabras compuestas por guión, y para el análisis de minería de textos es preferible separar estas palabras para tener un análisis más detallado. No obstante, en español las delimitaciones típicas corresponderían principalmente signos de puntuación, y en general signos que no sean letras [24].

Una vez identificados estas delimitaciones, se reemplaza cada una por un espacio en blanco [3], el cual finalmente termina siendo el único parámetro para desmenuzar los documentos en todas sus palabras, permitiendo así el proceso de tokenización.

Stopwords

La mayoría de los archivos de texto poseen una gran cantidad de palabras, lo que implica una alta dimensionalidad a la hora del análisis. Ante esto, siempre se aplican filtros que eliminan palabras que a priori no tienen ninguna influencia respecto de la clasificación de un texto por ejemplo. Ejemplo de estas palabras son las preposiciones, artículos ó pronombres.

Para disminuir la dimensionalidad del problema, se aplica un filtro de "Stopwords" (ó palabra vacía en español) una vez ya realizada la tokenización, eliminando palabras que no aportaran a caracterizar cada documento [4]. Un punto importante es que estas palabras tienen una alta ocurrencia dentro de los documentos, lo que conlleva una eliminar una cantidad importante de variables irrelevantes.

Es importante destacar que por cada idioma existen distintas stopwords, y no existe un diccionario estandarizado de estas palabras [46]. Para este experimento se confeccionó uno que consistió en una recopilación de distintas listas de stopwords encontradas en internet.

Stemming

Una vez que ya se ha tokenizado un documento y se han filtrado las stopwords, es probable que todavía se tenga una gran cantidad de atributos, no obstante, muchos de estos atributos que están considerados como palabras distintas en realidad están representando un mismo concepto. Por ejemplo, con la presencia de sinónimos, existirían 2 atributos distintos que en realidad hacen referencia al mismo término, así como también puede ocurrir que existan 2 verbos con conjugaciones distintas, o la utilización del plural y singular de una palabra. Para solucionar este problema se aplica el proceso llamado "*Stemming"* [46].

El proceso de stemming tiene como objetivo estandarizar los tokens, transformando cada palabra a su raíz lingüística, considerando que en un texto pueden aparecer variantes morfológicas de cada palabra.

Al realizar esta transformación, se tiene acceso a una mejor descripción del contenido del documento, al finalmente calcular la métrica de importancia en base a las ocurrencias totales de la misma raíz lingüística, lo que al hacerlo con cada palabra de manera individual podría no traer los mejores resultados [4].

Algunos de los efectos de la aplicación de stemming es la reducción del número total de atributos dentro del texto (o reducción del tamaño del diccionario [24]) y el incremento de la frecuencia de ocurrencia de algunos atributos [46].

Al igual que con el uso de stop words, no hay un algoritmo de stemming estandarizado, variando este dependiendo del idioma. Finalmente, con la utilización de este algoritmo, los documentos de textos están listos para ser ingresados a otras etapas de la minería de texto.

Metodologías de Aplicación de Minería de Datos

Habiendo ya revisado las etapas claves para entender el procesamiento de texto, es posible tener un mejor entendimiento sobre las metodologías aplicadas a la utilización de noticias para la predicción de alguna variable de interés, siendo en este caso el cambio en las expectativas del PIB.

Una metodología parecida a la propuesta había sido utilizada por Kroha [29], en donde se trata de predecir la tendencias a largo plazo de los mercados a través de la correlación existente entre el pasado y un conjunto de noticias. De esta forma se implementó un modelo que tratara de aprender este lenguaje correlacionado con la tendencia para luego utilizarlo para predicción.

En este trabajo se utiliza un supuesto en donde se asume que todas las noticias que pertenecen a un intervalo de tiempo, en donde existe una tendencia clara del mercado (por ejemplo, a la baja), presentan una característica en común y por ende son todas de la misma clase. Los resultados que obtuvieron consideran un "acurracy" de clasificación del 70%, no obstante se recalca que en primera instancia las noticias no están en un formato apto para su procesamiento, lo que mermó los resultados finales.

Las metodologías propuestas en varios papers [31][48][30] tratan como una problemática de suma importancia la clasificación de los textos previa al entrenamiento, lo cual se traduce de otra forma en la creación de atributos a utilizar. Este proceso se podría traducir en cómo decir que un texto es positivo o negativo, lo cual en primera instancia queda a criterio del usuario. No obstante se proponen diversas metodologías para esto.

Una de las más usadas corresponde a la confección de una lista de palabras claves creada por un grupo de expertos. Si bien esta metodología mostraría un poco de subjetividad al estar sujeta a la opinión y experiencia de expertos, son estas las que muestran los mejores resultados para la clasificación [36] [48] [40].

Al realizar la creación previa de los atributos, es posible asociar de antemano ciertas palabras a las etiquetas a utilizar, por ejemplo, relacionar palabras de manera positiva al crecimiento del precio de una acción. Esto ayuda bastante al rendimiento de un modelo, no obstante el costo de realizar una diccionario de este tipo es alto y no se puede aplicar en todos los ámbitos, al necesitar de tener una alta expertiz en el tema, y además que los documentos a utilizar deben tener fuerte relación ó pertenecer al rubro bajo el cual se quiere hacer un pronóstico.

La otra alternativa consiste en que los atributos (ó palabras) a utilizar estén determinados de manera automática, esto es, la aplicación del preprocesamiento descrito anteriormente más la posterior utilización de un proceso de selección de atributos.

Al ser de una menor dificultad de replicación, y considerando que puede ser aplicada en cualquier ámbito, esta metodología también es utilizada para la predicción de tendencias de crecimientos del precio de acciones [30][17][15].

Parecido a la problemática sobre los atributos a utilizar, otro factor en el cual difieren estos trabajos tiene relación con la etiqueta que se utiliza para predecir, ya que al igual que antes, puede ser definida de manera manual o de manera automática [30]. Esto se ejemplifica al momento de determinar si una noticia es positiva o no para un propósito. Por ejemplo, en [29] se utiliza un etiquetamiento automático al suponer que todas las noticias que pertenecen a un período donde hubo una tendencia de crecimiento positiva de las acciones, tienen una correlación positiva sobre este hecho.

La otra forma en la que esto se puede realizar es mediante la creación manual de etiqueta (ó "label" en inglés). En [40] se crearon 39 etiquetas distintas, y asignadas de manera manual a cada noticia, para predecir la volatilidad del precio de acciones.

Si bien se ha visto que existe una vasta cantidad de bibliografía enfocada en el uso de minería de textos para la predicción del precio de las acciones, no se ha encontrado bibliografía que utilice estas metodologías para determinar cambios en las expectativas económicas.

5.2.2 Metodología

La idea general consiste en entregar a un modelo de clasificación un input basado en las noticias de cada mes, donde la variable objetivo es una binaria, valiendo 1 en el caso de que la expectativa haya crecido, y 0 en caso contrario.

Un esquema de esta metodología se da a continuación:

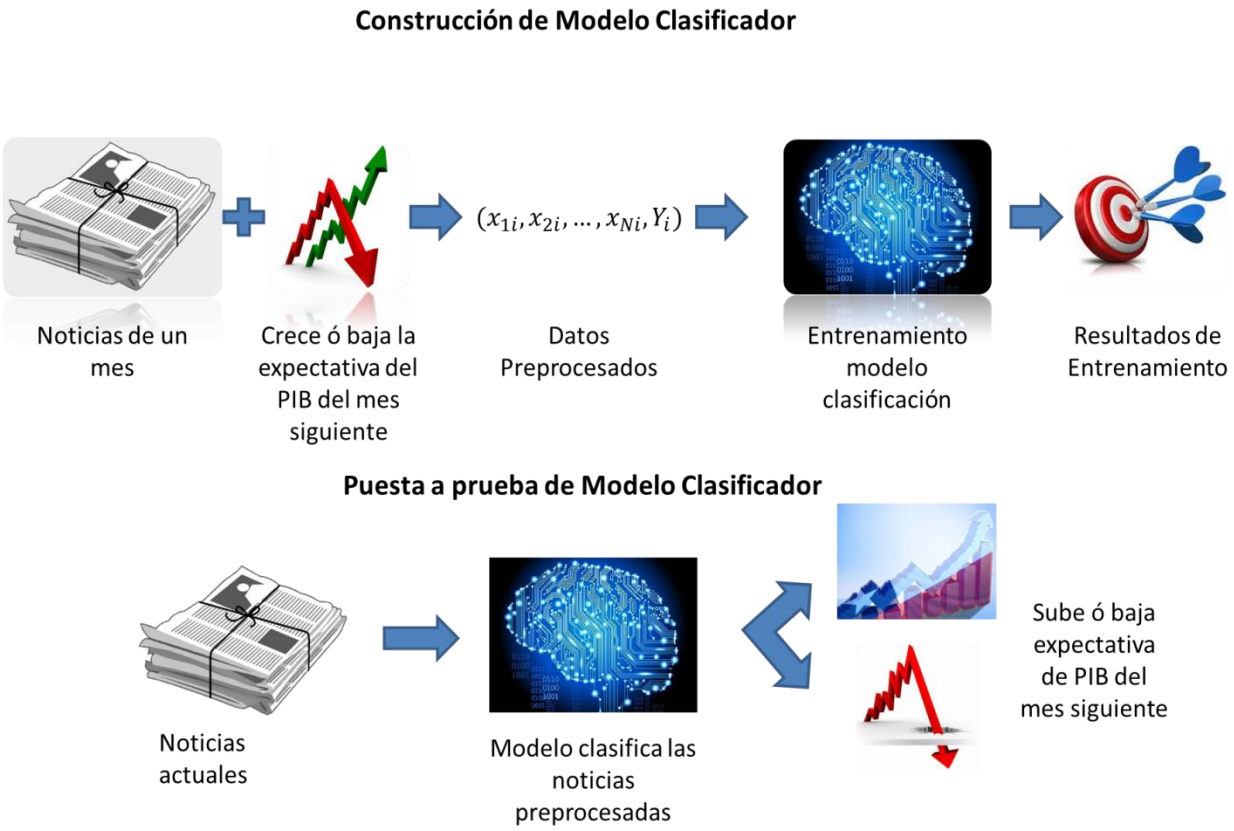


Ilustración 14: Metodología de Minería de Textos para predicción de cambio de expectativa del PIB
Fuente: Elaboración Propia

La construcción de un modelo, que pueda determinar en base a las noticias de un mes si la expectativa del PIB crecerá o bajará respecto de la del mes anterior, es la clave para probar la hipótesis previamente planteada, dado que se estaría probando que las noticias de un mes si contienen información sobre como variarán las expectativas económicas en el futuro.

Si bien la hipótesis planteada hace referencia a que las noticias de un mes contienen información sobre como cambiará la economía, se asume que esta información se refleja en el cambio de las expectativas del PIB del mes siguiente. En otras palabras, las noticias de un mes permitirían saber si las expectativas de crecimiento del PIB del mes siguiente suben o bajan respecto de las que fueron publicadas en el mismo mes que las noticias con las cuales se trabaja.

Como se aprecia en la figura, la primera parte de la metodología consiste en la construcción de un modelo clasificador que permita predecir si las expectativas bajarán o subirán en base a las noticias de un mes. Para lograr esto, el primer paso consiste en el preprocesamiento de las noticias para su posterior utilización en el entrenamiento del modelo.

5.2.2.1 Noticias Utilizadas

Las noticias utilizadas en este experimento fueron obtenidas de una base de datos del resumen de prensa de Chilectra. Las noticias son publicadas en la intranet de la empresa, donde cada día aparecen las noticias más relevantes respecto a temas como:

- Energía
- Telecomunicaciones
- Economía
- Internacional
- Política
- Cultural
- Servicios Públicos
- Nuevas Tecnologías
- Interés General
- Medio Ambiente

En total se utilizaron 81.419 noticias, publicadas entre Octubre de 2002 y julio de 2014. Se utilizaron los titulares de las noticias, dado que esta base del resumen de prensa asocia los titulares a fotografías del diario donde apareció la noticia, dejando disponible como dato a trabajar solo los titulares.

Las noticias provienen de distintos medios de publicación, como revistas, diarios, páginas de internet, así como también recopilaciones de noticias de canales de televisión. Todas las publicaciones utilizadas corresponden a medios chilenos, independiente de que traten noticias internacionales.

Los titulares de las noticias fueron agrupados por mes, dejando así en total 143 agrupaciones de titulares de noticias, con un promedio de 499,2 noticias por mes, existiendo un máximo de 968 noticias en un mes, y un mínimo de 213.

Con las noticias agrupadas, se armó la estructura básica para probar la hipótesis, la cual consiste en un atributo y una variable objetivo. El atributo inicial es el texto que representa la agrupación de los titulares de un mes, y la variable objetivo corresponde a si la expectativa del PIB sube o no en el siguiente mes.

No obstante, a priori se cree los pares "noticias-cambio de expectativa mes siguiente" no contienen información suficiente para que un modelo pueda captar la relación. Ante esto se decidió incluir más atributos que pudieran aportar más información.

5.2.2.2 Variables a Incluir

Para aportar más atributo iniciales, se agregó al par inicial las noticias de meses anteriores, incluyendo de esta forma noticias de hasta 1 año de antigüedad. En específico, en el registro de un mes, que contiene una variable objetivo sobre si sube o baja la expectativa del PIB en "t+1", se utilizaron los grupos de noticias de los periodos (ó meses) "t", "t-1", "t-3", "t-4", "t-5", "t-6" y "t-12".

De esta forma se estaría aportando más variables explicativas para comprobar el cambio de las expectativas, y no relegando la tarea de explicar solamente a las noticias del mes anterior. Para aportar más datos incluso, se utilizaron otras variables no textuales, que corresponderían a las siguientes:

- Cambio de las expectativas de meses anteriores
- Expectativas del PIB publicadas en el mismo mes

La primera de las variables listadas correspondería al dato de la etiqueta del registro de anterior, es decir, si llamamos Y_t a la etiqueta la variable del mes t, y X_t al cambio de expectativa en el mismo mes, entonces se cumpliría que Y_t es igual a X_{t+1}.

Tomando en cuenta la explicación anterior, se incluyeron varios rezagos de la etiqueta para un registro, teniendo así como variables explicativas si en el mismo mes la expectativa del PIB creció o no, y así como también esta misma información pero de meses anteriores. Suponiendo que un registro corresponde al mes "t", se incluyeron los siguientes rezagos del cambio de las expectativas del PIB:

- "t", "t-1", "t-2", "t-3", "t-4", "t-5" y "t-6"

Por último, se consideró de relevancia incluir información sobre los crecimientos del IMACEC al considerar que son de gran importancia para predecir las expectativas de crecimiento. Ante esto, se incluyó dos variables relacionadas con el IMACEC:

- Crecimiento respecto al año anterior
- Crecimiento Desestacionalizado

La primera variable considera el crecimiento del IMACEC publicado para un mes, respecto del publicado para el mismo mes, pero el año anterior. La segunda variable toma en cuenta la serie desestacionalizada del IMACEC, y luego se calcula el crecimiento del IMACEC destacionalizado para un mes versus el del mes anterior.

Dado que el IMACEC de un mes es publicado con 2 meses de retraso, las variables incluidas se asignaron a los meses de publicación y no al mes al que hace referencia el IMACEC.

El incluir variables externas a las obtenidas por el preprocesamiento de los títulos de noticias no nulifica las aspiraciones del experimento por demostrar la hipótesis. Si bien es cierto que se estaría suponiendo que hay variables externas que pueden ayudar a predecir los cambios en las expectativas del PIB, de todas formas hay una influencia por parte de las noticias, lo que demostraría que las noticias si contienen información que pueden aportar a la predicción de esta variable objetivo.

Con la incorporación de todas estas variables, se plantea recién se procede a realizar el experimento para entrenar el modelo clasificador.

5.2.2.3 Determinación de las etiquetas

Para determinar si la expectativa del PIB creció ó bajo, y tomando en cuenta que el Banco Central publica las expectativas de crecimiento para el año actual y dos años más adelante, se realizó el siguiente cálculo:

$$EX_{i,j} = 1 * CR_{i,j} * CR_{i,j+1} * CR_{i,j+2}$$

Donde:

- $EX_{i,j}$: Expectativa del mes "i" en el año "j"
- $CR_{i,j}$: Crecimiento esperado del PIB para el año "j" publicado por el banco central en el mes "i"

Luego,

$$EX_{i+1,j} - EX_{i,j} > 0$$

Entonces se estable que la expectativa del mes "i+1" crece respecto de la del mes "i". Al replicar este cálculo para todos los meses, se determina la etiqueta que cada mes tendrá.

5.2.2.4 Entrenamiento del Modelo Clasificador

El siguiente paso, con los pares ya armados, corresponde al proceso de entrenamiento del modelo clasificador, desarrolando toda esta metodología en el software "*RapidMiner 5",* el cual se enfoca mite el desarrollo de procesos de análisis de datos, teniendo un claro enfoque en la minería de datos, permitiendo aplicar todas las etapas del proceso KDD. A su vez, el computador utilizado fue el mismo con el cual se corrieron los experimentos de pronóstico.

El proceso comienza con el preprocesamiento de los atributos de texto para convertirlos en el Vector Space Model, los cuales pasan por cuatro etapas específicamente:

1. Tokenización
2. Conversión de Mayúsculas a minúsculas
3. Filtrado de Stopwords
4. Stemming

La segunda etapa es la única que no fue mencionada previamente, y consiste en transformar todas las letras en mayúsculas a minúsculas, de manera de que no se diferencia entre las palabras que estén presentes con distintas tamaño.

Respecto de la utilización de una métrica de importancia para cada palabra, se realizaron pruebas utilizando TF-IDF, Frecuencia, y Binario.

Con el conjunto de atributos y sus respectivas etiquetas listo, se utilizó un primer filtro de variables establecido por la eliminación de aquellas que tuviesen una correlación lineal mayor a 0,85.

Posterior a la utilización de este filtro, se pasó al entrenamiento del modelo, aunque con la utilización de una metodología embebida para la selección de atributos. Para esto se utilizó el operador "Optimize selection" de RapidMiner, el cual utiliza una optimización evolutiva de atributos.

Este tipo de algoritmos se basa en la búsqueda de soluciones imitando los postulados de la evolución biológica. En estos se mantiene un conjunto de variables que representan posibles soluciones, las cuales se mezclan, y compiten entre sí, de tal manera que las más aptas son capaces de prevalecer a lo largo del tiempo, evolucionando hacia mejores soluciones cada vez.

El parámetro a optimizar en este caso corresponde al desempeño del modelo clasificador, la cual está dada por la métrica de exactitud, que describe la capacidad del modelo para clasificar de buena forma ambas clases. La exactitud se describe en la siguiente ecuación:

$$Exactitud = \frac{V.Positivos + V.Negativos}{V.Positivos + F.Positivos + V.Negativos + F.Negativos}$$

Donde "V." es la abreviación de Verdaderos, y "F." de Falsos.

Esta métrica es la utilizada para definir el desempeño del modelo, y por ende es la que se utiliza como parámetro de término para el algoritmo de selección evolutiva de atributos.

Además de esta métrica, también existen otras que describen el desempeño del modelo, que son la precisión y la exactitud.

La precisión refleja el desempeño que tiene el modelo respecto de la clasificación de una clase. Esto se traduce en que porcentaje de los registros que el modelo clasificó como una clase, eran efectivamente de esa clase. La fórmula es:

$$Precisión = \frac{V.Positivos}{V.Positivos + F.Positivos}$$

Por otro lado, la exhaustividad el porcentaje de clasificaciones correctamente realizadas respecto del total verdadero existente en una clase. Su fórmula se define como:

$$Exhaustividad = \frac{V.Positivos}{V.Positivos + F.Negativos}$$

El algoritmo utilizado para clasificar fue el de Support Vector Machine, uno de los métodos más utilizados para tareas de clasificación de textos [27][35]. Esta decisión fue realizada en base a la gran utilización que tiene este método, además que tiene la característica de que no requiere de realizar mayores configuraciones al momento de variar la cantidad de variables que ingresan al algoritmo, a diferencia de las redes neuronales artificiales, las cuales tienen que estar diseñadas en base a la entrada de datos.

Para la implementación de este método se utilizó el operador "Support Vector Machine (PSO)", el cual utiliza un enfoque híbrido que combina las SVM con la optimización por enjambre de partículas[45]. La optimización por ejambre de partículas es una heurística que optimiza un problema al, iterativamente, tratar de mejorar una solución candidata respecto a una métrica de calidad.

[45] Particle Swarm Optimization, en inglés, que se representa por la sigla PSO

Esta métrica, al igual que para el método de selección de atributos, es la exactitud de clasificación.

Este método fue elegido dado que este tipo de optimización no requiere que el problema sea diferenciable, como es requerido para los métodos clásicos de optimización [28]. Luego, la optimización por ejambre de partículas puede ser aplicada a problemas que son irregulares, con mucho ruido, y sobre todas las cosas, que cambian a través del tiempo, factor que tiene influencia tomando en cuenta el enfoque de selección de atributos.

Debido a que se desconoce a priori la relación que puede existir entre las variables, se estableció que es posible que esta sea no lineal, por lo cual se decidió una función de Kernel polinomial de grado 3, posibilitando de esta forma que se puedan captar relaciones de distinto tipo.

El experimento fue configurado utilizando una validación cruzada, la cual se basa en entrenar y probar el modelo con distintos datos. Esto se logra al hacer una primera partición de los datos, dejando un porcentaje de estos para entrenamiento y luego otro para validar, y una vez realizada esta primera iteración, se realiza otra partición, de tal forma que el conjunto que previamente se utilizó para validar pasa a formar parte del de entrenamiento, y una parte del conjunto de entrenamiento previo queda excluida para ser utilizada para validar. De esta forma, y al iterar varias veces, se entrena un modelo que tiene resultados más robustos.

Para esta validación cruzada, se utilizaron 5 particiones, dejando así cada entrenamiento utilizando un 80% de los datos, y validando en un 20% restante. Esta validación está incluida en el proceso de "optimize selection" de Rapidminer, lo que implica que por cada iteración de este algoritmo de selección de atributos se ejecutan 5 validaciones, cada una acompañada de un entrenamiento del modelo clasificador, y una vez terminadas estas 5 validaciones, se obtiene un resultado de exactitud promedio, el cual es el que determina si se debe realizar o no otra iteración para mejorar los atributos utilizados.

En anexos se incluyen capturas de pantalla de la implementación de la metodología en Rapidminer.

5.2.2.5 Puesta a Prueba

La parte final de la metodología consiste en poner a prueba el modelo construido con datos (no noticias en este caso) no vistos previamente para medir su verdadera capacidad predictiva. Para esto, de los 143 registros totales, se dejaron 12 de lado para poner a prueba el modelo, los que corresponden a fechas no secuenciales, dado que el dejar los últimos datos para la prueba consistía en una muestra con 11 periodos donde no crecen las expectativas, y 1 registro donde si crece.

Para realizar esta prueba se requiere de procesar las noticias que el modelo no ha visto, realizando el preprocesamiento ya mencionado. No obstante, existe una diferencia fundamental entre ambos procesos, dado que se deben de filtrar todos los tokens que no aparecen en la fase de entrenamiento. Esto se realiza dado que el modelo ha sido entrenado para clasificar en base a una cantidad determinada de variables, y si se le entregan variables desconocidas el modelo no puede clasificar. Luego, el preprocesamiento está limitado a una lista de tokens extraída de la construcción del modelo clasificador.

La consideración anterior justifica el hecho de que solamente se hayan dejado 12 registros para la prueba, dado que se quería lograr una masa crítica de noticias que pudiera captar gran cantidad de variables, para que posteriormente no aparecieran muchas variables desconocidas.

Cabe destacar que esta prueba es la que determinara la verdadera capacidad del modelo para predecir los cambios futuros en las expectativas del IMACEC, donde se utilizará la exactitud como métrica de medida.

5.2.3 Resultados y Análisis

Construcción del Modelo

El preprocesamiento convirtió a las noticias en un total de 13.246 atributos, ó palabras procesadas, que además cuentan con otras variables correspondientes a las relacionadas directamente al PIB e IMACEC. En cada uno de los experimentos estas palabras fueron atribuidas a distintos valores para cada registro según la métrica de representación de importancia seleccionada.

Los procesos de filtrado por correlación redujeron la cantidad de variables a las siguientes cantidades, según las distintas métricas de importancia utilizadas para representar la importancia de las palabras:

- Binaria: 6.993
- Frecuencia: 6.274
- TF-IDF: 6.339

El resultado posterior a este corresponde a las capacidades predictivas (ó clasificación en este caso) de cada experimento. A continuación se muestran los resultados:

Métrica	V. Positivos	V. Negativos	F. Positivos	F. Negativos
Binario	43	38	28	22
Frecuencia	38	36	28	29
TF-IDF	37	40	26	28

Tabla 64: Clasificación hecha por los experimentos en construcción del modelo

Métrica	Precisión Positivos	Precisión Negativos	Exhaustividad Positivos	Exhaustividad Negativos	Exactitud
Binario	60,56%	63,33%	66,15%	57,58%	61,68%
Frecuencia	56,25%	56,72%	55,38%	57,58%	56,50%
TF-IDF	58,73%	58,82%	56,92%	60,61%	58,79%

Tabla 65: Métricas de rendimiento de experimentos en construcción del modelo

El utilizar la métrica binaria obtuvo los mejores resultados para el entrenamiento y validación del modelo, no obstante la magnitud de la exactitud es bastante baja, dado que un posible 40% de error en la clasificación dista de poder predecir de manera robusta el próximo cambio en el IMACEC. No obstante, es posible encontrar en la web distintas opiniones de analistas de minería de datos [53] que remarcan que no necesariamente existe un nivel aceptable de exactitud, y que este umbral debe fijarse en base a cada ámbito, por lo que los resultados no son desechables necesariamente en primera instancia.

Las métricas de precisión y exhaustividad muestran consistencia en los experimentos, mostrando que los modelos no tienden a clasificar bien una clase y otra no.

Puesta a Prueba

Los tres modelos generados fueron puestos a prueba para clasificar datos no vistos previamente. Los resultados fueron los siguientes:

Métrica	V. Positivos	V. Negativos	F. Positivos	F. Negativos
Binario	3	5	1	3
Frecuencia	2	4	4	2
TF-IDF	3	4	2	3

Tabla 66: Clasificación hecha por los experimentos en puesta a prueba

Métrica	Precisión Positivos	Precisión Negativos	Exhaustividad Positivos	Exhaustividad Negativos	Exactitud
Binario	75,00%	62,50%	50,00%	83,33%	61,68%
Frecuencia	50,00%	50,00%	66,67%	33,33%	56,50%
TF-IDF	60,00%	57,14%	50,00%	66,67%	58,79%

Tabla 67: Métricas de rendimiento de experimentos en puesta a prueba

Nuevamente los resultados de clasificación no son óptimos, no obstante se presenta una concordancia respecto de los resultados de entrenamiento/validación obtenidos previamente, a excepción de la exhaustividad obtenida en la clase negativa por el modelo que utiliza frecuencia, dado que es considerablemente más baja que lo obtenido en la validación.

El modelo que utiliza la métrica binaria es el que obtiene mejores resultados.

Una particularidad en el desempeño de estos modelos, es que su función de clasificación no otorga los mismo ratios a los distintos registros, algo que se temía que pudiera ocurrir dado que probablemente estos nuevos datos presentaban palabras claves que no aparecían en las noticias de entrenamiento, por lo cual no existía la posibilidad de que el modelo las detectará.

