UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENÍERA MATEMÁTICA

ANÁLISIS ESTADÍSTICO DE DATOS GEO-REFERENCIADOS: ESTUDIO DE EFECTOS AMBIENTALES EN ASMA Y NEUMONÍA EN CHILE

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL MATEMÁTICO

JOSÉ MANUEL CERECEDA CÁCERES

PROFESOR GUÍA:
RODRIGO ASSAR CUEVAS

MIEMBROS DE LA COMISIÓN:
AXEL OSSES ALVARADO
JORGE AMAYA

SANTIAGO DE CHILE
ABRIL 2016

Resumen

El presente estudio tiene como objetivo principal la construcción de una plataforma GIS (Sistema de Información Geo-referenciada) para el análisis estadístico de procesos temporales y geo-referenciados, dicha plataforma consiste en la visualización geográfica de una zona de estudio, en la cual se pueden realizar análisis estadísticos de variables asociadas a la región. Asimismo, se describe y profundiza el modero *ARIMA*, integrado como metodología estadística en dicha plataforma. Para la realización del estudio se seleccionaron cinco comunas de la región Metropolitana, Santiago de Chile, incluyendo variables clínicas y ambientales. Datadas en orden temporal, los ingresos hospitalarios por asma y neumonía quedan contrastados respecto a condiciones ambientales de Ozono, material particulado y precipitaciones.

Como herramientas para la construcción de la plataforma, se utilizaron los lenguajes `R` y `PHP`, como interfase de esta se construye un mapa interactivo. Con ello provee a un usuario, con o sin conocimientos previos en estadística, información exploratoria de los datos almacenados y resultados estadísticos que comparan o relacionan variables dentro de la región. El mapa interactivo, contiene la distribución geográfica de la región y distribución de estaciones meteorológicas que pueden ser seleccionadas para obtener información, así como la posibilidad de realizar interacciones usuario-plataforma, que mediante acciones del usuario producen análisis y gráficas mostradas en el GIS. Las aplicaciones GIS suelen ser utilizadas en distintas disciplinas e investigaciones científicas, por ejemplo, pueden ser utilizadas para la gestión de los recursos, para la evaluación del impacto ambiental, para la planificación urbana, en la sociología, entre otros.

Dentro de los principales aportes de este trabajo de titulo, cabe mencionar que el formato de la plataforma GIS construida facilita la incorporación de aspectos socio-culturales, económicos y ambientales que contribuyen a la toma de decisiones de una manera más eficaz. Para futuras investigaciones, se propone implementar el modelo *ARIMA* para series bivariadas en `R`, lo que permitiría realizar un Test de independencia dos de series temporales.

Palabras claves: Series temporales; ARIMA; GIS.

Abstract

The main objetive of this work is to create a web application of interactive Geografic information sistema (GIS), thais application is able to do statistical analysis of temporal series and geographic data bases. On this platform the ARIMA model is described and deepened, integrated as statistical methodology. For this study we selected five 'counties' in Santiago, Chile. Clinical and environmental temporal series were included on database, particularly hospital incomes (asthma and pneumonia) are contrasted with environmental variables (ozono, particular matherial and precipitations).

R and PHP languages were used to built the application. An interactive map is constructed as platform interface, this provides to the user exploratory variables stored in database and is easy to understand even for users without knowledge in statistics. The interactive map contains the geographic distribution of regions and the distribution of meteorological stations that can be selected to obtain charts and statistical analysis. GIS applications are often used in different disciplines and scientific research, for example, can be used for resource management for the environmental impact assessment for urban planning, sociology, among others.

Among the main contributions of this work, the format of GIS platform built facilitates the incorporation of social, cultural, economic and environmental aspects, this set of variables contributes to making decision. For future research, it is proposed to implement the ARIMA model for bivariate series in R, which allow for a test of independence two time series.

Key words: Temporal series, ARIMA, GIS.

Para mis padres Rosalba y Victor, mis hermanas Victoria, Valentina y Nathaly.

Thanks

Agradezco a Camila Martínez, por el acompañamiento y apoyo en este proceso, sin ella esto no sería posible, a Rodrigo Assar que me condujo y permitió que este trabajo fuese posible, y a todas las personas con las que tuve la oportunidad de cruzarme durante mi camino en la facultad.

