UNIVERSIDAD
DE SANTIAGO
DE CHILE

UNIVERSIDAD DE SANTIAGO DE CHILE
FACULTAD DE CIENCIA
Departamento de Matemática y Ciencia de la Computación

Modelo de Predicción de la Demanda de Gas Natural para el corto y largo plazo en la Región Metropolitana

Alejandra Lobos Márquez

Profesor Guía: Felipe Elorrieta

Trabajo de titulación presentado a la Facultad de Ciencia en cumplimiento de los requisitos para optar al título de Ingeniera Estadística

Santiago - Chile
2019

Resumen

El objetivo principal de este trabajo es predecir y modelar la demanda de gas natural, con esta información se pretende tener conocimiento lo más certero posible de la cantidad de gas que se consumirán en los siguientes días, meses y año. Para esto se utilizan técnicas de Series Temporales con el fin de pronosticar distintos elementos propios de la empresa AGESA S.A. para obtener, una visión anticipada del mercado y proponer planes de acción frente a éste.

De acuerdo a las necesidades de la empresa se ajustaron dos modelos, uno diario y el otro horario, para la demanda de gas natural en Santiago.

En cuanto a la *serie horaria*, se optó por utilizar imputación de Hot-Desk, para asignarle valores a la información perdida. Además para cada modelo se usaron variables exógenas como la temperatura, meses del año, días del año y número de clientes consumiendo. Se buscó una consolidación del ajuste de serie temporal y la influencia de variables exógenas para anticipar quiebres en el Mercado que suelen ocurrir con los cambios climáticos.

El modelo que mejora la predicción de la demanda de gas corresponde al modelo Arimax diario, el cual obtiene una deviacion máxima del 5 % respecto a los datos reales.

Y para finalizar, se aplicó a través de Shiny en R-studio, una aplicación web amigable a los operadores de AGESA S.A.

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción

Aprovisionadora Global de Energía S.A (AGESA) es una empresa dedicada a la compra y aprovisionamiento de gas natural, además de la comercialización de hidrocarburo a clientes del segmento mayorista, como las generadoras eléctricas y la exportación de gas natural. Nació en 2016 tras la separación de los negocios mayorista y minorista de Metrogas.

AGESA importa gas natural argentino a través de los gaseoductos entre Argentina y Chile, además de gas natural proveniente de distintos proveedores internacionales en formato GNL (gas natural licuado) que es almacenado en el terminal de Quintero.

La empresa AGESA requiere predecir y modelar la demanda de gas natural, con el objetivo de presupuestar y optimizar las posibilidades de compras de gas natural que se negocian con un año de anticipación.

1.1. Planteamiento del Problema

1.1.1. Enfoque de AGESA

La dificultad del proyecto radica en la necesidad predecir la demanda de gas.
Esta información es fundamental para la toma de decisiones, y generar contratos de compra y venta lo más beneficioso posible para AGESA.

Predicción de la Demanda de Gas Natural

Hay que considerar que la demanda de gas en Chile presenta comportamientos difíciles de anticipar con exactitud, considerando que depende de la temperatura, la cual es una variable en constante cambio y no existe como predecirla de forma óptima, lo que hace más difícil predecir el consumo de gas que depende de esta variable. Para solucionar esto se realiza una serie de predicciones mediante métodos estadísticos a fin de reducir el margen de error en los pronósticos de la demanda de gas.
Por lo tanto nuestra principal tarea es crear métodos de predicción lo más certeros posibles para la comercialización total de gas en la Región Metropolitana.

1.2. Objetivos

1.2.1. Objetivo General

Modelar y predecir la demanda de gas a corto y largo plazo en la Región Metropolitana.

1.2.2. Objetivos Específicos

- Imputar datos que carecen de registro.
- Crear una base de datos con variables influyentes en el consumo de gas.
- Analizar variables categóricas y optimizar la cantidad de categorías.
- Predecir y modelar el comportamiento de la demanda de gas.
- Realizar aplicación web que permita visualización sencilla a los trabajadores de AGESA.

1.3. Metodología

El primer contacto para el desarrollo de los objetivos, parte de la investigación de los datos disponibles, registros y condiciones de la empresa, de la cual se empieza a extraer información para desarrollar una base de datos que permita realizar análisis estadísticos.

Luego se aplican técnicas de imputación (Medina, (2017)[4]), para las variables con carencias de datos, una vez completada la base de datos, se procede a analizar las variables y determinar la relación de estas con la demanda de gas y las respectivas categorizaciones de las variables de tipo factor.

Para predecir la demanda de gas natural se genera un modelo ARIMAX (Andrews, Dean, Swain y Cole, (2013) [8]), el cual permite aplicar una serie de tiempo (Gras, (2001)[9]) pero que incorpora variables exógenas.
Implementado una aplicación web a través del software R-studio, a través del paquete "Shiny.[el] cual permitiría mostrar los resultados de forma sencilla a los funcionarios de AGESA.

1.4. Enfoque Estadístico

1.4.1. Predicción de la Demanda de Gas Natural

Para la predicción de la demanda se debe encontrar e identificar posibles factores estacionales dentro del comportamiento de la Demanda de Gas Natural, así como también anticipar tendencias y posibles cambios en está. De esta manera, podremos asesorar y recomendar métodos predictivos adecuados que sean lo más certeros posibles, estables y de un aceptable manejo futuro para un uso y actualización sostenida en el tiempo.

Además cabe mencionar que la dependencia con la temperatura en nuestro problema es fundamental ya que el comportamiento de la demanda se ve muy influenciado por las temperaturas y por lo tanto existe un compor-

tamiento estacional en los datos claramente definidos con un periodo anual, por ejemplo, en el consumo de gas suelen ocurrir cambios repentinos debido a una ola de bajas o altas temperatura, en donde el método para conseguir la predicción que capte este cambio en la demanda es a través de series de tiempo, que considera la tendencia climática con mayor peso a las observaciones más recientes.

La información disponibles para analizar y predecir el comportamiento de la demanda de gas natural son:

- Rescom: corresponde a la demanda residencial de gas natural.
- Temperatura: registro de temperatura por hora.
- Fecha: corresponde a toda la información referida al momento (tiempo) de registro, es decir, día de la semana, mes, año, orden de la semana, día del año que le corresponde, estación del año, etc.

Al aplicar series de tiempo (ver más en. Gras, (2001)[9]) se utiliza previamente el proceso de imputación, lo que permite tener una base de datos completa y periodos equidistantes, la información imputada por ausencia de datos debe ser concordante con los otros datos disponibles, para esto existen diversas técnicas de imputación posibles que serán aplicadas y comparadas para obtener la más óptima con el resto de la información disponible.
La presencia de datos faltantes, es la situación a la que permanentemente se enfrentan investigadores y tomadores de decisiones. En donde se debe disponer de un archivo de datos completos es ideal, pero al no tenerlo se debe aplicar métodos de imputación apropiados para lograrlo. Durante las últimas décadas se han desarrollado procedimientos que tienen mejores propiedades estadísticas que las opciones tradicionales como la eliminación de datos (listwise), el pareo de observaciones (pairwise), el método de medias y el "hot-deck". Los algoritmos de imputación múltiple (IM) se pueden aplicar utilizando paquetes comerciales y de acceso gratuito, pero imputar información no debe entenderse como un fin en sí mismo. Sus implicaciones en el análisis secundario de datos deben evaluarse con cautela, y este trabajo concluye que no existe el método de imputación ideal. Cada situación es

diferente, y la tasa de no respuesta o falta de registro y su distribución espacial cambia entre bases de datos, por lo que no es conveniente adoptar —a priori— el mismo procedimiento de imputación para todas las variables, en todas las bases de datos (Medina, (2007)[4]), (ver más en. Arce, (2019)[2]).

Con la base de datos completa, se procede a aplicar el "test de Tukey". El Test HSD (Honestly-significant-difference) de Tukey es un test de comparaciones múltiples. Permite comparar las medias de los t niveles de un factor, después de haber rechazado la Hipótesis nula de igualdad de medias mediante la técnica ANOVA. Es, por lo tanto, un test que trata de perfilar y especificar, una Hipótesis alternativa genérica como la de cualquiera de los Test ANOVA.
Se basa en la distribución del rango estudentizado que es la distribución que sigue la diferencia del máximo y del mínimo de las diferencias entre la media muestral y la media poblacional de t variables normales $N(0, 1)$ independientes e idénticamente distribuidas.
Se establece así un umbral, como en otros métodos por ejemplo; el Test LSD. Se calculan todas las diferencias de medias muestrales entre los t niveles del factor estudiado. Las diferencias que estén por encima de ese umbral se considerarán diferencias significativas, las que no lo estén se considerarán diferencias no significativas (Abdi, 2010[1]).

Esta técnica es aplicada para la relación del comportamiento de variables de tipo factor, por ejemplo; al analizar los días de semana, nos damos cuenta que no es lo mismo la demanda de un día sábado que un día miércoles, para saber todas las relaciones que existes, es necesario aplicar comparación múltiple de medias y en caso de que no existan diferencias significativas poder clasificar la variable eliminando categorías que tengan igual comportamiento, es decir que si la demanda no varía entre el día lunes, martes, miércoles, jueves y viernes; basta con tener la categoría de día hábil.

Además se requiere separar la demanda del consumo de gas natural residencia, comercial e industrial. Para esto se incorpora información de los clientes obtenida de la facturación histórica.

Para la predicción de la demanda se empleará un modelo ARIMAX ya que permite realizar series de tiempo considerando la dependencia con otras variables.
Los modelos ARIMAX incorporan información de una o más variables exógenas para explicar el comportamiento de una variable $Y_t\, t \in z$, que se comporta como una serie temporal y pueden ser vistos como un modelo de regresión lineal múltiple con uno o más términos autorregresivos y uno o más términos media móvil. En econometría los modelos ARIMAX forman parte de los denominados modelos dinámicos.

Toda esta información permitirá facilitar la toma de decisiones comerciales para el beneficio de AGESA, y manteniendo a todos los clientes con el servicio de gas continuo y permanente.

Capítulo 2

Marco Teórico

2.1. Series de Tiempo

Una serie temporal se define como una secuencia de datos medidos en determinados momentos y ordenados cronológicamente.
Para su análisis, se utilizan métodos que ayudan a estudiar su comportamiento, y a su vez, extrapolar a fin pronosticar valores de lo investigado (Brockwell, P. (2002)[11]).

Es una familia de variables aleatorias Y_t donde t denota el tiempo, tales que para cualquier valor de $t_1, t_2, ..., t_n$ existe la distribución de probabilidad conjunta correspondiente a las variables aleatorias $Y_{t1}, Y_{t2}, ..., Y_{tn}$.

Las series temporales se pueden clasificar en discretas, si es observada en instantes aislados en el tiempo, y en continuas, si ésta es observada continuamente en el tiempo.

2.1.1. Componentes Básicos de una Serie de Tiempo

Una serie de tiempo se puede denotar como:

$$Y_t = T_t + S_t + \epsilon_t$$

Donde T_t es la tendencia, S_t es la componente estacional y ϵ_t es la componente aleatoria.

- **Componente de Tendencia:** Se define como un cambio a largo plazo que se produce en la relación al nivel medio, o el cambio a largo plazo de la media. Se identifica con un movimiento suave de la serie a largo plazo.
- **Componente de Estacionalidad:** Corresponde a un comportamiento periódico de la serie. Son oscilaciones que se produce alrededor de la tendencia, en forma repetitiva.
- **Componente Aleatorio:** Esta componente no responde a ningún patrón de comportamiento si no que es el resultado de factores fortuitos o aleatorios que inciden de forma aislada en una serie de tiempo.

2.1.2. Clasificación de Series de Tiempo

En primera instancia las series de tiempo se pueden clasificar en series **estacionarias** y **no estacionarias**.

Una serie **estacionaria** se caracteriza por presentar un comportamiento relativamente estable a lo largo del tiempo con una media y varianza constante. Gráficamente los valores de la serie oscilan alrededor de una media que es invariable, es decir, los datos observados no se incrementan o decrecen con el pasar del tiempo y la variabilidad con respecto a esa media también permanece constante, dicho de otra forma el **ruido** de la serie no se incrementa.

Por otro lado una serie **no estacionaria** presenta tendencia o varianza inestable en el tiempo. Los cambios en la media determinan una tendencia a crecer o decrecer a largo plazo, por lo que la serie no oscila alrededor de un valor constante y la inestabilidad de la varianza se refleja en una oscilación creciente a lo largo del tiempo.

2.1.3. Proceso Estocástico y Estacionariedad

Para poder analizar la estacionariedad de una serie de tiempo primero se definirá un concepto base, los procesos estocásticos. (Brockwell, P. (2002). [11]).

Sea T un conjunto arbitrario, un proceso estocástico es una familia $Z = Z(t), t \in T$ tal que para cada $t \in T, Z(t)$ es una variable aleatoria.

Desde un punto de vista intuitivo, un proceso estocástico se describe como una secuencia de datos que evolucionan en el tiempo. Las series temporales se definen como un caso particular de los procesos estocásticos.

Un proceso estocástico se dice que es estacionario si su media y su varianza son constantes en el tiempo y si el valor de la covarianza entre dos periodos depende solamente de la distancia o rezago entre estos dos periodos de tiempo y no del tiempo en el cual se ha calculado la covarianza.

Proceso débilmente estacionario

Una serie de tiempo Y_t se dice débilmente estacionaria o estacionaria de segundo orden, si satisface las siguientes condiciones:

- $E(Y_t) = \mu_t \quad \forall\, t \in\ T$ Es decir, la media del proceso Y_t es constante.
- $V(Y_t) = \sigma_t^2 < \infty \quad \forall\, t \in\ T$ Es decir, la variabilidad del proceso es constante y finita.
- $Cov(Y_t, Y_{tk}) = \gamma_k$ Es decir, la covarianza no depende de t.

Proceso estrictamente estacionario

Un proceso es estrictamente estacionario si la distribución conjunta de $(Z_t, ..., Z_k)$ y $(Z_{t+h}, ..., Z_{k+h})$ es la misma para todo $k \geq 0$ y $h \in\ Z$.

2.1.4. Ruido Blanco

Un **ruido blanco** es un caso simple de los procesos estocásticos, donde los valores son independientes e idénticamente distribuidos a lo largo del tiempo, con media cero e igual varianza.

Un ruido blanco debe cumplir las siguientes condiciones:

- $E(Y_t) = 0$
- $V(Y_t) = \sigma_t^2$
- $Cov(Y_t, Y_{t-k}) = 0$

2.1.5. Función de Autocovarianza

La función de Autocovarianza de una serie de tiempo Y_t de orden k, denotada por γ_k, es la Covarianza de la serie separada en k periodos, es decir:

$$\gamma_k = Cov(Y_t, Y_{t-k}) = E[(Y_t - \mu)(Y_{t-k} - \mu)]$$

Donde $\mu = E(Y_t) = E(Y_{t-k})$.

El conjunto de valores de γ_k para $k \geq 0$ configuran la Función de Autocovarianza.

La cual satisface las siguientes propiedades:

- $\gamma_0 = V(Y_t)$
- $\gamma_k = \gamma_{-k}$ La función es par.
- $|\gamma_k| \leq \gamma_0$ para $k = 1, 2, ...$

2.1.6. Función de Autocorrelación

La autocorrelación de orden k para un periodo Y_t, denotada por ρ_k, es la correlación de la serie separada en k periodos [11], es decir:

$$\rho_k = Corr(Y_t, Y_{t-k}) = \frac{Cov(T_t, Y_{t-k})}{\sqrt{V(Y_t)V(Y_{t-k})}} \tag{2.1}$$

Y si el proceso es estacionario, entonces la autocorrelación de orden k es:

$$\rho_k = \frac{\gamma_k}{\gamma_0} \quad \forall k \geq 0 \tag{2.2}$$

El gráfico de los valores de autocorrelación se llama correlograma.

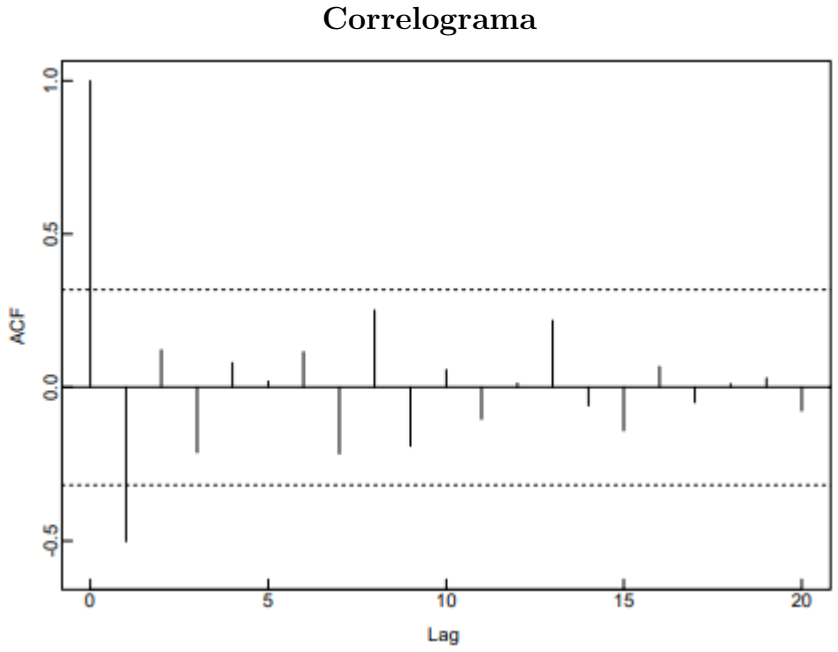


Figura 2.1: Fuente: Brockwell (2002). Introduction Time Series and Forecasting [11].

El correlograma de una serie puede ayudar a determinar si la serie es estacionaria de segundo orden:

- Si los valores del gráfico decaen rápidamente a cero se dice que el proceso es **estacionario**.

- Si los valores decaen lentamente a cero, el proceso es **no estacionario**.

2.1.7. Función de Correlación Cruzada

Considere dos procesos conjuntamente estacionarios x_t e y_t para $t = 0, \pm 1, \pm 2, ...$ La covarianza cruzada de orden k entre x_t e y_t está definida como (Castaño, E. (1998) [12]):

$$\gamma_{xy}(k) = E[(X_t - \mu_t)(y_{t+k} - \mu_y)] \quad \forall k = 0, \pm 1, \pm 2, ... \tag{2.3}$$

Como función de k, $\gamma_{xy}(k)$ es llamada la **Función de Covarianza Cruzada** entre x_t e y_t.

La estandarización de $\gamma_{xy}(k)$ produce la **Función de Correlación Cruzada** (CCF) representada matemáticamente por la siguiente expresión:

$$\rho_{xy}(k) = \frac{\gamma_{xy}(k)}{\sigma_x \sigma_y} \quad \forall k = 0, \pm 1, \pm 2, ... \tag{2.4}$$

donde, σ_x y σ_y son las desviaciones estándar de los procesos x_t e y_t.

La **FCC** no sólo mide la fortaleza en la relación lineal de dos procesos sino que también su dirección. Esta propiedad puede ser de gran utilidad para identificar causalidad entre variables, por lo que es una función que se debe estudiar tanto para valores positivos como negativos de k. Para valores negativos de k, la **FCC** describe la influencia lineal de los valores pasados de y_t sobre x_t. Para valores positivos de k, la **FCC** indica la influencia lineal de los valores pasados de x_t sobre y_t.