Keywords más relevantes

A pesar de que no se lograron resultados satisfactorios en la metodología, se decidió incluir un listado con las palabras más influyentes respecto a la baja y aumento en el cambio de las expectativas de crecimiento del PIB. Estas palabras son las utilizadas por el modelo con métrica binaria de importancia.

Las palabras a listar son los términos usados por el modelo final, no obstante en algunos casos se hará aclaraciones a que hacen referencia en caso de ser necesario.

Keywords que aumentan la expectativa de crecimiento del PIB

- **"Acreedor"**
- **"Ernst"** (Stemming de "Ernst & Young")
- **"Ideal"**
- **"Marcas"**
- **"Aplau"** (Stemming de "Aplausos", "Aplaudieron", etc.)
- **"Convocato"** (Stemming de "Convocatoria")

Keywords que disminuyen la expectativa de crecimiento del PIB

- **"opin"** (Stemming de "Opinión", "Opina", etc.)
- **"bachelet"**
- **"nev"** (Stemming de "Inevitable")
- **"mk"** (Stemming de "MK2" y "MK3", siglas que hacen referencia a las reformas en el mercado de capitales)
- **"pseg"** (hace referencia a la empresa PSEG Generación y Energía Chile Ltda., ahora del grupo SAESA)

Del primer conjunto de Keywords se destaca el término "Aplau", considerando que es una acepción positiva hacia una noticia, y tomando en cuenta el conjunto de noticias utilizados, hace casi siempre referencia a una medida que es considerada buena por el sector empresarial.

Por otra parte, la aparición de la empresa Ernst & Young en las noticias demuestra una correlación positiva respecto de las expectativas del PIB.

Los términos negativos hacen referencia a cambios en el mercado o a situaciones que parecen merecer de opiniones de expertos. En este último apartado, al analizar la base de datos de noticias, se observó que los títulos que contienen la palabra opinión hacen referencia a editoriales o columnas publicadas en diarios respecto a una situación en particular que afecta a la economía y/ó mercado eléctrico. A su vez, se podría establecer una relación entre las reformas y la aparición de nombres de Presidentes de la república como una señal de que se avecinan cambios en el mercado, lo que parecería tener un efecto negativo en las expectativas del PIB.

Discusión de resultados

La aplicación de esta metodología no llevó a la construcción de un modelo certero para el cambio de las expectativas del PIB en base a los títulos de noticias disponibles. Sin embargo, con el resultado de este experimento no es posible rechazar la hipótesis planteada en un principio (*"El conjunto de noticias publicadas en un período contiene información sobre cómo cambiará el estado de la economía"*).

En [29] se obtiene un *acurracy* de un 75% para la predicción de una tendencia de crecimiento en el precio de las acciones, y aun así no se considera el resultado como óptimo.

Este primer acercamiento metodológico a la prueba de esta hipótesis puede ser mejorado en trabajos futuros mediante la incorporación de más noticias, incluir los cuerpos de estas, y tratar de ingresar al procedimiento un filtro de keywords realizada por expertos.

De obtenerse resultados mejores resultados de *acurracy* en un experimento futuro, la herramienta generada sería una ayuda no solamente destinada a mejorar la proyección de la energía, sino que permitiría adelantarse al usuario al cambio de expectativas macroeconómicas para el país, factor que podría ser utilizado por la empresa para analizar distintos escenarios de inversión en base a lo predicho por la herramienta, y tomar decisiones en base a la relación beneficio/costo que tendría cada posible situación futura.

6. Conclusiones

La utilización de una metodología de pronóstico que combinara la perspectiva exógena (modelos econométricos) y endógena (autoregresivos) permitió tener resultados satisfactorios respecto a la disminución del error. Éstos no solamente avalan la mejora en los errores, sino que además implicaron disminuir la dependencia que tenían los buenos resultados del modelo anterior respecto de la calidad del pronóstico del IMACEC.

El efecto de estos términos autoregresivos se refleja en que, a diferencia del modelo previo de la empresa, la tendencia de crecimiento no solamente está explicada por el IMACEC, sino que también por el mismo pasado de la serie. Luego, la falta de expectativas de crecimiento disminuye su importancia, aunque sigue siendo relevante.

La transformación aplicada a la serie de potencia permitió obtener mejores resultados en el pronóstico de la serie, sobre respecto de los crecimientos de las máximas anuales de potencia, indicador que para la empresa es muy importante estimar de buena manera, considerando las inversiones que puede implicar.

Un caso que no se mencionó previamente en esta tesis tiene que ver con lo ocurrido el año 2014, donde ocurrió un crecimiento de la máxima anual de potencia de un 8,64%, evento bastante anormal considerando la situación de la economía chilena en ese momento[46]. Al aplicar la transformación a esta serie, el crecimiento de la máxima anual quedaría en 7,05%, que si bien sigue siendo alto, es más cercano al margen de error que considera la proyección de las series.

Ante situaciones como esta, se aconsejó a tener en consideración cuando la máxima anual de potencia ocurre un día Viernes, dado que este es el día que posee una mayor diferencia respecto de los otros. Por otro lado, una explicación más concreta de este efecto, es realizar el supuesto que las mismas condiciones se hayan dado en otro día de la semana, en específico el de históricamente más demanda. Al hacer esto, y en caso de que la máxima se haya dado un viernes, se puede suponer

[46] El crecimiento del PIB en el año 2014 fue de un 1,9% , el menor en 5 años.

que la potencial demanda máxima pudo haber sido un 1,83% más grande, valor suficiente para cambiar en al menos un 1% el crecimiento de la máxima anual respecto al siguiente año.

La razón de esto se debe a que dentro de las pruebas realizadas, no se pudo construir un modelo de configuración única que a través de los distintos horizontes pudiera captar bien los máximos de potencia anuales[47]. Se cree que en este período de invierno la relación entre las variables de entrada y la dependiente es distinta que en el resto del año, dificultando la estimación del pronóstico mediante un solo modelo.

Respecto del desempeño de las metodologías de inteligencia artificial utilizadas, se cree que la utilización de redes neuronales tiene una ventaja sobre la utilización de SVR, dado que las arquitecturas de red que obtuvieron los mejores resultados replican en cierta medida un modelo multivariado de regresión no lineal, tomando en cuenta que existe una sola capa oculta.

Luego la comunicación entre esta capa y la de salida es directamente comparable a una regresión, mientras que las entradas que recibe cada una de las neuronas de la capa oculta representaría esta misma relación de regresión pero con las variables de entrada. Es en esta última etapa donde se cree que la red termina generando una especie de preprocesamiento propio a las variables para estimar la función deseada.

Se destaca que en el trabajo realizado se ha realizado un acercamiento para desmitificar el concepto de modelo "Caja Negra" (en referencia a que se desconoce su funcionamiento interno) que es generado por las redes neuronales al realizar el análisis de sensibilidad.

Esta metodología permitió verificar la importancia que tiene cada variable explicativa sobre la dependiente, y en caso de realizar un análisis de sensibilidad con distintas tasas de cambio, se podría llegar a determinar si la relación es lineal o no.

La evaluación del modelo de minería de textos abre un debate sobre lo que puede ser considerado o no como un buen desempeño en

[47] El modelo que finalmente se dejó en la empresa es uno que prioriza minimizar los errores en los periodos de invierno, aunque de igual forma permite hacer modificaciones para estimar otras fechas.

la clasificación. Lo que abre la posibilidad de que tener un 60% de exactitud sea un buen resultado al tener tantas variables que influyen en el cambio de las expectativas de crecimiento de la economía.

Este factor podría implicar que se tiene un 60% de certeza si, en el largo plazo, las proyecciones realizadas por el modelo tenderán a una mayor sobreestimación ó subestimación. La información resultante de esto puede llevar a tomar acciones a los planificadores de Chilectra respecto del pronóstico para enfrentar de mejor forma las demandas futuras.

6.1 Trabajos Futuros

El no poder obtener resultados satisfactorios mediante la metodología de text mining no refuta la hipótesis planteada. El acceso a mayor capacidad de procesamiento permitiría la aplicación de condiciones claves a la hora de generar variables de interés para la metodología, como lo es la utilización de n-gramas.

Considerando la vital (y lógica) importancia que tiene la actividad económica cómo indicador de la demanda energética, se propone que la clave para mejorar este modelo consiste en poseer una visión más clara del futuro de la economía del país. Ante esto puede ser beneficioso estudiar la evolución de las expectativas, así como también la existencia de ciclos dentro de la economía, con tal de poder obtener un valor esperado de crecimiento a futuro que sea construido en base distintos factores explicativos.

A su vez, otra implicancia de la dependencia del IMACEC presente en el modelo podría llevar a la implementación de un método de pronóstico que no incluyera este tipo de variables (económicas), y tal vez reemplazarlas solamente por la entrada de más variables regresivas (ya sean de otras variables explicativas o de la misma serie a pronosticar).

Otro factor a tomar en cuenta con la metodología planteada corresponde a variar la transformación de la potencia respecto a la cantidad de datos históricos a tomar en cuenta para calcular la relación histórica de los valores de potencia en la semana.

Esto surge debido a que el utilizar todos los datos históricos supone que la relación entre las demandas se mantiene igual a lo largo del tiempo, cosa que no necesariamente es verdad, ya que al igual que como puede ocurrir para el modelo predictivo[48], la relación entre las demandas de los días de la semana, ó dicho de otra forma, el comportamiento de la demanda dentro de los días de la semana puede cambiar al considerar nuevos hábitos de los clientes. En este caso se debe ver el efecto que tendría y, mediante opinión de expertos, si representa la realidad las relaciones calculadas.

Observando los resultados del modelo de potencia, si bien se obtuvieron menores errores respecto de la metodología de la empresa, se recomienda que a futuro se implementen dos modelos de potencia para satisfacer las necesidades de la empresa respecto al pronóstico de esta variable. Uno de estos modelos debiese de enfocarse en sólo pronosticar los meses de invierno (ó el máximo anual) y otro enfocarse en el resto de las estaciones del año.

Si bien se logró una mejor precisión en los pronósticos, la confección de intervalos de confianza en estos permite a la empresa visualizar distintos escenarios de demanda. Éstos pueden ser analizados por la empresa para medir sus costos y beneficios, y en base a esto se deberían de tomar decisiones basadas en una política que tenga como objetivo, por ejemplo, minimizar la volatilidad de los resultados ó defenderse ante un escenario muy adverso.

La realización de este tipo de análisis se puede homologar con la programación estocástica, que según Prekopa [49] se define como "La ciencia que ofrece soluciones para problemas formulados en conexión con sistemas estocásticos, en los que el problema numérico resultante a resolver es un problema de Programación matemática de tamaño no lineal".

Haciendo la analogía, el sistema estocástico estaría formado por la incertidumbre de la demanda futura. Luego, con los intervalos de confianza calculados en este trabajo, sería posible definir una distribución de probabilidad de distintos escenarios.

---

[48] En relación a que la relación entre las variables explicativas y la dependiente puede cambiar a lo largo del tiempo, y se realiza el supuesto de que es estable cuando se hace la proyección.

[49] Prekopa, A. Stochastic Programming. Kluwer Academic Publishers, 1995.

Vistos estos escenarios, la empresa debería de evaluarlos respecto a las consecuencias que traería cada uno, y tomar decisiones que optimicen algún criterio como la protección ante escenarios muy adversos.

7. Bibliografía

1. Achen, C. H. (2001). Why Lagged Dependent Variables Can Suppress the Explanatory Power of Other Independent Variables. Ann Arbor, 1001, 48106-41248.

2. Azevedo, A. I. R. L. (2008). KDD, SEMMA and CRISP-DM: a parallel overview.

3. Bergo, A. (2001). Text categorization and prototypes. Document available on the Internet at http://www. illc. uva. nl/Publications/ResearchReports/MoL-2001-08. text. pdf.

4. Bramer, M. (2007). Principles of data mining.

5. Brown, R. G. (1956). Exponential Smoothing for Predicting Demand. Cambridge, Mass., Arthur D. Little.

6. Ceballos, L., & González, M. (2012). Indicador de Condiciones Económicas. Economía Chilena, 15(1), 105-117.

7. Chapman, S. N. (2006). Planificación y Control de la Producción.

8. Christensen, G., Rouhi, A., & Soliman, S. (1989). A new technique for unconstrained and constrained linear LAV parameter estimation. Electrical and Computer Engineering, Canadian Journal of, 14(1), 24-30.

9. Chumacero, R. A., Paredes M, R., & Sánchez C, J. M. (2000). Regulacion para Crisis de abastecimiento: Lecciones del racionamiento electrico en Chile. Cuadernos de Economía, 323-338.

10. Cox, D. R., & Stuart, A. (1955). Some quick sign tests for trend in location and dispersion. Biometrika, 80-95.

11. Crone, S. F. (2005). Stepwise selection of artificial neural network models for time series prediction. Journal of Intelligent Systems, 14(2-3), 99-122.

12. Del Carpio Huayllas, T. E., & Ramos, D. S. (2010). Electric Power Forecasting Methodologies of Some South American Countries: A Comparative Analysis. Latin America Transactions, IEEE (Revista IEEE America Latina), 8(5), 519-525.

13. Escandón, A., Morales, J. V., & Gajardo, P. (2005). Indicador mensual de actividad económica, IMACEC base 1996: nota metodología.

14. Fletcher, T. (2009). Support vector machines explained. Online]. http://sutikno. blog. undip. ac. id/files/2011/11/SVM-Explained. pdf.[Accessed 06 06 2013].

15. Fung, G. P. C., Yu, J. X., & Lam, W. (2003). Stock prediction: Integrating text mining approach using real-time news. Paper presented at the Computational Intelligence for Financial Engineering, 2003. Proceedings. 2003 IEEE International Conference on.

16. G. E. P. Box y G. M. Jenkins. (1976) Time Series Analysis: Forecasting and Control. Holden-Day, San Francisco.

17. Gidófalvi, G., & Elkan, C. (2001). Using news articles to predict stock price movements. Department of Computer Science and Engineering, University of California, San Diego.

18. González, J. R. H., & Hernando, V. J. M. (1995). Redes neuronales artificiales: fundamentos, modelos y aplicaciones.

19. González, M. I. (2006). Cómo diagnosticar y corregir el problema de la endogeneidad: el número de hijos tenidos en la predicción de las preferencias de fecundidad en Costa Rica. Población y Salud en Mesoamérica, 4(1).

20. Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. The Journal of Machine Learning Research, 3, 1157-1182.

21. Han, J., Kamber, M., & Pei, J. (2011). Data mining: concepts and techniques: concepts and techniques.

22. Hanke, J. E., & Wichern, D. W. (2006). Pronósticos en los negocios.

23. Haque, M. T., & Kashtiban, A. (2000). Application of neural networks in power systems; a review. Trans. Eng. Comput. Technol, 6.

24. Hotho, A., Nürnberger, A., & Paaß, G. (2005). A Brief Survey of Text Mining. Paper presented at the Ldv Forum.

25. Hyndman, R. J., & Athanasopoulos, G. (2014). Forecasting: principles and practice.

26. Isasi Viñuela, P., & Galván León, I. (2004). Redes de Neuronas Artificiales. Un Enfoque Práctico, Editorial Pearson Educación SA Madrid España.

27. Joachims, T. (2002). Learning to classify text using support vector machines: Methods, theory and algorithms.

28. Kennedy, J. (2010). Particle swarm optimization. Encyclopedia of Machine Learning, 760-766.

29. Kroha, P., Baeza-Yates, R., & Krellner, B. (2006). Text mining of business news for forecasting. Paper presented at the Database and Expert Systems Applications, 2006. DEXA'06. 17th International Workshop on.

30. Lavrenko, V., Schmill, M., Lawrie, D., Ogilvie, P., Jensen, D., & Allan, J. (2000). Language models for financial news recommendation. Paper presented at the Proceedings of the ninth international conference on Information and knowledge management.

31. Mittermayer, M.-A. (2004). Forecasting intraday stock price trends with text mining techniques. Paper presented at the System Sciences, 2004. Proceedings of the 37th Annual Hawaii International Conference on.

32. Mittermayer, M.-A., & Knolmayer, G. (2006). Text mining systems for market response to news: A survey.

33. Nahmias, S., & Olsen, T. L. (2015). Production and operations analysis.

34. Niu, D., & Wang, J. (2009). Combination of Text Mining and Corrective Neural Network in Short-term Load Forecasting. Journal of Computers, 4(12), 1188-1194.

35. Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. Paper presented at the Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10.

36. Peramunetilleke, D., & Wong, R. K. (2002). Currency exchange rate forecasting from news headlines. Australian Computer Science Communications, 24(2), 131-139.

37. Prékopa, A. (2013). Stochastic programming (Vol. 324). Springer Science & Business Media.

38. Reed, R. D., & Marks II, R. (1999). Neural Smithing: Supervised Learning in Feedforward Artificial Neural Networks.

39. Ringwood, J. V., Bofelli, D., & Murray, F. T. (2001). Forecasting electricity demand on short, medium and long time scales using

neural networks. Journal of Intelligent and Robotic Systems, 31(1-3), 129-147.

40. Seo, Y.-W., Giampapa, J., & Sycara, K. (2002). Text classification for intelligent portfolio management.

41. Seo, Y.-W., Giampapa, J. A., & Sycara, K. (2004). Financial news analysis for intelligent portfolio management.

42. Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. Statistics and computing, 14(3), 199-222.

43. Soliman, S. A.-h., & Al-Kandari, A. M. (2010). Electrical load forecasting: modeling and model construction.

44. Turney, P. D. (2002). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. Paper presented at the Proceedings of the 40th annual meeting on association for computational linguistics.

45. Urrutia, M., & Sánchez, A. (2008). Generación de Energía Eléctrica en un Modelo para Proyectar el IMACEC. Economía Chilena, 11(2), 99-108.

46. Weiss, S. M., Indurkhya, N., & Zhang, T. (2010). Fundamentals of predictive text mining.

47. Wilkins, A. S. (2014). To Lag or Not to Lag? Re-evaluating the Use of Lagged Dependent Variables in Regression Analysis.

48. Wüthrich, B., Permunetilleke, D., Leung, S., Lam, W., Cho, V., & Zhang, J. (1998). Daily prediction of major stock indices from textual www data. HKIE Transactions, 5(3), 151-156.

49. Chilectra S.A. (2014) Memoria Chilectra 2014

50. http://www.cne.cl/electricidad/f_sector.html

51. Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. Econometrica: Journal of the Econometric Society, 424-438.

52. http://www.hacienda.cl/glosario/imacec.html

53. http://www.dataminingblog.com/what-is-a-good-classification-accuracy-in-data-mining/

8. Anexos

8.A Retropropagación Estándar

Dar una explicación del algoritmo general, espcificando objetivos y cosas por el estilo. Luego poner los pasos para la ejecución del modelo.

Paso Nº1: Inicialización aleatoria de ponderadores.

La red se inicializa con ponderadores aleatorios y pequeños (se quiere evitar la saturación de las funciones de transferencia).

Paso Nº2: Propagación hacia adelante.

La red calcula su salida de acuerdo a los ponderadores existentes.

- Evaluación de la(s) neurona(s) de la(s) capa(s) oculta(s). Para la neurona j de la capa oculta h:

$$net_{pj}^{h} = \sum_{i=1}^{N} w_{ji}^{h}\, x_{pi} + \theta_{pj}^{h}$$

$$y_{pj} = f_{j}^{h}\left(net_{pj}^{h}\right)$$

Donde:

w_{ji}^{h}: Ponderador de la conexión entre la neurona i y la neurona j de la capa oculta h

x_{pi}: Componente i del vector de entrada p

θ_{pj}^{h}: sesgo de la neurona j de la capa oculta h

f_{j}^{h}: Función de transferencia de la neurona j de la capa oculta h

Evaluación de la(s) neurona(s) de la capa de salida. Para la neurona k de la capa de la salida:

$$net_{pj}^{o} = \sum_{j=1}^{L} w_{kj}^{o}\, x_{pi} + \theta_{k}^{o}$$

$$y_{pk} = f_k^o\left(net_{pk}^o\right)$$

Donde:

w_{kj}^o: Ponderador de la conexión entre la neurona j y la neurona k de la capa de salida

y_{pk}: Componente k del vector de entrada p

θ_k^o: sesgo de la neurona k de la capa de salida

f_k^o: Función de transferencia de la neurona k de la capa de salida

Paso 3: Cálculo del error (neuronas de la capa de salida)

$$E = \frac{1}{2}\sum_p \sum_k (d_{pk} - y_{pk})^2$$

Donde:

d_{pk} : Salida deseada de la neurona k dado el patrón de entrenamiento p

y_{pk} : Salida efectiva de la neurona k dado el patrón de entrenamiento p

Paso 4: Retropropagación

Se calcula la derivada del error con respecto a los ponderadores de las conexiones.

$$\frac{\partial E}{\partial w_{kj}} = \sum_p \frac{\partial E_p}{\partial w_{kj}} = \sum_p \frac{\partial E_p}{\partial net_{pk}} \frac{\partial net_{pk}}{\partial w_{kj}} = \sum_p \delta_k \frac{\partial E_{pk}}{\partial w_{kj}}$$

Para los nodos de la capa de salida

$$\frac{\partial E}{\partial w_{kj}} = \sum_p \delta_k \frac{\partial net_{pk}}{\partial w_{kj}} = \sum_p \frac{\partial E_p}{\partial y_{pk}} \frac{\partial y_{pk}}{\partial net_{pk}} \frac{\partial net_{pk}}{\partial w_{kj}} = \sum_p -(d_{pk} - y_{pk}) f_k' y_{pj}$$

Para los nodos de la capa oculta, el término δ se obtiene de manera indirecta:

$$\delta_j = \frac{\partial E_p}{\partial net_{pj}} = \sum_k \frac{\partial E_p}{\partial net_{pk}} \frac{\partial net_{pk}}{\partial net_{pj}} = \sum_k \delta_k \frac{\partial net_{pk}}{\partial net_{pj}} = \sum_k \delta_k \frac{\partial net_{pk}}{\partial y_{pj}} \frac{\partial y_{pj}}{\partial net_{pj}}$$
$$= \sum_k \delta_k w_{kj} f_j'$$

Paso 5: Actualización de ponderadores. Método de descenso del gradiente

Por definición el error crece más rápido en la dirección del gradiente del error. Para minimizar el error, los ponderadores deben ser ajustados en el sentido opuesto al gradiente.

$$\Delta w_{ji} = -\eta \frac{\partial E}{\partial w_{ji}}$$

Donde:

- η: Tasa de aprendizaje

8.B Confección de Variable Laboralidad

Por su definición y el objetivo que persigue la variable Laboralidad, existen variadas formas para su creación. La opción elegida fue la obtención de una que represente el número de días laborales equivalentes que posee un mes, en relación a la demanda de energía que necesita el sistema; para eso la referencia utilizada es el comportamiento del día miércoles, el que por estar en medio de la semana es más regular. De esta forma el valor mensual de la serie siempre será un número menor al total de días del mes (no necesariamente entero).

Como la demanda diaria de energía aumenta su nivel en el tiempo, la comparación directa entre días de meses o años distintos pierde sentido, por lo que se hace necesaria la introducción de un indicador comparable. Éste será calculado como el ratio entre la demanda diaria y la demanda del día miércoles perteneciente a esa misma semana:

$$RD = \frac{Demanda\ Diaria}{Demanda\ Miércoles}$$

De esta forma es posible calcular y asignar a cada día un indicador que elimina el efecto de aumento en el nivel de demanda y mantiene la diferencia en el nivel entre distintos días. Eso sí, se debe aplicar un tratamiento especial a aquellos días miércoles que no sean del todo representativos, como por ejemplo feriados o bajo el efecto de otro feriado; en estos casos lo que se hace es utilizar un promedio de los miércoles que no se encuentren afectados de las semanas adyacentes.

Además, es posible estimar el indicador RD de días futuros, basándose en el promedio histórico de los días de la semana en cada mes. La excepción son sólo los casos atípicos como feriados y afectados por feriados, pero en estos casos puede usarse igualmente el promedio de los casos atípicos históricos.

Finalmente al obtenerse valores históricos y futuros del indicador RD , se puede obtener la variable Laboralidad buscada como la adición mensual de estos valores.

Creación

El trabajo comienza con los datos históricos diarios de la demanda de energía del sistema:

1. Se identifican los días feriados.

2. Se calcula para cada día el promedio de la demanda entre los dos mismos días de las semanas anteriores, los dos mismos días de las semanas posteriores y del día correspondiente. Si alguno de esos días era feriado se excluye del promedio. De esta forma se obtiene un nivel promedio representativo del día, que será preliminar pues aún pueden estar afectados por eventos los días utilizados.

3. Se calcula la diferencia porcentual entre el promedio anterior y la demanda registrada para cada día.

4. Si la diferencia porcentual es negativa (el promedio es superior) y de magnitud mayor a 4%, entonces el día pasa a ser considerado como atípico (preliminar).

5. Se vuelve a calcular el promedio de los días de las semanas adyacentes, pero esta vez se excluyen los atípicos. De esta forma el nuevo nivel promedio estándar es más representativo.

6. Se calcula la diferencia porcentual entre el promedio anterior y la demanda registrada para cada día.

7. Si la diferencia porcentual es negativa (el promedio es superior) y de magnitud mayor a 4%, entonces el día pasa a ser considerado como atípico.

8. Si la magnitud de la diferencia porcentual es mayor a 4% (el promedio puede ser mayor o menor), entonces el día pasa a ser considerado como fuera del estándar.

9. El indicador RD histórico se calcula con las demandas reales de los días, utilizando el promedio estándar en caso de que el miércoles correspondiente esté fuera del estándar (de forma de no ensuciar la serie con demandas muy bajas o altas en los días miércoles base).

Resta estimar los futuros valores del indicador RD:

1. Se calcula el promedio por día y mes del indicador RD, dejando de lado todos aquellos días fuera del estándar o con el miércoles correspondiente fuera del estándar. Sólo se consideran los tres últimos años para este promedio.

2. Se identifican los feriados futuros y se consideran atípicos.

3. En base al comportamiento histórico de los feriados, se determinan cuáles serán los días afectados por cada feriado en el futuro y se consideran atípicos.