Tabla de contenido

Introducción

La presente investigación tiene dos principales objetivos; por una parte, la elaboración de una plataforma interactiva para el análisis estadístico sobre procesos temporales y geo-referenciados, y por otra, la descripción de el modelo *ARIMA* para series temporales.

En primera instancia, se confecciona una plataforma interactiva capaz de generar la visualización tanto geográfica como de resultados estadísticos (gráficas y análisis de modelos), que pueda ser comprendida por un usuario sin conocimientos exhaustos sobre estadística. En segunda instancia, se estudia en profundidad el modelo *ARIMA* y su aplicación a series temporales, con el fin de proporcionar herramientas y análisis a la plataforma construida.

En relación al aspecto metodológico, se explican los distintos modelos estadísticos sobre procesos temporales que se utilizan en la plataforma y que contribuyen al entendimiento y proyecciones de los sucesos. Asimismo, se explicitan los análisis estadísticos más tradicionales, como modelos lineales y descomposición clásica de una serie temporal, y profundizase en el modelo *ARIMA*. Finalmente, se describe la construcción de la plataforma interactiva, en términos de los lenguajes utilizados (especialmente `PHP` y `R` [5]), librerías citadas y su estructura.

Con respecto a la información obtenida para realizar la investigación, se obtuvieron variables correspondientes a ingresos hospitalarios y condiciones ambientales de la Región Metropolitana, Santiago de Chile. Más específicamente, la investigación se limita a la comprensión y análisis sobre las siguientes variables: Asma, Neumonía (enfermedades), ozono, material particulado (contaminación), en las comunas de Las Condes, La Reina, La Florida, Independencia, Cerrillos, Pudahuel, y precipitaciones en estaciones meteorológicas de Chile (en particular de la región Metropolitana). Todas las variables en cuestión están datadas en un orden temporal, razón por la cual se escogieron los modelos estadísticos mencionados anteriormente para poder analizarlas.

Con los datos obtenidos, además de una componente temporal contienen una componente geográfica, se construyó una plataforma GIS (Sistema de información geo-referenciada) interactiva. Con una presentación atractiva y sencilla para mostrar los análisis realizados. La plataforma GIS es una aplicación que permite observar la zona geográfica en estudio y está enfocada a que un usuario sin conocimientos científicos previos, pueda realizar distintas acciones con las variables en juego y finalmente poder adquirir información relevante a partir

de ellas.

Durante el desarrollo de esta investigación se utilizó el software estadístico `R`, integrado a la plataforma interactiva via `PHP` permitió realizar cálculos estadísticos y visualizaciones gráficas, así como realizar simulaciones y aplicaciones en el contexto del modelo *ARIMA*.

En el primer capítulo de este trabajo de titulo se introducen y describen brevemente conceptos y lenguaje básico utilizado. En el segundo capítulo se describe en primer lugar la parametrización para un proceso temporal $\{X_t\}_t$ según el modelo *ARIMA* y, en segundo lugar los materiales para la construcción de la plataforma interactiva, lenguajes y datos utilizados. En el capítulo tercero, se describe el ajuste paramétrico mencionado en el capitulo anterior. Finalmente, en el capítulo cuatro se muestra, por un lado, el resultado de la construcción de la plataforma GIS, y por otro, las parametrizaciones y predicciones sobre las series temporales en el contexto del modelo *ARIMA*.

Este trabajo de titulo fue realizado íntegramente en Assar-Lab, y la plataforma interactiva se encuentra en `http://www.assar-lab.cl`.

Alcances

El conjunto de la plataforma interactiva busca ser una herramienta estadística para instituciones o empresas que puedan adquirir y/o exponer resultados estadísticos y/o gráficos sobre datos de interés, establecidos en una región geográfica. En otras palabras, esta plataforma busca ser un servicio web para análisis estadístico de datos geo-referenciados.

Cobra importancia que la aplicación GIS, que contiene en un inicio objetos con una dimensión geográfica (regiones, edificios, ríos, lagos), que exhibe variables que contienen además una dimensión temporal.

En el contexto de estudio sobre la dinámica de un sistema, es natural preguntarse acerca de la dependencia del estado actual respecto a instantes anteriores. El modelo *ARIMA* provee una aproximación probabilista para el pronóstico de series temporales basado en los momentos previos.