El gráfico de la **FCC** contra k, llamado **Correlograma Cruzado** (Figura 6.2), es útil para visualizar estas relaciones.

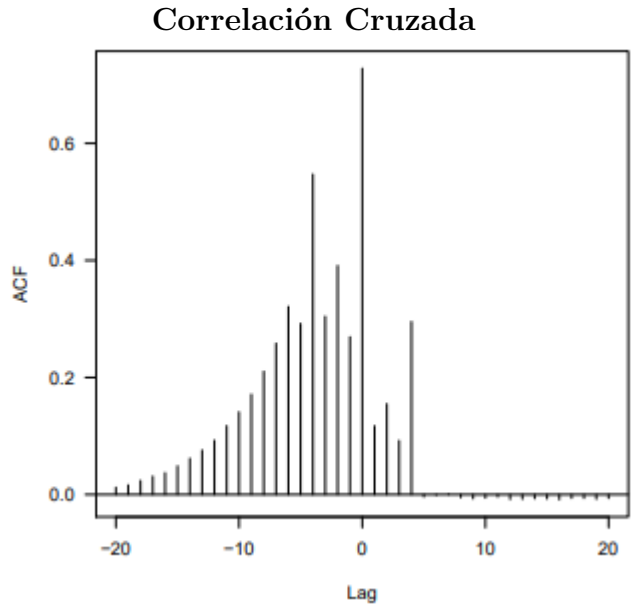


Figura 2.2: Fuente: Castaño, E. y Martínez, J. (1998), Uso de la función de correlación cruzada en la identificación de modelos ARMA.[12].

Dada una realización de n periodos del proceso estacionario bivariante x_t, y_t, la **FCC** es estimada con la función de correlación cruzada muestral (**FCCM**):

$$\hat{\rho}_{xy}(k) = \frac{\hat{\gamma}_{xy}(k)}{\hat{\sigma}_x \hat{\sigma}_y} \quad \forall k = 0, \pm 1, \pm 2, ... \tag{2.5}$$

donde:

$$\hat{\gamma}_{xy}(k) = \begin{cases} \frac{1}{n} \sum_{t=1}^{n-k} (x_t - \overline{x})(y_{t+k} - \overline{y}) & , \text{si } k \geq 0 \\ \frac{1}{n} \sum_{t=1-k}^{n} (x_t - \overline{x})(y_{t+k} - \overline{y}) & , \text{si } k < 0 \end{cases} \tag{2.6}$$

con $\hat{\sigma}_x = [\hat{\gamma}_{xx}(0)]^{\frac{1}{2}}$, $\hat{\sigma}_y = [\hat{\gamma}_{yy}(0)]^{\frac{1}{2}}$, $\overline{x}$ e $\overline{y}$ son las desviaciones estándar y medias muestrales de las series x_t e y_t respectivamente.

Con los supuestos de normalidad, es decir, que la serie x_t es **ruido blanco** y que las series x_t e y_t son incorrelacionadas, Bartlett (1985) probó que:

$$Var[\hat{\rho}_{xy}(k)] \approx (n-k)^{-1} \tag{2.7}$$

Por lo tanto, cuando la serie x_t es **ruido blanco** y hay normalidad, podemos contrastar la hipótesis de que las dos series tienen **correlación cruzada** nula comparando $\hat{\rho}_{xy}(k)$ con su error estándar aproximado $\sqrt{(n-k)^{-1}}$.

2.1.8. Modelo AR

Sea un proceso estocástico X_t con $t \in T$, y $T \in \mathbb{Z}$ Se dice autoregresivo de orden p, (AR(p)) (Brockwell, P. (2002) [11]) si:

- $X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + ... + \phi_p X_{t-p} + \epsilon_t \quad t \in T,\ p \geq 1$ Donde $\epsilon_t \sim RB(0, \sigma^2)$
- $Cov(\epsilon_t, X_{t-j}) = E(\epsilon_t X_{t-j}) = 0$
- $\phi_1, ..., \phi_p$ son coeficientes fijos.

Equivalentemente se define el polinomio:

$$\Phi(B) = 1 - \phi_1 B - \phi_2 B^2 - ... - \phi_p B^p \tag{2.8}$$

$$\Phi(B) X_t = \epsilon_t \tag{2.9}$$

De esta manera el proceso es invertible, ya que $\Phi(B)$ es finito. Para que un modelo AR sea estacionario se debe cumplir $|\phi| < 1$ que quieres decir que todas las raíces están fuera del circulo unitario.

2.1.9. Modelo MA

Sea ϵ_t una sucesión de **ruido blanco**, $E(\epsilon_t) = 0$ y $V(\epsilon_t) = \sigma^2$ consideremos el proceso X_t (Brockwell, P. (2002) [11]) definido por: $X_t = \epsilon_t - \theta_1 \epsilon_{t-1} - ... - \theta_q \epsilon_{tq}$. Este modelo se conoce como modelo de medias móviles, desde ahora MA(q).
Otra forma de escribir el proceso:

$$X_t = \epsilon - \theta_1 B \epsilon_t - ... - \theta_q B^q \epsilon_t = (1 - \theta_1 B - ... - \theta_q B^q)\epsilon_t \tag{2.10}$$

$$X_t = \theta(B)\epsilon_t \tag{2.11}$$

Como $\theta(B)$ es finita, se puede decir que el proceso es estacionario. Para que el proceso sea invertible, se debe cumplir que todas las raíces del polinomio deben estar fuera del circulo unitario $|\theta| < 1$.

2.1.10. Modelo ARMA

Sea X_t un proceso definido por la ecuación (Brockwell, P. (2002) [11]):

$$X_t - \phi_1 X_{t-1} - ... - \phi_p X_{t-p} = \epsilon_t - \theta_1 \epsilon_{t-1} - ... - \theta_q \epsilon_{t-q} \tag{2.12}$$

El cual es una combinación de los procesos AR(p) y MA(q).

Este modelo mixto se denota como $X_t \sim$ ARMA(p,q) proceso autoregresivo de medias móviles, donde:

- $\epsilon_t \sim RB(0, \sigma^2)$
- $\phi_1, ..., \phi_p,\ \theta_1, ..., \theta_q$ son coeficientes fijos.

Otra forma equivalente de escribir el modelo ARMA(p,q) es:

$$\Phi(B)X_t = \Theta(B)\epsilon_t \tag{2.13}$$

X_t será causal e invertible, si las raíces de $\Phi(Z) = 0$ y $\Theta(Z) = 0$ están fuera del circulo unitario.

Suponiendo que se cumplen estas condiciones, se podrán escribir los procesos MA(∞) y AR(∞).

2.1.11. Representaciones MA(∞) y AR(∞)

Representación MA(∞) de Wold

La descomposición en MA(∞) o representación de Wold esta definida por:

$$X_t = \sum_{j=0}^{+\infty} \Psi_j \epsilon_{t-j} \tag{2.14}$$

$$X_t = \Psi(B)\epsilon_t \tag{2.15}$$

Sea el proceso ARMA(p, q) $\Psi(B)X_t = \Theta(B)\epsilon_t$

$$X_t = \frac{\Theta(B)}{\Phi(B)}\epsilon_t \Rightarrow \Psi(B) = \frac{\Theta(B)}{\Phi(B)} \tag{2.16}$$

Expansión AR(∞)

Por otro lado se puede encontrar la representación de AR(∞) que esta definida por:

$$\sum_{j=0}^{+\infty} \Pi_j X_{t-j} = \epsilon_t \tag{2.17}$$

$\Pi(B)X_t = \epsilon_t$ Sea el proceso ARMA(p, q)

$$\Phi(B)X_t = \Theta(B)\epsilon_t \tag{2.18}$$

$$\frac{\Phi(B)X_t}{\Theta(B)} = \epsilon_t \quad \Rightarrow \quad \Pi(B) = \frac{\Phi(B)}{\Theta(B)} \tag{2.19}$$

2.1.12. Teorema de Proyección

Espacio de Hilbert

Un espacio de Hilbert es un espacio con producto interno completo, donde un producto interno completo, cumple con:

$$< X, X >= ||X||^2 \quad \Rightarrow \quad ||X|| = \sqrt{< X, X >} \tag{2.20}$$

Este espacio se dice completo si toda la sucesión de Cauchy tiene límite en el mismo espacio.

Sucesión de Cauchy

X_n es una sucesión de Cauchy si

$$||X_n - X_m||_{n,m\to\infty} \to 0 \tag{2.21}$$

La definición de los espacios de Hilbert se hace necesaria para tratar los casos de pasado infinito.

Por otra parte, para un teorema de proyección, es necesario tener un espacio con producto interno, el cual define la geometría del espacio.

Sea $X_1, ..., X_n$ un proceso estacionario en sentido débil. Interesa predecir X_{n+1}.

Para resolver esta pregunta se debe definir un teorema de proyección.

Teorema de Proyección

Sea $\mathbb{H}$ un espacio de Hilbert y sea M un subespacio cerrado de Hilbert.

Sea $X \in \mathbb{H}$ existe un único elemento $\hat{X} \in M$ tal que:

$$||X - \hat{X}|| \leq ||X - Y|| \quad \forall\, Y \in M \tag{2.22}$$

Además, $< X - \hat{X}, Y >= 0 \quad \forall\, Y \in M$

Mejor Predictor Lineal

De acuerdo con el teorema de proyección, existe un único elemento $\hat{X}_{n+1} \in M$ tal que:

$$< X_{n+1} - \hat{X}_{n+1}, X_t >= 0 \quad \forall\, t = 1, ..., T \tag{2.23}$$

Sea $M = S_p\{X_1, ..., X_T\}$ y el espacio del Hilbert es L^2.

$\Rightarrow$ Como M es un subespacio generado y $\hat{X}_{n+1} \in M$

$$\hat{X}_{n+1} = \alpha_1 X_1 + ... + \alpha_n X_n \tag{2.24}$$

El espacio $M \in L^2$, entonces $< X, Y >= E(XY) \Leftrightarrow E(X) = E(Y) = 0$.

En el caso general se busca predecir $\hat{X}_{n+1}$ en base a $M = \{X_n, X_{n+1}, X_{n+2}, ...\}$.

El valor de $\hat{X}_{n+1}$ corresponde a:

$$\hat{X}_{n+1} = E(X_{n+1}|X_1, ..., X_n) \tag{2.25}$$

También se puede denotar como:

$$E_M(X) = E(X|M) = P_M X \tag{2.26}$$

2.1.13. Estimación para Modelos ARMA

Ecuación de Yule-Walker

Las ecuaciones de Yule-Walker, también conocidas como ecuaciones de los momentos, consisten en aplicar un esquema para estimar los parámetros de un proceso.

Sea $X_t \sim$ ARMA(p, q)

$$(B)X_t = \Theta(B)\epsilon_t \ / * X_{tj} \ / * E() \tag{2.27}$$

El resultado permite escribir los parámetros del modelo, en función de las autocorrelaciones, que por lo general son conocidas.

Función de Autocorrelación Parcial

La autocorrelación parcial mide el exceso de correlación debida a X_{t-k} y se obtiene como la correlación que existe entre X_t y X_{t-k} después de eliminar el efecto de todas las variables aleatorias que están entre ellas, se denota por φ_{kk} y esta dada por:

$$\varphi_{kk} = Corr(X_t, X_{t-k}|X_{t-1}, X_{t-2}, ..., X_{t-k+1}) \tag{2.28}$$

$$\varphi_{kk} = \frac{Cov(X_t - \hat{X}_t, X_{t-k} - \hat{X}_{j-k})}{\sqrt{V(X_t - \hat{X}_t)}\sqrt{V(X_{t-k} - \hat{X}_{j-k})}} \tag{2.29}$$

A continuación veremos dos métodos para obtener la **Función de Autocorrelación Parcial** (FACP):

- **Algoritmo Durbin-Levinson**
  Los coeficientes $\phi_{n1}, ..., \phi_{nn}$ se pueden calcular de forma recursiva con el algoritmo de Durbin-Levinson:

$$\phi_{nn} = \{\gamma(n) - \sum_{j=1}^{n-1} \phi_{n-1}, j\gamma_{1-j}\}V_{n-1}^{-1} \tag{2.30}$$

$$\begin{bmatrix} \phi_{n1} \\ \vdots \\ \phi_{nn} \end{bmatrix} = \begin{bmatrix} \phi_{n-1,1} \\ \vdots \\ \phi_{n-1,n-1} \end{bmatrix} - \phi_{nn} \begin{bmatrix} \phi_{n-1,1} \\ \vdots \\ \phi_{n-1,n-1} \end{bmatrix} \tag{2.31}$$

$$V_n = V_{n-1}(1 - \phi_{nn}^2) \tag{2.32}$$

donde, $\phi_{11} = \frac{\gamma(1)}{\gamma(0)} = \rho(1)$

$$V_0 = \gamma(0) \tag{2.33}$$

$$\hat{X}_{n-1} = \sum_{j=1}^{n} \phi_{nj} X_{n+1-j} = \phi_{n1} X_n + \phi_{n2} X_{n-1} + ... + \phi_{nn} X_1 \tag{2.34}$$

- **Algoritmo de Innovaciones**

  Los coeficientes $\theta_{n1}, ..., \theta_{nn}$ se pueden calcular de forma recursiva con el algoritmo de innovaciones:

$$V_0 = \Gamma(1, 1) \tag{2.35}$$

$$\theta_{n,n-k} = V_n^{-1}(\Gamma(n+1, k+1) - \sum_{j=0}^{k-1} \theta_{k,k-j}\theta_{n,n-j}V_j) \tag{2.36}$$

$$\text{Con } V_n = \Gamma(n+1, n+1) - \sum_{j=0}^{n-1} \theta_{n,n-j}^2 V_j \tag{2.37}$$

$$\hat{X}_{n+1} = \sum_{k=1}^{n} \theta_{nk}(X_{n+1-k} - \hat{X}_{n+1-k}) \tag{2.38}$$

2.1.14. Identificación de Modelos

Para poder identificar los modelos de la series de tiempo, se debe observar la **Función de Autocorrelación** (FAC) y la **Función de Autocorrelación Parcial** (FACP).

2.1.15. Diferenciación

Cuando una serie tiene tendencia es un proceso **no estacionario**, para poder solucionar ese problema se ocupa el operador de diferencia:

$$\bigtriangledown^d = (1 - B)^d \tag{2.39}$$

Modelo	FAC	FACP
$AR_{(p)}$	Infinito. Decrecimiento como mezcla de exponenciales y senoidales	Finito. p primeros coecientes no nulos, el resto cero
$MA_{(q)}$	Finito. q primeros coecientes no nulos, el resto cero	Infinito. Decrecimiento como mezcla de exponenciales y senoidales
$ARMA_{(p,q)}$	Infinito. Decrecimiento hacia cero desde q	Infinito. Decrecimiento hacia cero desde p

Cuadro 2.1: Fuente: Brockwell (2002). Introduction Time Series and Forecasting [11]

$$\text{tal que} \quad Y_t = (1-B)^d X_t \tag{2.40}$$

Se dice que el proceso Y_t es de orden d, si se tuvo que diferenciar la serie d veces para que fuese un proceso estacionario.

Para saber si se debe diferenciar o no, se debe estudiar la variabilidad de la serie original y la serie diferenciada. En general, no debería observarse un aumento sustantivo en la varianza de la serie diferenciada.

2.1.16. Modelo ARIMA

Sea $Y_t = (1-B)^d X_t$ y suponga que $Y_t \sim$ ARMA(p, q) es decir (Brockwell, P. (2002) [11]):

$$\Phi(B)X_t = \Theta(B)\epsilon_t \tag{2.41}$$

Entonces podemos reescribir el proceso como:

$$\Phi(B)(1-B)^d X_t = \Theta(B)\epsilon_t \tag{2.42}$$

Este proceso se conoce como Autoregresivo Integrado de Medias Móviles y se denota como $X_t \sim$ ARIMA(p, d, q) donde p corresponde al orden de la parte autorregresiva estacionaria, d es él número de veces que se diferencia la serie

y q es el orden de la parte de medias móviles.

Donde:

- $\Phi(B) = 1 - \phi_1 B - \phi_2 B^2 - ... - \phi_p B^p$
- $\Theta(B) = 1 - \theta_1 B - \theta_2 B^2 - ... - \theta_q B^q$

2.1.17. Modelo SARIMA

El modelo **SARIMA**, también llamado **ARIMA Estacional**, permite la aleatoriedad en la forma estacional de ciclo a ciclo. Este modelo se escribe de la siguiente forma (Brockwell, P. (2002) [11]):

$$\Phi(B)\Phi_P(B^s)(1-B)^d(1-B^s)X_t = \Theta(B)\Theta_Q(B^s)\epsilon_t \tag{2.43}$$

Donde:

- $\Phi(B) = 1 - \phi_1 B - \phi_2 B^2 - ... - \phi_p B^p$
- $\Theta(B) = 1 - \theta_1 B - \theta_2 B^2 - ... - \theta_q B^q$
- $\Phi_P(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - ... - \Phi_P B^P s$
- $\Theta_Q(B^s) = 1 - \Theta_1 B^s - \Theta_2 B^{2s} - ... - \Theta_q B^Q s$

2.1.18. Estrategia de Predicción

1. Obtener la mayor cantidad de información posible acerca del fenómeno a analizar.
2. Graficar datos originales, fijándose en:
   - Tendencias.
   - Variaciones estacionales.
   - Cambios estructurados.
   - Varianza (estable).

3. "Limpiar datos", eliminar datos extremos. Transformación de datos originales.

4. Determinar si la serie posee variación estacional, ya sea aditiva, multiplicativa u otra. Además revisar su tendencia.

5. Ajustar modelo.

6. Estudiar si el modelo es el adecuado, si no es así volver al punto 4 y 5.

7. Calcular predicciones.

2.1.19. Análisis Espectral

El análisis espectral es una herramienta que permite analizar las características cíclicas de los datos. Su objetivo es determinar las frecuencias que influyen mayoritariamente en la variabilidad de la Serie de Tiempo.

Densidad Espectral

Sea X_t un proceso estocástico estacionario, donde:

- $E(X_t)$: Esperanza del proceso.
- $\gamma(k)$: Función de autocovarianza del proceso.

Se define la función generadora de autocovarianzas como:

$$\Gamma(Z) = \sum_{k=-\infty}^{+\infty} \gamma(k) Z^k \qquad \frac{1}{r} < |Z| < r \tag{2.44}$$

$$\text{Y donde} \quad \sum_{k=-\infty}^{+\infty} |\gamma(k)| < \infty \tag{2.45}$$

Luego, la densidad espectral se define como:

$$f(\lambda) = \frac{1}{2\pi}\Gamma(e^{-ik\lambda}) = \frac{1}{2\pi} \sum_{k=-\infty}^{+\infty} \gamma(k) e^{-ik\lambda} \tag{2.46}$$

Donde $e^{-ik\lambda} = cos(k\lambda) - i\ sen(k\lambda); \quad \text{con} \quad i = \sqrt{-1}$.

Ya que Coseno y Seno tienen periodo 2π, entonces $f(\lambda)$ también está en el intervalo $[-\pi, \pi]$.

Propiedades

- $f(\lambda)$ es una función real de λ.
- $f(\lambda)$ es simétrica entorno al cero.
- $f(\lambda) \geq 0$.
- $f(\lambda) = f(\lambda + 2k\pi)$.