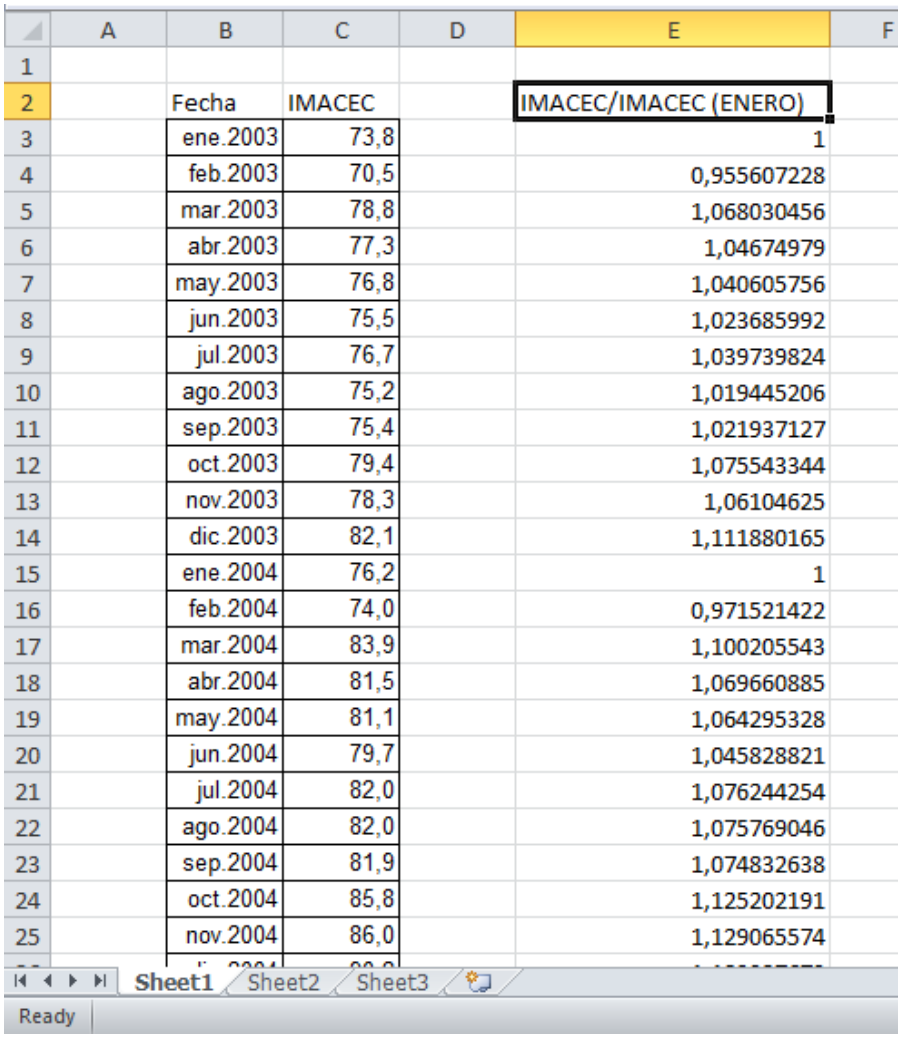
4. Se asigna el valor promedio calculado a cada día en el futuro dependiendo del tipo de día y el mes, excepto por los atípicos.

5. Para estos últimos, lo que se realiza es un análisis del comportamiento histórico de los feriados y los días que afectan, considerando los promedios que se alcanzan y los días afectados, y asignando un valor representativo.

Finalmente de la adición mensual de los valores RD se obtiene la serie Laboralidad completa.

8.C Aplicación de Metodología de Pronóstico de IMACEC

La primera etapa es calcular una curva de estacionalidad típica del IMACEC, realizando esto mediante el cálculo del cociente entre el valor de la serie en el mes de Enero y el de los otros meses del mismo año. Esta operación se ve reflejada en la siguiente imagen.



	Fecha	IMACEC	IMACEC/IMACEC (ENERO)
3	ene.2003	73,8	1
4	feb.2003	70,5	0,955607228
5	mar.2003	78,8	1,068030456
6	abr.2003	77,3	1,04674979
7	may.2003	76,8	1,040605756
8	jun.2003	75,5	1,023685992
9	jul.2003	76,7	1,039739824
10	ago.2003	75,2	1,019445206
11	sep.2003	75,4	1,021937127
12	oct.2003	79,4	1,075543344
13	nov.2003	78,3	1,06104625
14	dic.2003	82,1	1,111880165
15	ene.2004	76,2	1
16	feb.2004	74,0	0,971521422
17	mar.2004	83,9	1,100205543
18	abr.2004	81,5	1,069660885
19	may.2004	81,1	1,064295328
20	jun.2004	79,7	1,045828821
21	jul.2004	82,0	1,076244254
22	ago.2004	82,0	1,075769046
23	sep.2004	81,9	1,074832638
24	oct.2004	85,8	1,125202191
25	nov.2004	86,0	1,129065574

Sheet1 Sheet2 Sheet3

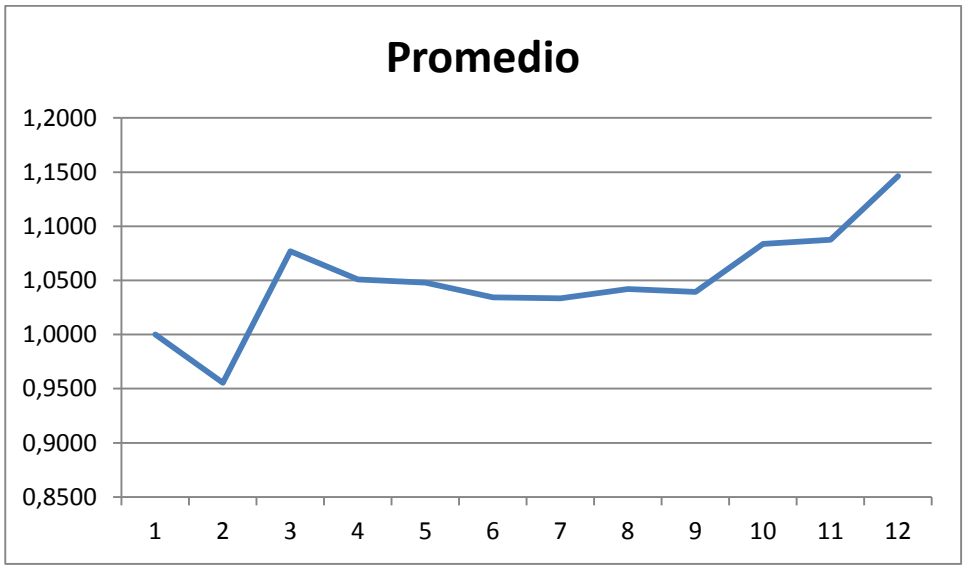
Ready

Como se aprecia en la figura, la columna de "IMACEC/ [IMACEC(ENERO)]" comienza con el primer valor igual a 1, y luego este se repite al comenzar el siguiente año. Una vez calculada por completo

esta columna para todos los registros a utilizar, se procedió a calcular el valor promedio de "IMACEC/ [IMACEC(ENERO)]" para un mismo mes a través de los años.

2003	2004	2005	2006	2007	Promedio
1,00000	1,0000	1,0000	1,0000	1,0000	1,0000
0,95561	0,9715	0,9579	0,9681	0,9527	0,9554
1,06803	1,1002	1,0773	1,0776	1,0857	1,0770
1,04675	1,0697	1,0621	1,0427	1,0438	1,0509
1,04061	1,0643	1,0497	1,0639	1,0539	1,0481
1,02369	1,0458	1,0407	1,0395	1,0389	1,0344
1,03974	1,0762	1,0452	1,0433	1,0153	1,0334
1,01945	1,0758	1,0697	1,0493	1,0370	1,0421
1,02194	1,0748	1,0582	1,0511	1,0376	1,0392
1,07554	1,1252	1,0870	1,1082	1,0887	1,0837
1,06105	1,1291	1,1084	1,1066	1,0933	1,0876
1,11188	1,1839	1,1696	1,1631	1,1574	1,1463

Los valores que aparecen en la columna "Promedio" corresponderían a una curva de estacionalidad promedio calculada para un caso. Esta curva tiene la siguiente forma:



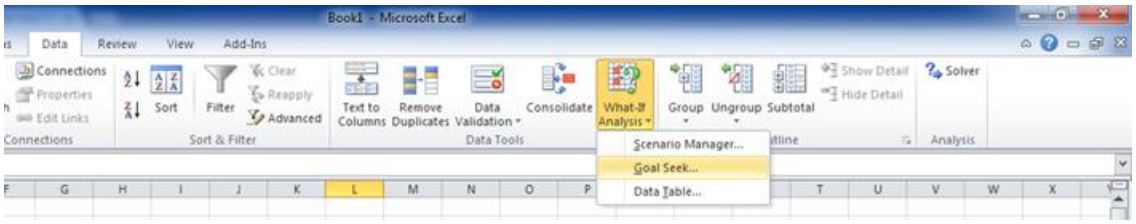
El siguiente paso para la proyección consiste en utilizar la función de Excel *"Goal Seek"* con los crecimientos del PIB y también en conjunto con "curva de forma" calculada, representada por el promedio de las relaciones "IMACEC/ [IMACEC(ENERO)]" (a partir de ahora llamada *curva de forma*).

La utilización de esta función de Excel se basa en responder la siguiente interrogante:

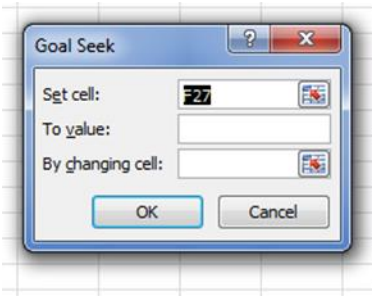
"¿Cuál debería de ser el valor del IMACEC de los siguiente meses con tal de calzar con las expectativas de crecimiento del PIB?".

Esta pregunta podría tener muchas respuestas, dado que en un año donde faltan varias cifras que publicar del IMACEC las soluciones son infinitas, no obstante, con el cálculo de la curva de forma se ha limitado el número de respuestas dado que se fijó una relación entre los el valor del IMACEC del mismo año.

Hecho esto se procedió a utilizar la función *Goal Seek*, la cual busca el valor del IMACEC que cumple con las condiciones de relación. Esta función se encuentra en la sección de Datos en Excel, y luego en "What If Analysis" (esto fue hecho con una versión en Inglés de Excel), mostrando esto en la siguiente figura:



Esto abrirá un pequeño cuadro como el siguiente:



Para la utilización de esto primero se fijaron unas celdas que expresan el crecimiento del IMACEC total de un año respecto del año anterior (el cociente entre las sumas del IMACEC).

Luego para la celda que indica el crecimiento del IMACEC respecto del año en el cuál faltan datos, se selecciona para la opción "*Set Cell*" y

en la opción "*To Value"* se ingresa el valor de la expectativa de crecimiento correspondiente, para finalmente seleccionar la celda que contiene el primer dato faltante en la opción "*By changing cell"*.

La función "*Goal Seek"* de Excel realiza una búsqueda lineal [50] [51] para poder encontrar los argumentos de la función determinada anteriormente. En un inicio, el algoritmo cambia de manera positiva y negativa el valor de la celda a cambiar ("*By changing cell"*). Al recalcular el valor de la función objetivo, el algoritmo vuelve a iterar en la dirección que se acercó más al valor buscado, y a medida que avanza en las iteraciones varía en distintas cantidades el valor a cambiar hasta lograr llegar al valor deseado de la función objetivo.

Al hacer esto, la herramienta buscará el valor del siguiente IMACEC que falta (es decir, el primero que requiere ser pronosticado) para que el crecimiento sea el correspondiente a la expectativa. Notar que con la curva de forma estimada, al fijar la relación entre los valores del IMACEC del año, cuando se realiza un cambio respecto del primer IMACEC, también lo está haciendo para los restantes del año.

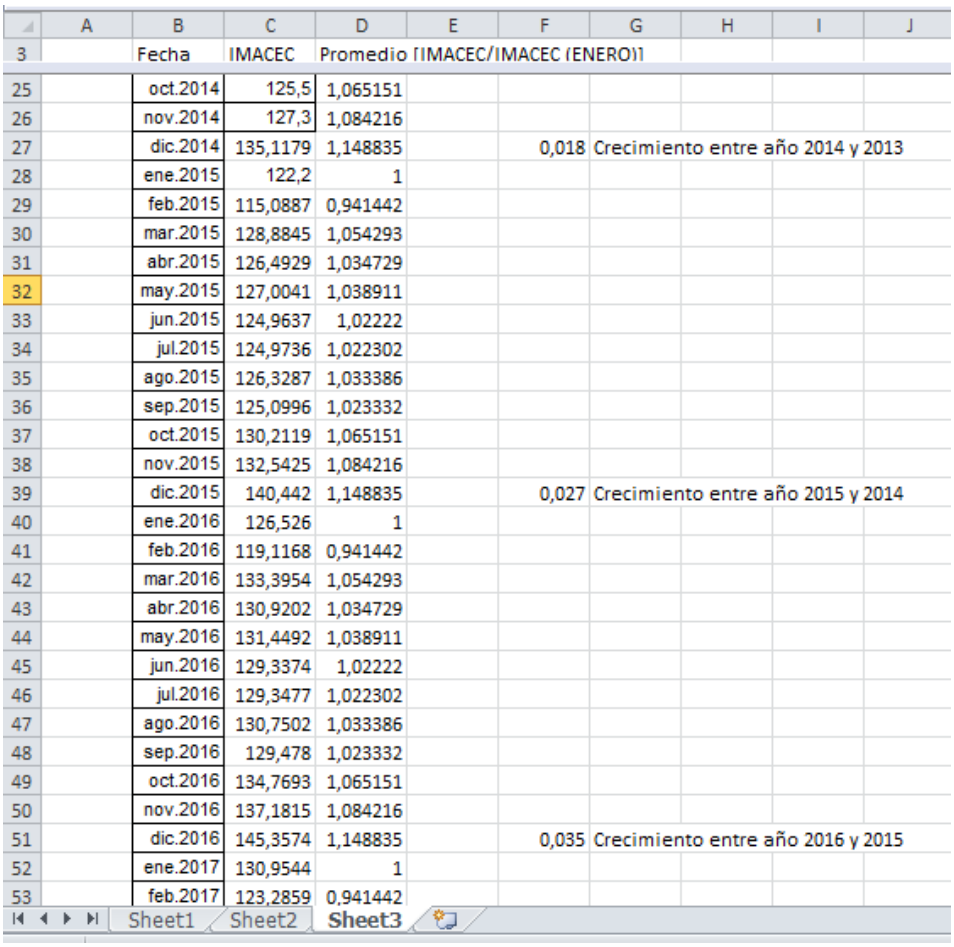
Este procedimiento termina por entregar el crecimiento del IMACEC de tal forma que calce con las expectativas del PIB, además de la curva de forma calculada. Luego, este procedimiento se repite para los siguientes años del horizonte de pronóstico, donde se utilizan las expectativas respectivas de crecimiento. No obstante, dado que el Banco Central sólo pública expectativas de crecimientos para el año actual y los dos años siguientes, y para este ejercicio se requiere hacer un pronóstico de 5 años, la expectativa de crecimiento correspondiente al último año se repite hasta completar el horizonte necesitado.

A continuación se muestra una imagen de Excel con el resultado final de este procedimiento.

---

[50] https://support.microsoft.com/en-us/kb/100782

[51] Knuth, Donald (1997). "Section 6.1: Sequential Searching,". Sorting and Searching. The Art of Computer Programming 3 (3rd ed.). Addison-Wesley. pp. 396–408



	A	B	C	D	E	F	G	H	I	J
3		Fecha	IMACEC	Promedio [IMACEC/IMACEC (ENERO)]						
25		oct.2014	125,5	1,065151						
26		nov.2014	127,3	1,084216						
27		dic.2014	135,1179	1,148835		0,018	Crecimiento entre año 2014 y 2013			
28		ene.2015	122,2	1						
29		feb.2015	115,0887	0,941442						
30		mar.2015	128,8845	1,054293						
31		abr.2015	126,4929	1,034729						
32		may.2015	127,0041	1,038911						
33		jun.2015	124,9637	1,02222						
34		jul.2015	124,9736	1,022302						
35		ago.2015	126,3287	1,033386						
36		sep.2015	125,0996	1,023332						
37		oct.2015	130,2119	1,065151						
38		nov.2015	132,5425	1,084216						
39		dic.2015	140,442	1,148835		0,027	Crecimiento entre año 2015 y 2014			
40		ene.2016	126,526	1						
41		feb.2016	119,1168	0,941442						
42		mar.2016	133,3954	1,054293						
43		abr.2016	130,9202	1,034729						
44		may.2016	131,4492	1,038911						
45		jun.2016	129,3374	1,02222						
46		jul.2016	129,3477	1,022302						
47		ago.2016	130,7502	1,033386						
48		sep.2016	129,478	1,023332						
49		oct.2016	134,7693	1,065151						
50		nov.2016	137,1815	1,084216						
51		dic.2016	145,3574	1,148835		0,035	Crecimiento entre año 2016 y 2015			
52		ene.2017	130,9544	1						
53		feb.2017	123,2859	0,941442						

Sheet1 Sheet2 Sheet3

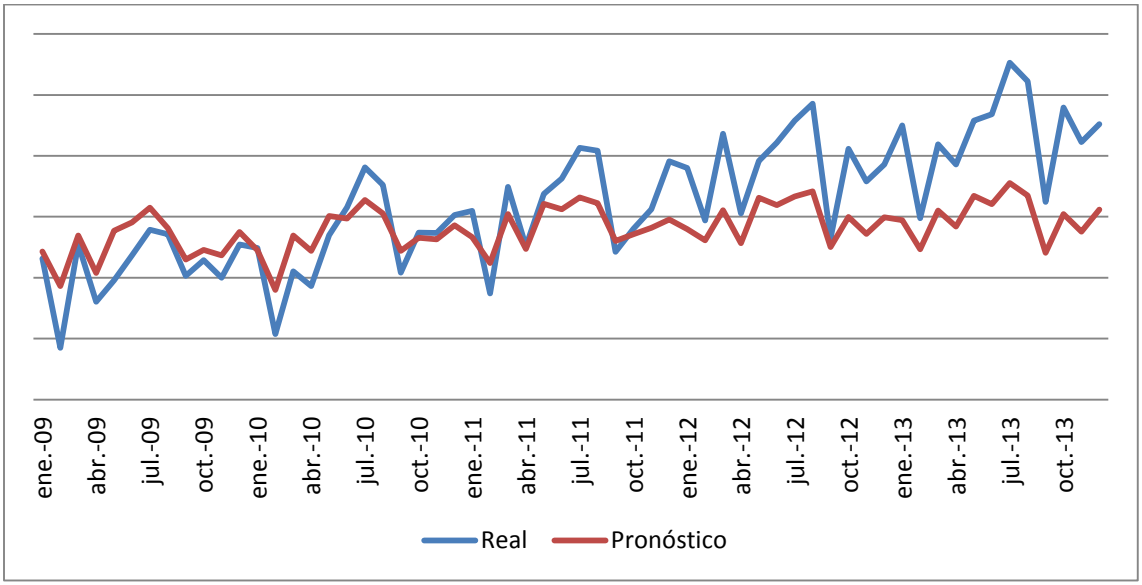
Se aprecia que en la primera columna a la izquierda está la fecha y posteriormente el valor del IMACEC. En la siguiente están los valores de la "curva forma" calculada, y más a la derecha están las expectativas de crecimiento del PIB.