Visualización y geo-referenciación de la región

A nivel mundial distintas instituciones contribuyen con estudios y generación de conocimiento en el área de salud y calidad de vida, las cuales han utilizado GIS dentro de sus servicios, donde la estadística contribuye al estudio y entendimiento de procesos sociales y ambientales, por ejemplo en consumo de alcohol [8], o en datos epidemiológicos, que también es integrado con otras variables socio económicas, tales como la edad de la población [9]. Como se ha evidenciado en numerosos trabajos, estas herramientas contribuyen de manera

significativa a gobiernos locales o nacionales [10].

De acuerdo a las condiciones de Chile, existen iniciativas incipientes en el desarrollo de herramientas GIS, tales como aplicaciones web destinadas a la búsqueda de depósitos mineros, aplicaciones acerca de la distribución del virus VIH, como puede verse en el sitio `http://www.ArcGIS.com` (ESRI) y también en `http://www.deis.cl` (MINSAL), el servicio de información pública acerca de la localización exacta, fecha y grado de sismos como puede verse en *Centro Sismológico Nacional, Universidad de Chile* (`http://www.sismologia.cl`) y esfuerzos que contribuyen con modelos aplicados a la hidrodinámica y dispersión de contaminantes, como puede es el caso de *Modelación Ambiental* (`http://www.modelacion.cl`).

Entre la amplia gama de aplicaciones mencionadas, esta plataforma, escrita en lenguaje de programación `PHP`, tiene por objetivo ser utilizado como aplicación web, para usuarios con y sin conocimientos estadísticos previos.

Capítulo 1

Preliminares

En este capítulo se describen conceptos básicos para que el lector identifique el lenguaje utilizado durante los siguientes capítulos, las descripciones están basadas en la referencia [4]. Para facilitar la lectura de los capítulos posteriores dedicados a series temporales, se menciona e introduce el contraste de hipótesis sobre una población o variable.

Del mismo modo, se mencionan e introducen los lenguajes involucrados dentro de la plataforma interactiva y las librerías que contribuyen a su materialización.

1.1. Conceptos básicos

En esta primera sección se describen definiciones para realizar análisis estadístico. Se introducen los conceptos de variable aleatoria necesarios para definir formalmente una serie temporal. Así mismo se introduce el contraste de hipótesis utilizado para poder establecer algún grado de veracidad sobre una hipótesis.

1.1.1. Espacio muestral, $\sigma-$ álgebra, medida

El conjunto de elementos sobre el cual se desea conocer u obtener conclusiones (universo o población), habitualmente es de gran tamaño y no se cuentan con los recursos suficientes para obtener la información de todo el conjunto. Es por esto que se hace necesario extraer un subconjunto (muestra) de casos de dicha población, es decir, la información que otorga la muestra es parcial. Teniendo presente este contexto, se presentarán algunas definiciones que formalizan la idea de muestra y variables sobre ella.

En primer lugar, frente a un experimento aleatorio se construirá el conjunto de todos los eventos posibles que se pueden obtener, al que se denotará por Ω. Se considerará, una familia de subconjuntos del espacio muestral Ω, que en adelante se denotará por $\mathcal{F}$, es decir, que si

$A \in \mathcal{F} \Rightarrow A \subset \Omega$

Definición 1.1 *(***Espacio muestral)** *Se define al espacio muestral Ω como el conjunto de todos los posibles resultados individuales de un experimento aleatorio.*

Definición 1.2 *(*$\sigma-$ **álgebra)** *Dado un espacio Ω, la familia de subconjuntos de Ω, $\mathcal{F}$ se dirá $\sigma-$ algebra si cumple que:*

- *El conjunto vacío está en $\mathcal{F}$ ($\phi \in \mathcal{F}$).*
- *El complemento de todos los conjuntos en $\mathcal{F}$, también se encuentran en $\mathcal{F}$, es decir, si $E \in \mathcal{F} \Rightarrow E^c \in \mathcal{F}$.*
- *Si $A_1, A_2, ...$ es una sucesión contable de conjuntos en $\mathcal{F}$, entonces la unión contable $\bigcup_{i \in \mathbb{N}} A_i$ también se encuentra en $\mathcal{F}$.*