A partir de $f(\lambda)$ se puede escribir:

$$\gamma(k) = \int_{-\pi}^{\pi} f(\lambda) e^{-ik\lambda} d\lambda \tag{2.47}$$

Otra forma de escribir la densidad espectral, es a través de un proceso estacionario

$$X_t = \sum_{j=0}^{+\infty} \Psi_j \epsilon_{t-j} \tag{2.48}$$

$$\text{Donde} \quad \Psi(Z) = \sum_{j=-\infty}^{+\infty} \Psi_j B^j \quad \Rightarrow \quad f(\lambda) = \frac{\sigma^2}{2\pi} | \sum_{j=-\infty}^{+\infty} \Psi_j e^{-ij\lambda} |^2 \tag{2.49}$$

Periodograma

Si se quiere estimar el espectro a partir de la información que se tiene disponible se utiliza el periodograma, su característica es que corresponde a un estimador asintóticamente insesgado con varianza no consistente.

$$I(\lambda) = \frac{1}{2\pi n} | \sum_{t=1}^{n} X_t e^{-i\lambda t|^2} \tag{2.50}$$

Al graficar el periodograma versus la frecuencia se obtiene un gráfico que muestra los diversos "peacks" los cuales aportan a la variabilidad de la serie.

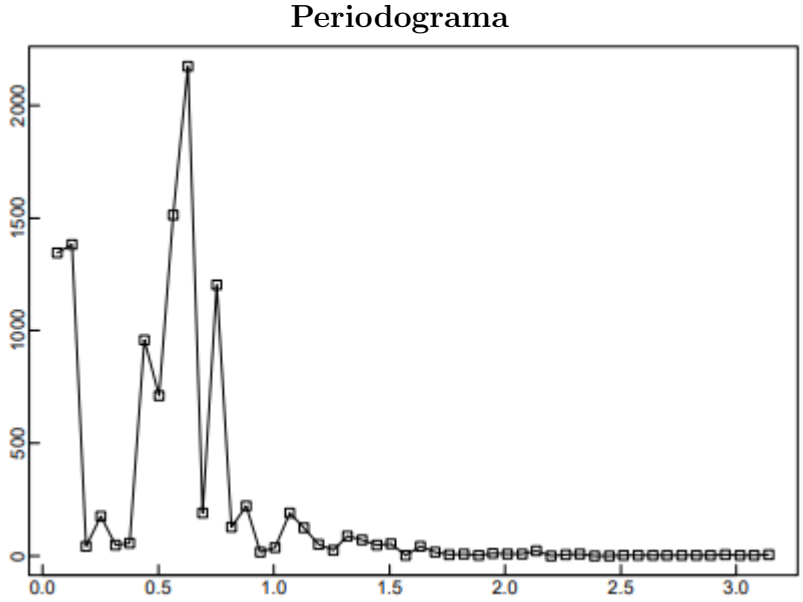


Figura 2.3: Fuente: Fuente: Brockwell (2002). Introduction Time Series and Forecasting. [11].

Análisis Espectral en una Serie de Tiempo

- Graficar la serie y observar la existencia de tendencia.
- Estudiar la variabilidad de los datos.
- Si la serie presenta una tendencia la clara (vector de medias no constante), aplicar una transformación a los datos, como Logarítmica, Exponencial, entre otros.
- Si la serie presenta un comportamiento **no estacionario**, se procede a diferenciarla.
- Estimar el periodograma y graficar estimadores con el fin de observar un comportamiento más suave de la serie y de esta manera poder identificar aquellos "peaks" más pronunciados que realmente aportan a la variabilidad.

2.2. Modelo ARIMAX

Proviene del acrónimo en inglés Autoregressive Integrated Moving average with eXogenous variables.

Éste modelo se puede seccionar en tres partes: la primera parte es Autoregresiva (AR) que plantea una combinación lineal con la serie misma en periodos de tiempo anterior, la segunda de Medias Móviles (MA) que estima los valores a través de una relación en una sucesión de errores ponderados y de períodos anteriores y por último la tercera parte (X) son las observaciones anteriores de una serie exógena o dicho de otra forma, de datos conocidos de una variable independiente a través del tiempo. Adicionalmente, al considerar la serie diferenciada, que se espera que sea estacionaria se obtiene el modelo ARIMAX.

La estimación de parámetros ARIMA es análogo al modelo tradicional mientras que las variables independientes incluidas se estiman mediante regresión, quedando finalmente el modelo de la forma:

$$(1 - \sum_{i=1}^{p} \phi_t B^i) y_t = (1 + \sum_{j=1}^{q} \theta_j B^j)\epsilon_j + (1 + \sum_{i=1}^{k} \beta_i B^i) x_i \tag{2.51}$$

O bien también podemos reescribirlo de ésta manera:

$$y_t = \beta_0 + \beta_1 X_{1,t} + \beta_2 X_{2,t} + ... + \beta_k X_{k,t} + \frac{(1 - \theta_1 B - \theta_2 B^2 - \theta_q B^q)}{(1 - \phi_1 B - \phi_2 B^2 - \phi_q B^q)} \epsilon_t \tag{2.52}$$

2.3. Análisis y diagnóstico de Modelos

2.3.1. Normalidad de los Residuos

Existen dos formas de comprobar éste supuesto, una es mediante la representación gráfica de histograma, el cual debe seguir la forma de una campana de Gauss, es decir que se ajuste a la distribución Normal.
Otra forma es mediante el gráfico de probabilidad normal, la cual contrasta cuantiles de una variable contra cuantiles de distribución normal. Cuando más cerca estén los cuantiles de la variable a los cuantiles de la distribución

normal (línea diagonal continua) más cerca está las variables de estar normalmente distribuida.

Por otra parte está el contraste inferencial, entre los test más conocidos para poder evaluar dicho supuesto es Shapiro- Wilk y Kolmogorov-Smirnov cyta dócima corresponde a:

$$H_0 : \epsilon_i \sim N(0, \sigma^2) \quad versus \quad H_1 : \epsilon_i \not\sim N(0, \sigma^2) \tag{2.53}$$

Donde cada test en particular presenta su estadístico para analizar dicha dócima, finalidad de estos test es no rechazar H_0, ya que de esta forma se podría decir que los residuos distribuyen normal.

Si es que esto no se cumple se podría realizar algunas transformaciones convenientes para solucionar el problema y poder cumplir con el supuesto. Entre la más conocida está la familia de Box-Cox, que soluciona los problemas de falta de normalidad y de heterocedasticidad, cuya definición es la siguiente:

$$Z_{\lambda,i} = \begin{cases} \frac{y_i \lambda^{-1}}{\lambda \dot{y}^{\lambda-1}} & si \quad \lambda \neq 0 \\ \dot{y} * ln(y_i) & si \quad \lambda = 0 \end{cases} \tag{2.54}$$

Donde $\dot{y} = (y_1 y_2 * ... * y_n)^{1/2}$

Homocedasticidad

Se habla de homocedasticidad cuando la varianza de las perturbaciones aleatorias condicional a los valores de las variables independientes se mantiene constante $V(i|x_i) = \sigma^2$ Este supuesto es muy importante, ya que si se cumple homocedasticidad (Figura 2.54) signica que el modelo está bien especicado, no debería existir un patrón denido entre los residuos y la variable dependiente. Si no se cumple el supuesto se le denomina heterocedasticidad (Figura 2.53), provocando que la varianza de los estimadores este sesgada, por lo que las pruebas t y F no serían tan conables.

Para la detección de heterocedasticidad existen métodos grácos e inferenciales para evaluar el supuesto.

Grácamente se tiene la relación de los residuos del modelo en el eje y versus el indicador de la observación correspondiente en el eje x Por otra parte dentro de los métodos formales que se utilizan para probar el supuesto de homocedasticidad se encuentra el Contraste de Breusch-Pagan, Contraste de White y El test de Bartlett.

En caso de que se esté en presencia de heterocedasticidad (Figura 2.9) se puede aplicar la transformación de Box-Cox antes mencionada, ya que proporciona una solución tanto para el cumplimiento de la normalidad y homocedasticidad.

No Autocorrelación de los errores

Este supuesto quiere decir que los errores no dependan de sus valores anteriores, es decir, que la estimación debe ser aleatoria. Si es que no se cumple este supuesto implica que la estimación de no son eciente ya que no serán de mínima varianza, por otra parte el $\hat{\sigma^2}$ será sesgado. Una de las formas de detectar la autocorrelación es a través de test de Durbin Watson, cuya dócima corresponde a:

H_0: No hay correlación de los errores
v/s
H_1: Hay correlación de los errores

La estadística para este test se muestra a continuación, cabe mencionar que está estadística es comparada con la distribución $(D = 2(1 - \hat{p}))$, la cual fue conocida para los autores del test en base a simulación.

$$d = \frac{\sum_{i=2}^{n}(\epsilon_i - \epsilon_{i-1})^2}{\sum_{i=1}^{n}\epsilon_i^2} \sim D = 2(1 - \hat{p}) \tag{2.55}$$

A continuación se presenta una tabla que indica la decisión que se debe considerar respecto a este test.

H_0	Decisión	Criterio
No autocorrelación positiva	Rechazar	$0 < d < d_l$
No autocorrelación positiva	No tomar decisión	$d_l < d < d_u$
No correlación negativa	Rechazar	$4\text{-}d_l < d < 4$
No correlación negativa	No tomar decisión	$4\text{-}d_u < d < 4 - d_l$
No autocorrelación, positiva o negativo	No rechazar	$d_u < d < 4 - dl$

2.3.2. Test de Dickey-Fuller

David Dickey y Wayne Fuller (1979) desarrollaron esta prueba para determinar la estacionaridad de un modelo, planteándolo como un contraste de no estacionariedad, ya que la hipótesis nula es la presencia de una raíz unitaria en el proceso generador de datos de la serie analizada. En otras palabras esta prueba conrma si una raíz unitaria está presente en el modelo Auto regresivo. Para realizar esta prueba se asume que la serie se puede aproximar por un proceso AR(1) con tres variantes: con media igual a cero, con media distinta de cero y con tendencia lineal. Inicialmente se asume que Yt sigue un modelo AR(1), es decir:

$$Y_t = \phi_1 Y_{t-1} + \epsilon_t$$

$$Y_t Y_{t-1} = (\phi_1 - 1) Y_{t-1} + \epsilon_t$$

$$\triangle Y_t =_{t-1} + \epsilon_t$$

con $\epsilon_t \sim RB(0, \sigma^2)$, además $\rho = \phi_1 - 1$. Por lo tanto la raíz unitaria equivale a $\phi_1 = 1$ o $\rho = 0$.

Luego, la prueba de Dickey-Fuller puede ser estimada de tres distintas formas, las cuales serán:

- Caso1: si Y_t es un camino aleatorio, es decir, con media cero.
  El modelo será de la forma:

$$_t =_{t-1} + \epsilon_t \tag{2.56}$$

  Esto no contiene tendencia ni intercepto.

- Caso2: Si Y_t es un camino aleatorio con intercepto, es decir, con media distinta de cero.
  El modelo será de la forma:

$$\triangle Y_t = \mu + \rho Y_{t-1} + \epsilon_t \tag{2.57}$$

  Este modelo será de la forma:

$$\triangle Y_t = \mu + \rho Y_{t-1} + \epsilon_t \tag{2.58}$$

  Este no contiene tendencia, pero si intercepto.

- Caso3: Si Y_t es un camino con tendencia e intercepto, es decir, con media distinta de cero tendencia.
  El modelo será de la forma:

$$\triangle Y_t = \mu + \beta_t + \rho Y_{t-1} + \epsilon_t \tag{2.59}$$

  Este contiene tendencia e intercepto.

En cada uno de estos casos la hipótesis nula será que existe una raíz unitaria (serie no estacionaria) y la hipótesis alternativa es $\rho < 0$, que representa la estacionalidad de la serie Y_t,con media distinta de cero y una tendencia determinística.

H_0= La serie de tiempo no es estacionaria ($\rho = 0$) y presenta raíz unitaria.
Por lo que $\rho = 0$, $\mu = 0$ y $\beta = 0$
v/s
H_1= La serie de tiempo es estacionaria y no presenta raíz unitaria.
Por lo que $\rho < 0$, $\mu \neq 0$ y $\beta \neq 0$

2.3.3. Test de Ljung-Box

Ésta prueba tiene su nombre debido a Greta M. Ljung y George E. P. Box y tiene como fin, para un modelo de la serie de tiempo, comprobar la existencia de autocorrelación residual.

El test de Ljung-Box puede definirse de la siguiente manera (Ljung, G. M. (1978)):

H_0: Los datos se distribuyen de forma independiente.

v/s

H_1: Los datos no se distribuyen de forma independiente.

El estadístico Chi-Cuadrado del test se descompone de la siguiente forma:

$$Q = n(n+2)\sum_{k=1}^{h}\frac{\rho_k^2}{n-k} \tag{2.60}$$

Con ${}^{2}_{1-\alpha,h}$ es la $\alpha - cuantil$ de la distribución Chi-Cuadrado con h grados de libertad.

Al probar la hipótesis sobre un modelo integrado, como por ejemplo un ARIMA, los grados de libertad quedan sujetos al orden de los parámetros Auto Regresivos (p) y de Medias Móviles (q) mediante la siguiente expresión: Grados de libertad= m-p-q

2.3.4. Error Medio Absoluto Porcentual

El Error Porcentual Medio MAPE por sus siglas en inglés Mean Absolute Percentage Error, es un indicador del desempeño del pronóstico de demanda que mide el tamaño del error en términos absolutos porcentuales.

El hecho que se estime una magnitud del error porcentual lo hace un indicador frecuentemente la magnitud de la variable a estudiar y el margen de error aceptable dentro del contexto.

Puede utilizarse para medir tanto el ajuste en la muestra de entrenamiento como el poder predictivo con datos desconocidos en la muestra de validación.

Su fórmula matemática está dada por la siguiente expresión (Krajewski, L. (1998)):

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\frac{|Y_t - F_t|}{|Y_t|} \tag{2.61}$$

2.3.5. Error Cuadrático Medio de Predicción

Corresponde un indicador de rendimiento residual, utilizado en el presente informe para determinar la distancia entre las predicciones que se realizan sobre la variable de interés y los datos observados.

Mide el tamaño del error, elevado al cuadrado, para evitar que la media muestral se vea afectada por los distintos signos que pueda haber en el vector residual.

Su utilización se hará sobre el último 10 por ciento de los datos para verificar la capacidad predictiva del modelo.

Básicamente es un Error Cuadrático Medio, pero aplicado a datos desconocidos para el Modelo Predictivo, su fórmula matemática está dada por la siguiente expresión:

$$ECMP = \frac{1}{n}\sum_{t=1}^{n}\frac{(Y_t - F_t)^2}{Y_t} \tag{2.62}$$

2.4. Imputación

Definición

En estadística, la imputación es la sustitución de valores no observado o perdidos.

2.4.1. Imputación Hot-Deck

El procedimiento Hot Deck es un proceso de duplicación. Cuando falta un valor, se duplica un valor ya existente en la muestra para reemplazarlo. Su principal propósito es reducir el sesgo debido a la no respuesta. Existen diferentes variantes del método Hot Deck:

Imputación aleatoria Hot Deck (Imputación Hot Deck por muestreo aleatorio simple):

Se asigna aleatoriamente un valor recogido en la muestra de la variable a imputar. Conserva la distribución de los respondientes pero no considera si es factible la imputación ni la correlación con otras variables. Es un método estocástico.
Por lo general el procedimiento Hot Deck tiene un proceso de clasificación asociado a él. Todas las unidades de la muestra están clasificadas en grupos disjuntos de forma que las unidades sean lo más homogéneas posibles dentro de los grupos. A cada valor que falte, se le asigna un valor del mismo grupo. Así la suposición que se está utilizando es que dentro de cada grupo de clasificación la no respuesta sigue la misma distribución que los que responden. Las variables de clasificación han de estar correlacionadas con los valores que falten y con los valores de los que contestan. Si esto no se mantiene, el procedimiento Hot Deck puede llevar a resultados erróneos. Teniendo en cuenta lo anterior podemos encontrar otras variantes como:
Imputación aleatoria Hot Deck por grupos Imputa con un valor recogido de la muestra perteneciente al grupo. Es un método estocástico.

Imputación Hot Deck secuencial

Se usa cuando la muestra tiene algún tipo de orden dentro de cada grupo de clasificación. Cada valor faltante se reemplaza por el registro sin valor missing, perteneciente al mismo grupo e inmediatamente anterior a él; si el primer registro tiene un dato faltante, este es reemplazado por un valor inicial que puede obtenerse de información externa. Las desventajas de este método son: 1. Si es necesario imputar muchos registros se tiende a emplear el mismo valor, llevando a una perdida de precisión de las estimaciones. 2. Es difícil estudiar la precisión de las estimaciones.
Imputación Hot Deck: Vecino más cercano Es un procedimiento no paramétrico basado en la suposición de que los individuos cercanos en un mismo espacio tienen características similares. Es un método de imputación determinístico. Para aplicarlo se requiere definir una medida de distancia. Por ejemplo, consideremos $x_i = (x_{i1}, ..., x_{iK})^T$ los valores de las K covariables para la unidad i en la cual el valor y_i es faltante. Si estas variables están clasificadas por

grupos, una métrica adecuada sería:

$$d(i,j) = \begin{cases} 0 & \text{si } i,j \text{ está en el mismo grupo} \\ 1 & \text{si } i,j \text{ está en diferentes grupos} \end{cases} \tag{2.63}$$

Pero otras posibles métricas son:

- Máxima desviación: $d(i,j) = max_K|x_{iK} - x_{jk}|$
- Distancia de Mahalanobis:$d(i,j) = (x_i - x_j)^T S_{xx}^{-1}(x_i - x_j)$ donde S_{xx}^{-1} es una estimación de la matriz de covarianzas de x_i.
- Distancia Euclidiana: $d(i,j) = \sqrt{\sum_{k=1}^{K}(x_{ik} - x_{jk})^2}$

Un posible peligro al usar el método Hot Deck es la duplicación del mismo valor muchas veces. Esto ocurre cuando en los grupos de clasificación hay muchos valores faltantes y pocos valores registrados. Resulta mejor cuando se trabaja con tamaños de la muestra grandes para así poder seleccionar valores que reemplacen a las unidades faltantes

2.5. Comparación Múltiple Método Tukey

2.5.1. Introducción

Cuando el análisis de varianza (ANOVA) da un resultado significativo, esto indica que al menos un grupo difiere de los otros grupos, sin embargo, para analizar el patrón de diferencia entre los grupos, se sugiere realizar comparaciones, y el método más comúnmente utilizado es la comparación de dos grupos (el método llamado "comparaciones de a pares").

Una técnica de comparación de a pares fácil y frecuentemente utilizada fue desarrollada por Tukey bajo el nombre de la prueba de diferencia honestamente significativa (**hsd**).

La principal idea de la **hsd** es calcular la diferencia honestamente significativa (es decir, la **hsd**) entre dos grupos, utilizando una distribución estadística definida por el investigador y llamada distribución q. Esta distribución da la distribución de muestreo exacta de la mayor diferencia entre un conjunto de grupos originados en la misma población.
Todas las diferencias de pares se evalúan utilizando la misma distribución de muestreo utilizada por la mayor diferencia, esto hace que el enfoque **hsd** sea bastante conservador.