8.D Detalle de Pronóstico con Métodos de Inteligencia Artificial

8.D.a Aplicación de SVR

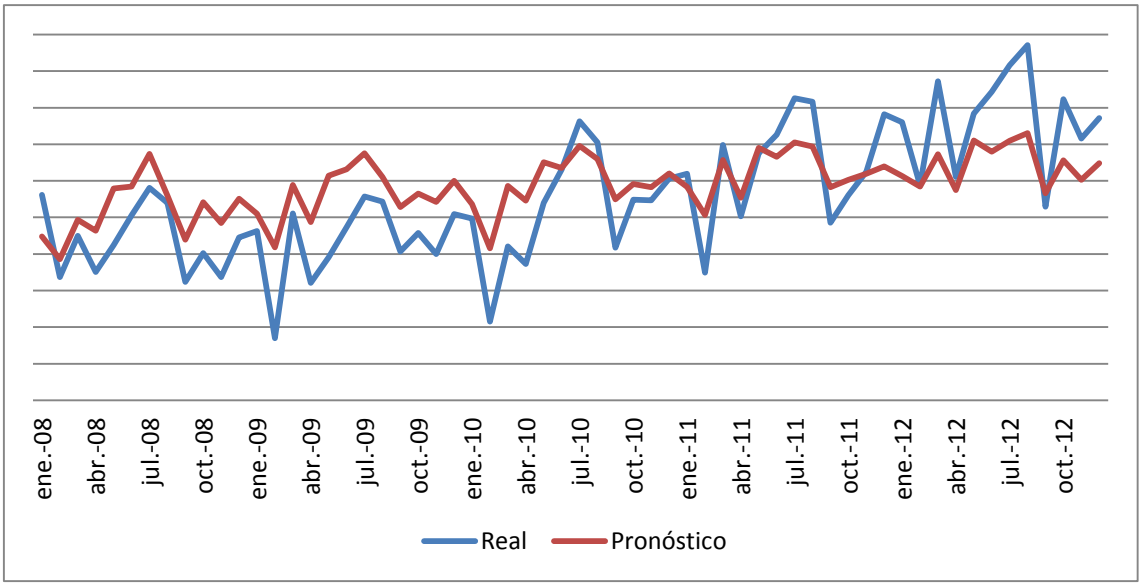
Demanda de Energía en el Sistema

2009-2013



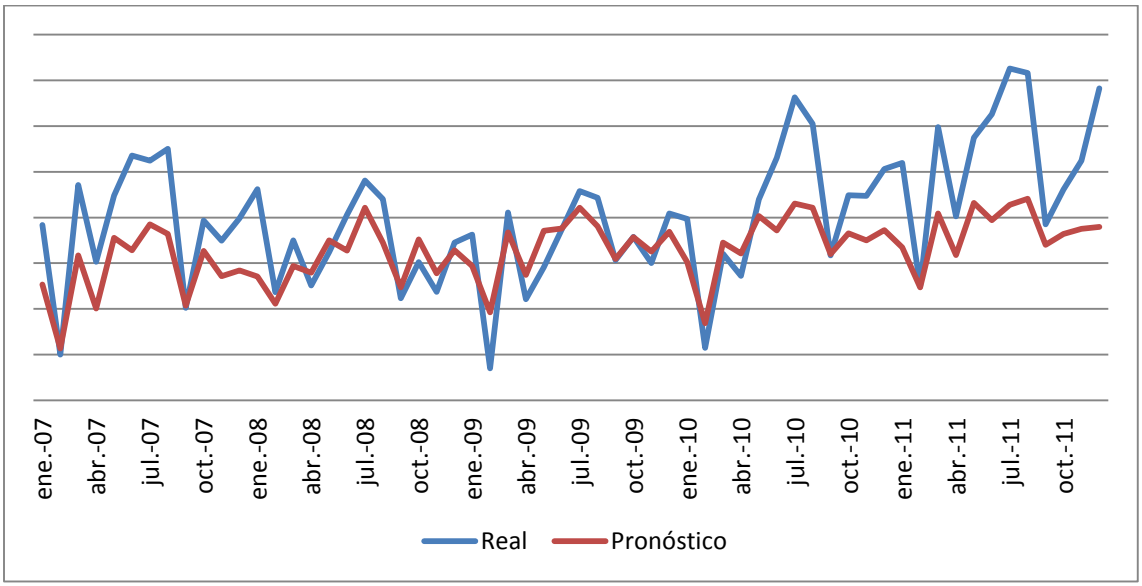
MAPE: 5,33%

2008-2012



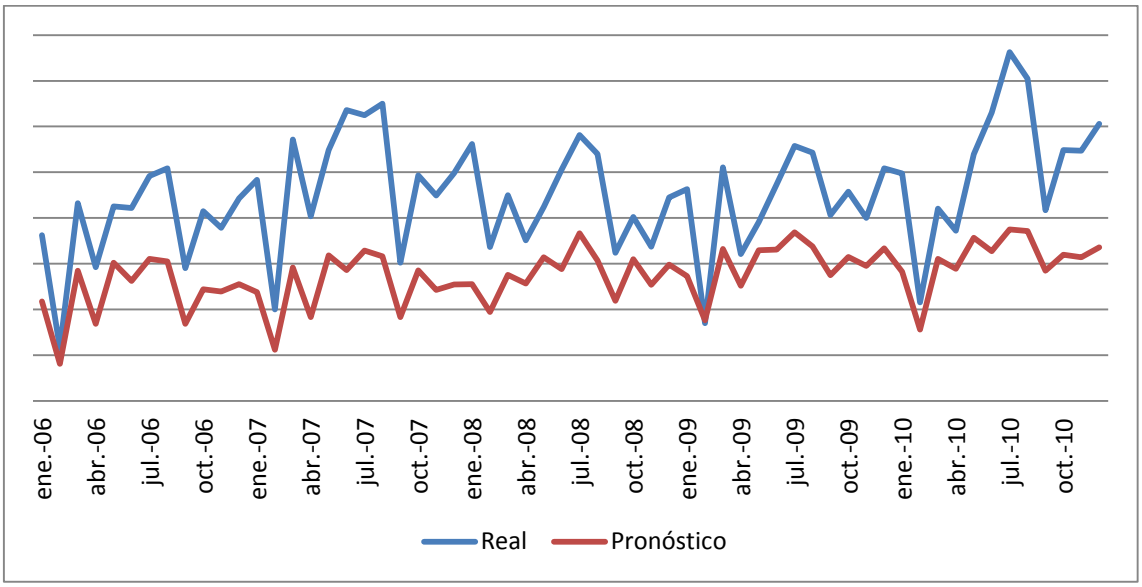
MAPE: 4,42%

2007-2011



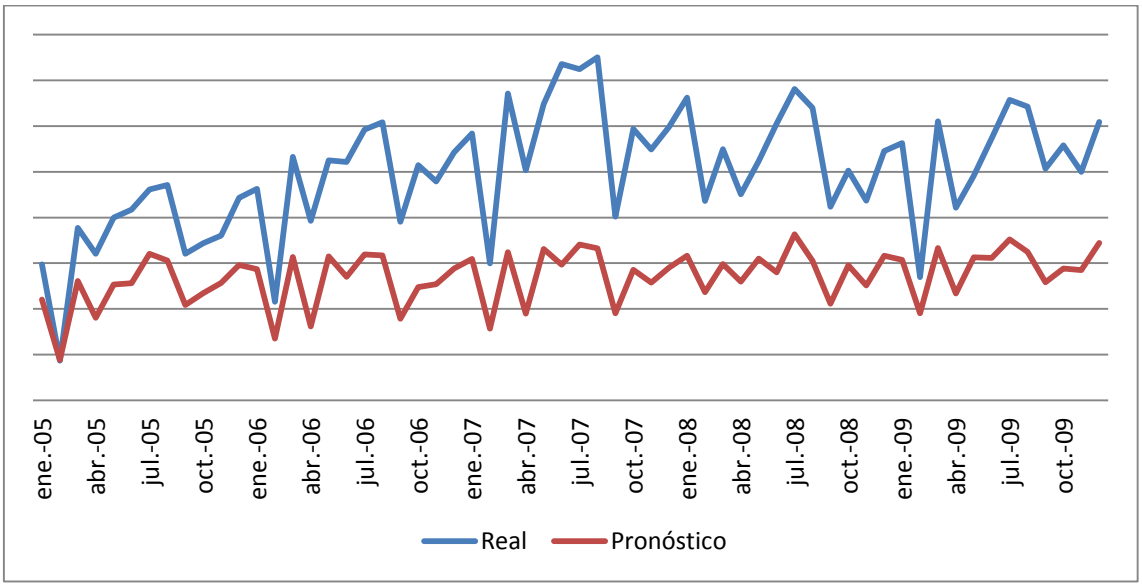
MAPE: 3,96%

2006-2010



MAPE: 7,70%

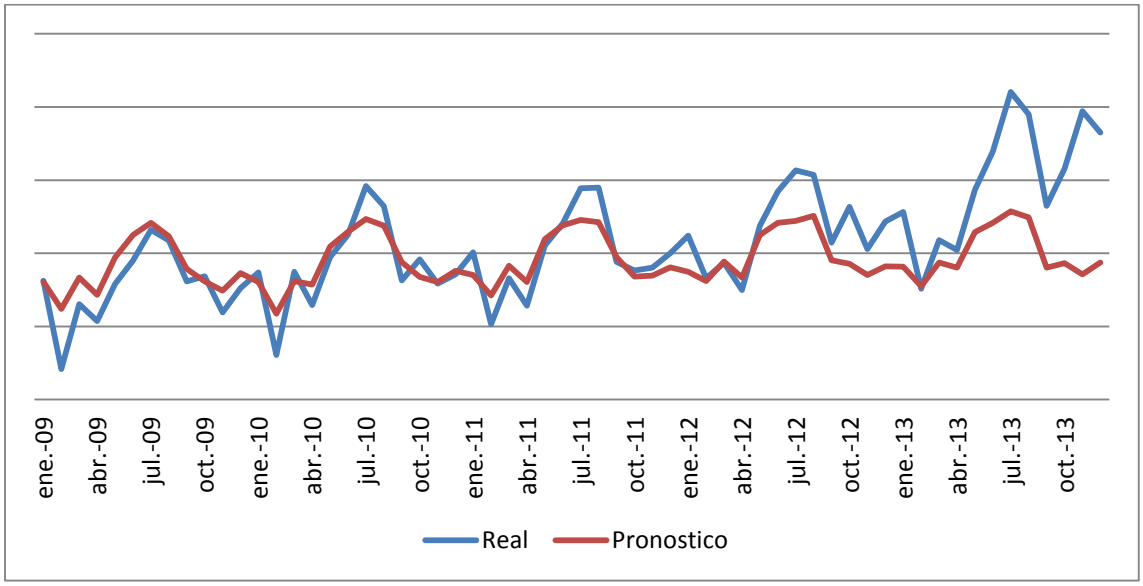
2005-2009



MAPE: 10,21%

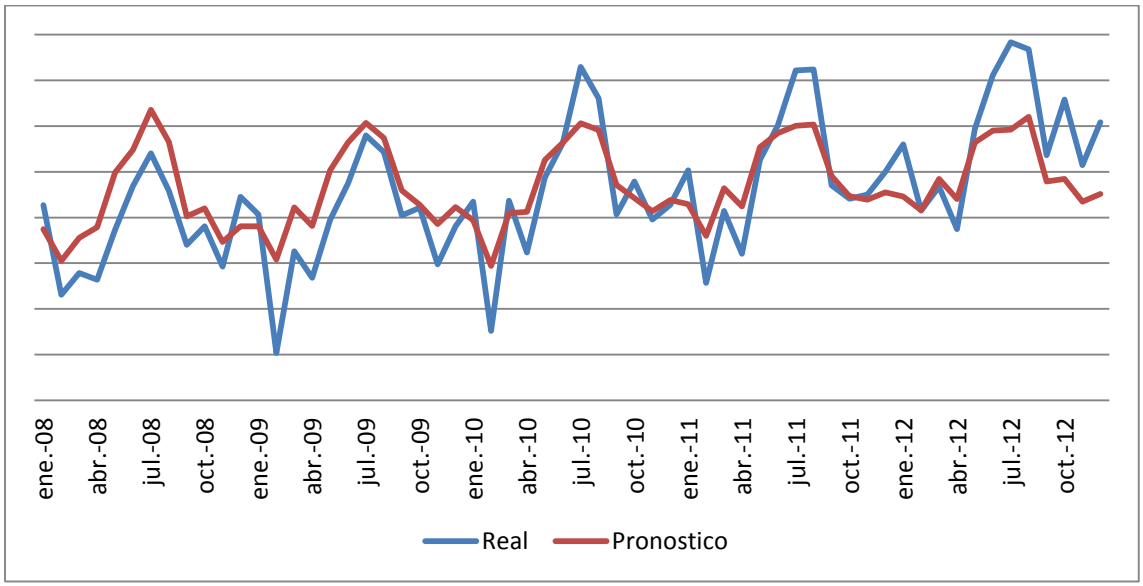
Demanda de Energía Residencial

2009-2013



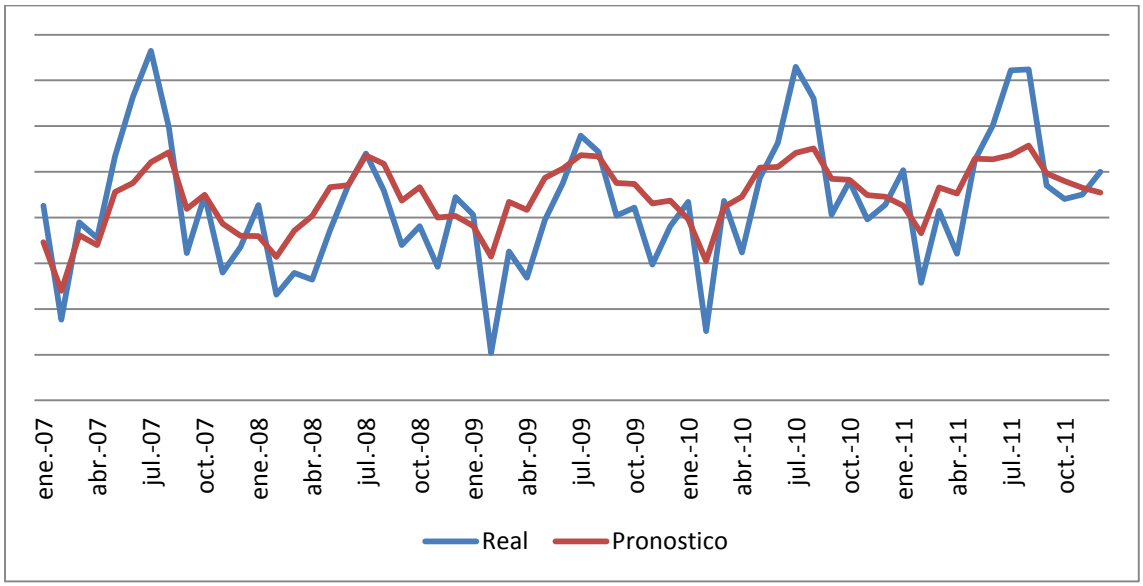
MAPE: 6,39%

2008-2012



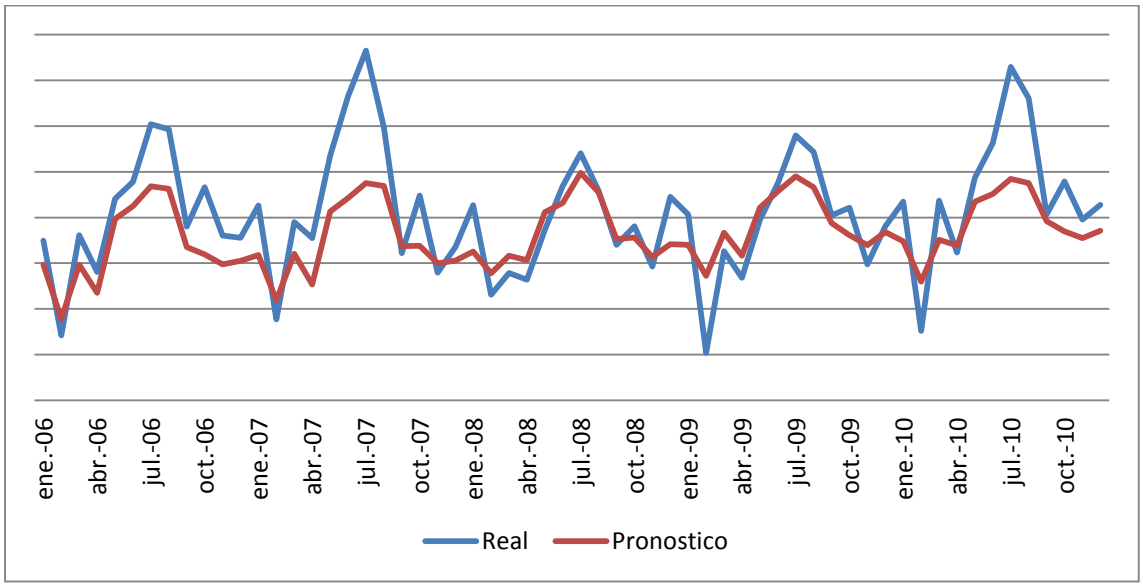
MAPE: 5,03%

2007-2011



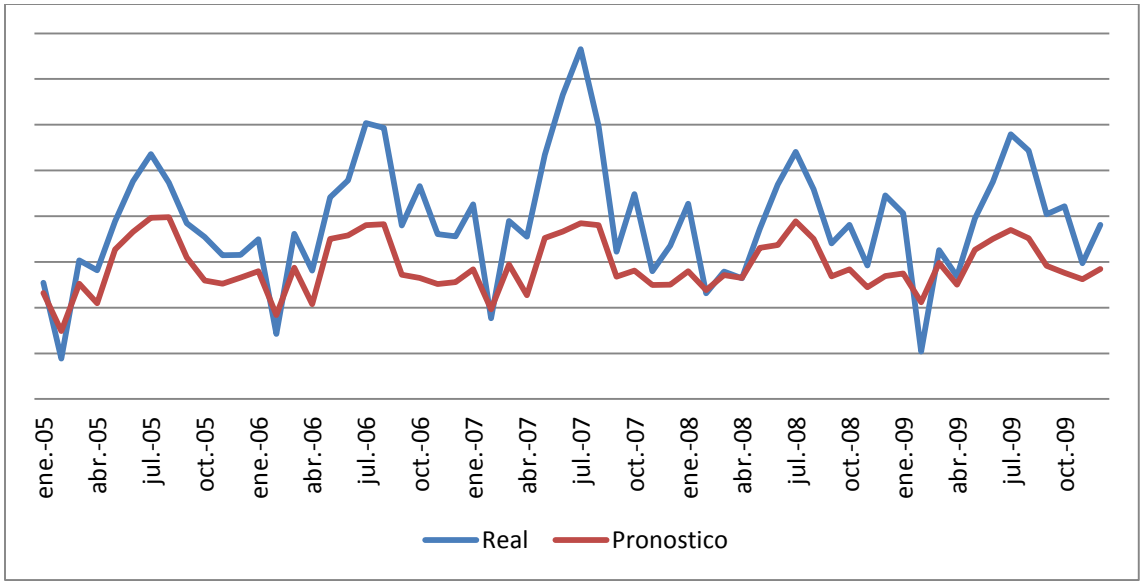
MAPE: 5,46%

2006-2010



MAPE: 5,06%

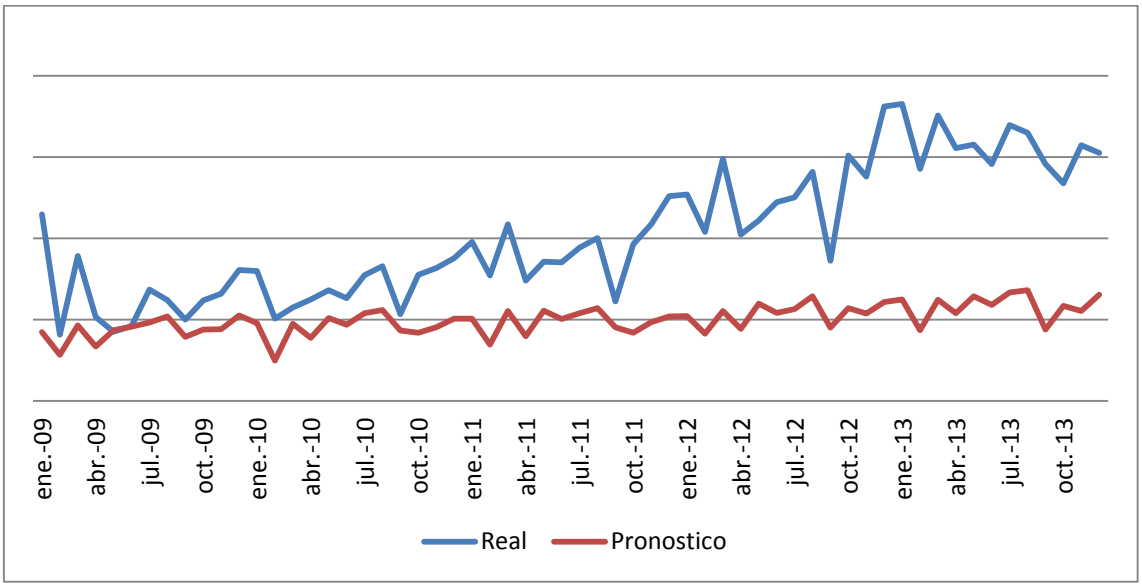
2005-2009



MAPE: 7,31%

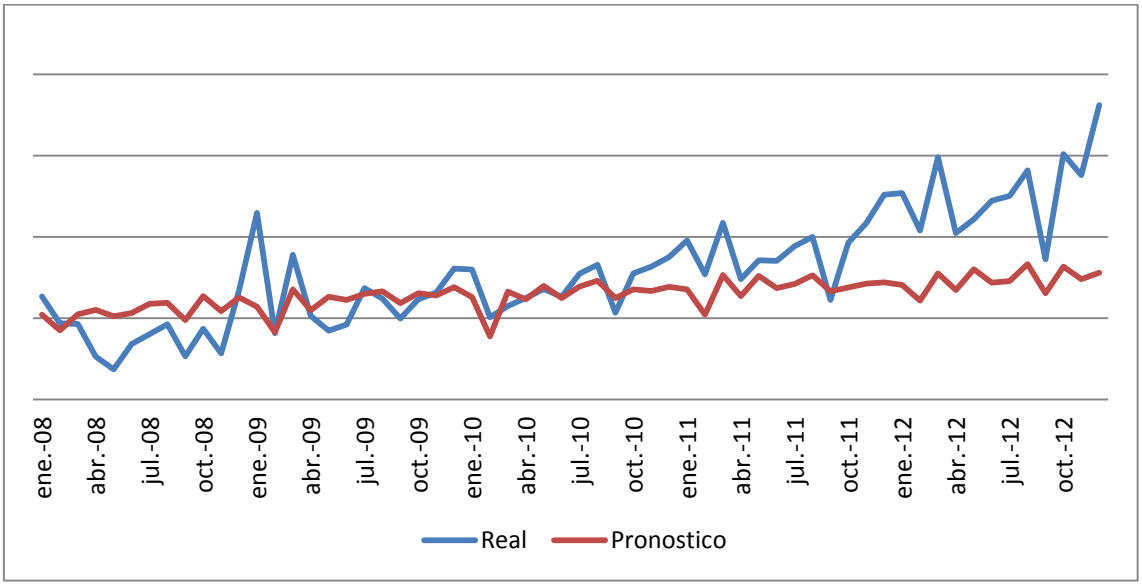
Demanda de Energía Comercial

2009-2013



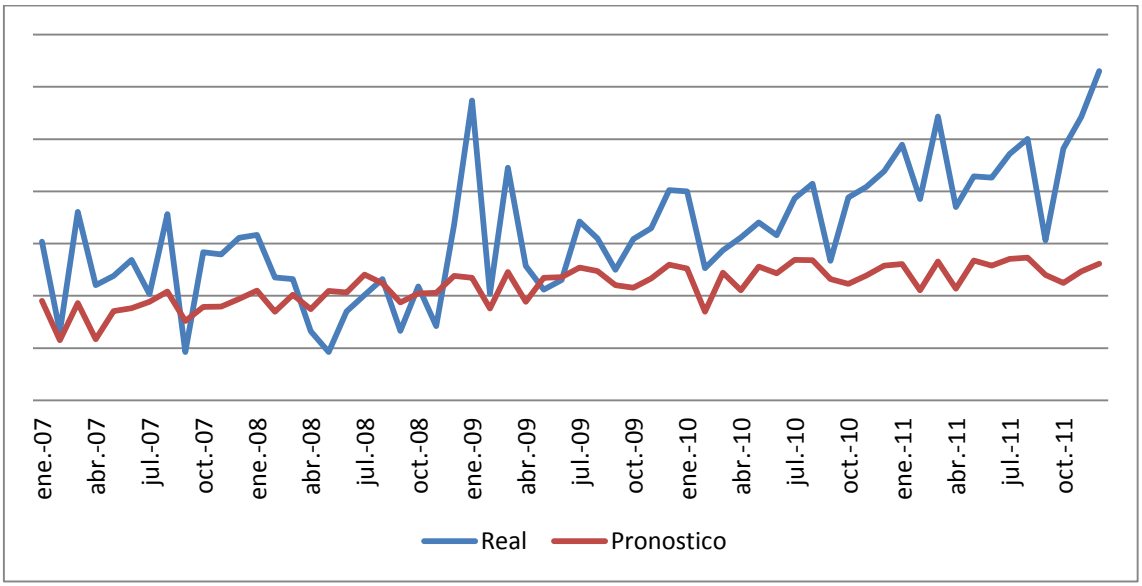
MAPE: 15,09%

2008-2012



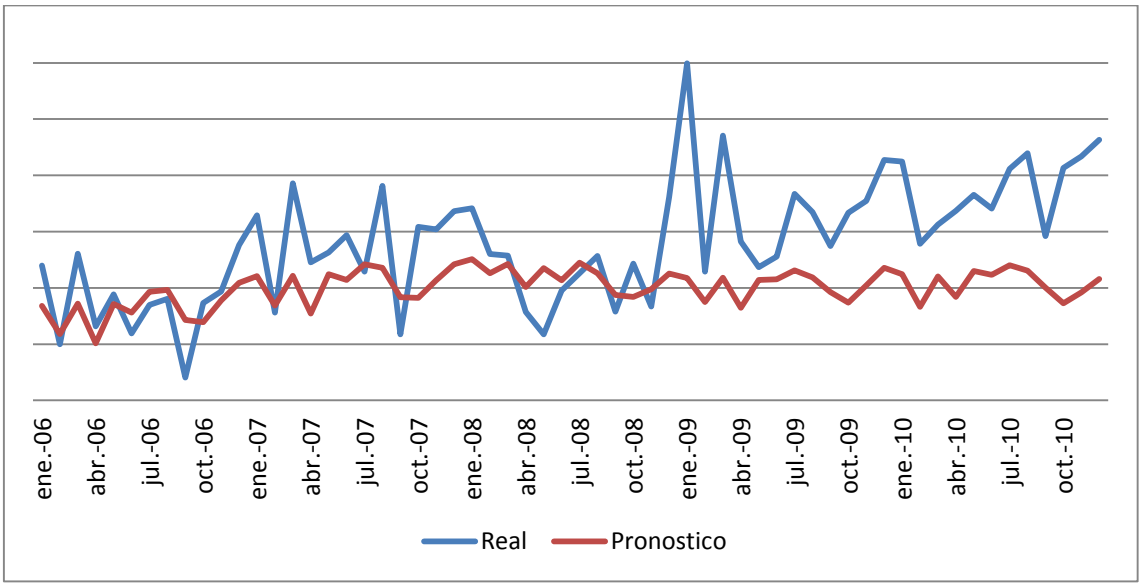
MAPE: 7,33%

2007-2011



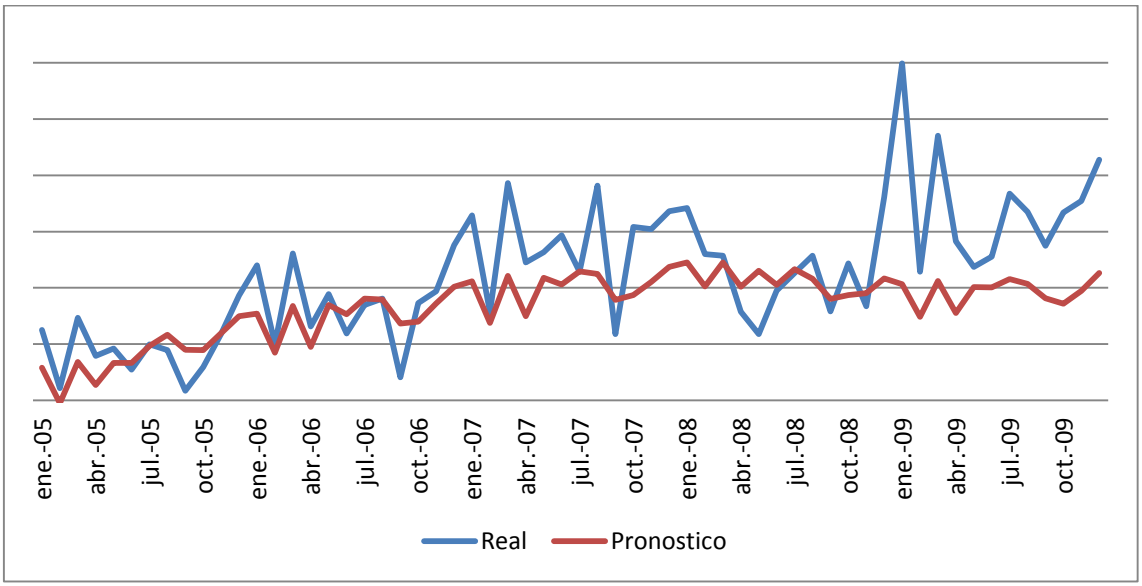
MAPE: 7,61%

2006-2010



MAPE: 6,82%

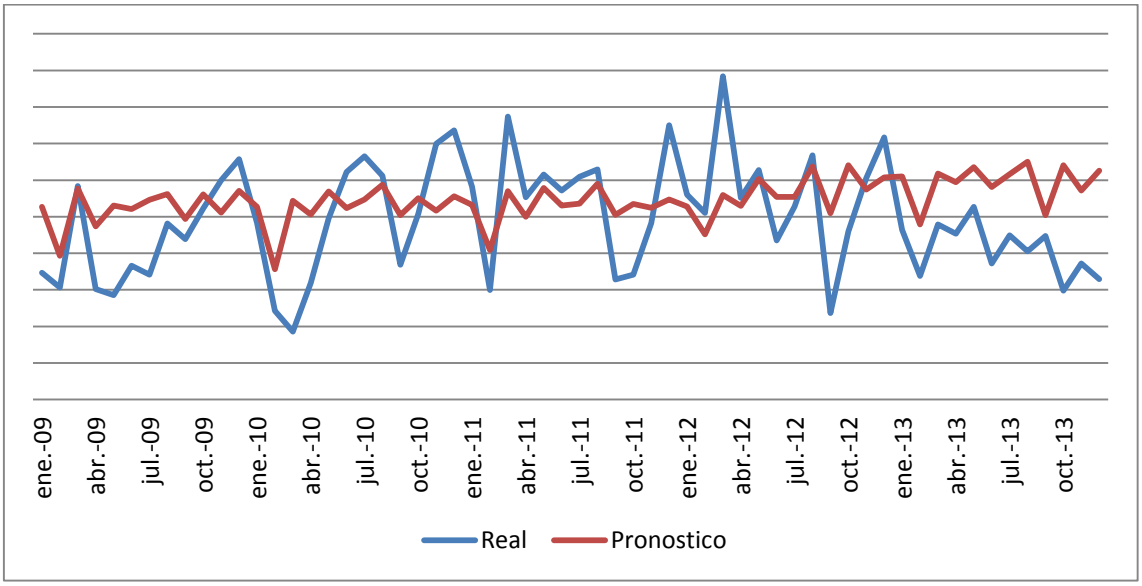
2005-2009



MAPE: 5,37%

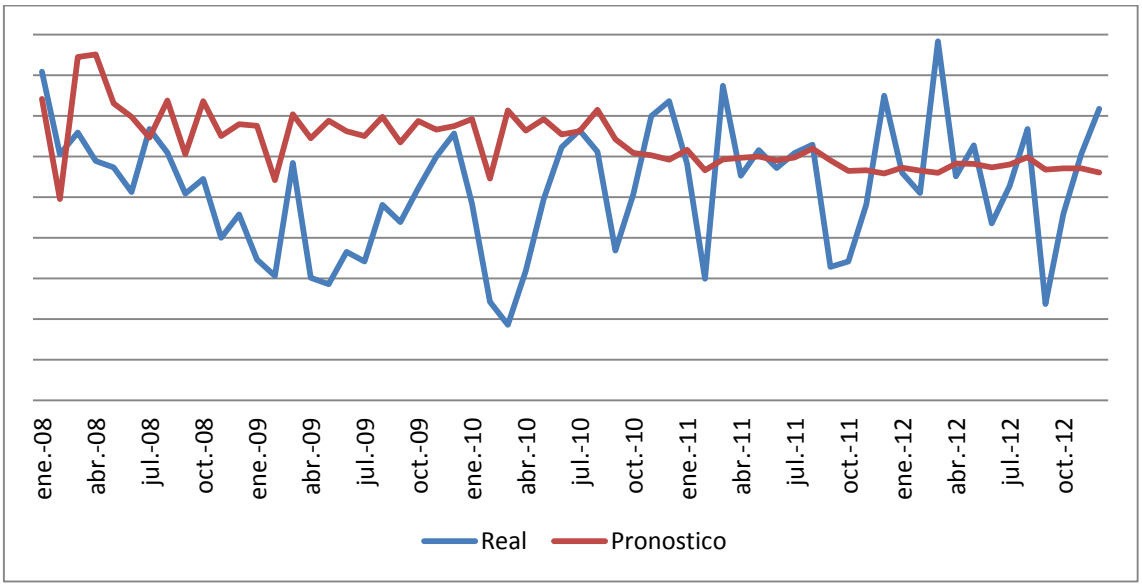
Demanda de Energía Industrial

2009-2013



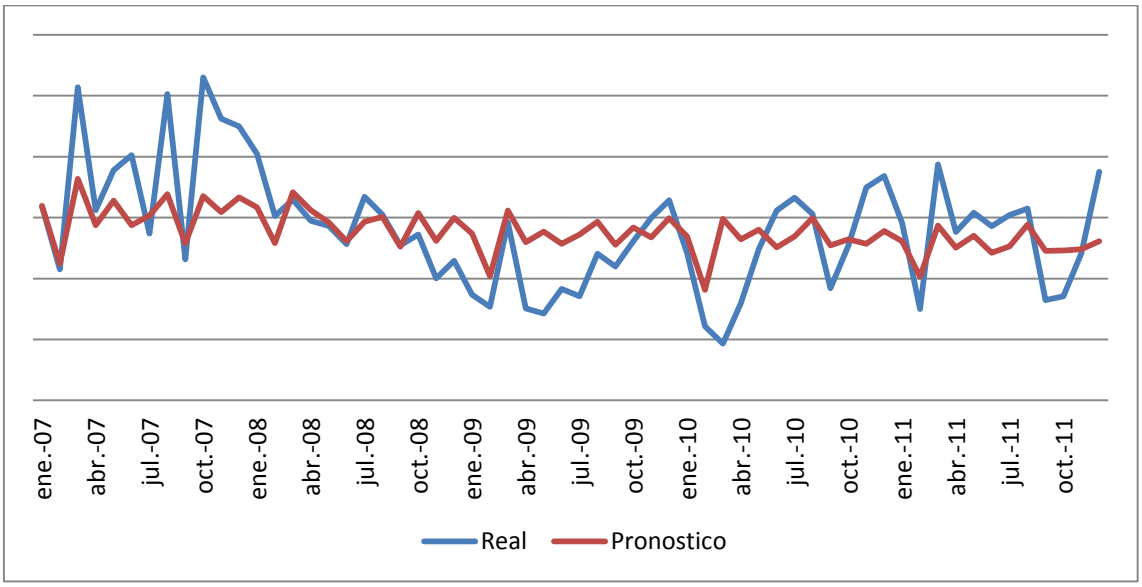
MAPE:5,20%

2008-2012



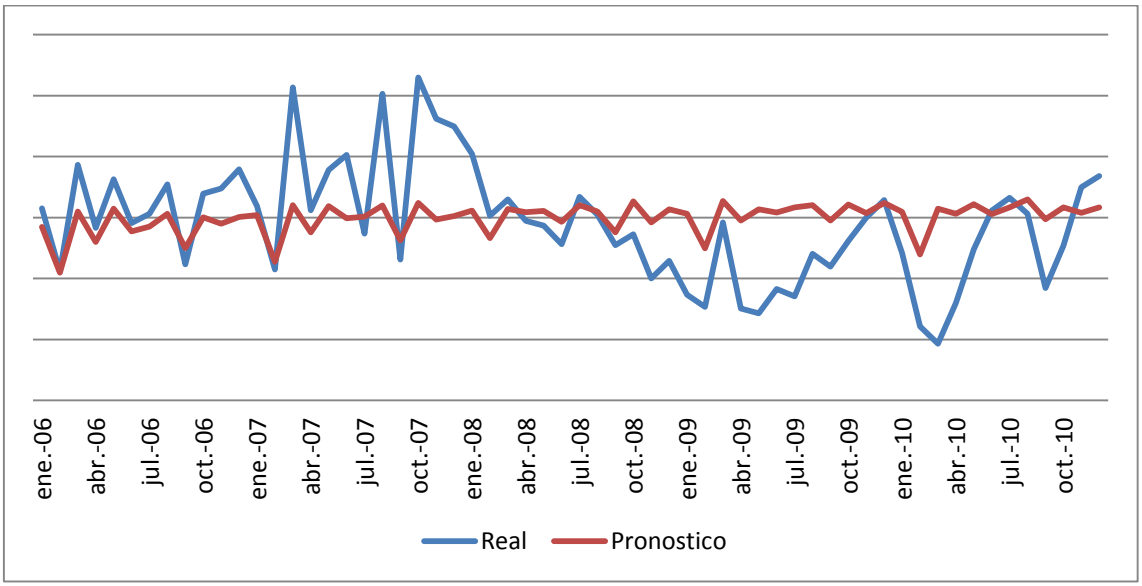
MAPE: 6,57%

2007-2011



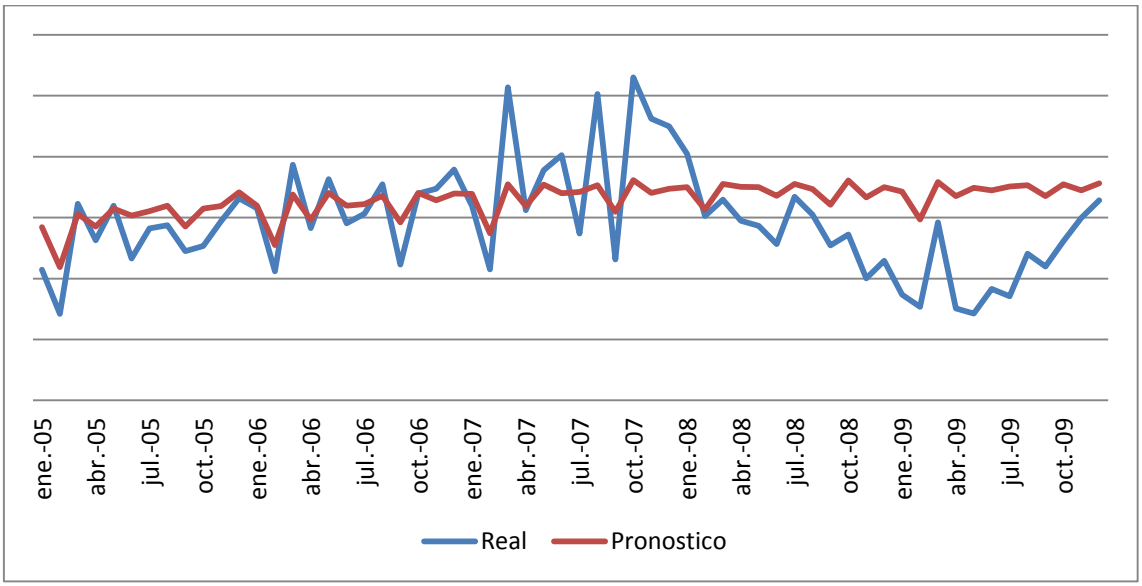
MAPE:4,73%

2006-2010



MAPE: 5,42%

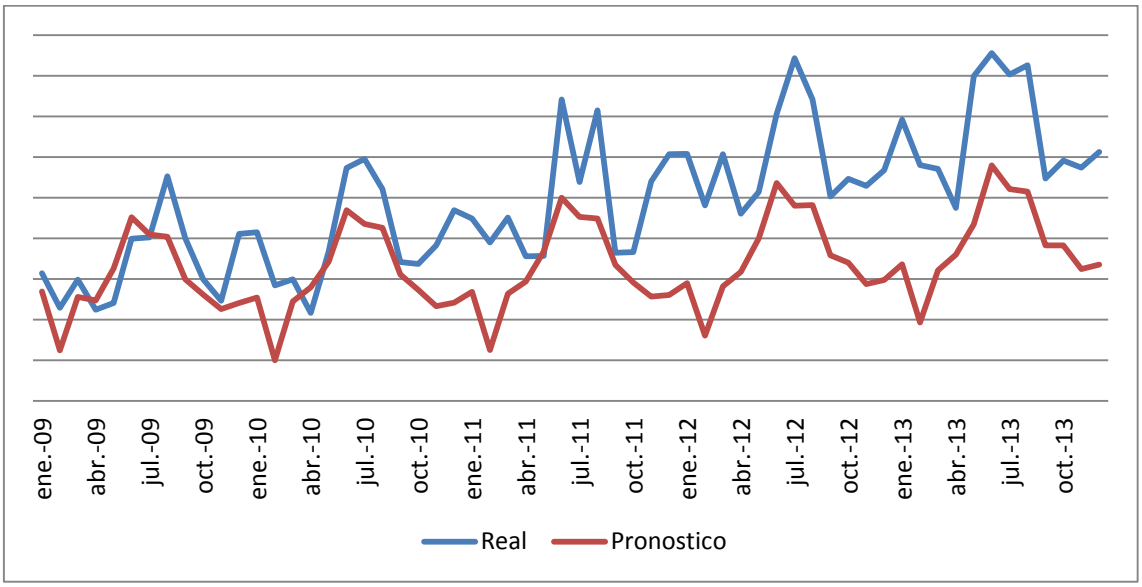
2005-2009



MAPE: 5,45%

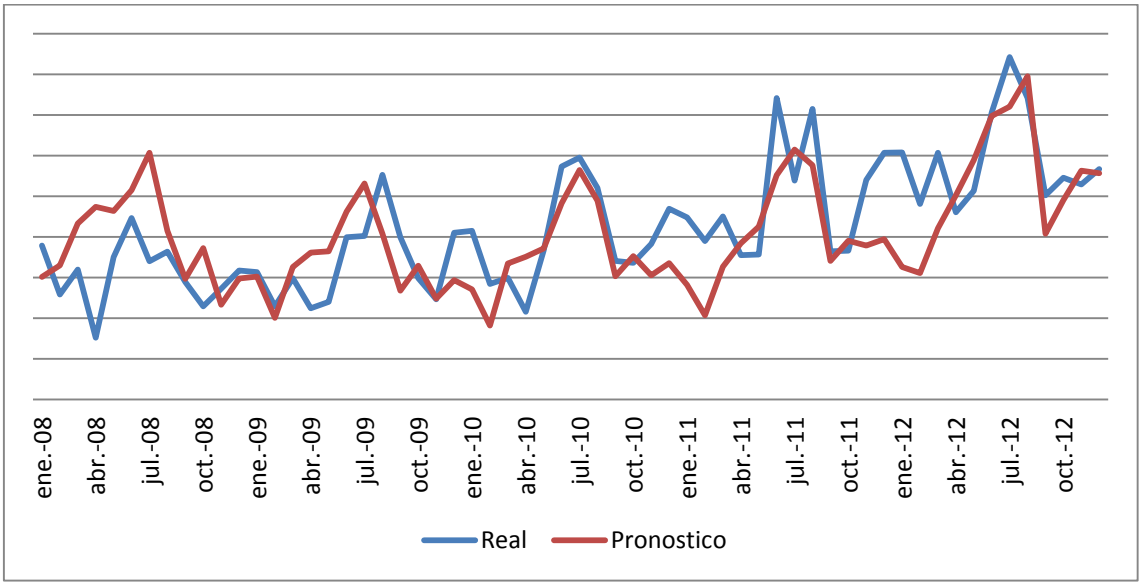
Demanda de Potencia Máxima en el Anillo

2009-2013



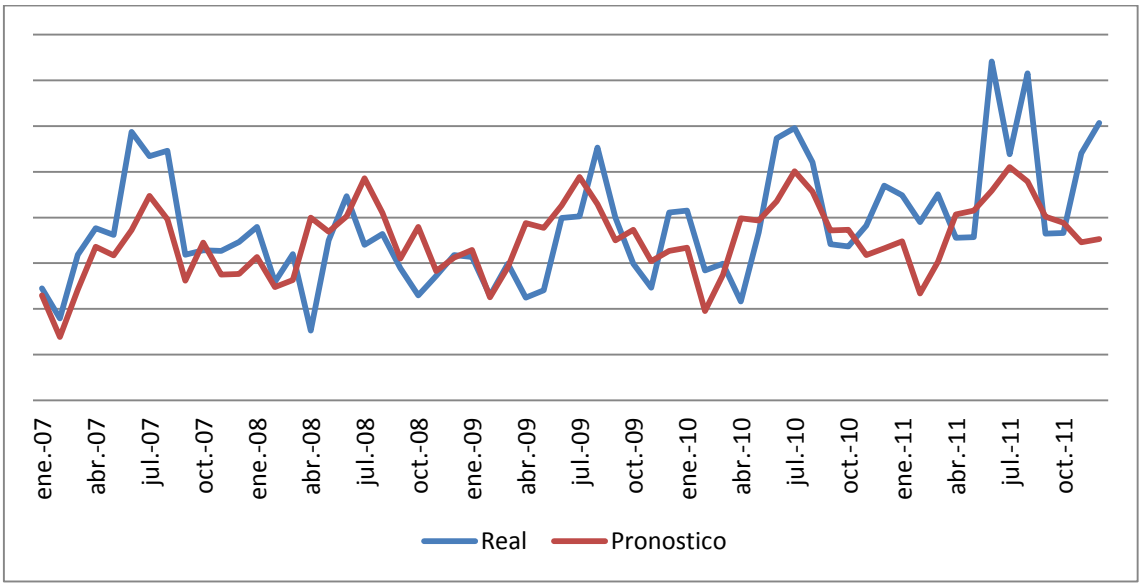
MAPE: 10,71%

2008-2012



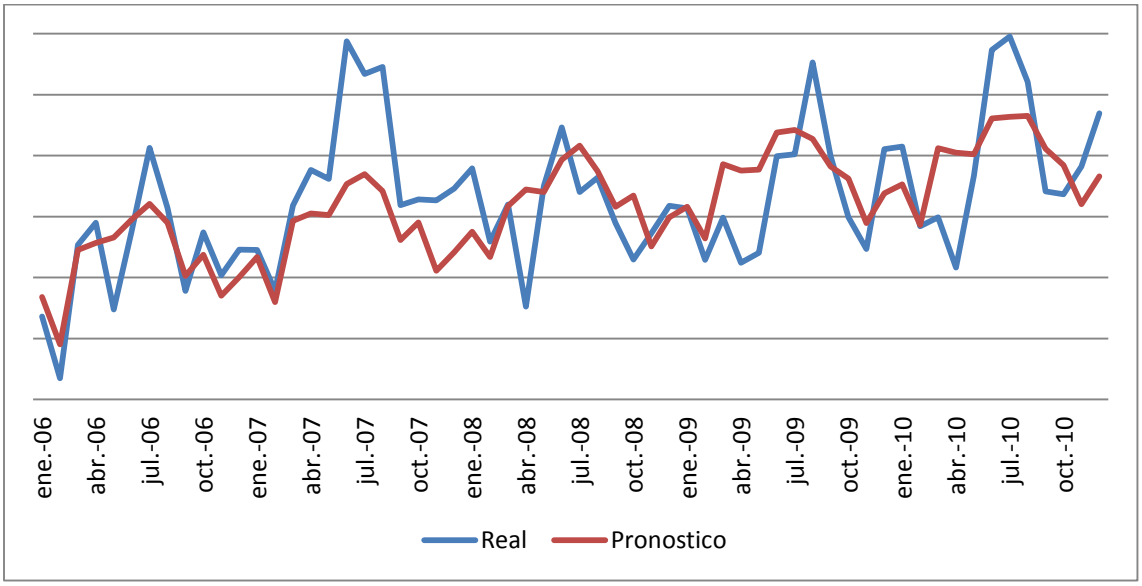
MAPE: 9,85%

2007-2011



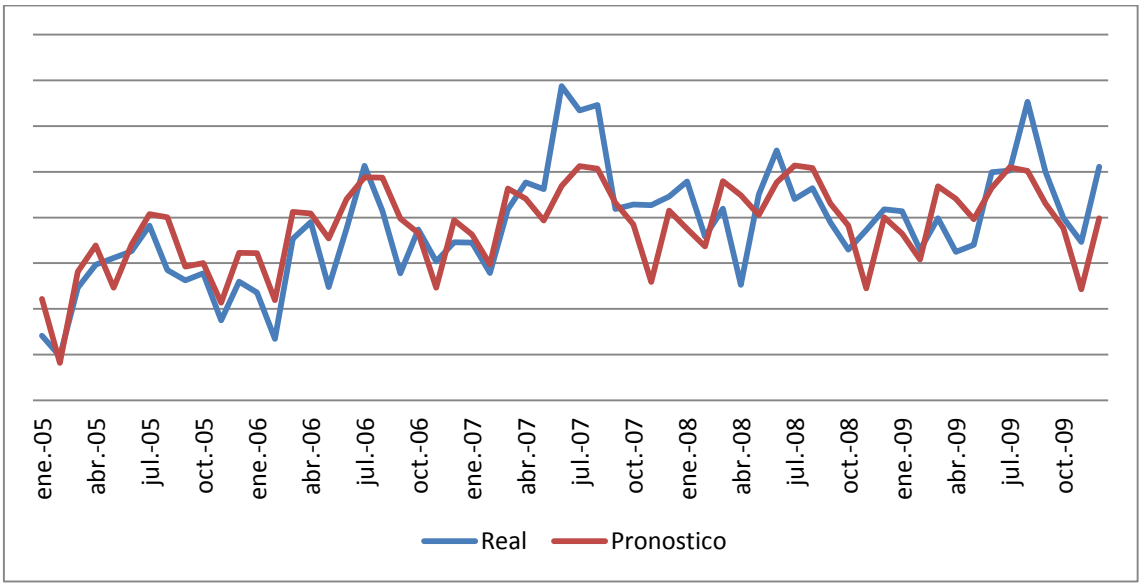
MAPE: 8,63%

2006-2010



MAPE: 6,52%

2005-2009



MAPE: 7,58%

8.D.b Resultados con Distintos Kernels

Demanda de Energía en el Sistema

El mejor modelo con Kernel Lineal corresponde a la siguiente configuración de SVR:

- Costo: 0.5
- Epsilon: 0.1

Los resultados obtenidos fueron:

Métricas	Horizontes					
	2009-2013	2008-2012	2007-2011	2007-2010	2007-2009	Promedio
MAPE Entrenamiento	3,29%	3,29%	4,39%	5,82%	7,32%	**4,82%**
MAPE Validación	5,07%	6,44%	6,72%	6,95%	9,99%	**7,04%**

El mejor modelo con Kernel Polinomial corresponde a la siguiente configuración de SVR:

- Grado: 3
- Costo: 3
- Epsilon: 0.1

Los resultados obtenidos fueron:

Métricas	Horizontes					
	2009-2013	2008-2012	2007-2011	2007-2010	2007-2009	Promedio
MAPE Entrenamiento	3,97%	4,87%	4,67%	6,94%	8,74%	**5,84%**
MAPE Validación	7,96%	8,28%	12,26%	10,82%	12,62%	**10,39%**

Demanda de Energía Residencial

El mejor modelo con Kernel Lineal corresponde a la siguiente configuración de SVR:

- Costo: 0.5
- Epsilon: 0.1

Los resultados obtenidos fueron:

Métricas	Horizontes					
	2009-2013	2008-2012	2007-2011	2007-2010	2007-2009	Promedio
MAPE Entrenamiento	4,20%	3,92%	6,74%	5,37%	6,86%	**5,42%**
MAPE Validación	3,55%	7,87%	7,61%	9,95%	9,27%	**7,65%**

El mejor modelo con Kernel Polinomial corresponde a la siguiente configuración de SVR:

- Grado: 3
- Costo: 0.7
- Epsilon: 0.1

Los resultados obtenidos fueron:

Métricas	Horizontes					
	2009-2013	2008-2012	2007-2011	2007-2010	2007-2009	Promedio
MAPE Entrenamiento	5,00%	5,01%	4,37%	6,58%	5,45%	**5,28%**
MAPE Validación	6,25%	6,48%	7,46%	7,23%	15,80%	**8,64%**

Demanda de Energía Comercial

El mejor modelo con Kernel Lineal corresponde a la siguiente configuración de SVR:

- Costo: 0.5
- Epsilon: 0.1

Los resultados obtenidos fueron:

Métricas	Horizontes					
	2009-2013	2008-2012	2007-2011	2007-2010	2007-2009	Promedio
MAPE Entrenamiento	4,85%	4,81%	5,71%	4,90%	5,73%	**5,20%**
MAPE Validación	6,01%	8,53%	7,70%	9,33%	16,57%	**9,63%**

El mejor modelo con Kernel Polinomial corresponde a la siguiente configuración de SVR:

- Grado: 3
- Costo: 5.5
- Epsilon: 0.1

Los resultados obtenidos fueron:

Métricas	Horizontes					
	2009-2013	2008-2012	2007-2011	2007-2010	2007-2009	Promedio
MAPE Entrenamiento	5,66%	5,58%	7,45%	7,78%	6,79%	**6,65%**
MAPE Validación	9,12%	13,91%	13,00%	12,75%	21,44%	**14,04%**

Demanda de Energía Industrial

El mejor modelo con Kernel Lineal corresponde a la siguiente configuración de SVR:

- Costo: 5.5
- Epsilon: 0.1

Los resultados obtenidos fueron:

Métricas	Horizontes					
	2009-2013	2008-2012	2007-2011	2007-2010	2007-2009	Promedio
MAPE Entrenamiento	2,54%	2,48%	3,02%	6,13%	7,19%	**4,27%**
MAPE Validación	2,66%	2,95%	10,34%	15,04%	21,09%	**10,42%**

El mejor modelo con Kernel Polinomial corresponde a la siguiente configuración de SVR:

- Grado: 3
- Costo: 0.5
- Epsilon: 0.1

Los resultados obtenidos fueron:

Métricas	Horizontes					
	2009-2013	2008-2012	2007-2011	2007-2010	2007-2009	Promedio
MAPE Entrenamiento	3,33%	3,28%	3,38%	3,25%	3,80%	**3,41%**
MAPE Validación	4,88%	5,42%	7,28%	7,45%	13,64%	**7,73%**

Demanda de Potencia Máxima en el Anillo

El mejor modelo con Kernel Lineal corresponde a la siguiente configuración de SVR:

- Costo: 1
- Epsilon: 0.1

Los resultados obtenidos fueron:

Métricas	Horizontes					
	2009-2013	2008-2012	2007-2011	2007-2010	2007-2009	Promedio
MAPE Entrenamiento	3,13%	3,13%	5,09%	6,99%	7,01%	**5,07%**
MAPE Validación	6,08%	6,16%	11,01%	11,18%	11,90%	**9,27%**

El mejor modelo con Kernel Polinomial corresponde a la siguiente configuración de SVR:

- Grado: 3
- Costo: 4
- Epsilon: 0.1

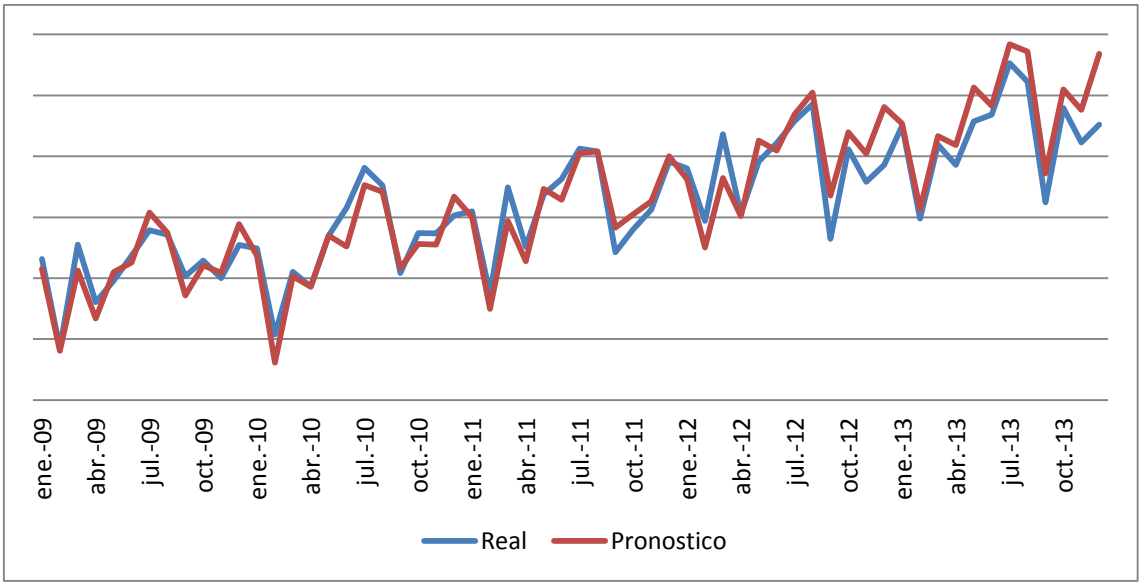
Los resultados obtenidos fueron:

Métricas	Horizontes					
	2009-2013	2008-2012	2007-2011	2007-2010	2007-2009	Promedio
MAPE Entrenamiento	4,41%	5,89%	7,37%	4,73%	9,07%	**6,29%**
MAPE Validación	9,89%	10,90%	11,19%	11,58%	13,57%	**11,43%**