Definición 1.3 *(***Medida)** *Consideraremos la aplicación $\nu : \mathcal{F} \to \mathbb{R}^+$, ν será medida si cumple que*

- $\nu(\phi) = 0$.
- *Si $A, B \in \mathcal{F}$ se tiene que $\nu(A \cup B) \leq \nu(A) + \nu(B)$*
- *Si $A_1, A_2, ...$ es una sucesión contable de conjuntos disjuntos en la $\sigma-$ álgebra $\mathcal{F}$, entonces $\nu(\bigcup_{i \in \mathbb{N}} A_i) = \sum_{i \in \mathbb{N}} \nu(A_i)$*

Si $\nu \in [0,1]$ entonces diremos que ν es medida de probabilidad.

Observación Se llamará espacio medible (o de probabilidad si $\nu \in [0,1]$) a la tripleta $(\Omega, \mathcal{F}, \nu)$

1.1.2. Variable aleatoria

Dentro del estudio sobre una población, existen variables que desean ser medidas pero que en la realidad no es posible conocer con certeza el valor que tomarán. Basta comprender la aleatoriedad producida por el lanzamiento de una moneda; se sabe que puede puede resultar cara o sello, pero no se sabe con certeza cual de los dos sucesos ocurrirá.

Definición 1.4 *(***Variable aleatoria)***Dado un espacio de probabilidad $(\Omega, \mathcal{F}, \mathbb{P})$ y un espacio medible (S, Σ), $X : \Omega \to S$ es una variables aleatoria si es una aplicación, $\mathcal{A} - \Sigma$ medible.*

Observación Si $S = \mathbb{R}$, se dría que X es una variable aleatoria continua, mientras que si $S = \mathbb{N}$ se dirá que X es variable aleatoria discreta.

Definición 1.5 *(***Serie temporal)** *Se considerará al conjunto T dotado de un orden total, al que se identifica con una variable temporal (tiempo). El conjunto $\{X_t\}_{t \in T}$ de variables aleatorias indexadas por T, se llamará serie temporal.*

Además se denotara por $\{x_t\}_{t\in\{1,...,n\}}$ *a la muestra observada de* $\{X_t\}_{t\in T}$.

1.1.3. Descripción paramétrica de una distribución

En lo que sigue, considérese una variable aleatoria X. Se describe su función de distribución F, como

$$F_X(x) = F(x) = \mathbb{P}\left(\{\omega : X(\omega) \leq x\}\right)$$

Considérese además un conjunto de muestras aleatorias simple $\Omega = \{x_1, x_2, ..., x_n\}$ de X tales que $x_1 \leq x_2 \leq ... \leq x_n$. Para estudiar la variable aleatoria se consideran los siguientes estadísticos con sus respectivos estimadores.

1. Cuantil: Dada una función de distribución $F(x)$, x_p es cuantil de orden p de F, con $p \in [0,1]$ si $F(x_p) = P(X \leq x_p) = p$.

   Suelen usarse en grupos que dividen la distribución en iguales cantidades. Por ejemplo, cuartiles que dividen la distribución en 1/4,1/2,3/4

2. Valor esperado y media ponderada: El valor esperado de una variable aleatoria corresponde a $\mu = \mathbb{E}(X)$. Se define la media de X como

   $$\overline{X} = \frac{1}{n}\sum_{\mathrm{i}=1}^{n} x_{\mathrm{i}}$$

   Notar que $\mathbb{E}(\overline{X}) = \mathbb{E}(X) = \mu$.

   Definición 1.6 *(***Estimador insesgado)** *Un estimador* $\hat{h}$ *de un parametro* h *se dirá insesgado si* $\mathbb{E}(\hat{h}) = h$

3. Varianza de X y Desviación estándar: La varianza de X está definida por

   $$\sigma^2 = Var(X) = \mathbb{E}\{(X - \mathbb{E}(X))^2\}$$

   considérese la varianza muestral como

   $$s^2 = \frac{1}{n}\sum_{\mathrm{i}=1}^{n}(X_{\mathrm{i}} - \overline{X})^2$$

   estimador de σ^2, se puede mostrar que $\mathbb{E}(s^2) = \frac{n-1}{n}\sigma^2$

   Definición 1.7 *(Asintóticamente insesgado) Un estimador* $\hat{h}$ *de un parámetro* h *se dirá asintóticamente insesgado si,* $\mathbb{E}(\hat{h}) \to h$ *si* $n \to \infty$.

mientras que el estimador corregido de σ^2 dado por

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})^2$$

es insesgado, es decir $\mathbb{E}(\hat{\sigma}^2) = \sigma^2$

1.1.4. Representaciones gráficas

Es natural al momento de realizar un muestreo sobre una población, obtener una primera inferencia sobre aquella, de manera resumida y fácil de interpretar. A continuación, se mencionarán los métodos y gráficos para la obtención de información cualitativa en primera instancia.