2.5.2. Notación

- Los datos a analizar comprenden los grupos **A**.
- Un grupo dado se denota **a**.
- El número de observaciones del grupo **a-ésimo** se denota como S.
- Si todos los grupos tienen el mismo tamaño se denota S_a.
- El número total de observaciones se denota por **N**.
- La media de los grupo **a** se denota M_{a+}.
- La fuente de error (dentro del grupo) se denota con $\mathcal{S}(\mathcal{A})$.
- El efecto del error (entre el grupo) es denotado $\mathcal{A}$.
- El error cuadrático medio se denota $MS_{S(A)}$.
- El efecto del error cuadrático medio se denota MS_A.

2.5.3. Diferencia Menos Significativa

La razón detrás de la técnica **hsd** proviene de la observación de que, cuando la hipótesis nula es cierta, el valor de las estadísticas q evalúa la diferencia entre los grupos a y a' es igual a:

$$q = \frac{M_{a+} - M_{a'+}}{\sqrt{\frac{1}{2} MS_{S(A)} (\frac{1}{S_a} + \frac{1}{S_{a'}})}} \tag{2.64}$$

y sigue, una distribución q de rango estudientizado con un rango de A y $N-A$ grados de libertad. Por lo tanto, la relación t se declararía significativa a un nivel α determinado, si el valor de q es mayor que el valor crítico para el nivel α obtenido de la distribución q y se denota $q_{A,\alpha}$ donde $v = N - A$ es el número de grados de libertad del error, y A es el rango (es decir, el número de grupos). Este valor se puede obtener a partir de una tabla de la distribución del rango estudentizado.
La ecuación (6.73) muestra que una diferencia entre las medias de los grupos a y a' será significativo sí:

$$|M_{a+} - M_{a'+}| > HSD = q_{A,\alpha}\sqrt{\frac{1}{2}MS_{S(A)}(\frac{1}{S_a} + \frac{1}{S_{a'}})} \tag{2.65}$$

Cuando hay un número igual de observaciones por grupo, la ecuación (6.74) puede ser simplificada como:

$$HSD = q_{A,\alpha}\sqrt{\frac{MS_{S(A)}}{S}} \tag{2.66}$$

Para evaluar la diferencia entre las medias de los grupos a y a', se debe tomar el valor absoluto de la diferencia entre las medias y compararlo con el valor de **hsd**, sí:

$$|M_{a+} - M_{a'+}| \geq HSD \tag{2.67}$$

Entonces la comparación se declara significativa en el nivel α elegido (generalmente 0,05 o 0,01). Luego se repite este procedimiento para todas las comparaciones $\frac{A(A-1)}{2}$ (Abdi,2010 [5]).

Capítulo 3

Análisis Descriptivo

3.1. Descripción de la base de datos

Los datos recopilados dan origen a una base de datos con información del consumo de gas natural por hora desde el 1 de enero del 2010 al presente con un registro mayor de 80.000 datos.

La base de datos cuenta con 13 variables, en donde se considera la fecha, hora, el día de semana, estación del año, el día del año que le corresponde, la semana del año que le corresponde, temperatura, humedad, consumo entre otras.

En la variable de temperatura se tiene ausencia de datos del 17 % aproximadamente, al igual que la humedad, por otra parte se tiene ausencia en el registro de consumo de gas natural por hora en un 1 % de los datos.

3.1.1. Comportamiento del consumo de gas por meses

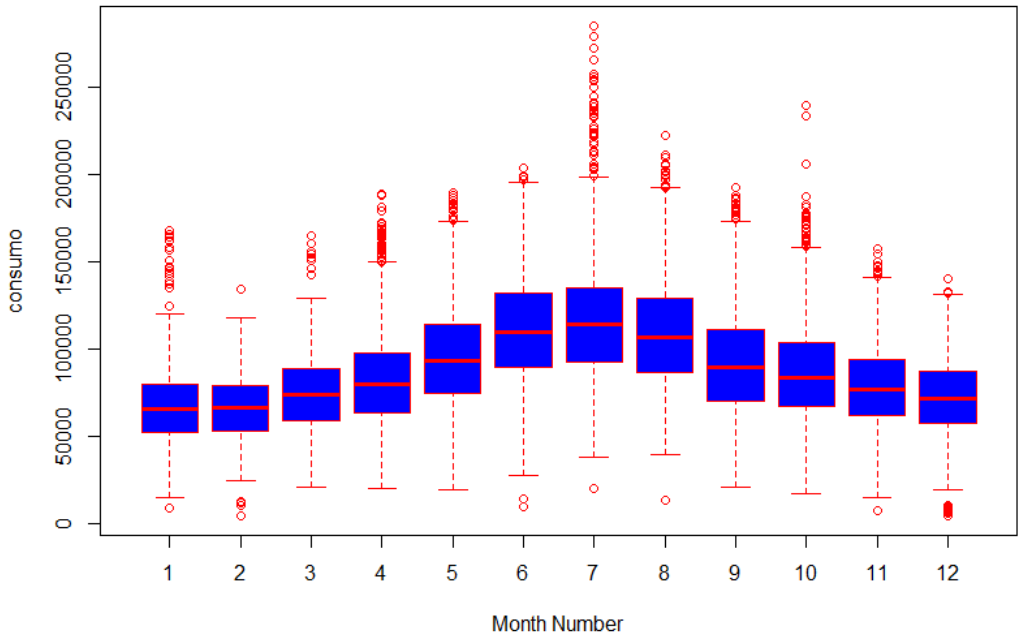


Figura 3.1: Consumo de gas natural por meses

Aplicando un gráfico de cajas (boxplot) en el software R-studio, podemos notar como varía el consumo de un mes a otro y la dispersión que poseen entre ellos, además de obtener el patrón de comportamiento de nuestros datos. En donde se observa claramente que los meses de temperaturas más bajas que son junio, julio y agosto corresponden a los meses con mayor consumo de gas natural, y en el caso opuesto los meses más cálidos que son diciembre, enero y febrero presentan el menor consumo de gas natural, lo que lleva a sospechar una alta dependencia entre el consumo de gas natural y la temperatura.

Además llama la atención la cantidad de datos fuera de línea que tiene el mes de julio indicando un consumo muy por encima de la media, esto puede aludir a los días más fríos del año.

3.1.2. Análisis y Relación de Variables

La primera herramienta que es aplicada corresponde a un gráfico de dispersión que permite observar la relación y tendencia de las variables continuas, enfocando especial observación en el consumo de gas natural (nodos) respecto a otra variable continua, ya que el consumo es la variable que nos interesa predecir.

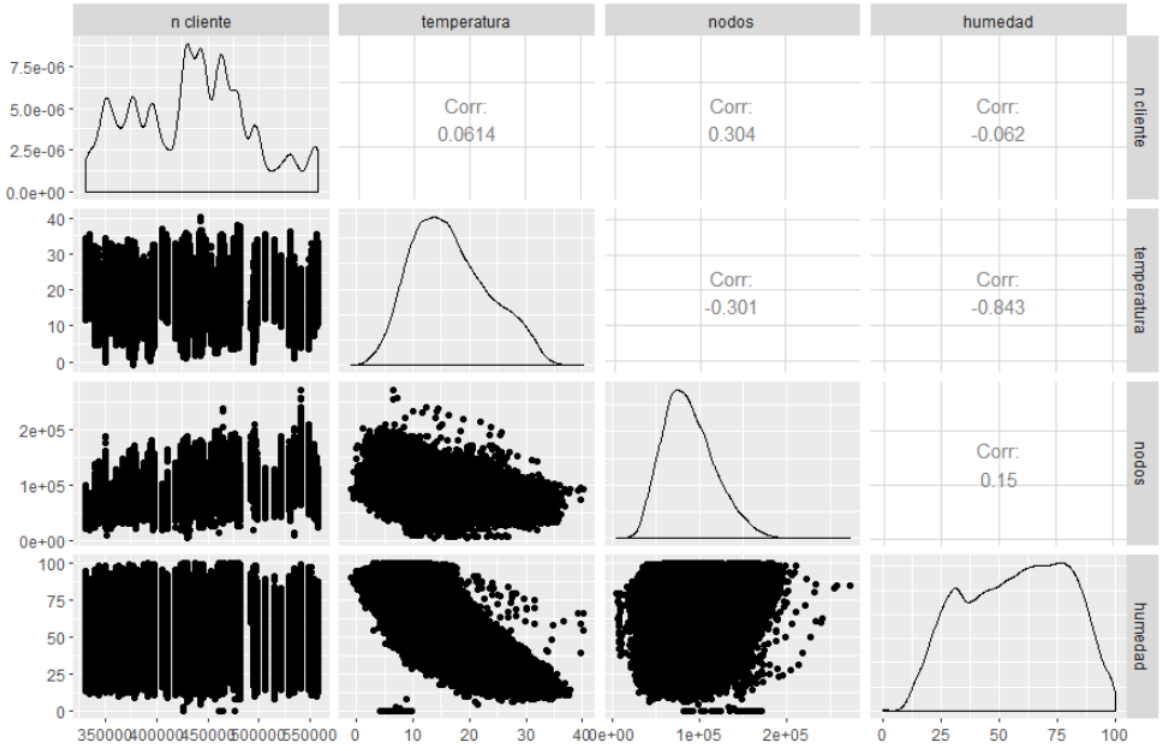


Figura 3.2: Gráfico de dispersión

El número de clientes está directamente relacionado con nodos que corresponde al consumo de gas natural, con una correlación de 0,304.
La temperatura se relaciona de modo inverso al consumo de gas natural con una correlación de -0,301.
Y por último tenemos una baja relación directa con la humedad, la cual tiene una correlación de 0,15.

Para analizar las variables de tipo factor se utilizara gráficos de caja.

Consumo v/s Tipo de día de la semana

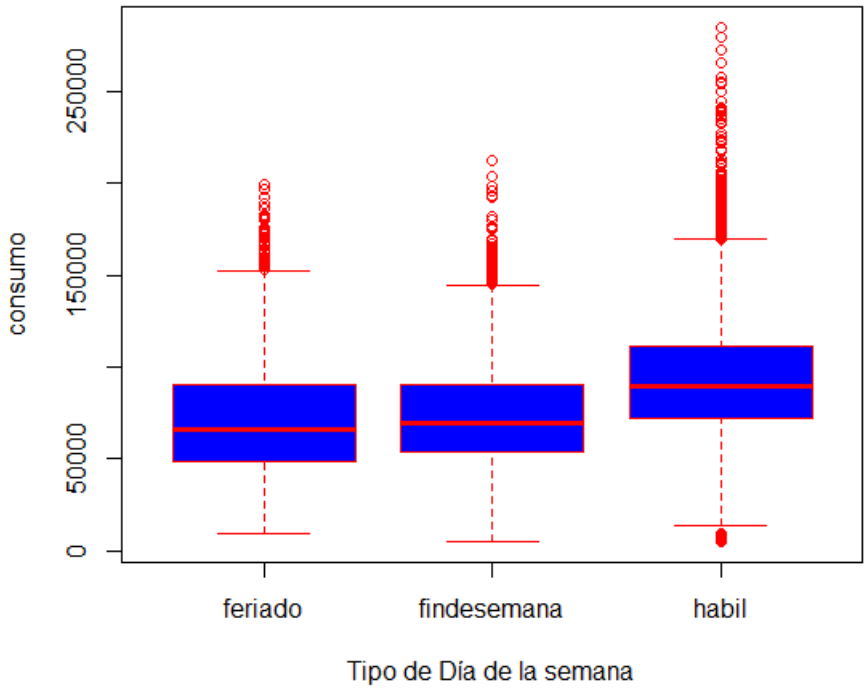


Figura 3.3: Gráfico de cajas entre el consumo de gas natural y tipos de día de la semana

Los días hábiles tienen un comportamiento relativamente diferente que un fin de semana y los festivos, pero los festivos son muy similares a los días de fin de semana, para investigar un poco más estas categorías, veremos que ocurre cuando se visualizan los meses del año en relación a los tipos de días.

Consumo de gas natural versus mes y categorías de días

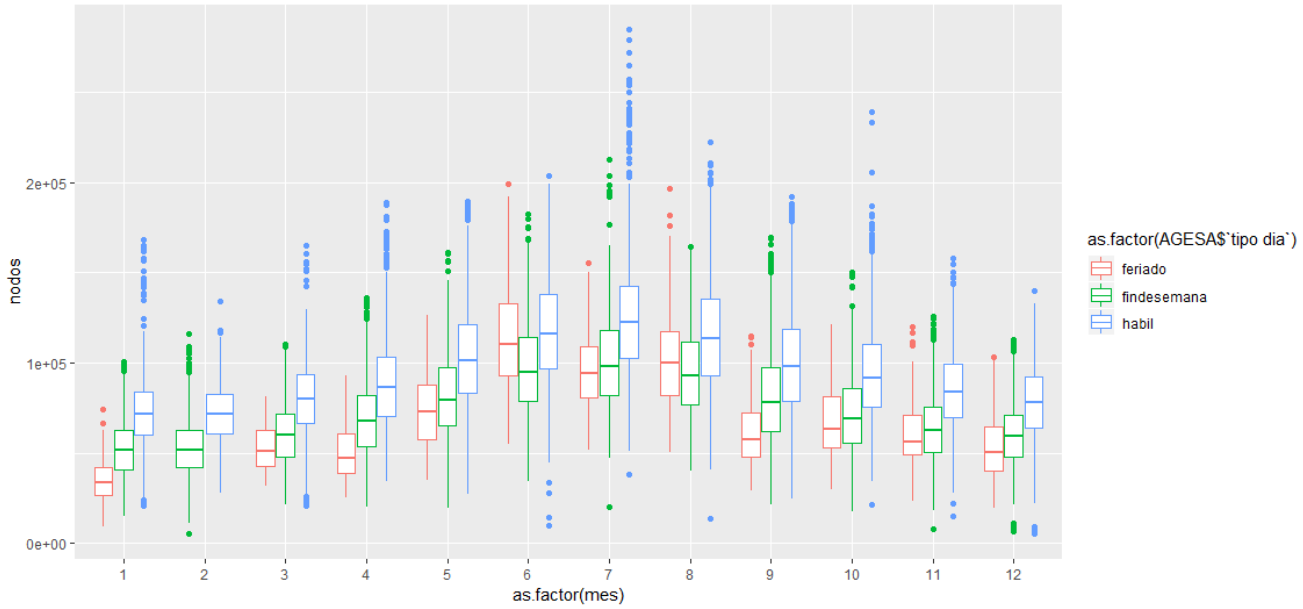


Figura 3.4: Gráfico de cajas, comparando los tipos de días en los meses del año.

Se aprecia que las categorías (feriado, día hábil, fin de semana), no en todos los meses se diferencian significativamente, lo que da a evaluar la aplicación del método de Tukey no sólo por tipo de día sino que también respecto a cada mes de año, ya que no poseen el mismo comportamiento entre ellos.

Consumo v/s día de la semana

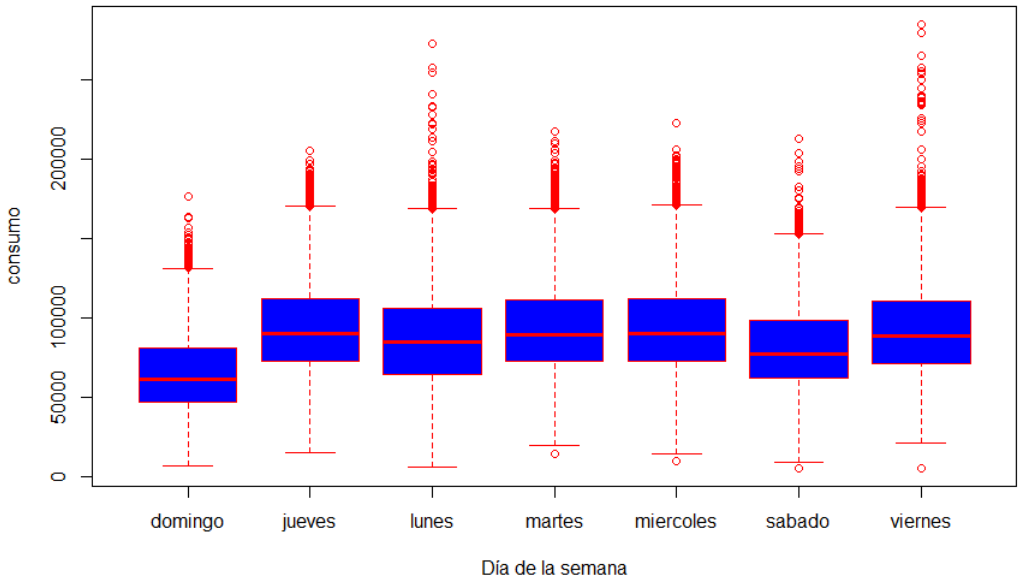


Figura 3.5: Gráfico de cajas de consumo v/s día de la semana

Se aprecia mucha semejanza entre los días hábiles y se observa un comportamiento relativamente distinto el día domingo y sábado. Además los días con más valores atípicos con el día lunes y viernes, por ende se evaluará si eso datos corresponden a días festivos.
Para evaluar si existe diferencia significativa entre los días de la semana se aplica el método de Tukey, considerando el gráfico anterior, estas categorías deben evaluarse por mes, incorporando posibles días interferiados para determinar si poseen un comportamiento distinto a los otros tipos de días designados.

Consumo de gas natural versus semana del año

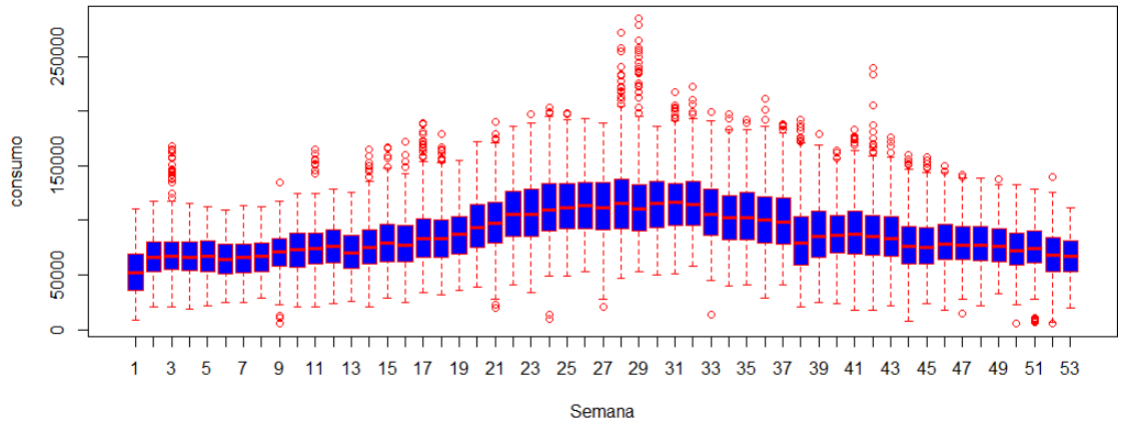


Figura 3.6: Gráfico de cajas de las semanas del año en relación con el consumo de gas natural

Este gráfico no muestra información distinta que el mismo gráfico por meses. Se visualiza el alza del consumo en temporada de menor temperatura y menor consumo en la temporada cálida del año.