8.D.c Aplicación de Redes Neuronales

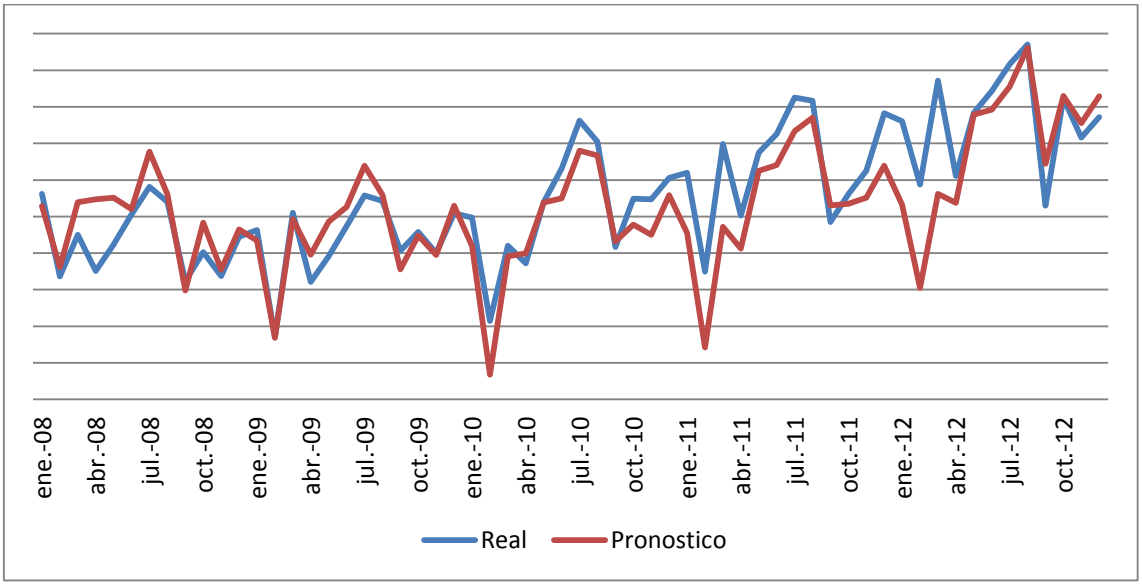
Demanda de Energía en el Sistema

2009-2013



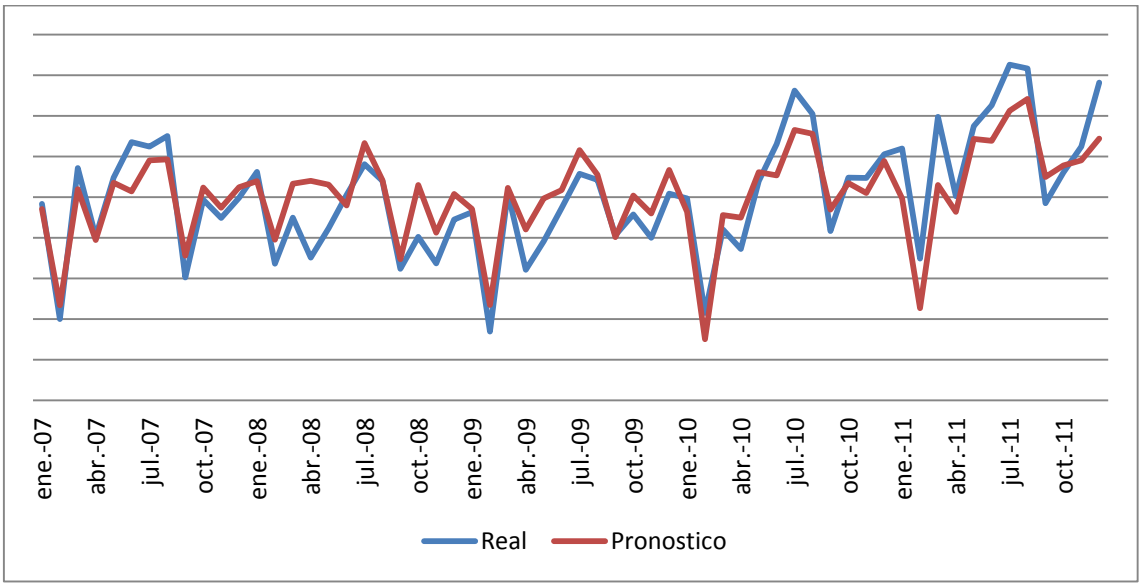
MAPE: 2,21%

2008-2012



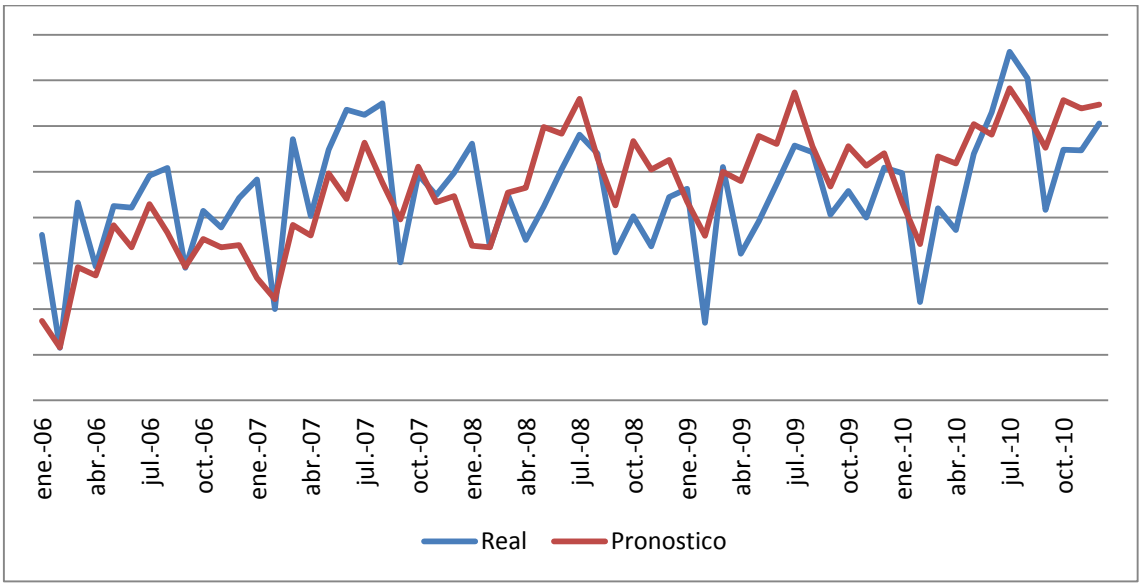
MAPE: 3,10%

2007-2011



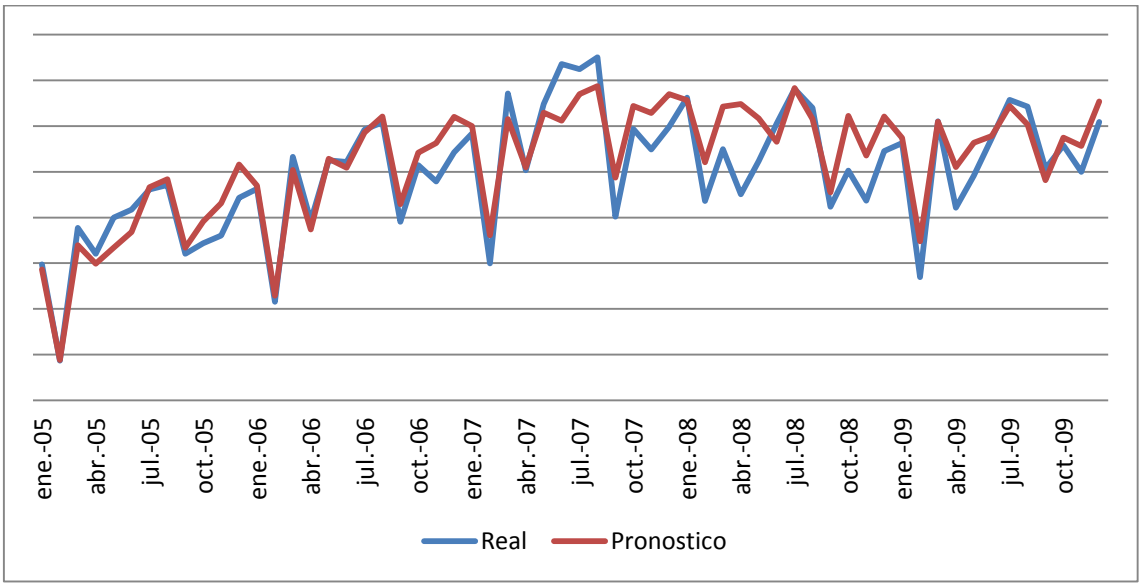
MAPE: 2,51%

2006-2010



MAPE: 4,01%

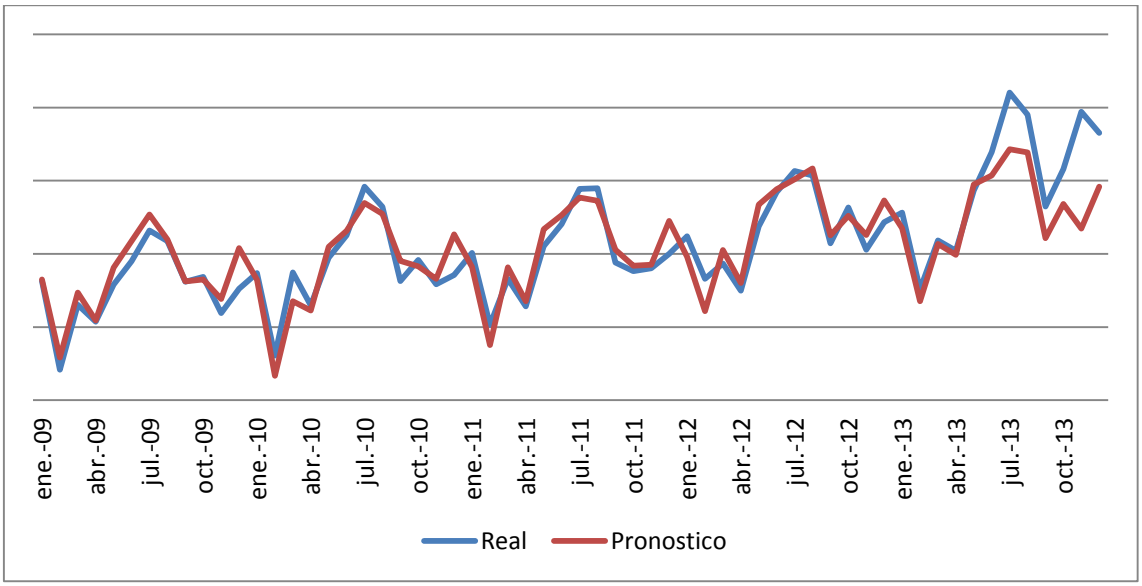
2005-2009



MAPE: 2,08%

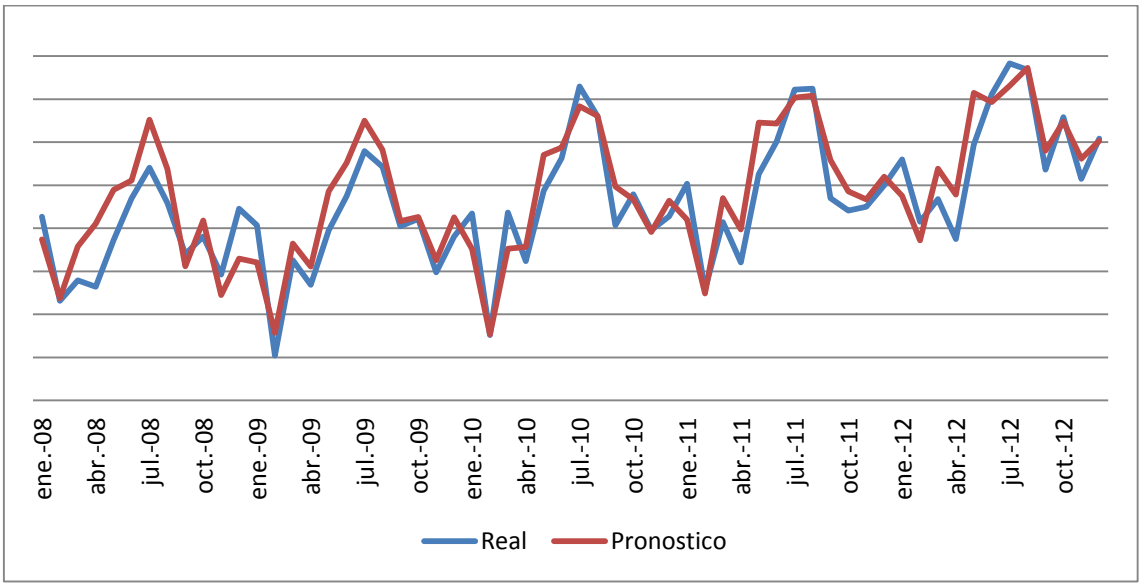
Demanda de Energía Residencial

2009-2013



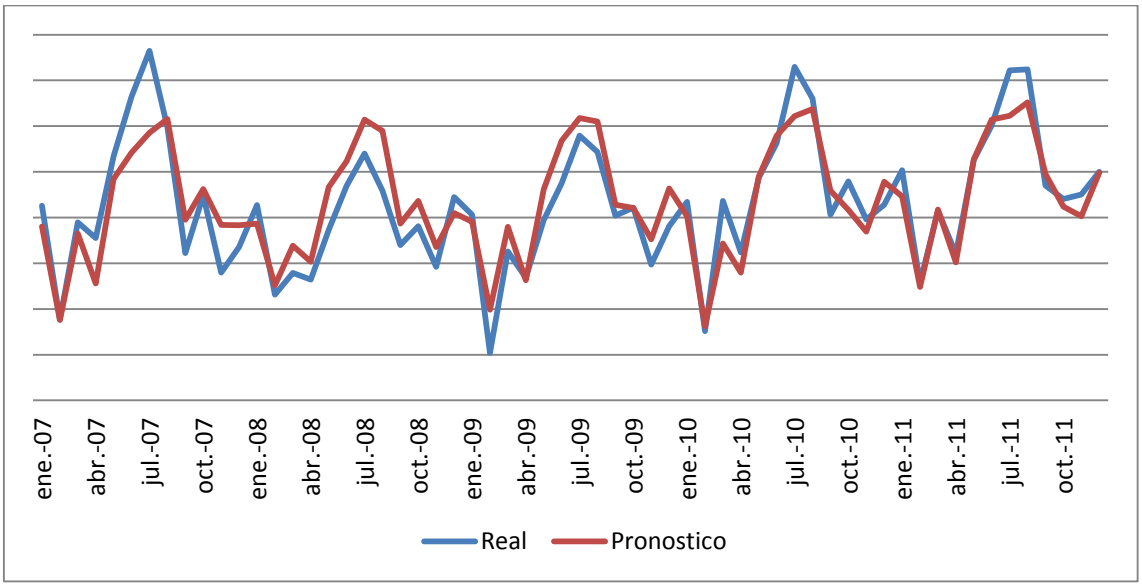
MAPE: 2,96%

2008-2012



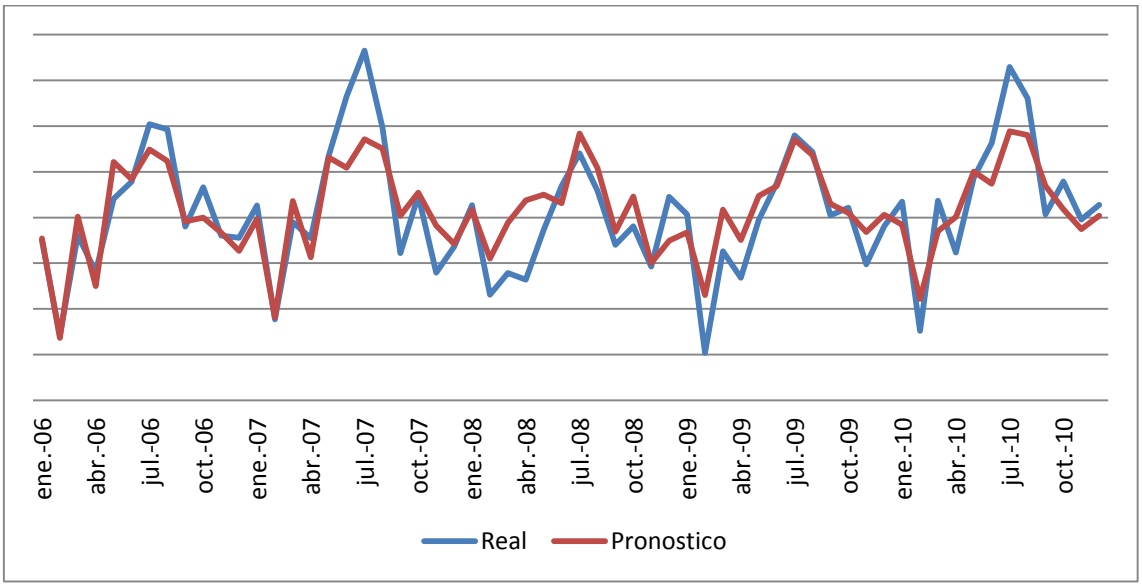
MAPE: 3,79%

2007-2011



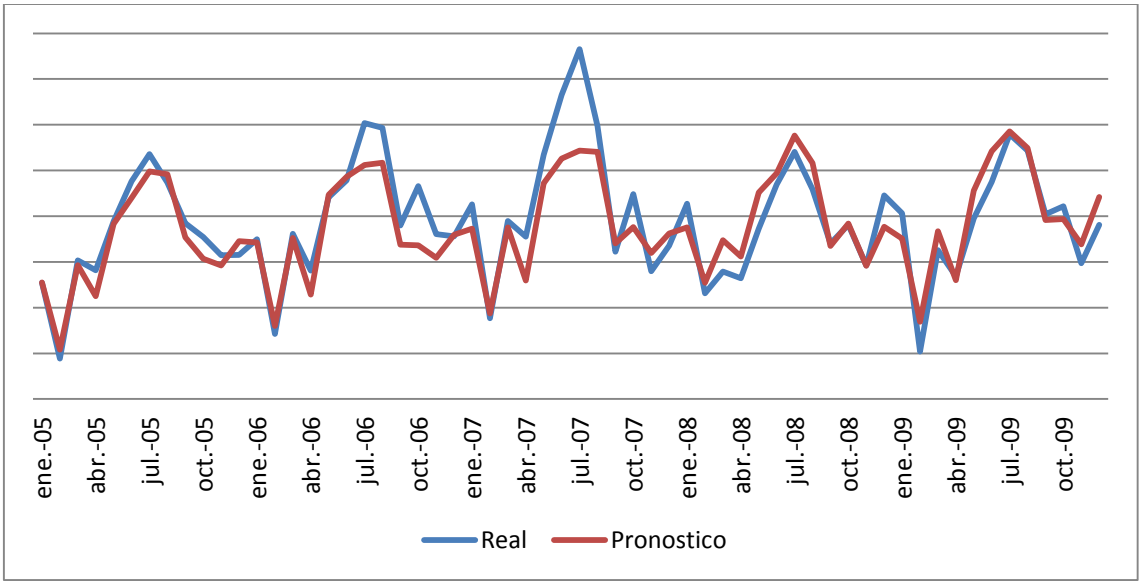
MAPE: 3,41%

2006-2010



MAPE: 3,65%

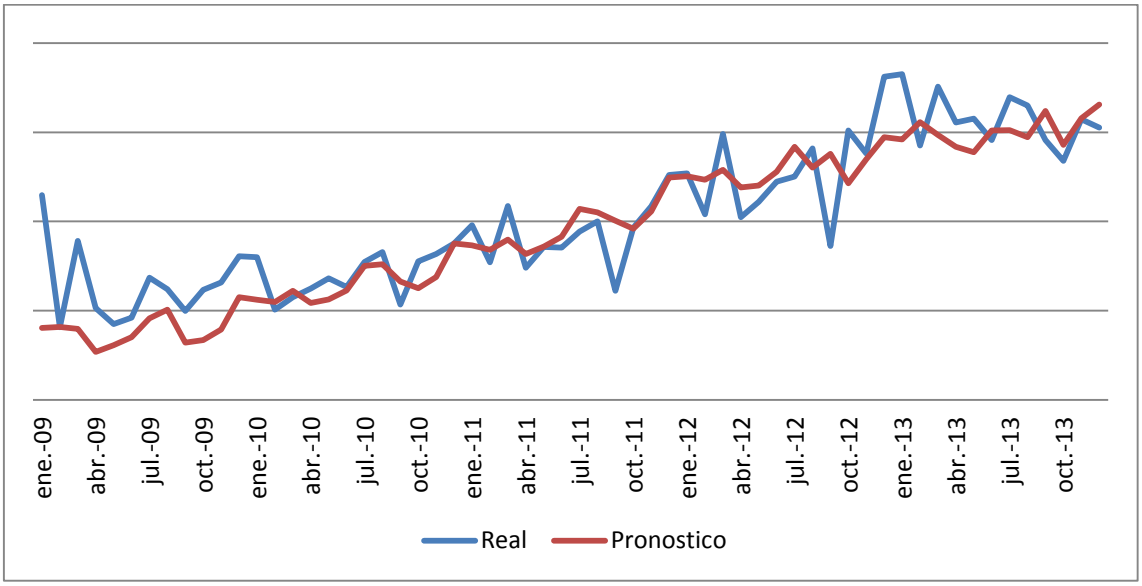
2005-2009



MAPE: 3,68%

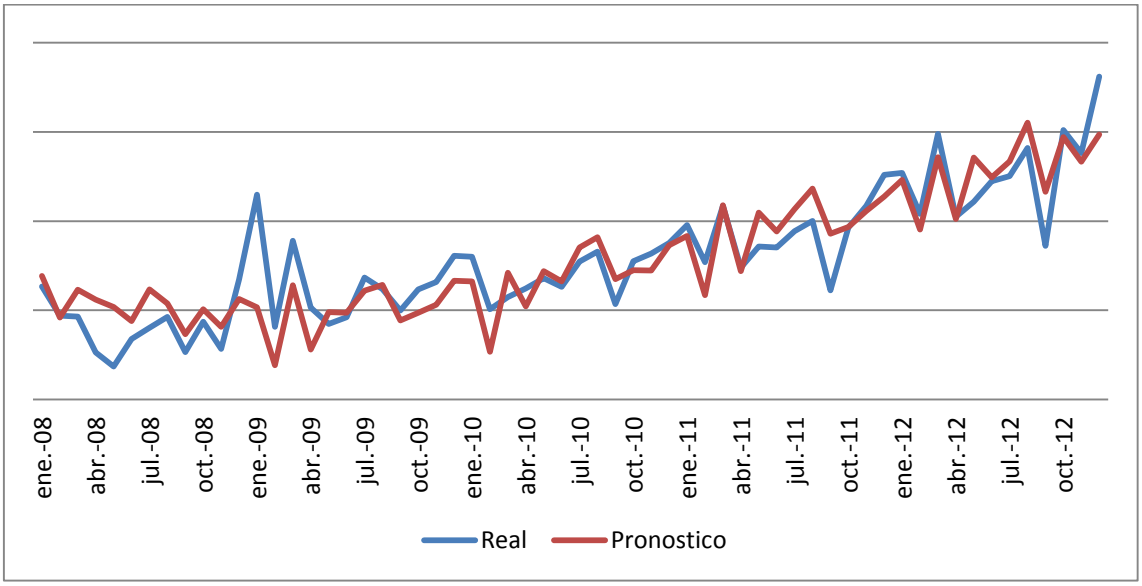
Demanda de Energía Comercial

2009-2013



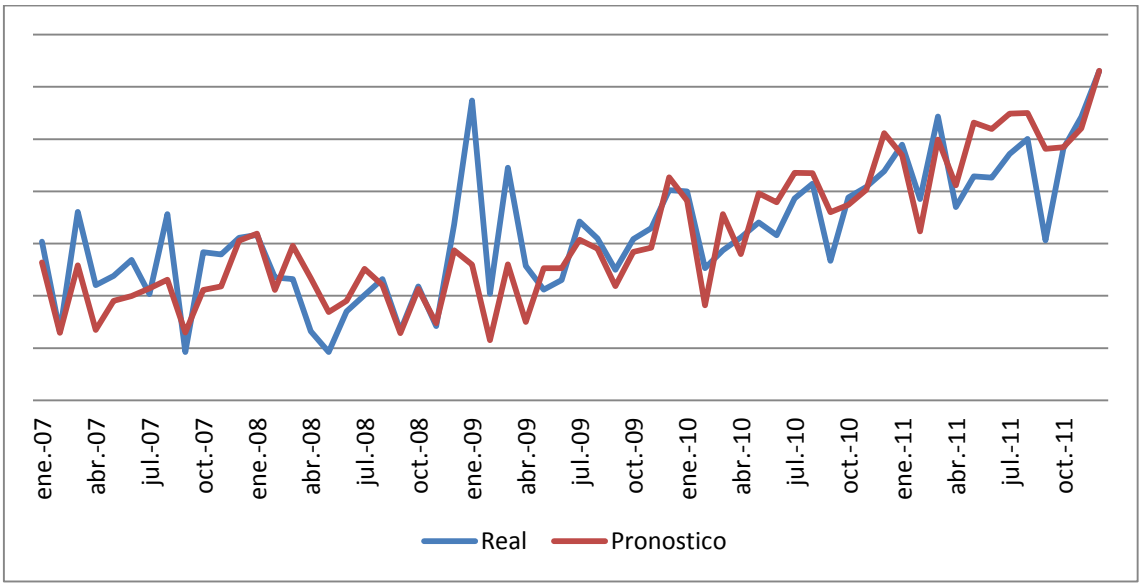
MAPE: 4,71%

2008-2012



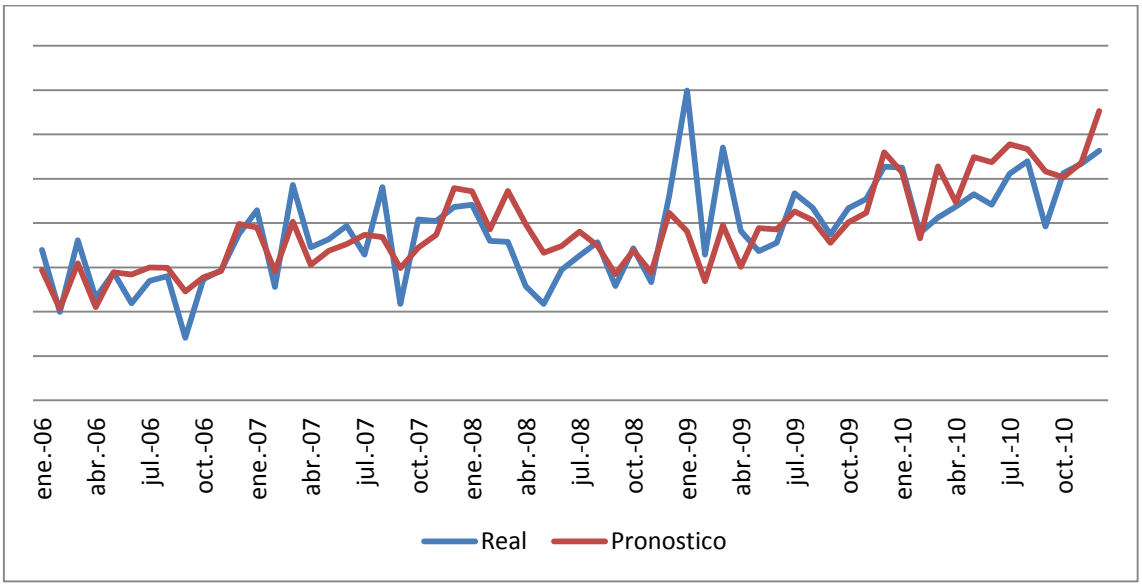
MAPE: 4,22%

2007-2011



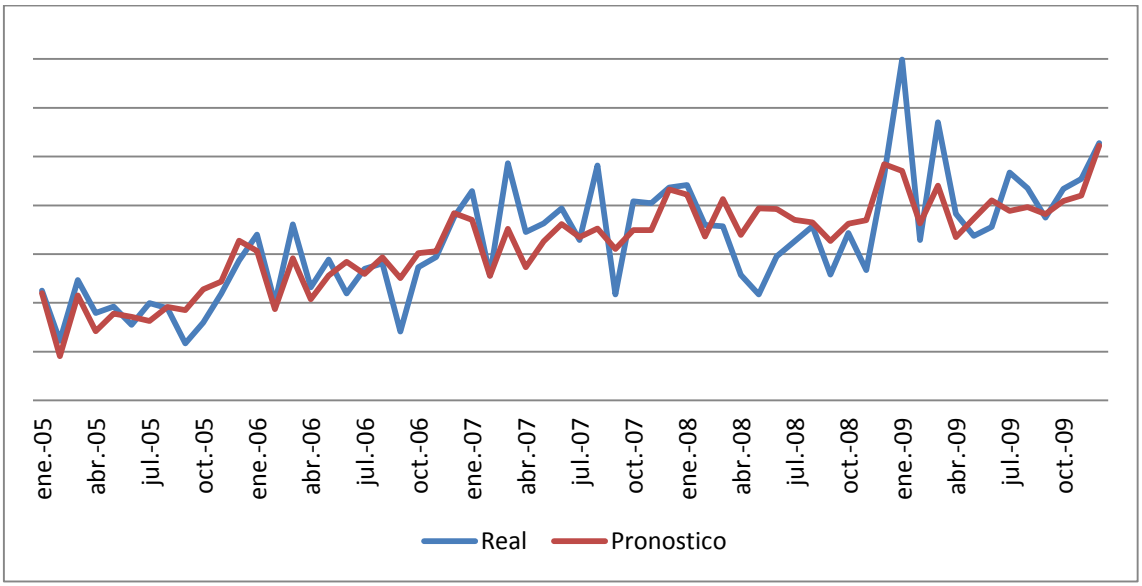
MAPE: 3,73%

2006-2010



MAPE: 3,80%

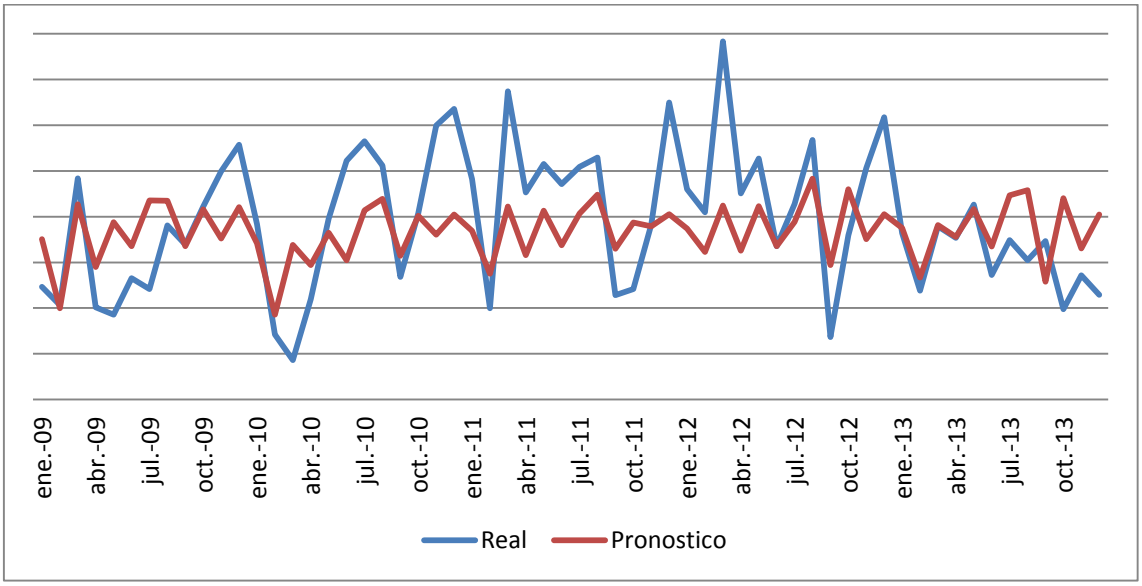
2005-2009



MAPE: 3,64%

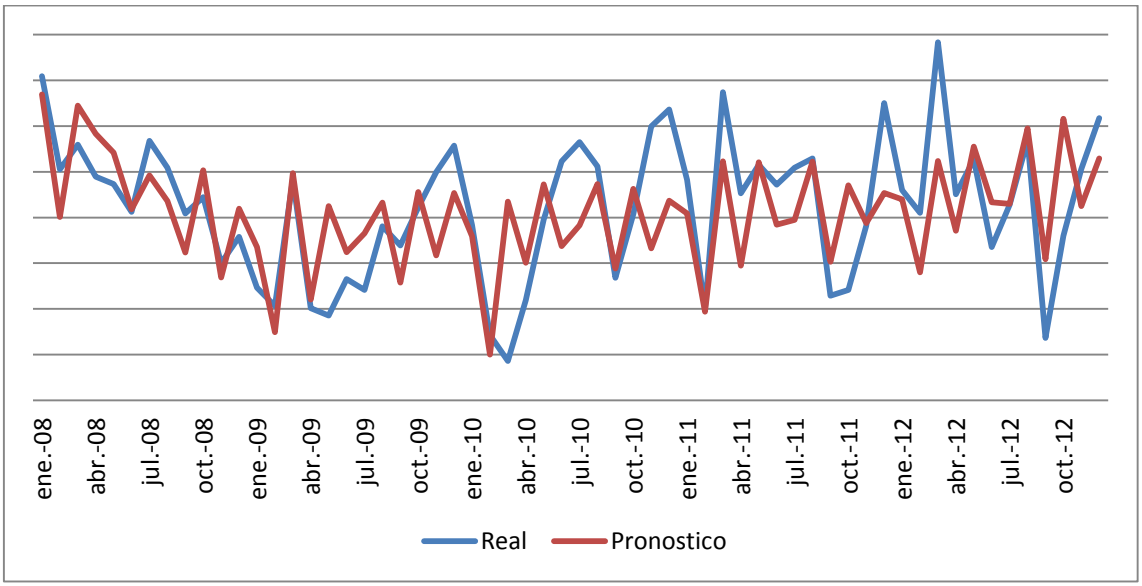
Demanda de Energía Industrial

2009-2013



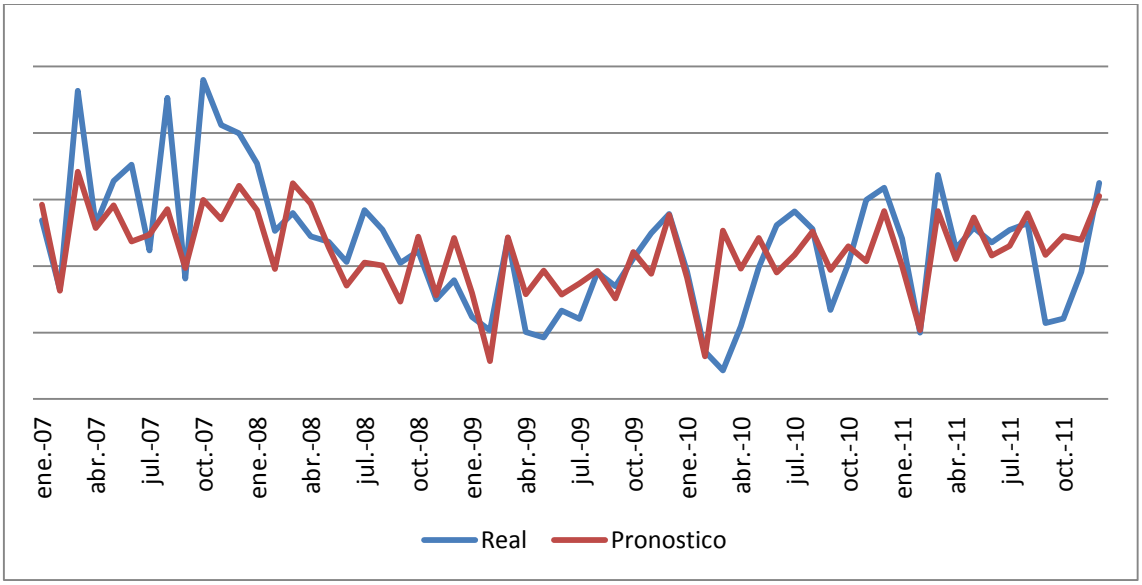
MAPE: 4,25%

2008-2012



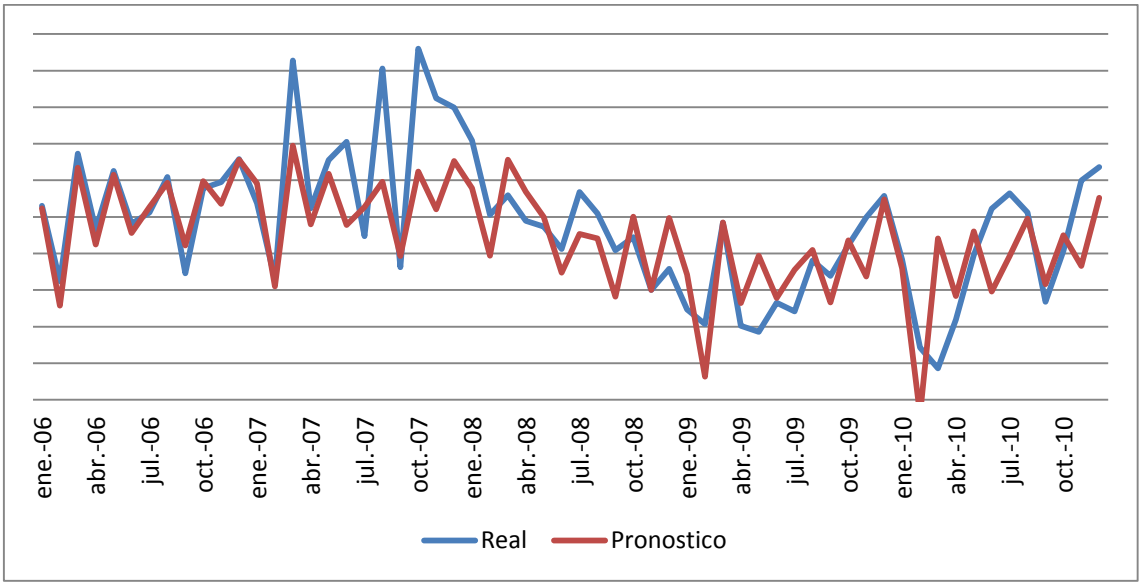
MAPE: 3,78%

2007-2011



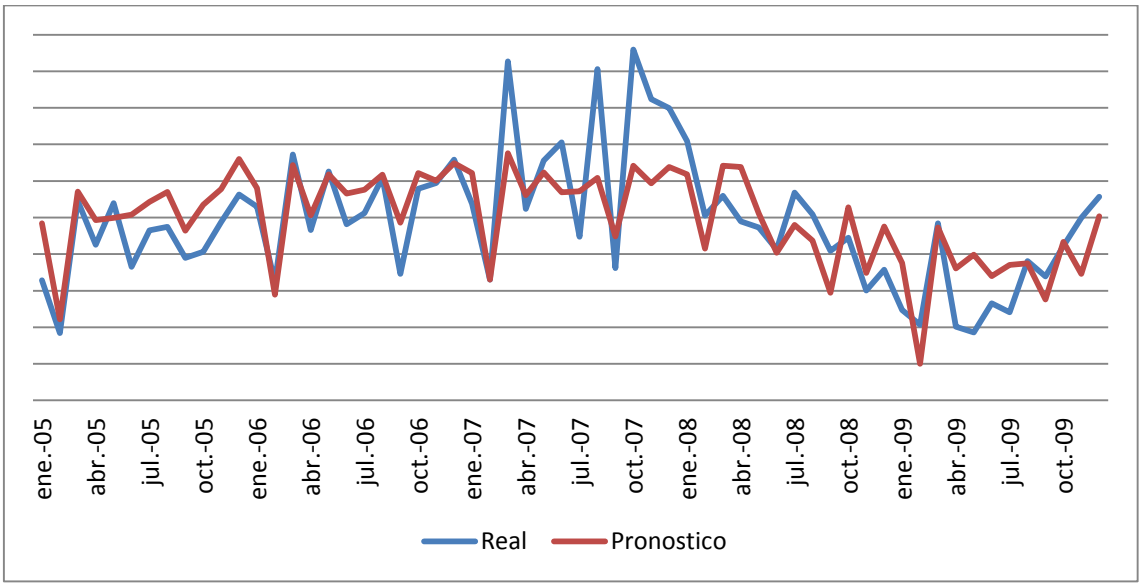
MAPE: 3,96%

2006-2010



MAPE: 3,62%

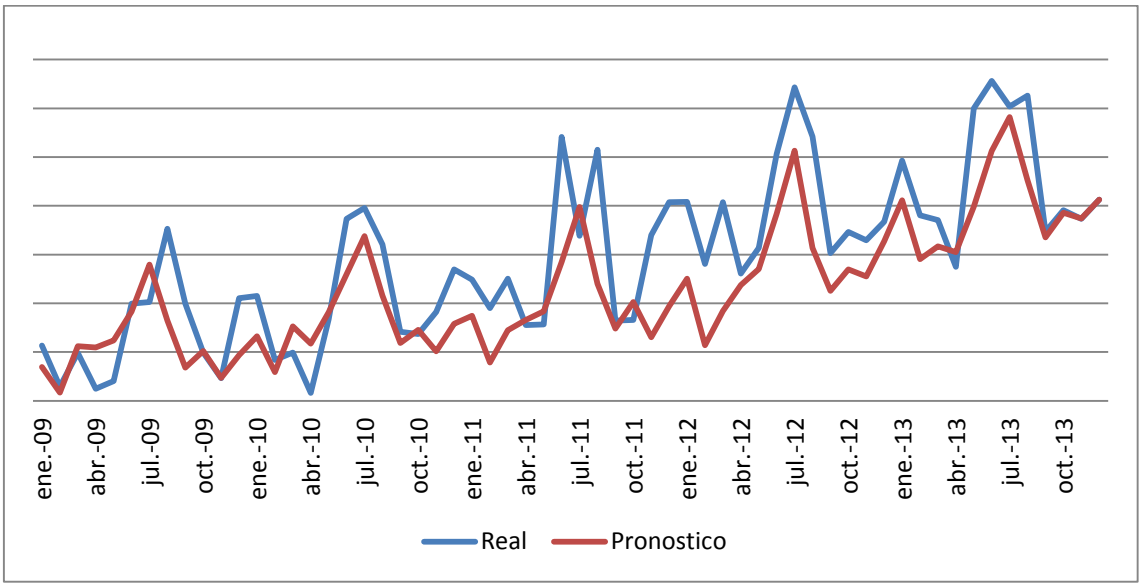
2005-2009



MAPE: 3,45%

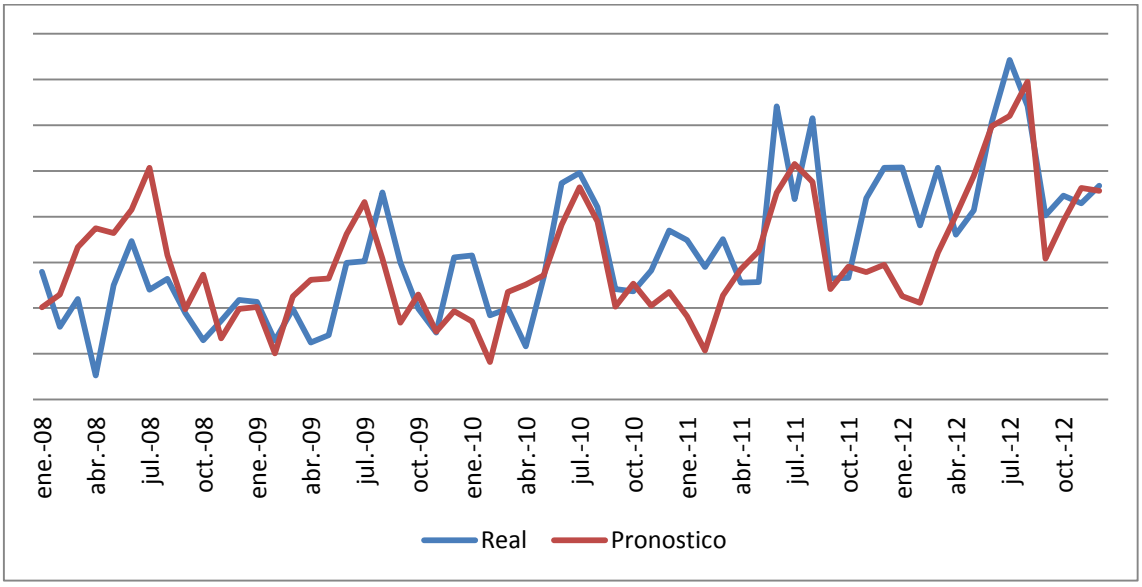
Demanda de Potencia Máxima en el Anillo

2009-2013



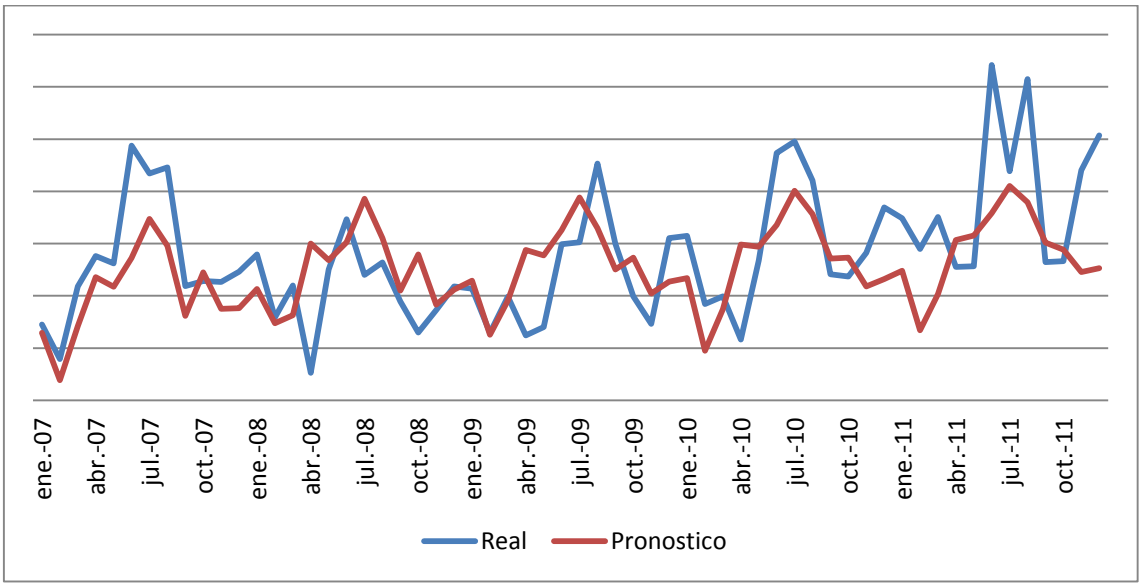
MAPE: 3,78%

2008-2012



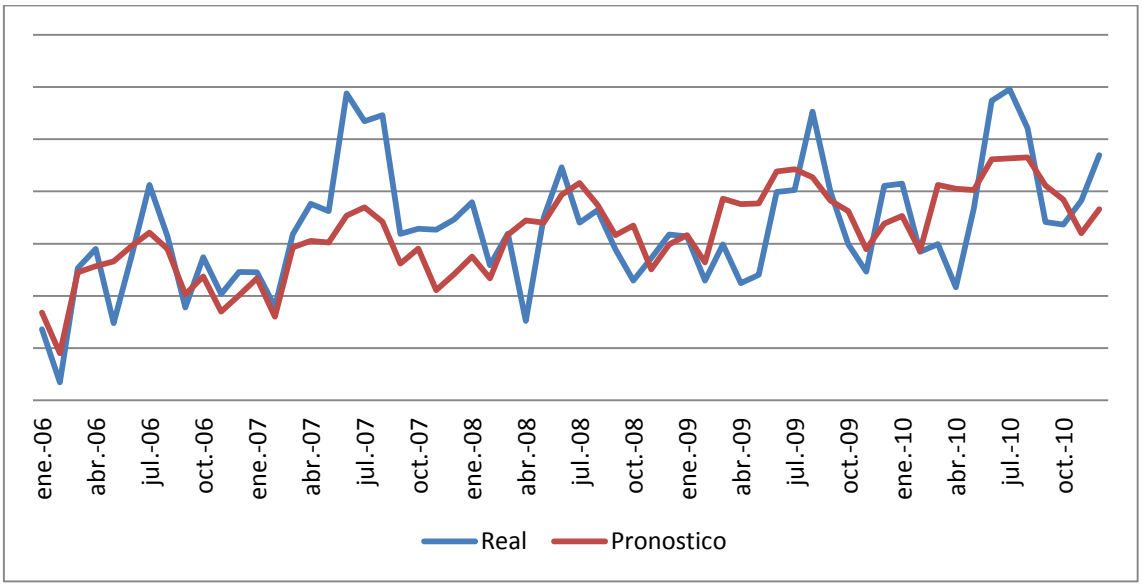
MAPE: 4,40%

2007-2011



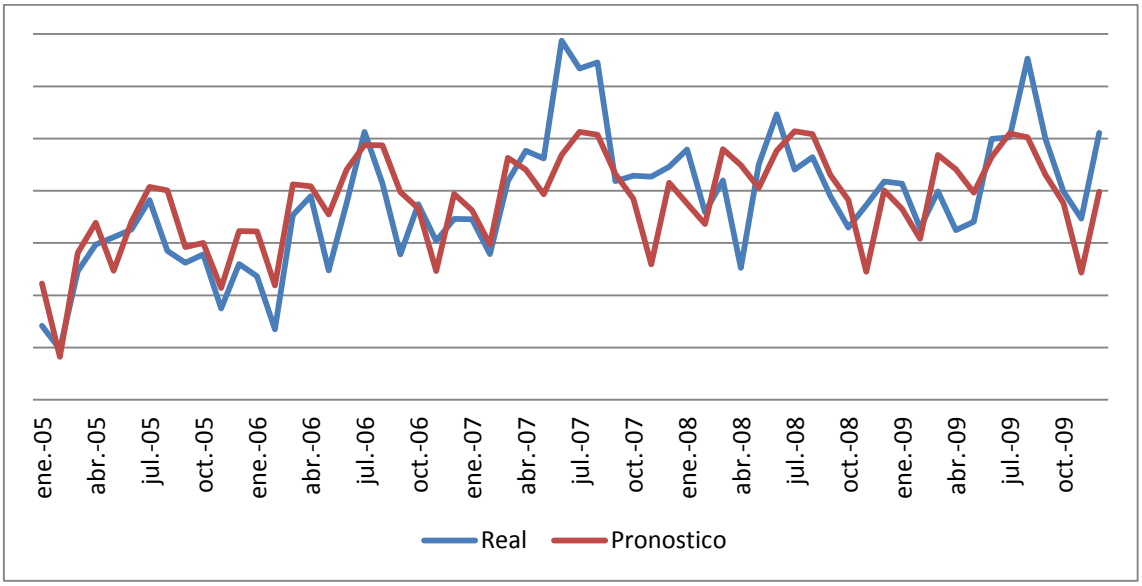
MAPE:3,98%

2006-2010



MAPE: 3,31%

2005-2009



MAPE: 3,21%

8.E Prueba de Cox-Stuart

La prueba de Cox-Stuart [10] consiste en realizar un análisis a los datos para comprobar si estos presentan una tendencia. Para esta prueba se consideran una sucesión de observaciones de una muestra aleatoria, sean estos datos $x_1, x_2, \dots, x_n$, arreglados en un orden particular, que en este caso corresponderían a su fecha de ocurrencia.

Estas muestras se agrupan en parejas (x_1, x_{1+c}), $(x_2, x_{2+c}), \dots, (x_{n-c}, x_n)$, donde c=n/2 si n es par, y c=n/2+1 si n es impar. La prueba consiste determinar la probabilidad, en estos pares ordenados, de que el valor de la derecha sea mayor al de la izquierda, así como también calcular la probabilidad de que el valor de la derecha sea menor al de la izquierda.

Luego, la hipótesis se basa en comparar estas probabilidades. De no existir una diferencia significativa, entonces no hay tendencia, en caso contrario si la hay.

Formalmente:

$$H_0: p(x_i < x_{i+c}) = p(x_i > x_{i+c}) \ \forall i$$

$$H_1: p(x_i < x_{i+c}) \neq p(x_i > x_{i+c}) \ \forall i$$

Donde H_0 implica que no existe tendencia, mientras que H_1 indica que existe una tendencia hacia abajo o hacia arriba.

8.F Prueba de Granger

El objetivo de este test en probar si una variable sirve para pronosticar a otra, y además si esta relación es unidireccional ó bidireccional.

La metodología consiste en comparar si el comportamiento actual y pasado de una serie "X" predice a otra serie "Y". Esto se evalua mediante el valor de el estadístico F. Si este tiene valor que implcan un p-valor menor al umbral establecido, entonces se establece que "X" predice a "Y".

El caso que interesa para este trabajo es cuando se tiene una relación bidireccional ("X" predice a "Y", y "Y" predice a "X") dado que representa un primer indicio de que hay endogeneidad entre ambas variables.

Esta prueba fue realizada utilizando dos variantes de las series en cuestión. La primera prueba se realizó utilizando las series originales de IMACEC y demanda de energía en el sistema, mientras que en un segundo caso se utilizaron estas series desestacionalizadas.

Los resultados se muestran a continuación:

Hipótesis Nula	Estadístico F	P-valor
ENERGIA no causa a la Granger IMACEC	33,07	2*E-12
IMACEC no causa a la Granger ENERGIA	15,53	8*E-7
ENERGIATEND no causa a la Granger IMACECTEND	414,99	0,0178
IMACECTEND no causa a la Granger ENERGIATEND	592,50	0,0034

De la tabla se puede ver que la relación entre todas estas variables es significativa a un nivel de confianza de un 98%, esto dado que la única prueba que no tiene un p-valor menor a 0,01 es si la energía desestacionalizada cuasa al IMACEC desestacionalizado.