En rigor, antes de realizar algún test estadístico específico, se realiza una primera inferencia a través de herramientas gráficas para establecer si existe semejanza con alguna distribución teórica conocida. Es por ésto, que nacen gráficos para poder establecer a priori cómo se comporta la muestra. Considere en lo que sigue, la variable aleatoria X y un conjunto de muestras aleatorias simple $\Omega = \{x_1, x_2, ..., x_n\}$ de X tales que $x_1 \leq x_2 \leq ... \leq x_n$. Se define la distribución empírica $\hat{F}$ como $\mathbb{P}(\text{escoger } x_i) = \frac{1}{n}$, de la siguiente forma

$$\hat{F}(x) = \frac{\#\{x_i \leq x\}}{n}$$

Así, la primera gráfica que entrega inferencia acerca de la distribución empírica es el histograma.

1. Histograma: Representación gráfica de una variable, la cual indica frecuencia de los valores representados.El histograma entrega una primera aproximación a la distribución de la variable.

2. Diagrama de caja (boxplot): Diagrama basado en cuartiles de una muestra. Generalmente, se visualiza el valor mínimo, máximo, los cuartiles y valores atípicos.

3. Gráfico Q-Q: Se conoce como el método gráfico de comparar una distribución de probabilidad de una muestra aleatoria, con una distribución teórica con la cual desea ser comparada.

1.1.5. Relaciones entre variables

Considere un conjunto de variables aleatorias $(Y, X_1, X_2, \ldots, X_k)$, con muestra aleatoria simple $(y_i, x_1^i, x_2^i, \ldots, x_k^i)_{i=1}^n$,

Una de las preguntas frecuentes dentro de una muestra multivariada $(Y, X_1, \ldots, X_k)$, es si existe alguna relación entre ellas. Considérese por ejemplo el hecho de describir Y en función

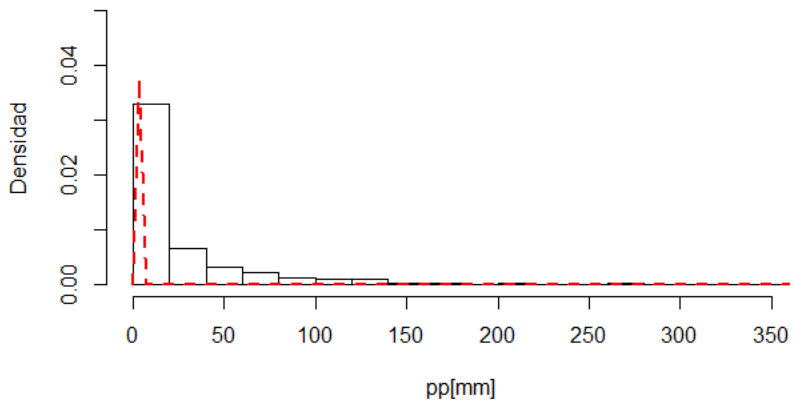

Figura 1.1: Histograma que indica la frecuencia de milímetros de precipitación caida mensualmente en la estación *Terraza oficinas centrales DGA*, la linea roja representa una distribución *Gamma(2,2)* que aproxima la frecuencia empírica entregada por la estación

de $(X_1, \ldots, X_k)$, es decir, si existe f tal que

$$Y = f(X_1, \ldots, X_k) \tag{1.1}$$

Para ésto, se desarrollan modelos paramétricos, los cuales se mencionarán a continuación, dando una breve explicación de relaciones utilizadas en estadística.