Consumo de gas natural v/s estación del año

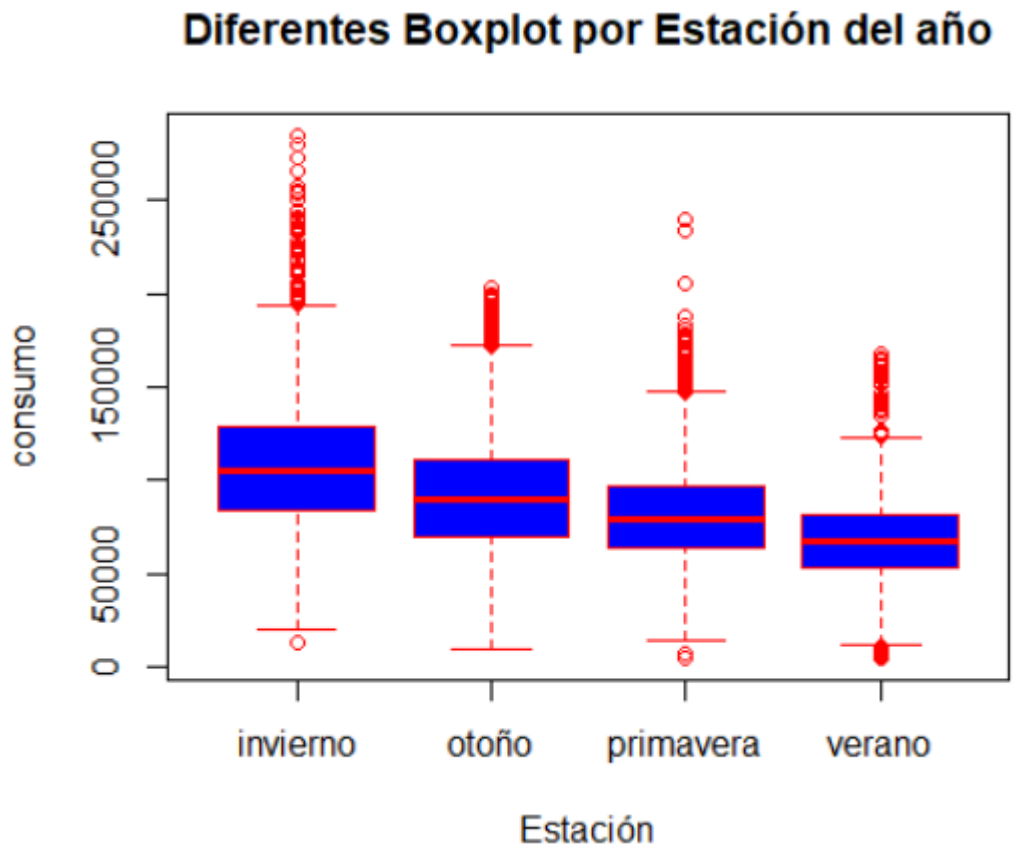


Figura 3.7: Gráfico de cajas del Consumo de gas natural en relación con la estación del año

Es bastante clara la influencia que tiene la temperatura respecto al consumo, mostrando un comportamiento prácticamente lineal entre las estaciones del año las cuales representan los cambios climáticos que se producen durante un año cronológico. Por otra parte se puede observar que entre otoño y primavera el cambio es menos significativo que las otras estaciones dado que corresponden a las estación de transición entre la temporada más fría y la más cálida.

Consumo de gas natural versus número de la semana por mes

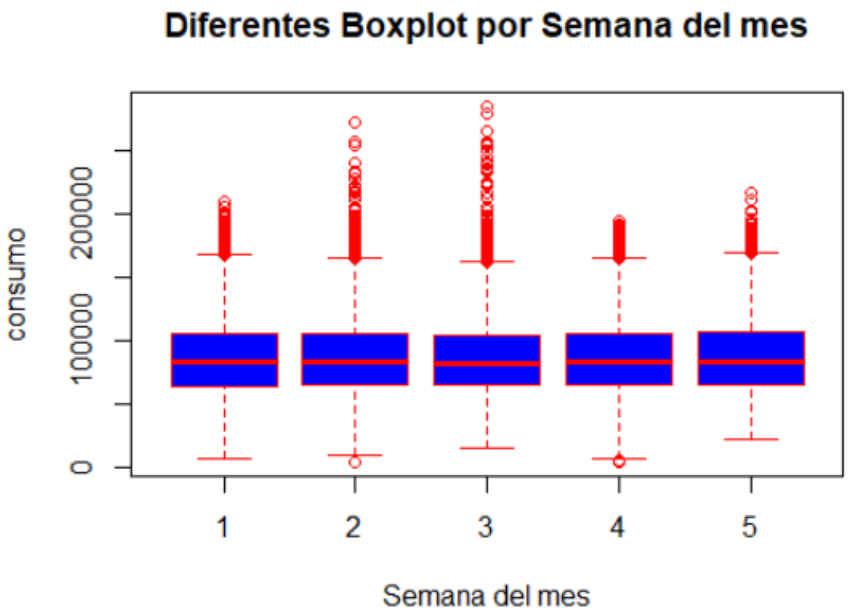


Figura 3.8: Gráfico de cajas del consumo de gas natural versus la semana del mes

Se aprecia que la media del consumo según el orden de la semana es casi idéntico pero con diferente dispersión y se observa mayor cantidad de datos lejanos a la media en la segunda y tercera semana de cada mes, con esto se descarta que exista una relación del consumo de gas respecto a la fecha de ingresos que tengan los consumidores, ya que si tuviese una dependencia se debería aprecia mayor consumo en la primera semana.

Consumo de gas natural v/s día del año

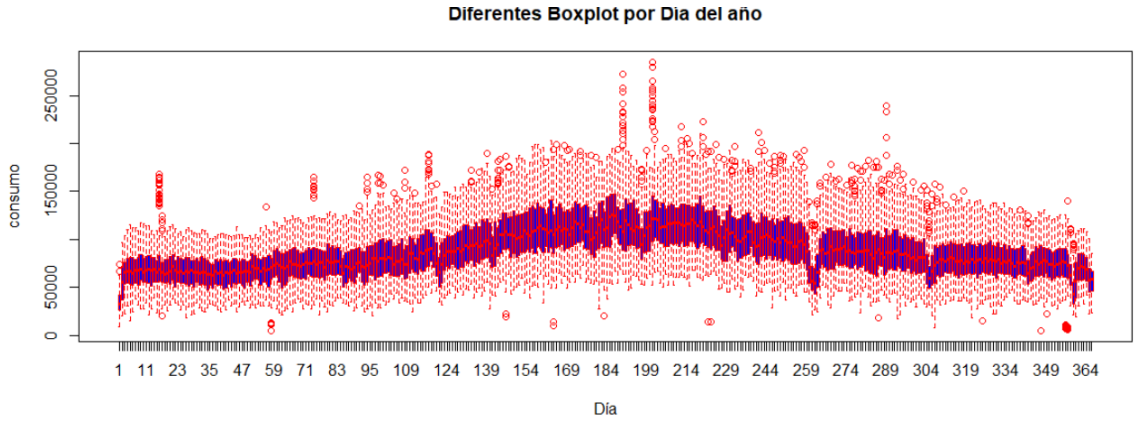


Figura 3.9: Gráfico de cajas para el consumo de gas natural respecto al consumo por día del año

Al observar el consumo de gas natural pero considerando cada día del año, podemos observar la misma tendencia mostrada anteriormente, pero con bajas específicas que corresponden a festividades las cuales tiene un comportamiento atípico, como por el ejemplo; el 1 de enero, navidad, entre otros.

3.1.3. Clasificación de Categorías para la Variable Días

Como vimos en los gráficos anteriores, la clasificación de tipo de día no es válida para todos los meses. Para esto se separa la base de datos por mes y luego se aplica el método de comparaciones múltiples de Tukey.

Contrastes de Tukey

Se utiliza el método de contrastes de Tukey, comparando de a pares de medias. Se plantea la siguiente hipótesis:

$$H_0 : \mu_i - \mu_l = 0 \text{ ; para i} \neq \text{l}$$
$$V/S$$
$$H_1 : \mu_i - \mu_l \neq 0 \text{; para i} \neq \text{l}$$

Donde:
μ_i: Transformación promedio del tipo de día i.
μ_l: Transformación promedio en el tipo de día l.
Quedando de la siguiente forma:

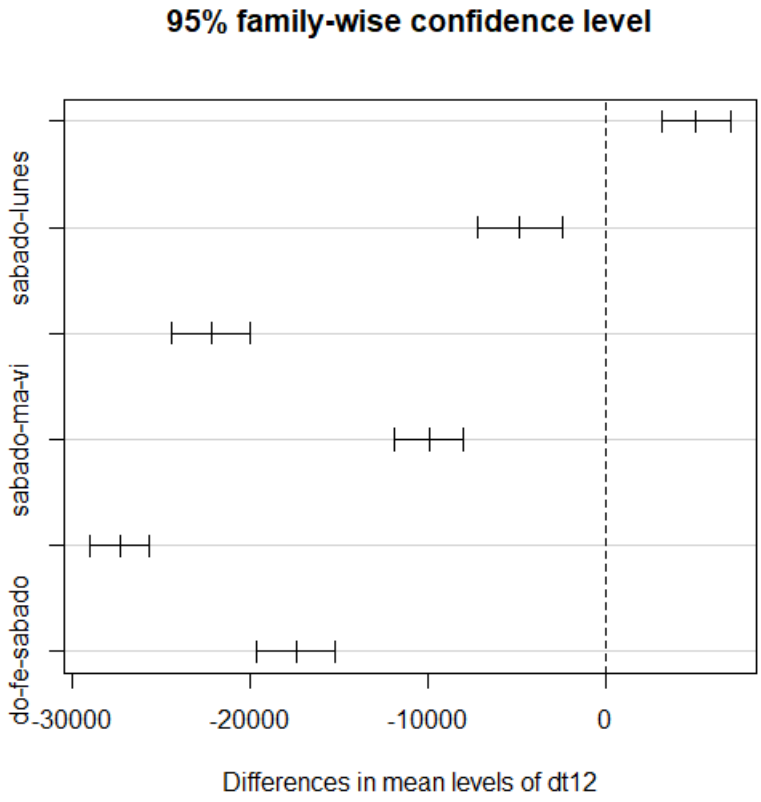


Figura 3.10: Gráfico de comparación múltiple Tukey para el mes de diciembre

Las categorías para el mes de diciembre son: Lunes, Martes a viernes, Sábado, Domingo-Feriado.
Esto quiere decir que no existe diferencia significativa en el consumo medio de gas natural en los días martes, miércoles, jueves y viernes, pero si se comportan distinto el día lunes y sábado de forma individual, no así el caso del domingo que es semejante a un festivo, durante el mes de diciembre.
Estas comparaciones se realizaron para cada mes utilizando el mismo procedimiento de clasificación.

Las categorías para los meses son:

Mes	**Categorías**				
Enero	Lunes	Sábado	Domingo	Martes a Viernes	Feriado
Febrero	Lunes	Sábado	Domingo	Martes a Viernes	
Marzo	Lunes	Sábado	Domingo-Festivo	Martes a Viernes	
Abril	Lunes	Sábado	Domingo	Martes a Viernes	Feriado
Mayo	Lunes	Sábado	Domingo-Festivo	Martes a Viernes	
Junio	Lunes	Sábado	Domingo-Festivo	Martes a Viernes	
Julio	Lunes	Sábado	Domingo-Festivo	Martes a Viernes	
Agosto	Lunes	Sábado-Feriado	Domingo	Martes a Viernes	
Septiembre	Lunes	Sábado	Domingo	Martes a Viernes	Feriado
Octubre	Lunes	Sábado	Domingo	Martes a Viernes	Feriado
Noviembre	Lunes	Sábado	Domingo-Festivo	Martes a Viernes	
Diciembre	Lunes	Sábado	Domingo-Festivo	Martes a Viernes	

Cuadro 3.1: Tabla de categorías para cada mes con método de comparación múltiple Tukey

3.1.4. Imputación de Datos Perdidos

Para poder imputar los datos se comienza visualizando el patrón de perdida de datos.
Gráficamente:

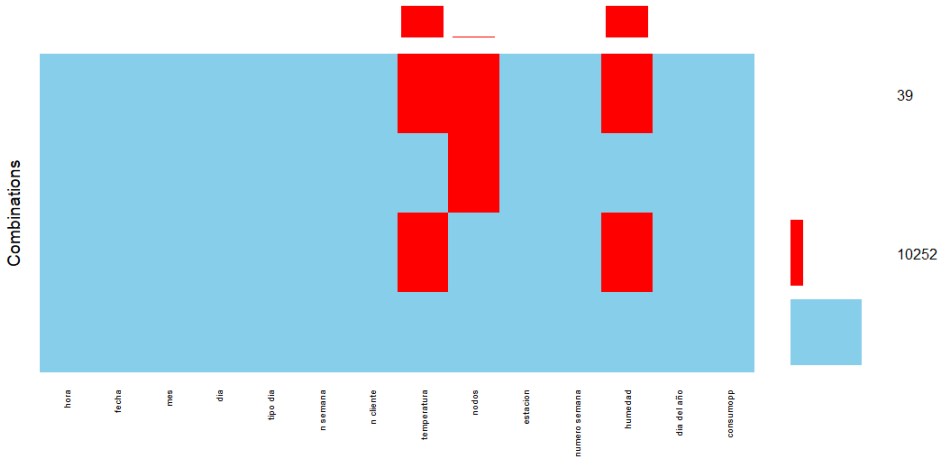


Figura 3.11: Gráfico de datos faltantes

En la parte superior de la figura 7.11 se observa la proporción de datos perdidos, y en el interior de gráfico se observa el lugar de donde se encuentran los datos perdidos en la base de datos, en color rojo se representan los datos faltantes, estos tienen igual patrón y cantidad de datos faltantes en temperatura, por otra parte en una baja proporción tenemos datos faltantes en la variable Rescom.
Revisando con más detalles el comportamiento de los datos faltantes se muestra el siguiente gráfico:

Rescom

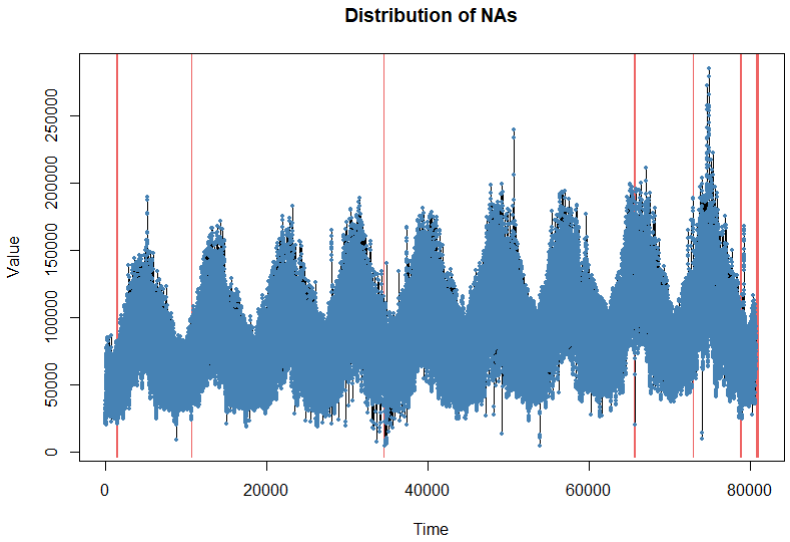


Figura 3.12: Gráfico de los datos faltantes en la variable Rescom

La pérdida de datos en este caso es baja y además tiene un comportamiento aleatorio.

Temperatura

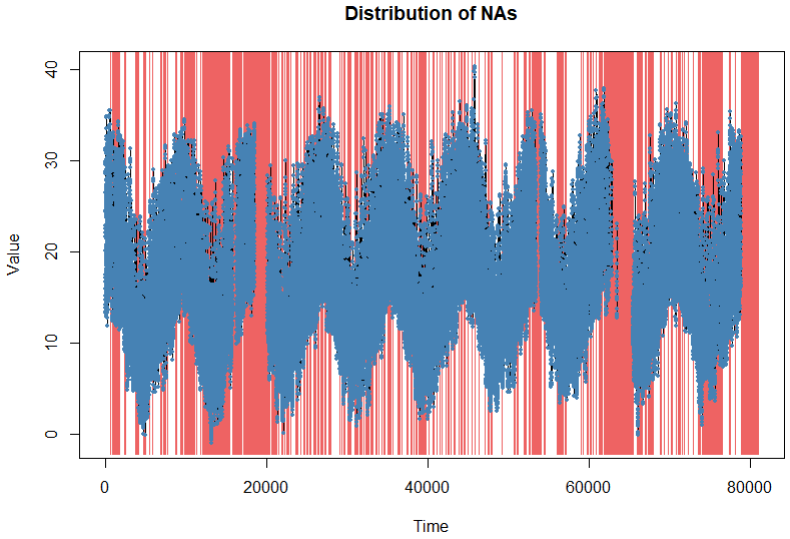


Figura 3.13: Gráfico de pérdida de datos de la variable temperatura

Se observan varios periodos continuos de perdida de datos, a pesar de esto no tiene un patrón determinado por ende se considera perdida aleatoria.

Para solucionar esto se aplicó la imputación por Hot-Desk, la cual asigna valor por semejanza, a pesar de tener un buen criterio de imputación no ha sido suficiente para tener buenos resultados razonables con tendencia del conjunto de datos, es por esto que se procede a realizar un segundo criterio, donde se seleccionan todos las temperaturas que se diferencian en más de 4 grados Celsius generados por los datos reales y la imputación, a ellos se le aplica por formula el promedio de la observación anterior y posterior, de esta forma se suaviza los datos generados que poseían una variación mayor a la habitual.
Quedando con la siguiente representación gráfica para los datos de forma horaria (desde el 2016, por fiabilidad de datos, dado que la empresa AGESA se crea en este año):

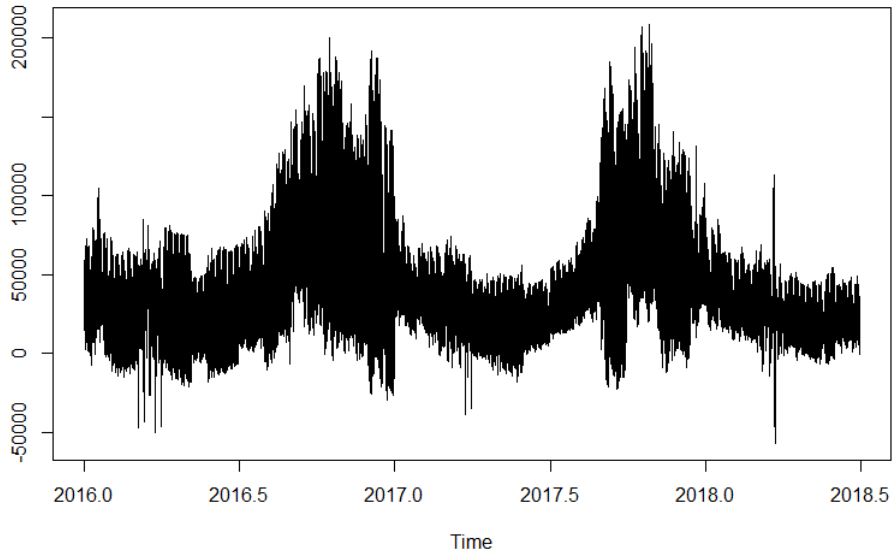


Figura 3.14: Rescom por hora

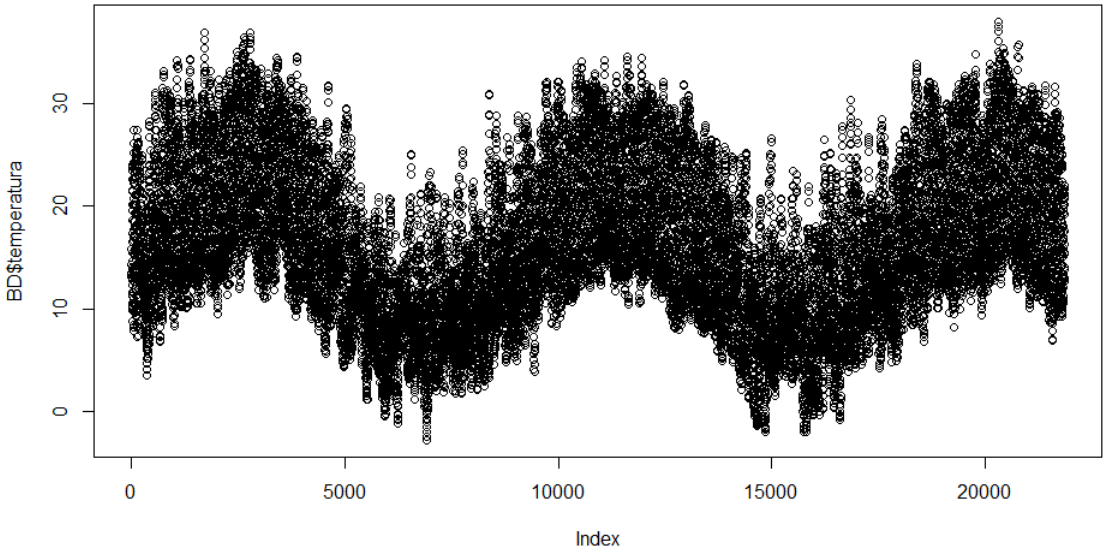


Figura 3.15: Temperatura por hora

Periodograma horario

Al aplicar este procedimiento se estudia la posibilidad de que la serie estudiada sea multiperiodica y conocer así sus periodos.