Los resultados de estas pruebas muestran una clara relación bidireccional entre las variables, lo que corresponde a la primera prueba de endogeniedad entre éstas.

8.G Prueba de Hausman[52]

El objetivo de esta prueba es probar si existe simultaneidad entre dos variables que se incluyen en un modelo de regresión.

Para comprobar esto en el caso que imcumbe esta tesis, se planteó un primer modelo de regresión sin rezagos, considerando que esto afecta al momento de realizar la prueba. Este modelo fue estimado utilizando el software EViews 7.

A continuación se muestra en una ilustración el resultado de la priemra estimación del modelo y las variables que este incluye:

Dependent Variable: ENERGIA
Method: Least Squares
Date: 09/07/15 Time: 11:50
Sample (adjusted): 2001M06 2013M12
Included observations: 151 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.017140	0.042529	-0.403011	0.6875
LABORALIDAD	0.179692	0.021029	8.544814	0.0000
TPOS	0.171507	0.016266	10.54363	0.0000
IMACEC	0.896027	0.028955	30.94509	0.0000
HSOL	-0.112356	0.033578	-3.346077	0.0010
HREL	0.002953	0.048146	0.061328	0.9512
PRECIO	-0.052023	0.020921	-2.486689	0.0140

R-squared	0.960103	Mean dependent var	0.541767
Adjusted R-squared	0.958441	S.D. dependent var	0.199894
S.E. of regression	0.040751	Akaike info criterion	-3.517443
Sum squared resid	0.239128	Schwarz criterion	-3.377569
Log likelihood	272.5670	Hannan-Quinn criter.	-3.460619
F-statistic	577.5520	Durbin-Watson stat	1.301559
Prob(F-statistic)	0.000000		

[52] Basado en González, M. I. (2006). Cómo diagnosticar y corregir el problema de la endogeneidad: el número de hijos tenidos en la predicción de las preferencias de fecundidad en Costa Rica. Población y Salud en Mesoamérica, 4(1).

Las variables utilizadas corresponden a las mismas utilizadas en el modelo de análisis de regresión pero sin los rezagos. Se puede ver que el IMACEC es sumamente relevante en este caso al tomar en cuenta que tiene el mayor valor en su estadístico t.

La prueba a realizar consiste en estimar el IMACEC mediante un modelo de regresión que incluya las variables previamente utilizadas en el modelo de energía, aunque también puede incluir otras variables externas.

De esta tarea se obtienen 2 resultados:

- Una serie estimada de IMACEC
- Residuos de la estimación de IMACEC

Al tener estas dos series nuevas, se vuelve a estimar el modelo de energía, pero reemplazando la variable IMACEC por estas dos series.

El resultado de esta prueba se determina por la importancia que tiene la serie de los residuos en esta nueva ecuación de regresión. De ser relevante (su p.valor asociado es menor a 0,05 por ejemplo), entonces con un nivel de confianza de un 5% se puede establecer que existe simultaneidad entre las dos variables, y en caso contrario esto no se puede establecer.

Siguiendo estos pasos, se muestra a continuación los resultados de la estimación del IMACEC en base a las otras variables explicativas.

Dependent Variable: IMACEC
Method: Least Squares
Date: 09/07/15 Time: 11:50
Sample (adjusted): 2001M06 2013M12
Included observations: 151 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
DUMMIE1	0.077694	0.173303	0.448309	0.6546
DUMMIE2	0.096978	0.133513	0.726358	0.4689
DUMMIE3	0.118336	0.141741	0.834874	0.4053
DUMMIE4	0.110469	0.106676	1.035554	0.3023
DUMMIE5	0.083414	0.116457	0.716265	0.4751
DUMMIE6	0.115033	0.115391	0.996897	0.3206
DUMMIE7	0.081931	0.133767	0.612492	0.5412
DUMMIE8	0.061492	0.128918	0.476989	0.6341
DUMMIE9	0.115296	0.091398	1.261470	0.2093
DUMMIE10	0.060006	0.131924	0.454851	0.6499
DUMMIE11	0.128620	0.137065	0.938391	0.3497
DUMMIE12	0.205528	0.163644	1.255944	0.2113
PRECIO	0.566662	0.030651	18.48787	0.0000
TPOS	-0.113147	0.122643	-0.922575	0.3579
HSOL	0.014348	0.166482	0.086185	0.9314
LABORALIDAD	0.239651	0.119114	2.011949	0.0462

R-squared	0.760226	Mean dependent var	0.486021
Adjusted R-squared	0.733585	S.D. dependent var	0.224776
S.E. of regression	0.116019	Akaike info criterion	-1.370209
Sum squared resid	1.817158	Schwarz criterion	-1.050497
Log likelihood	119.4507	Hannan-Quinn criter.	-1.240325
Durbin-Watson stat	0.102059		

La gran cantidad de variables explicativas tuvo como objetivo aumentar el valor de R-cuadrado, de manera de reducir la serie de residuos (y así probablemente rechazar la simultaneidad entre las variables).

Posterior a esto se procedió a estimar nuevamente la serie de energía con las variables nuevas, obteniendo los siguientes resultados.