Definición 1.8 *(***Modelo lineal)** *Un modelo lineal busca explicar la variable dependiente Y de forma lineal con respecto a las variables explicativas $X_1, X_2, ..., X_k$ típicamente continuas*

$$Y = \alpha + \sum_{i=1}^{k} \beta_i X_i + \varepsilon \tag{1.2}$$

donde β_i son los parámetros del modelo y ε una perturbación aleatoria. En particular puede considerarse $\varepsilon \sim N(0, \sigma^2)$.

Definición 1.9 *(***Modelo exponencial)** *Otro modelo que se puede encontrar en distintos experimentos, es aquel que está caracterizado por la relación exponencial de las variables, es decir, dado por la función*

$$Y = a + \exp\{b_1 X_1 + b_2 X_2 + \ldots b_k X_k\} \tag{1.3}$$

Diagrama mensual para la descripción de la climatología histórica de Santiago

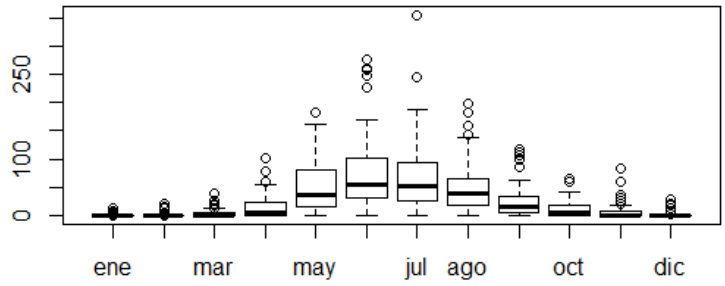

Figura 1.2: Diagrama de caja mensual para los datos de precipitación, estación Quinta Normal, Santiago de Chile. Altura 527 msnm.

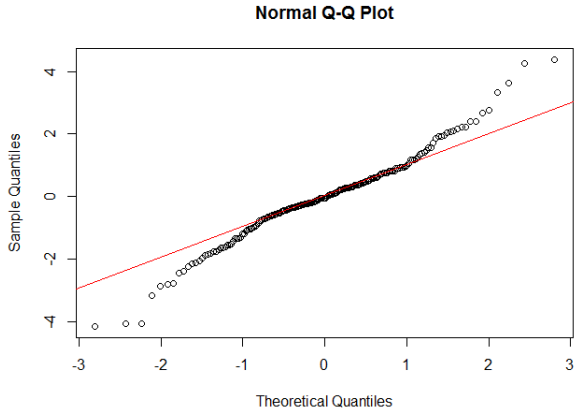

Figura 1.3: Gráfico que compara los cuantiles de la distribución normal, comparados con la generación aleatoria de 200 sucesos que siguen una distribución t de Student de 5 grados de libertad

Estimadores

Una vez fijado el modelo por el cual se desea un grupo de variables, se plantea encontrar el mejor ajuste perimétrico para que la curva coincida con los datos. A continuación se mencionan dos métodos para encontrar los parámetros.

Definición 1.10 *(***Estimador de máxima verosimilitud)** *Método para ajustar un modelo y sus parámetros. Si* X_{i} *siguen una distribución con densidad* f*, entonces la densidad*

conjunta

$$f(x_1, \ldots, x_k|\theta) = f(x_1|\theta) \cdots f(x_k|\theta)$$

con θ parámetros de la distribución, Luego se define la función de verosimilitud

$$L(\theta|x_1, \ldots, x_k) = f(x_1|\theta) \cdots f(x_k|\theta)$$

función de θ con la muestra $(x_1, \ldots, x_k)$ dada. Luego el estimador de máxima verosimilitud es aquel que maximiza $L(\theta|x_1, \ldots, x_k)$, es decir

$$\hat{\theta} = \underset{\theta \epsilon \Theta}{\operatorname{argmax}} L(\theta|x_1, \ldots, x_k)$$

En la práctica, de manera equivalente se plantea maximizar la función objetivo

$$l(\theta|x_1, \ldots, x_k) = \log\{L(\theta|x_1, \ldots, x_k)\} = \sum_{\mathrm{i}=1}^{k} \log\{f_{\mathrm{i}}(x_{\mathrm{i}}|\theta)\}$$