Gráficamente:

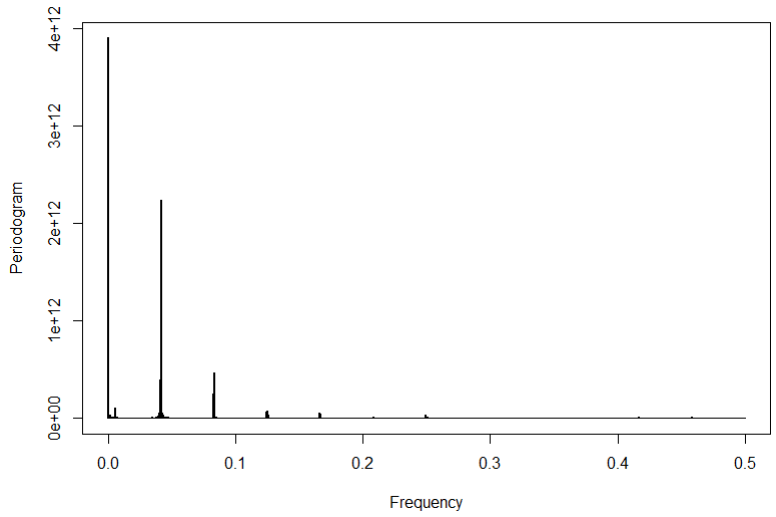


Figura 3.16: Periodograma Rescom por hora

Al verificar las frecuencias obtenidas con ayuda del software R. Se obtuvo periodos de 12 horas, 24 horas y anual.

Periodograma diario

Gráficamente:

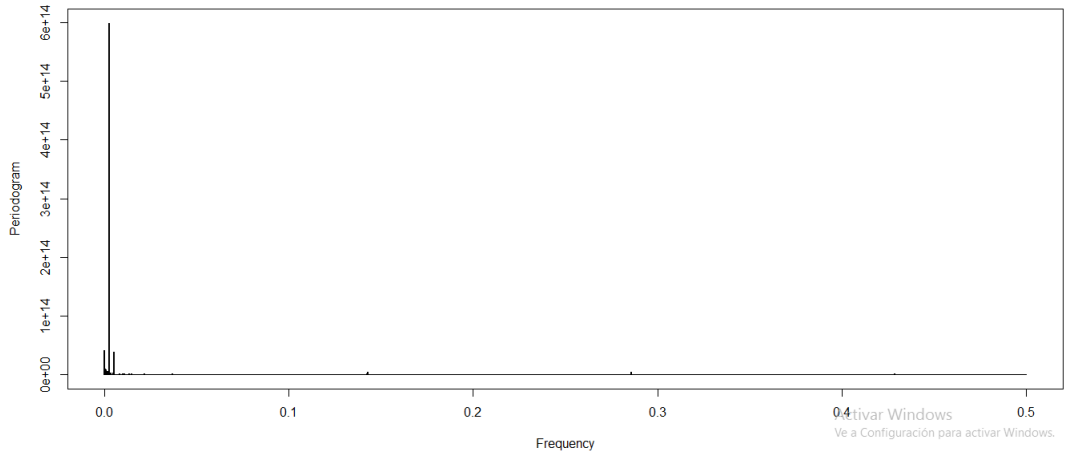


Figura 3.17: Periodograma Rescom diario

La frecuencia principal corresponde a 365 días.
Ver anexo (6.4)

Correlación Cruzada

Por otro lado se aplicó una correlación cruzada (CCF por sus siglas en inglés Cross Corre lation Function") para estudiar si es posible encontrar algún desfase temporal en las relaciones que tiene la temperatura con la Demanda de Gas natural.

Variable	Correlación Neta	Correlación Nueva	Lag	Diferencia
Temperatura	-0.301	-0,563	6	0,262

Ver graficas en Anexo (6.1)

En este caso se presentan 4 columnas con índices de correlación, la columna Correlación Neta hace referencia al nivel de asociación lineal de las variables independientes con la Demanda de Gas Natural en la misma hora mientras que la columna Correlación Nueva es la que indica el nivel de correlación con X horas de desfase, donde X es igual al número de lags indicado en la columna siguiente, y por último la columna Diferencia que muestra cuando mejora la correlación al aplicar el respectivo desface horario.

Capítulo 4

Modelos Arimax

4.1. Demanda de gas horaria

El consumo residencial de gas natural es analizado y transformado con función log, dado que la tendencia no es lineal y al observar la descomposición de la serie se aprecia que no es una serie aditiva, sino multiplicativa. La transformación de la demanda de gas natural se le aplico una translación de 56.868 unidades, ya que se producian valores no aplicables a la función logaritmica.

Se trabajo con el conjunto de datos compuesto desde octubre 2016 a marzo 2019, ya que la información horaria anterior no pudo ser verificada dado que AGESA no existía antes de esa fecha y esa información correspondía a Metrogas.

El modelo ARIMAX$(1,0,0)(0,0,0)_{8760}$ se puede escribir del siguiente modo:

$$\Phi(B)(1-B^{8760})X_t = \epsilon_t + \beta Z_t \tag{4.1}$$

donde:

- $\Phi(B) = 1 - 0,6678B$
- $\beta Z_t = 11,8067 - 0,066EFe + 0,015MFe + 0,0213FA + 0,0831MaFe + 0,2962JFe + 0,2488JuFe + 0,2423AgFe + 0,064SFe + 0,563NFe +$

$$0,0184DFe+0,0376EH+0,0154FH+0,0525MH+0,854AH+0,2055MaH+$$
$$0,2974JH + 0,2896JuH + 0,2522AgH + 0,1980SH + 0,1453OH +$$
$$0,0810NH+0,0711DH-0,0003Efinde-0,0016Ffinde+0,0357Mfinde+$$
$$0,0581Afinde + 0,1731Mafinde + 0,2808Jfinde + 0,2658Jufinde +$$
$$0,2593Agfinde + 0,1763Sfinde + 0,0835Ofinde + 0,406Nfinde +$$
$$0,0243Dfinde - 0,0052temperatura_{t-6}$$

Donde:

- H= Día hábil
- finde= Fin de semana
- Fe= Feriado
- E= Enero
- F=Febrero
- M= Marzo
- A= Abril
- Ma= Mayo
- J= Junio
- Jul= Julio
- Ag= Agosto
- S= Septiembre
- O= Octubre
- N= Noviembre
- D= Diciembre

Ajuste y Predicción del Modelo Arimax por hora

Para la predicción se trabaja con una base entrenamiento la cual corresponde al 90 % de los datos, y se pone a prueba comparando el 10 % de datos restantes para verificar que tan certera es la predicción.

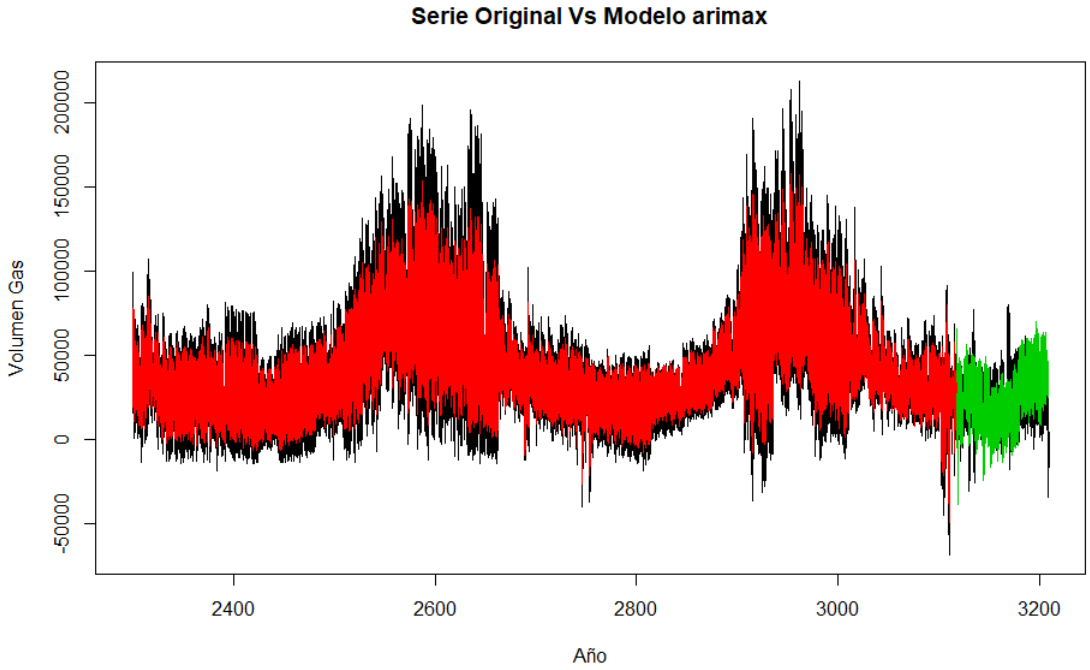


Figura 4.1: Datos reales con ajustes de modelo y predicción horaria

4.1.1. Demanda de gas diario

El consumo residencial de gas natural es analizado diariamente a partir de junio del año 2005. Analogo al procedimiento utilizado en la demandada horario.
El modelo ARIMAX$(1,0,0)(0,0,0)_{365}$ se puede escribir del siguiente modo:

$$\Phi(B)(1-B^{365})X_t = \epsilon_t + \beta Z_t \tag{4.2}$$

donde:

- $\Phi(B) = 1 - 0,94B$

- $\beta Z_t = 977782,9 + 131689,61 DomO + 90750,14 DomN + 52256,66 DomD - 128,023 DomF + 48382,41 DomM + 100919,29 DomA + 155531,48 DomMa + 239465,85 DomJ + 231804,2 DomJul + 266591,7 DomAg + 186419,12 DomS - 100715,27 FeE + 246663,85 FeO + 136003,4 FeN + 33060,19 FeD + 127791,9 FeM + 104308,54 FeA + 201755,29 FeMa + 448046,23 FeJ + 370181,04 FeJul + 386006,36 FeAg + 188636,8 FeS + 146002,3 HE + 289634,47 HO + 232105,21 HN + 196713,21 HD + 127455,5 HF + 169292,95 HM + 222159,9 HA + 315370,66 HMa + 406888,23 HJ + 429833,3 HJul + 446787,87 HAg + 377978,27 HS + 67957,794 SaE + 216351,3 SaO + 172505,87 SaN + 130709,7 SaD + 50863,73 SaF + 115756,47 SaM + 168846,13 SaA + 242478,54 SaMa + 322494,8 SaJ + 319301,53 SaJul + 343741,21 SaAg + 278230,7 SaS - 15206,1014 Tempmáx_t + 1,3 n\_clientes_t$

Donde:

- H= hábil
- Sa= Sábado
- Fe= Feriado
- Dom= Domingo
- E= Enero
- F=Febrero
- M= Marzo
- A= Abril
- Ma= Mayo
- J= Junio
- Jul= Julio
- Ag= Agosto
- S= Septiembre

- O= Octubre
- N= Noviembre
- D= Diciembre
- n_cliente= número de clientes

Ajuste de Modelo diaria

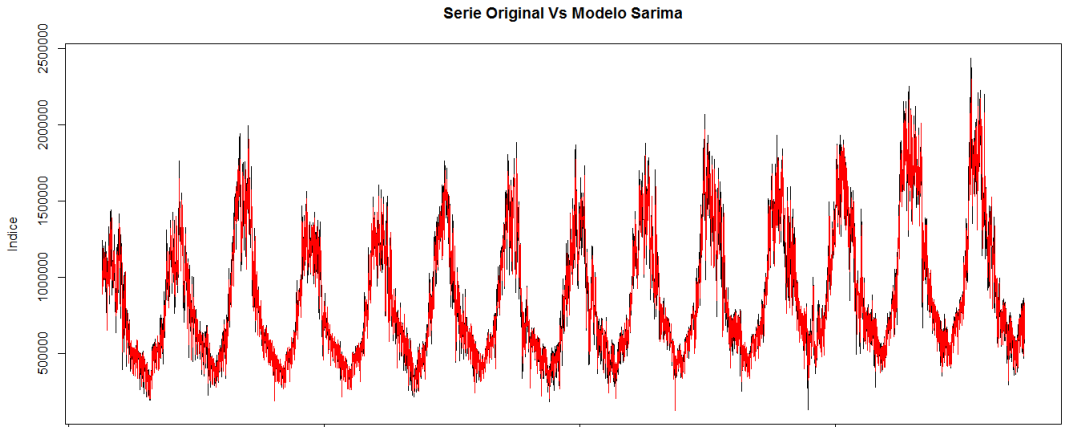


Figura 4.2: Ajuste Modelo Arimax por día

Predicción Modelo diario

Para la predicción se trabaja con una base entrenamiento la cual corresponde al 90 % de los datos, y se pone a prueba comparando el 10 % de datos restantes para verificar que tan certera es la predicción.

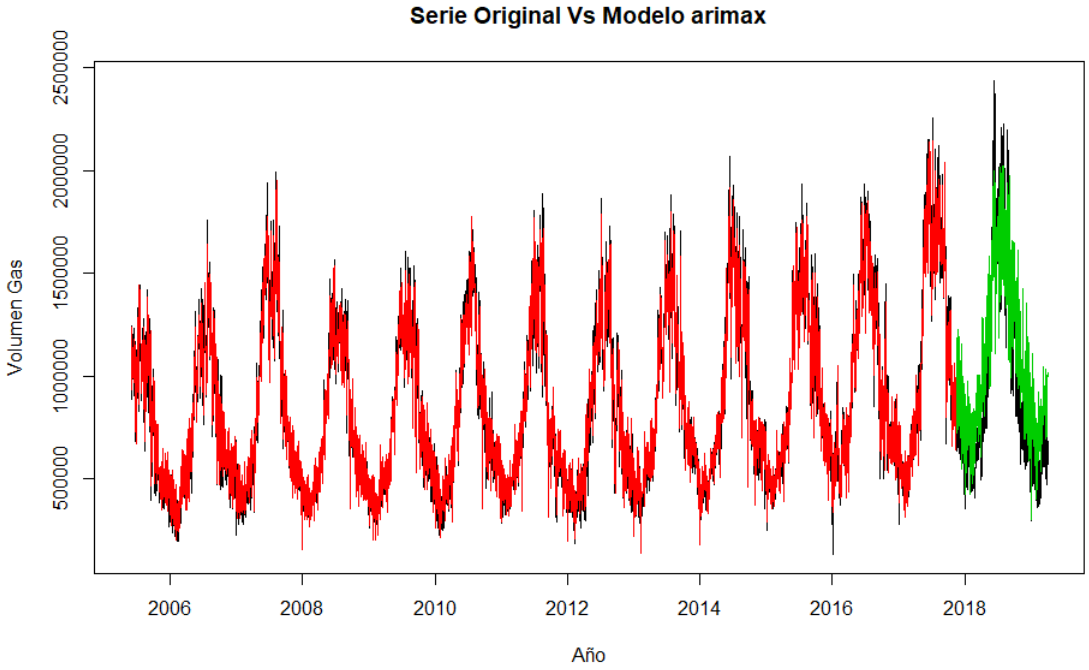


Figura 4.3: Datos originales, ajuste del modelo y predicción diaria

Modelo	N^o Variables	Coeficientes	AIC	MAPE aj.	MAPE pred.	ECMP
Día	4	50	119.645,5	0,0469276	0,06209655	38.349,74
Hora	3	37	-26.510,83	0,0384556	0,257892	5.867,98

Cuadro 4.1: Tabla comparación de modelos ARIMAX

Ver Anexo (6.2)

Supuestos de los modelos

Modelo	Autocorrelación	Homocedasticidad	Media Constante	Normalidad
ARIMAX Día	Si	Si	Si	Si
ARIMAX Hora	No	No	Si	No

Cuadro 4.2: Tabla de comparación de los supuesto en los modelos ARIMAX

Ver en Anexo (6.3)

Capítulo 5

Dashboard AGESA

Una vez seleccionado el modelo final (Modelo Arimax diario) se requiere de una herramienta que muestre visualmente los resultados alcanzados por el modelo y las métricas de desempeño.

Para ello se desarrolló e implementó un dashboard usando el paquete Shiny del software R-studio.

Shiny es una herramienta para crear fácilmente aplicaciones web interactivas (apps) que permiten a los usuarios interactuar con sus datos sin tener que manipular el código.