Dependent Variable: ENERGIA
Method: Least Squares
Date: 09/07/15 Time: 11:48
Sample (adjusted): 2001M06 2013M12
Included observations: 151 after adjustments

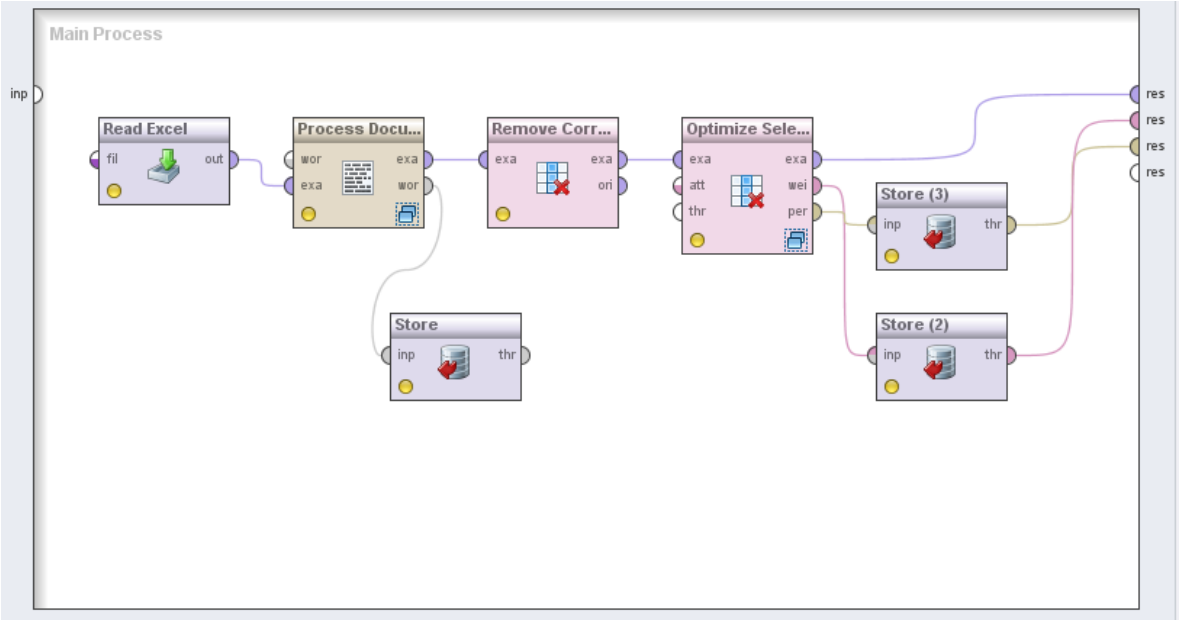
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.023570	0.037032	0.636470	0.5255
LABORALIDAD	0.293638	0.024075	12.19699	0.0000
TPOS	0.107210	0.016619	6.450978	0.0000
IMACECF	0.314678	0.084761	3.712530	0.0003
RESID05	0.951664	0.026094	36.47112	0.0000
HSOL	-0.056873	0.029910	-1.901507	0.0592
HREL	0.021920	0.041512	0.528042	0.5983
PRECIO	0.280080	0.049660	5.640004	0.0000

R-squared	0.970666	Mean dependent var	0.541767
Adjusted R-squared	0.969230	S.D. dependent var	0.199894
S.E. of regression	0.035064	Akaike info criterion	-3.811735
Sum squared resid	0.175820	Schwarz criterion	-3.651879
Log likelihood	295.7860	Hannan-Quinn criter.	-3.746793
F-statistic	675.9759	Durbin-Watson stat	1.085246
Prob(F-statistic)	0.000000		

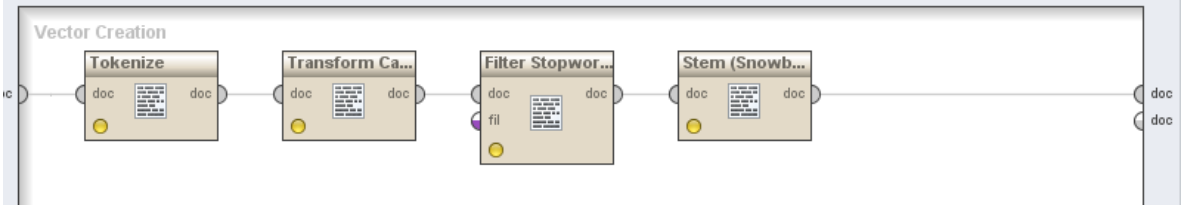
Se puede ver que la serie “RESID05” (representando los residuos de la estimación de IMACEC) es relevante, luego existe simultaneidad entre las dos variables.

8.H Configuraciones en Rapidminer

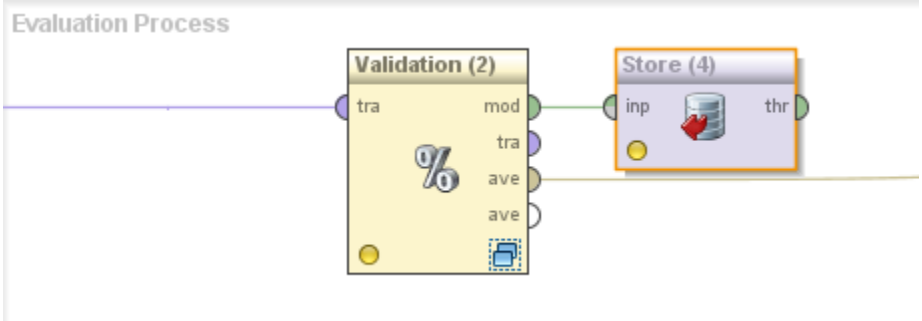
Proceso General de Entrenamiento



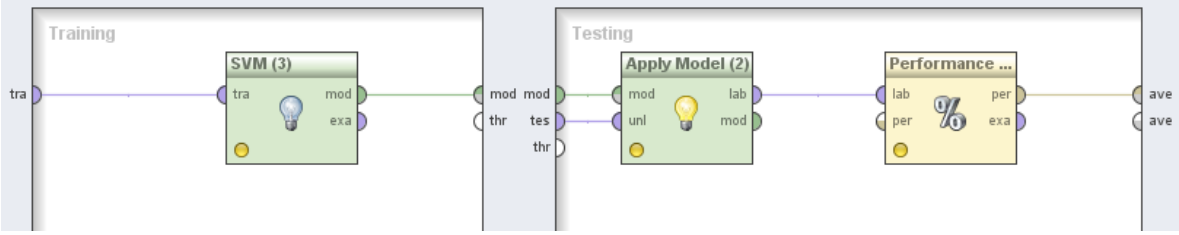
Detalle de Procesamiento de Noticias



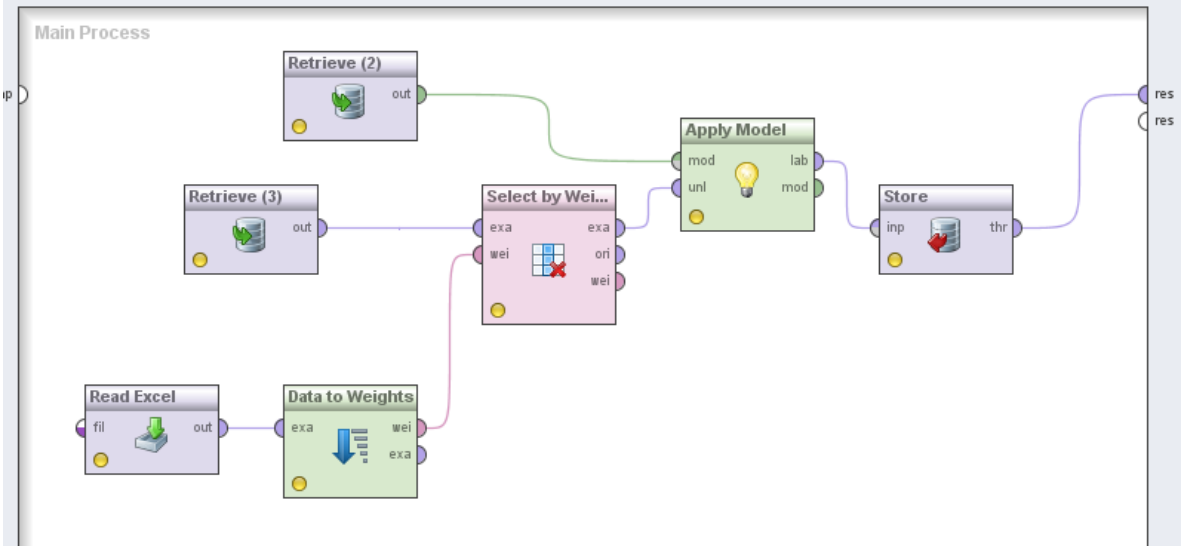
Evaluación Interna de optimización en la selección de atributos



Entrenamiento de modelo y posterior validación



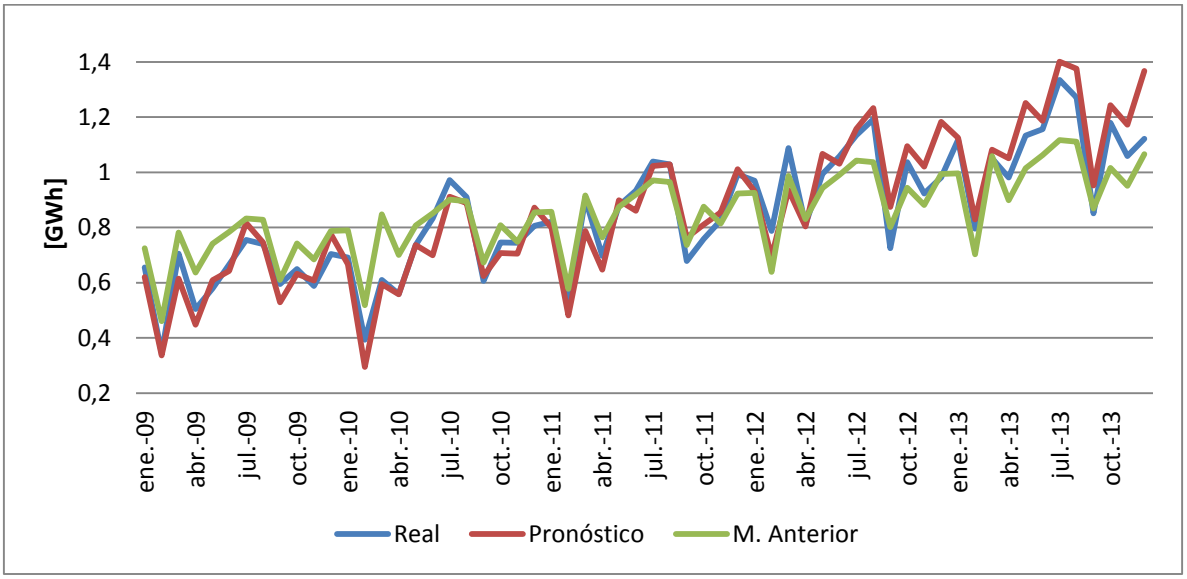
Proceso de Puesta a Prueba de Modelo Entrenado



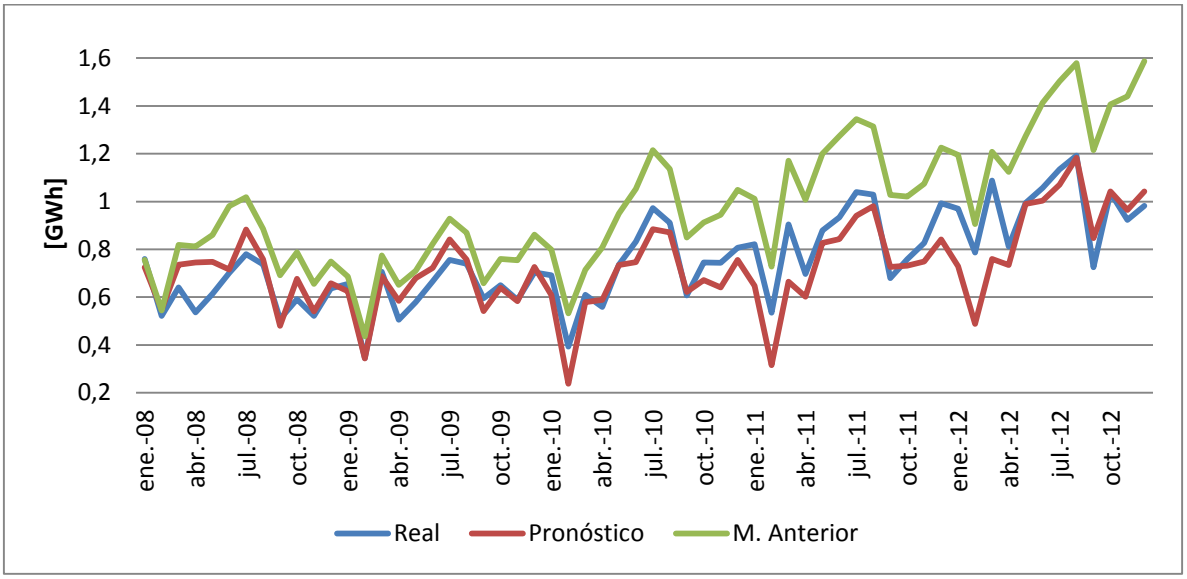
8.I Detalle Comparación de Modelos Finales

Demanda de Energía del Sistema

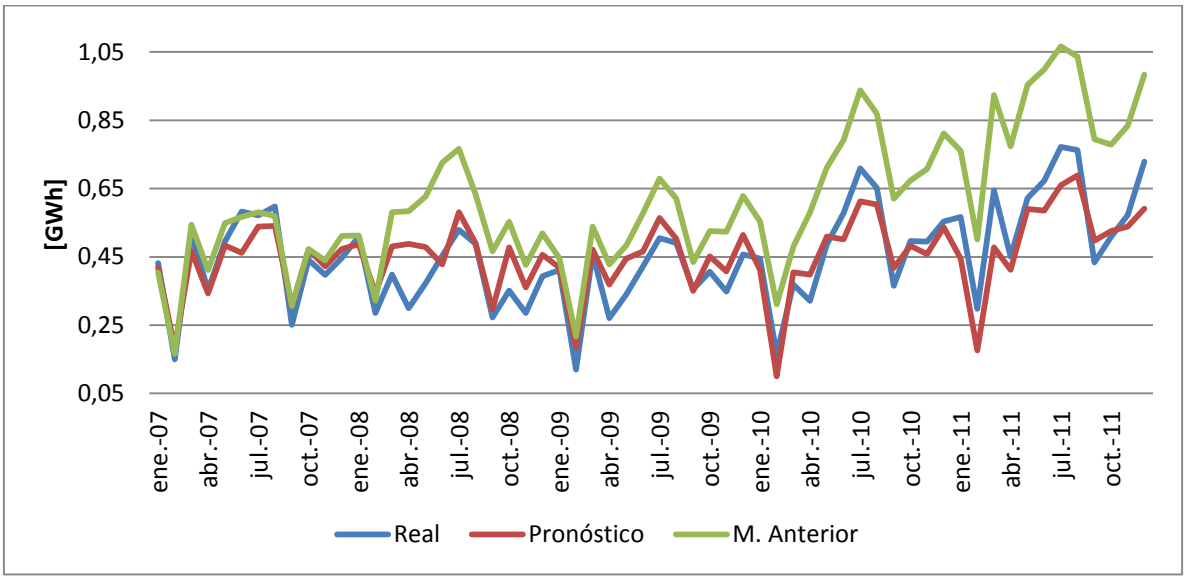
2009-2013



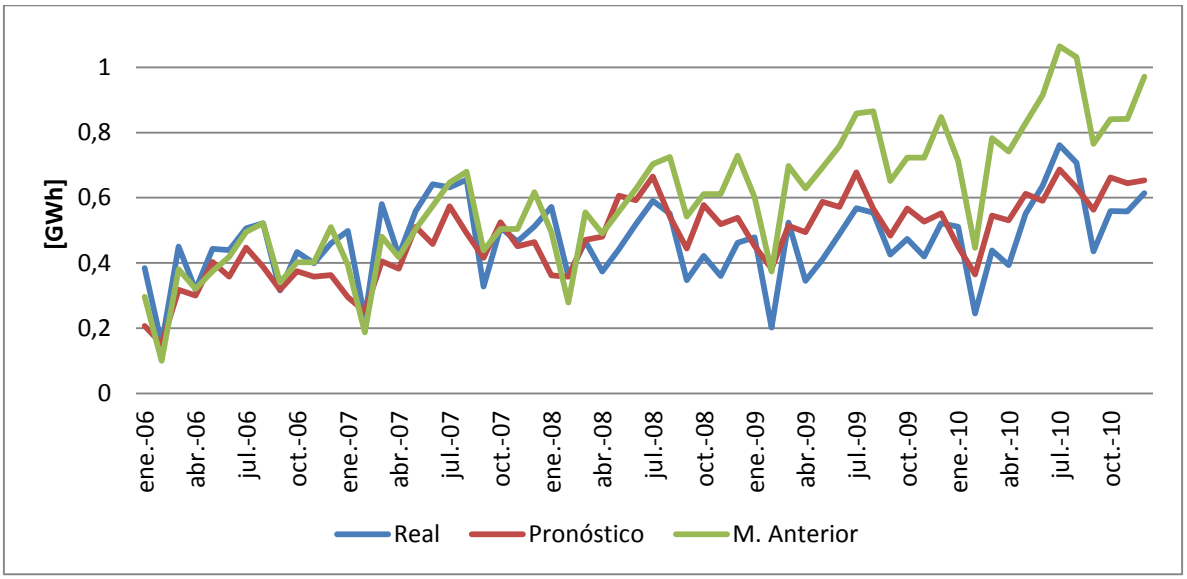
2008-2012



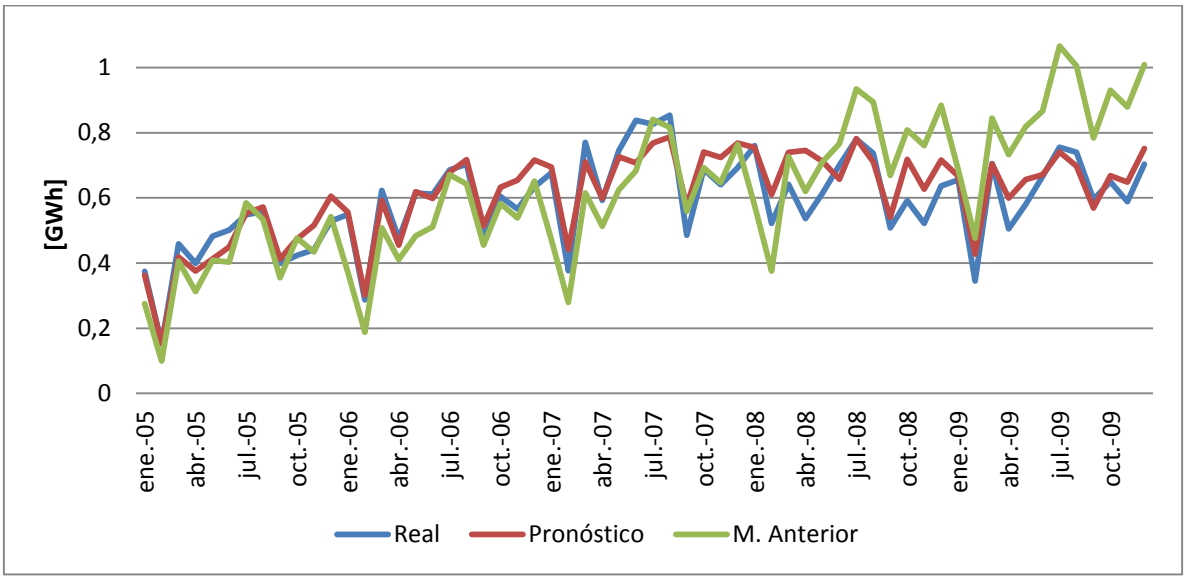
2007-2011



2006-2010

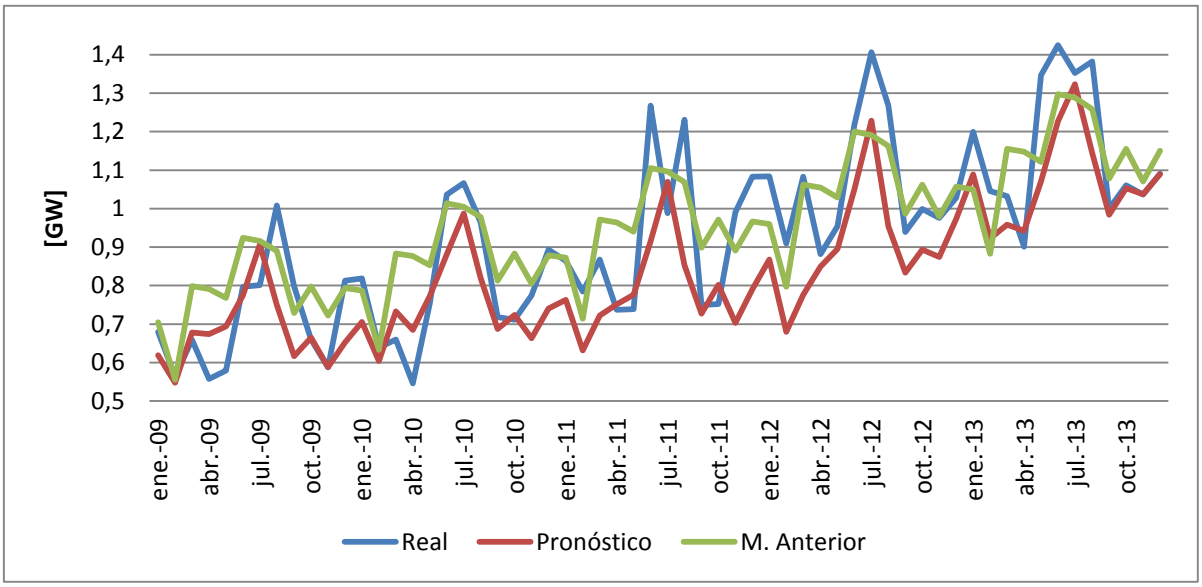


2005-2009

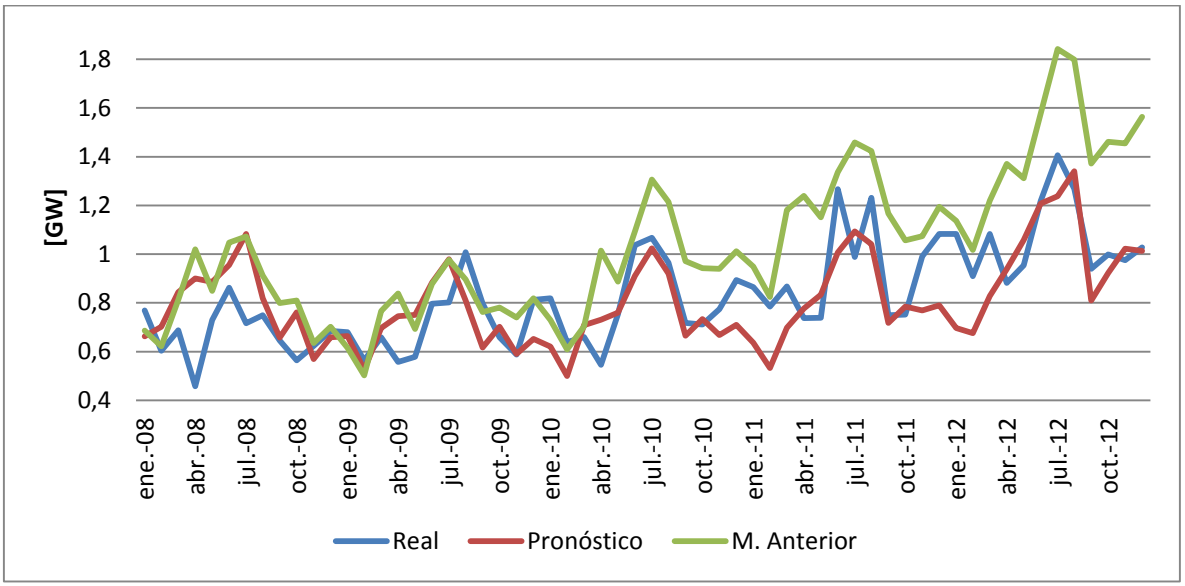


Demanda de Potencia Máxima en el Anillo

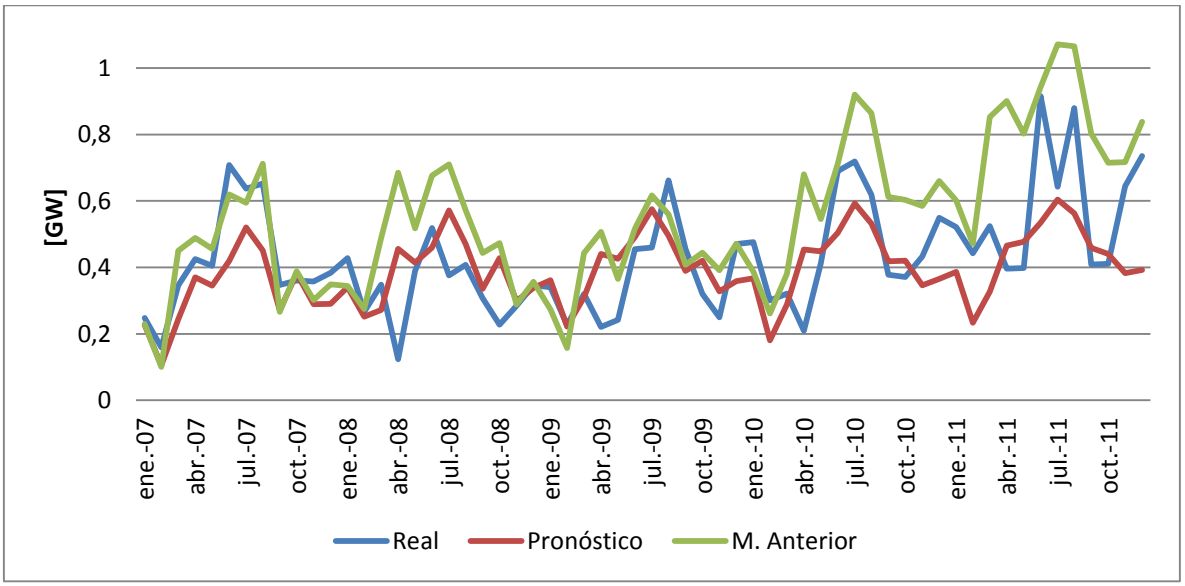
2009-2013



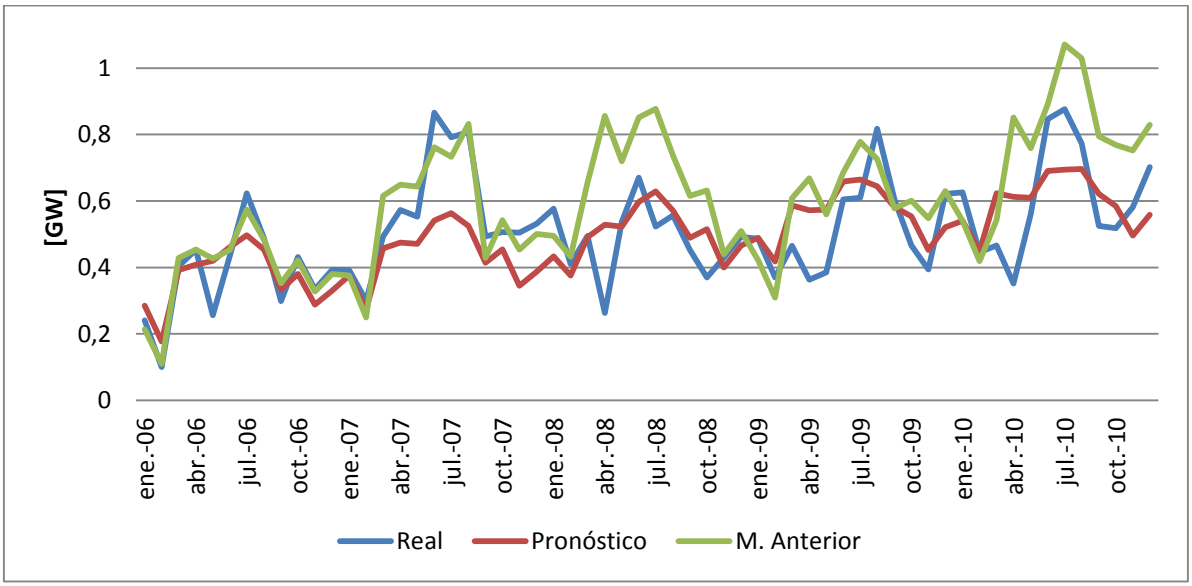
2008-2012



2007-2011



2006-2010



2005-2009