5.1. Visualización de Dashboard

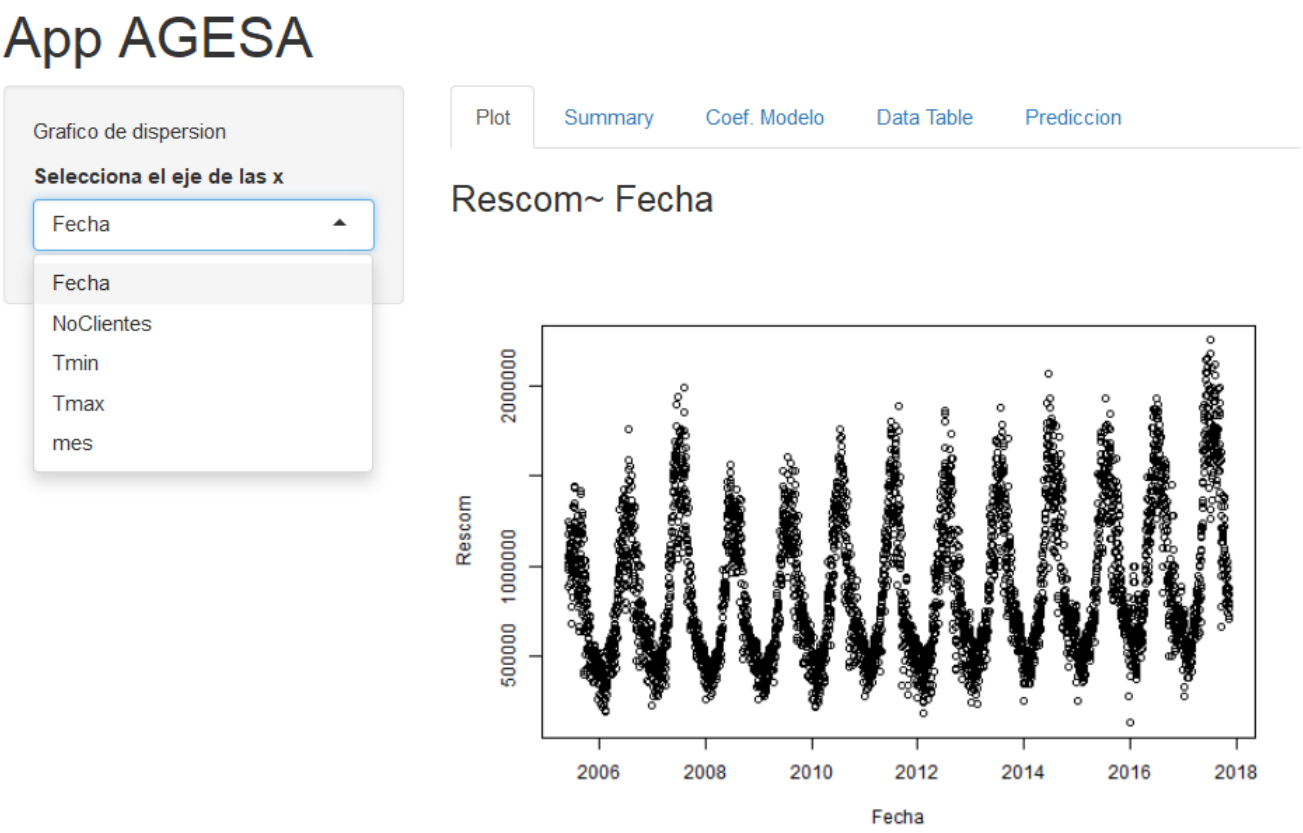


Figura 5.1: Pantalla principal de Dashboard AGESA

Esta imagen corresponde a la parte principal de la aplicación donde se puede visualizar gráficos de dispersión seleccionando la variable influyente que se quiera observar, además cuenta con varias pestañas;
donde:

- **Summary**: corresponde al resumen de las variables con su respectiva media, máximo, mínimo, etc.
- **Coef. Modelo**: corresponde a los coeficientes del modelo Arimax diario.

- **Data Table**: muestra la base de datos.
- **Predicción**: corresponde a los valores que predice el modelo en formato tabla.

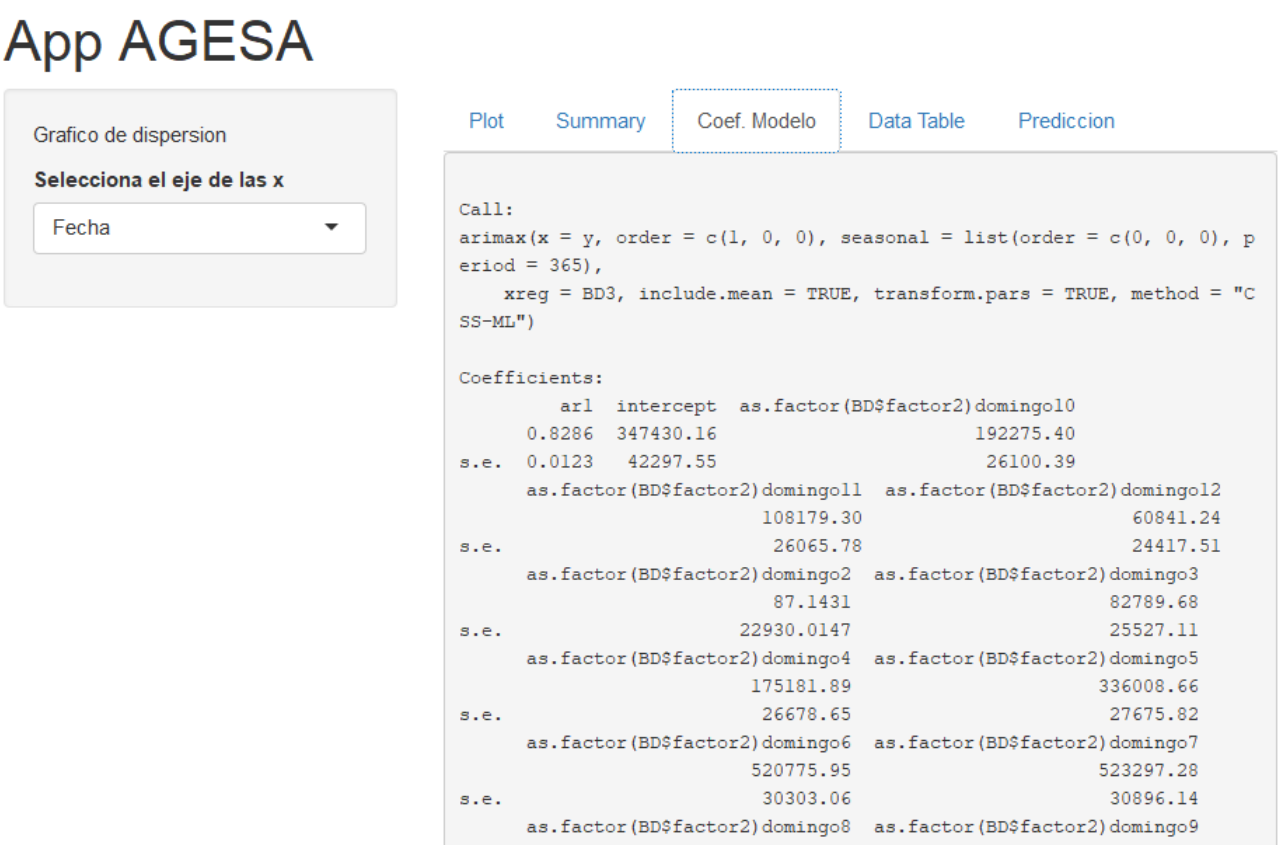


Figura 5.2: Ventana de visualización Coef. Modelo de Dashboard AGESA

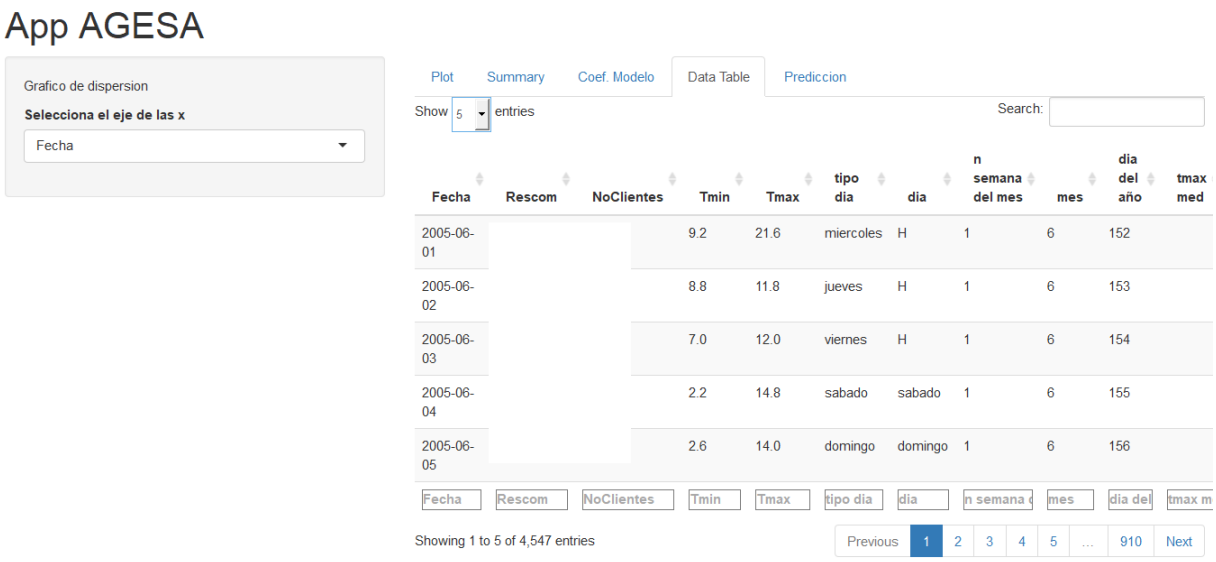


Figura 5.3: Ventana de visualización de Base de datos en Dashboard AGESA

En la Figura (5.3) por confidencialidad no se exponen los datos del consumo de gas natural.

Además la Figura (5.3) se aprecia la base de datos con todas las funciones disponibles para filtrar y buscar datos, además de tener posibilidades de observar de 5, 9, 28, 29, 30 y 31 días en la misma página.

Se deja un manual de uso para los trabajadores de AGESA que les permita añadir nuevas observaciones al modelo.
Ver Script del Dashboard en Anexo (6.5)

Capítulo 6

Conclusiones

La empresa AGESA disponía de modelos lineales no actualizados de forma constante, es por esto que se ejecutó una actualización y modernización del procedimiento utilizando herramientas más avanzadas de la estadística, aprovechando el comportamiento histórico se realizaron series de tiempo con variables exógenas.

Para modelar y predecir se aplicó modelo ARIMAX, tanto para determinar la demanda por hora y día, el objetivo de esto es determinar con más precisión la demanda diaria.

Al llevar a cabo la recolección de los datos, se obtuvo una base horaria, la cual contenía datos faltantes los cuales fueron imputados utilizando el método Hot-Desk para la variable Rescom y temperatura (variable exógena), además los datos de temperatura imputados que no seguían la tendencia de la serie se les aplica una segunda imputación, donde se condiciona los datos que se diferenciaban en más de 4 grados Celsius entre una hora y la siguiente,y luego se le aplica el promedio entre la hora anterior y posterior, con el fin de no perder información y ser concordantes con el resto de información disponible.

Para la demanda diaria si se tenía la información completa desde el año 2005 recolectada entre Metrogas y AGESA.

Los días de la semana tienen un papel fundamental en la predicción de la demanda, ya que depende del día es si el consumo tienen una tendencia al alza o baja. Los días festivos suelen ser los días de menor consumo de gas natural y se relación con la temperatura, pero estas relación no son siempre iguales durante el año. Para determinar las categorías de los tipos de días (hábiles, festivos, fin de semana, etc.) se utilizo el método de comparaciones múltiples Tukey, con el cual se pudo determinar que no todos los meses tienen las mismas categorías y es por este motivo que el modelo se incluye como variable exógena los tipos de días con la interacción del mes del año.
Tanto para la base diaria y horaria, se aplico el mismo procedimiento.

Luego para verificar si la máxima correlación de la temperatura con el consumo ocurre en el mismo instante se utiliza CCF, el cual en la base de datos diaria se cumple sin inconvenientes, pero en la base de datos horarias se obtuvo que la máxima correlación de la temperatura con el consumo ocurre con la temperatura que ocurrió hace 6 horas, esto produce un alza en la correlación de las variables de -0,301 al -0,563.

Con toda la información recolectada se realiza un modelo ARIMAX por hora y otro por día.
El modelo ARIMAX por hora no cumple todos los supuestos, pero si genera un ajuste de datos muy satisfactorio pero no una buena predicción, es decir que el poder predictivo del modelo no es el esperado, esto sucede porque la cantidad de periodos disponibles de datos históricas es poca.
Para solucionar esto se sugiere continuar con el registro de datos por hora para implementar a futuro un mejor modelo con unos años más de información y tener la precaución de no tener vacíos en este registro para que la información sea lo real posible sin necesidad de incurrir en imputación de datos.

El modelo ARIMAX diario si cumple con los supuesto y el poder predictivo es mejor que el modelo horario. Este modelo cuenta con más información histórica y la desviación de la predicción con los consumos reales no supera el 5 %, por ende cumple con los requerimientos solicitados por AGESA.

El mejor modelo y predicción corresponde al modelo Arimax diario. Para un manejo más amigable se desarrolló una aplicación web que permite observar las variables influyentes en el modelo diario, además de observar el resumen numérico, gráficas, coeficientes del modelo y predicción del mismo utilizando el paquete Shiny del software R-studio.

Bibliografía

[1] Abdi, H. Edelman, B. Valentin, D. Dowling, W. J. (2009). *Experimental Design and Analysis for Psychology*, Oxford University Press, Oxford.

[2] Arce, A. Canales, V. Lehmann, K. (2019). *Métodos de Imputación VIII EPF: Gastos diarios e ingresos de la actividad laboral y jubilaciones Nº 7*, INE, Santiago de Chile.

[3] Rouse, M. (2017). *Aprendizaje automático (machine learning)*, https://searchdatacenter.techtarget.com/es/definicion/Aprendizaje-automatico-machine-learning

[4] Medina, F. Galván, M. (2007). *Imputación de datos: teoría y práctica*, División de Estadística y Proyecciones Económicas 54 estudios estadísticos y prospectivos, Santiago de Chile.

[5] Abdi, H. Williams, L. J. (2010). *Tukey Honestly significant difference (HSD) test*. The University of Texas at Dallas, USA.

[6] Ruiz Usano, R. Framiñán Torres, J. M. Crespo, A. Muñoz, M. A, (2001). *Simulación continua y discontinua de un sistema de producción con inventario en proceso constante*, Congreso de Ingeniería de Organización, Sevilla.

[7] Lantz, B. (2013). *Machine Learning with R*, Birmingham, Inglaterra.

[8] Andrews, B. H. Dean, D. M. Swain, R. y Cole, C. (2013). *Building ARIMA and ARIMAX Models for Predicting Long-Term Disability Benefit Application Rates in the Public/Private Sectors*, University of Southern Maine.

[9] Gras, A. J. (2001). *Diseños de series temporales: técnicas de análisis*, Universidad de Barcelona, España.

[10] Garcoa, A. (2000). *Enfoque y practica para planificación y control de inventarios*, México.

[11] Brockwell, P. y Davis, R. (2002).*Introduction Time Series and Forecasting*. Department of Statistics, Colorado State University, USA.

[12] Castaño, E. y Martínez, J. (1998). *Uso de la función de correlación cruzada en la identificación de modelos ARMA*. Facultad de Ciencias, Universidad Nacional de Colombia, Bogotá, Colombia.

[13] Guerrero, V. (2002). *Pronósticos con restricciones en series de tiempo univariadas: aplicación al seguimiento del PIB de México en 2001*. México.

[14] Wu, X. y Kuman, V. (2009). *The top ten Algorithms in Data Mining*. Minnesota, USA.

Anexo

6.1. Función de Correlación Cruzada

6.1.1. Grafico función de correlación cruzada de temperatura con la demana diaria

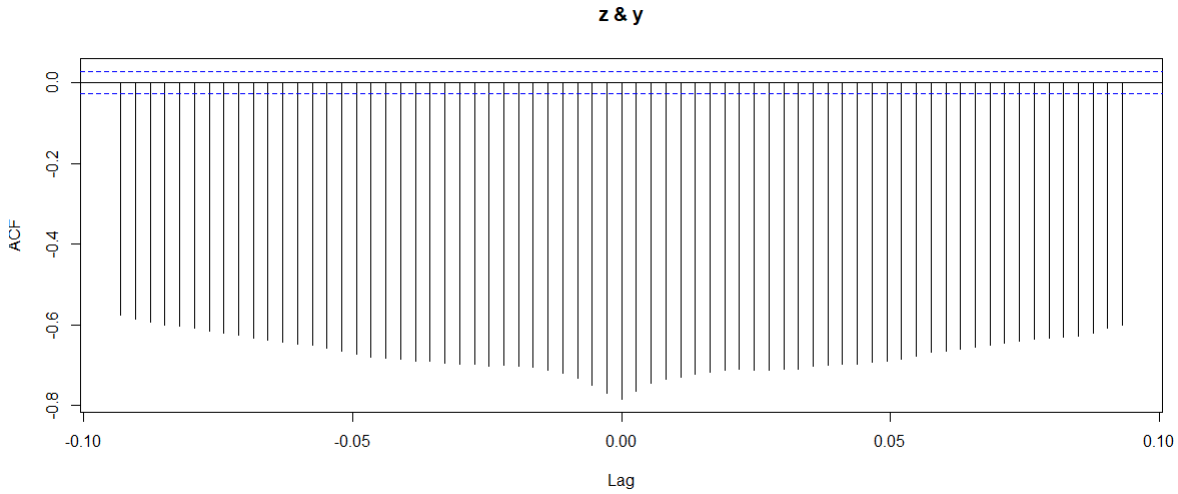


Figura 6.1: FCC temperatura diaria

6.1.2. Grafico función de correlación cruzada de temperatura con la demana horaria

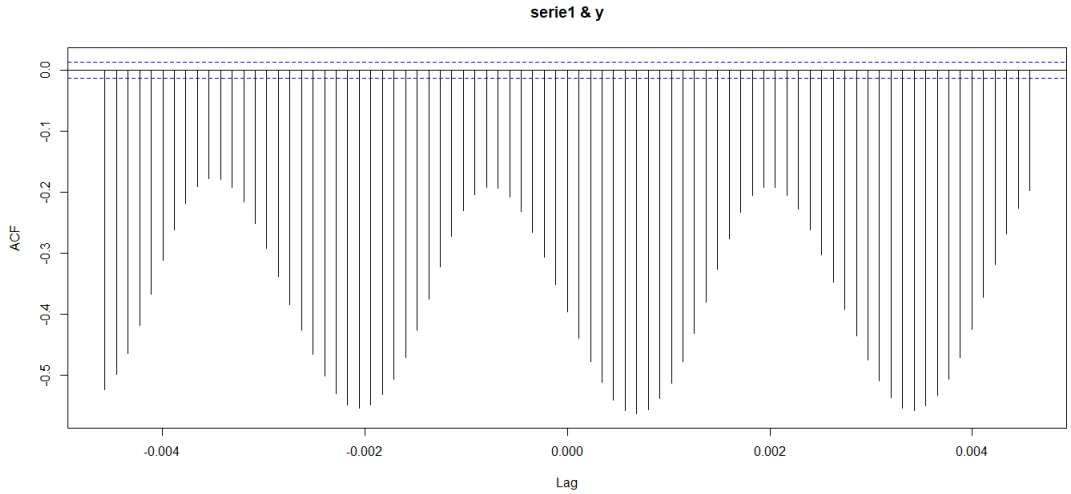


Figura 6.2: FCC temperatura hora

6.1.3. Grafico función de correlación cruzada de temperatura con la demana horaria correguido

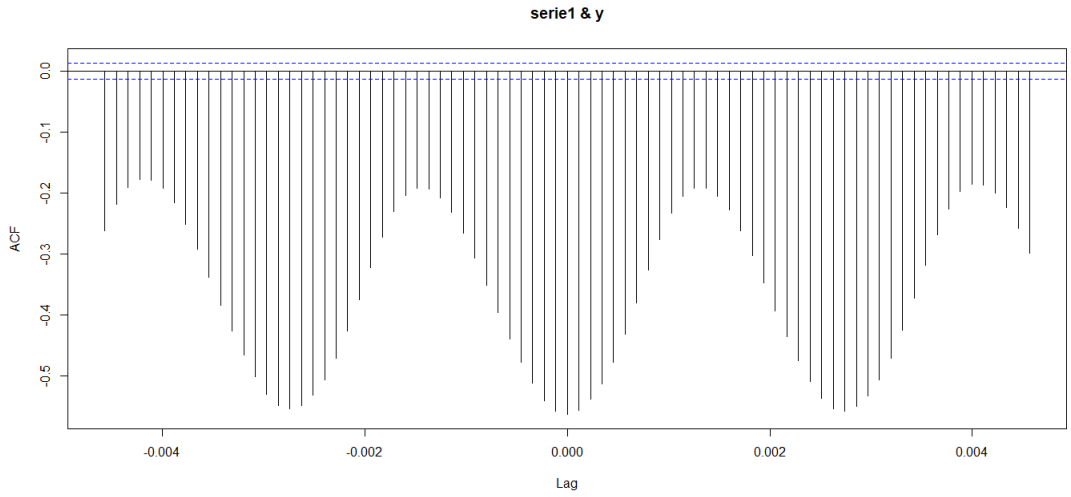


Figura 6.3: FCC temperatura hora

6.1.4. Descomposición serie horaria de la demanda de gas natural

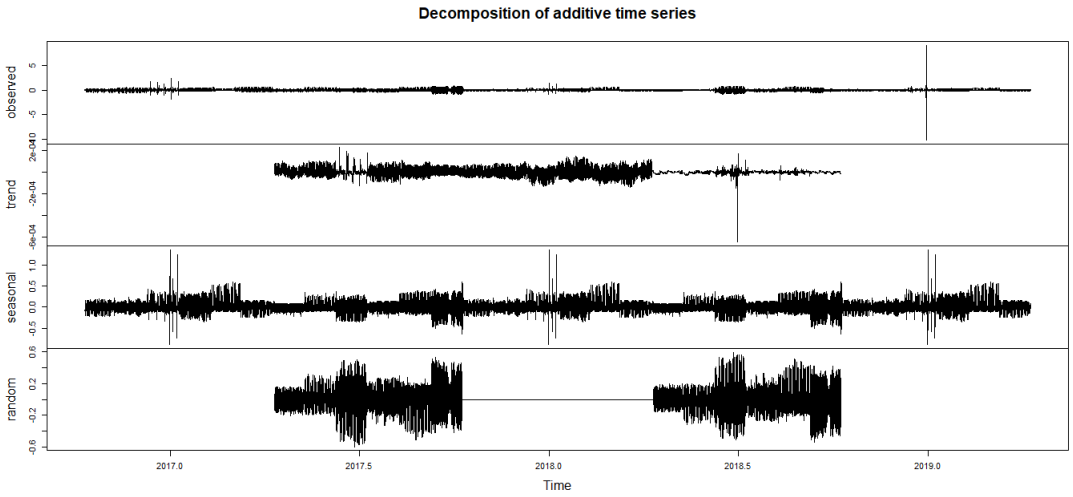


Figura 6.4: Descomposición de la demanda horaria

6.2. Modelos Arimax

6.2.1. Modelo Arimax por hora

```
Call:
arimax(x = serie1, order = c(1, 0, 0), seasonal = list(order = c(0, 0, 0), period = 8760),
    xreg = BD3, include.mean = TRUE, transform.pars = TRUE, method = "CSS-ML")

Coefficients:
         ar1  intercept  as.factor(BD$factor2)10findesemana  as.factor(BD$factor2)10habil  as.factor(BD$factor2)11feriado
      0.6678    11.8067                             0.0835                        0.1453                          0.0563
s.e.  0.0054     0.0278                             0.0297                        0.0287                          0.0417
      as.factor(BD$factor2)11findesemana  as.factor(BD$factor2)11habil  as.factor(BD$factor2)12feriado
                                  0.0406                        0.0810                          0.0184
s.e.                              0.0310                        0.0292                          0.0392
      as.factor(BD$factor2)12findesemana  as.factor(BD$factor2)12habil  as.factor(BD$factor2)1feriado
                                  0.0243                        0.0711                       -0.0660
s.e.                              0.0306                        0.0293                        0.0677
      as.factor(BD$factor2)1findesemana  as.factor(BD$factor2)1habil  as.factor(BD$factor2)2findesemana  as.factor(BD$factor2)2habil
                                -0.0003                       0.0376                         -0.0016                        0.0154
s.e.                             0.0322                       0.0298                          0.0326                        0.0300
      as.factor(BD$factor2)3feriado  as.factor(BD$factor2)3findesemana  as.factor(BD$factor2)3habil  as.factor(BD$factor2)4feriado
                             0.0150                          0.0357                        0.0525                          0.0213
s.e.                         0.0553                          0.0325                        0.0298                          0.0553
      as.factor(BD$factor2)4findesemana  as.factor(BD$factor2)4habil  as.factor(BD$factor2)5feriado
                                  0.0581                        0.0854                          0.0831
s.e.                              0.0320                        0.0300                          0.0452
      as.factor(BD$factor2)5findesemana  as.factor(BD$factor2)5habil  as.factor(BD$factor2)6feriado
                                  0.1731                        0.2055                          0.2962
s.e.                              0.0325                        0.0298                          0.0453
      as.factor(BD$factor2)6findesemana  as.factor(BD$factor2)6habil  as.factor(BD$factor2)7feriado
                                  0.2808                        0.2974                          0.2488
s.e.                              0.0322                        0.0299                          0.0518
      as.factor(BD$factor2)7findesemana  as.factor(BD$factor2)7habil  as.factor(BD$factor2)8feriado
                                  0.2658                        0.2896                          0.2423
s.e.                              0.0318                        0.0299                          0.0452
      as.factor(BD$factor2)8findesemana  as.factor(BD$factor2)8habil  as.factor(BD$factor2)9feriado
                                  0.2593                        0.2522                          0.0640
s.e.                              0.0325                        0.0298                          0.0415
      as.factor(BD$factor2)9findesemana  as.factor(BD$factor2)9habil
                                  0.1763                        0.1980  -0.0052
s.e.                              0.0317                        0.0302   0.0002

sigma^2 estimated as 0.01513:  log likelihood = 13292.41,  aic = -26510.83
```

6.2.2. Modelo Arimax por día

```
Call:
arimax(x = y, order = c(1, 0, 0), seasonal = list(order = c(0, 0, 0), period = 365),
    xreg = BD3, include.mean = TRUE, transform.pars = TRUE, method = "CSS-ML")

Coefficients:
         ar1  intercept  as.factor(BD$factor2)domingo10  as.factor(BD$factor2)domingo11  as.factor(BD$factor2)domingo12
      0.8324  363659.79                       200192.50                       113201.30                        62280.05
s.e.  0.0120   43394.36                        25872.82                        25791.63                        23854.25
      as.factor(BD$factor2)domingo2  as.factor(BD$factor2)domingo3  as.factor(BD$factor2)domingo4  as.factor(BD$factor2)domingo5
                           3032.835                       84843.78                      174193.48                      337292.60
s.e.                      21837.588                       25450.26                       27160.07                       29668.64
      as.factor(BD$factor2)domingo6  as.factor(BD$factor2)domingo7  as.factor(BD$factor2)domingo8  as.factor(BD$factor2)domingo9
                          523424.46                      526199.68                      510472.15                      317852.10
s.e.                       32122.23                       30823.53                       28721.31                       26535.74
      as.factor(BD$factor2)feriado1  as.factor(BD$factor2)feriado10  as.factor(BD$factor2)feriado11  as.factor(BD$factor2)feriado12
                          -94040.91                      315393.09                       186868.11                       45515.90
s.e.                       25262.92                       28102.32                        30877.93                       26832.97
      as.factor(BD$factor2)feriado3  as.factor(BD$factor2)feriado4  as.factor(BD$factor2)feriado5  as.factor(BD$factor2)feriado6
                          164552.9                       177801.43                      366805.31                      734169.90
s.e.                       44215.6                        33685.68                       30502.56                       35435.42
      as.factor(BD$factor2)feriado7  as.factor(BD$factor2)feriado8  as.factor(BD$factor2)feriado9  as.factor(BD$factor2)H1
                          673530.33                      629502.38                      319694.51              138014.473
s.e.                       34721.31                       31261.45                       30900.67                9740.936
      as.factor(BD$factor2)H10  as.factor(BD$factor2)H11  as.factor(BD$factor2)H12  as.factor(BD$factor2)H2  as.factor(BD$factor2)H3
                     352485.47                 253057.47                 204969.66                121175.30                  197896.5
s.e.                  24952.05                  24819.78                  22682.44                 20294.74                   24363.0
      as.factor(BD$factor2)H4  as.factor(BD$factor2)H5  as.factor(BD$factor2)H6  as.factor(BD$factor2)H7  as.factor(BD$factor2)H8
                     300849.88                 503710.66                 690445.59                725423.90                  682498.94
s.e.                  26033.65                  28843.76                  31074.27                 29999.29                   27771.78
      as.factor(BD$factor2)H9  as.factor(BD$factor2)sabado1  as.factor(BD$factor2)sabado10  as.factor(BD$factor2)sabado11
                     508942.29                    65022.038                    281029.91                  197840.38
s.e.                  25377.04                     9837.263                     25889.12                   25750.51
      as.factor(BD$factor2)sabado12  as.factor(BD$factor2)sabado2  as.factor(BD$factor2)sabado3  as.factor(BD$factor2)sabado4
                          139179.95                     49237.26                     149502.03                     243583.85
s.e.                       23729.09                     21926.44                      25514.07                      27113.46
      as.factor(BD$factor2)sabado5  as.factor(BD$factor2)sabado6  as.factor(BD$factor2)sabado7  as.factor(BD$factor2)sabado8
                          431073.82                     609913.04                     620558.99                     585367.17
s.e.                       29574.17                     32091.45                      30824.34                      28773.25
      as.factor(BD$factor2)sabado9
                          409571.56  -15601.853  1.3297
s.e.                       26467.57     348.859  0.1016

sigma^2 estimated as 6.989e+09:  log likelihood = -59772.74,  aic = 119645.5
```

6.3. Supuestos

6.3.1. Supuestos Modelo horario

```
        Box-Pierce test

data:  Mod$residual
X-squared = 0.7661, df = 1, p-value = 0.3814
```

```
> MannKendall(serie1)
tau = 0.000962, 2-sided pvalue =0.8311
```

```
        Kruskal-Wallis rank sum test

data:  serie1 by BD$RESCOMHORA
Kruskal-Wallis chi-squared = 21856, df = 21856, p-value = 0.4987
```

```
> bptest(Mod$residual~t)

        studentized Breusch-Pagan test

data:  Mod$residual ~ t
BP = 49.794, df = 1, p-value = 1.708e-12
```

```
  One Sample t-test

data:  Mod$residuals
t = -0.0030862, df = 21856, p-value = 0.9975
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.002164364  0.002157559
sample estimates:
    mean of x
-3.402499e-06
```

```
> SeasonalMannKendall(serie1)
tau = 0.114, 2-sided pvalue =< 2.22e-16
```

6.3.2. Supuestos Modelo diario

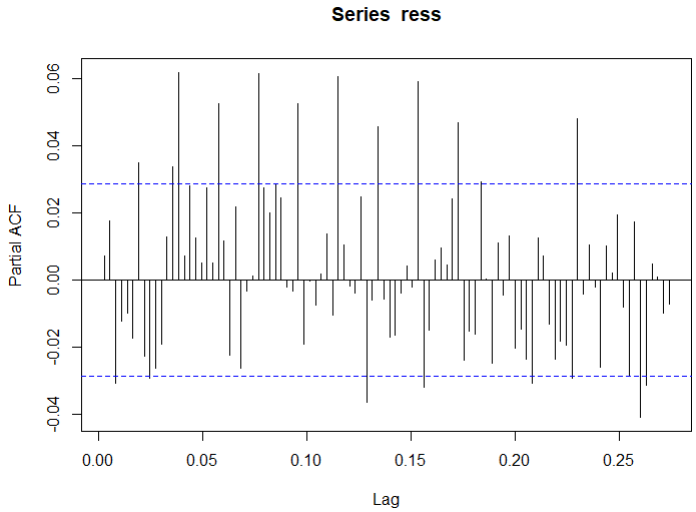
```
        studentized Breusch-Pagan test

data:  Mod$residual ~ t
BP = 0.88726, df = 1, p-value = 0.3462
```

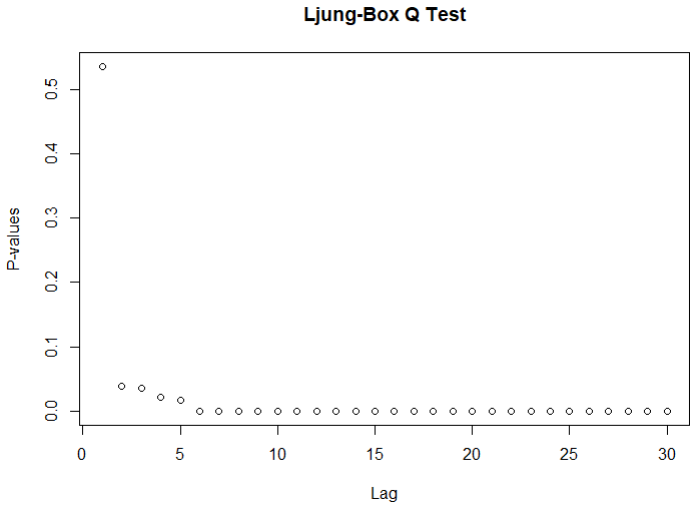
```
> MannKendall(z)
tau = 0.144, 2-sided pvalue =< 2.22e-16
> kruskal.test(z~BD$Rescom,data=BD)

        Kruskal-Wallis rank sum test

data:  z by BD$Rescom
Kruskal-Wallis chi-squared = 5051, df = 5040, p-value = 0.4538
```

```
> t.test(Mod$residuals)

        One Sample t-test

data:  Mod$residuals
t = -0.0073345, df = 5051, p-value = 0.9941
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -2765.993  2745.373
sample estimates:
mean of x
-10.30974
```

```
> SeasonalMannKendall(y)
tau = 0.413, 2-sided pvalue =< 2.22e-16
```

```
        Box-Pierce test

data:  Mod$residual
X-squared = 23.832, df = 10, p-value = 0.008059
```


Series ress
Partial ACF
0.06
0.04
0.02
0.00
-0.02
-0.04
0.00
0.05
0.10
0.15
0.20
0.25
Lag


Ljung-Box Q Test
P-values
0.5
0.4
0.3
0.2
0.1
0.0
0
5
10
15
20
25
30
Lag

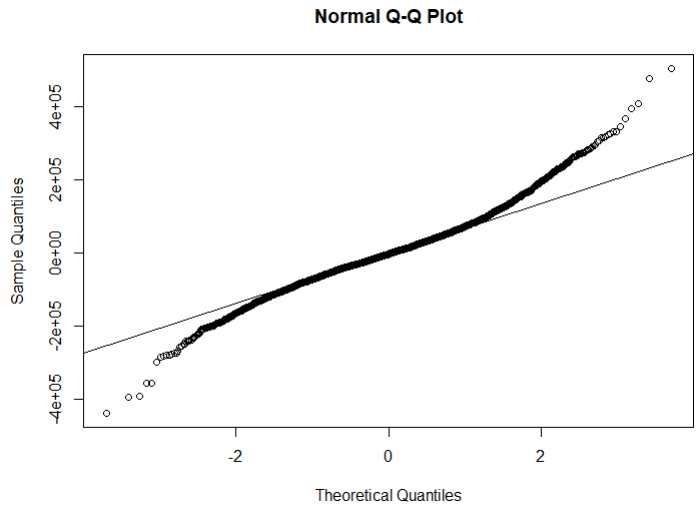
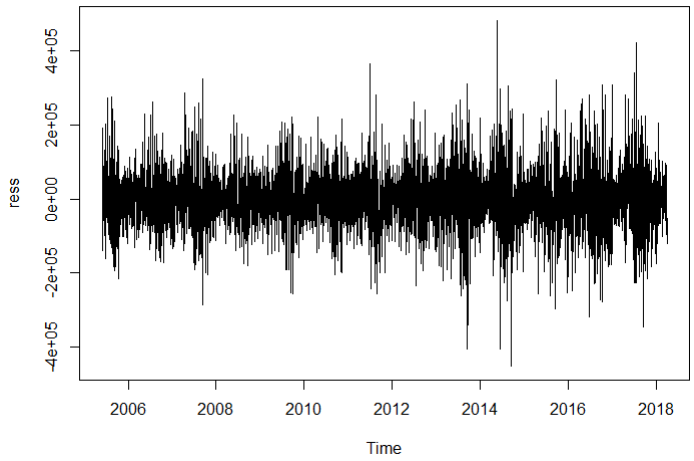
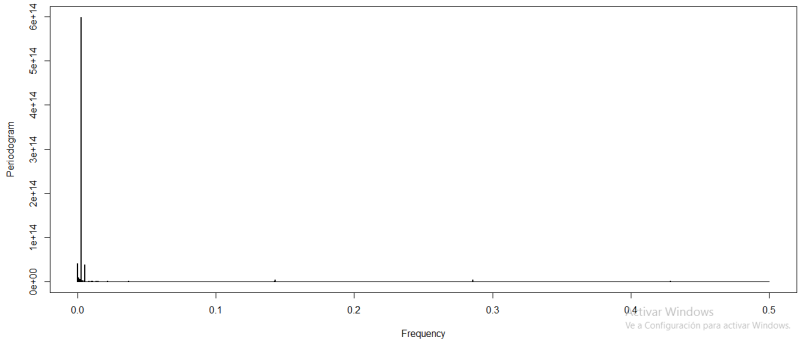

Normal Q-Q Plot
Sample Quantiles
4e+05
2e+05
0e+00
-2e+05
-4e+05
-2
0
2
Theoretical Quantiles

Gráfico residuos



6.4. Periodograma diario




```
> head(1/x$freq[order(x$spec,decreasing=T)])
[1]  365.7143 5120.0000  182.8571 2560.0000 1280.0000  512.0000
```

6.5. Script Aplicación AGESA

```
library(shiny)
shinyUI(
  pageWithSidebar(
    headerPanel("App AGESA"),
    sidebarPanel(
      p("Grafico de dispersion"),
      selectInput("x","Selecciona el eje de las x",
                                  choices = c("Fecha","NoClientes","Tmin",
                                              "Tmax","mes"))),
    mainPanel(
      tabsetPanel(
        tabPanel("Plot",
       h3(textOutput("output_text")),
       plotOutput("output_plot")
        ),
   tabPanel("Summary",verbatimTextOutput("summary")),
   tabPanel("Coef. Modelo", verbatimTextOutput("summary1")),
   tabPanel("Data Table", dataTableOutput("datatable")),
   tabPanel("Prediccion",dataTableOutput("datatable1"))
      )
    )
  )
)
```

```
library(shiny)
shinyServer(function(input, output) {
output$output_text=renderText(paste("Rescom~",input$x))
output$output_plot=renderPlot(plot(as.formula(paste("Rescom~",input$x)),
                                   data=BD))
output$summary=renderPrint({
  summary(BD)
})
output$datatable=renderDataTable({
  BD
},options = list(aLengthMenu=c(5,9,30,60,90),
                 iDisplayLenght=9)
)
output$summary1=renderPrint({
  summary(Mod)
})
output$datatable=renderDataTable({
  pred
},options = list(aLengthMenu=c(7,14,28,30,31),
                 iDisplayLenght=9)
)
  })
```