Definición 1.11 *(***Método mínimos cuadrados)** *El objetivo del método mínimos cuadrados es encontrar el estimador de parámetros para una relación que se desee afirmar. Consiste en minimizar la función objetivo*

$$\varepsilon = \sum_{\mathrm{i}=1}^{n} (y_{\mathrm{i}} - f(x_1^{\mathrm{i}}, x_2^{\mathrm{i}}, \ldots, x_k^{\mathrm{i}}))^2$$

explícitamente para la ecuación (1.2) *se tiene que*

$$\varepsilon = \sum_{\mathrm{i}=1}^{n} (y_{\mathrm{i}} - \alpha - \beta_1 x_1^{\mathrm{i}} - \beta_2 x_2^{\mathrm{i}} - \ldots - \beta_k x_k^{\mathrm{i}}))^2$$

luego , y en general si la función f cumple las condiciones, para obtener los parámetros del modelo se analiza el problema

$$\min_{\alpha_1, \beta_{\mathrm{i}}} \varepsilon^2 = \sum_{\mathrm{i}=1}^{n} (y_{\mathrm{i}} - \alpha - \beta_1 x_1^{\mathrm{i}} - \beta_2 x_2^{\mathrm{i}} - \ldots - \beta_k x_k^{\mathrm{i}})^2$$

Observación Cabe mencionar, que en el caso del modelo lineal las variables Y_{i} sean independientes y los errores $\varepsilon_{\mathrm{i}} \sim N(0, \sigma^2)$ el estimador por mínimos cuadrados coincide con el estimador de máxima verosimilitud.

1.2. Test de hipótesis

Una vez realizada una inferencia cualitativa de una muestra aleatoria, de manera similar, se busca establecer una inferencia más formal que permita cuantificar de manera fidedigna una primera inferencia o simplemente la intuición.

En concreto, para poder establecer alguna aseveración respecto al experimento en cuestión, se plantea, por una parte, una hipótesis de nulo efecto sobre la población (que se llamará hipótesis nula o H_0) y por otra, una hipótesis con la cual se desea contrastar la H_0, llamada hipótesis alternativa o H_1.

Dado que sobre el experimento se conoce una parte de la población total, es decir, una muestra, no se puede asegurar con certeza si la hipótesis nula (o alternativa) es verdadera, sin embargo, se puede establecer una regla de decisión para asegurar con algún grado de significancia, qué hipótesis considerar verdadera a partir de la muestra obtenida.

Definición 1.12 *(***Error tipo 1)** *Se llama error de tipo 1 o falso positivo al evento de rechazar la hipótesis nula H_0 dado que H_0 es verdadera. De otra forma, podemos llamar al error de tipo uno al evento $\{Escoger\, H_1 | H_0\, es\, verdadero\}$*

Definición 1.13 *(***Error tipo 2)** *De manera similar a la definición anterior, se llama error de tipo 2 o falso negativo al evento de aceptar H_0 dado que H_1 es verdadero.*

$$\mathbb{P}(Escoger\, H_1 \,|\, H_0\, es\, cierto) = \alpha$$
$$\mathbb{P}(Escoger\, H_0 \,|\, H_1\, es\, cierto) = \beta$$

Según las definiciones de errores anteriores, se rechazará la hipótesis nula H_O si $\alpha \geq \alpha_0$ para α_0 fijo. Convencionalmente, $\alpha_0 = 0.01$, 0.05 ó 0.1, que corresponde a la tolerancia que se permite al error.

En general, a partir de la muestra de la población, se extrae un estadístico (un valor que es función de la muestra) cuya distribución de probabilidad esté relacionada con la hipótesis. Se considera como *región de rechazo*, al conjunto de valores que es más improbable bajo la hipótesis, esto es, el conjunto de valores para el que rechazaremos la hipótesis nula si el valor del estadístico observado entra dentro de él.

1.2.1. Prueba de Shapiro-Wilk

La prueba de Shapiro-Wilk contrasta sobre el conjunto $(x_1, \ldots, x_n)$, si la v.a. es distribuida de forma normal. Es decir

$$H_0 : (X_\mathrm{i}) \sim N(\mu, \sigma^2) \qquad \forall \mathrm{i} \in \{1, \ldots, n\}$$
$$H_1 : \sim H_0$$

Para algún μ, σ, sin embargo, esta prueba no provee información acerca de μ, σ. Para esto se pueden utilizar otras pruebas, como el caso de la prueba *t de Student* que asume normalidad en la v.a. que se describe a continuación [6].

1.2.2. Prueba t de Student

Sea X variable aleatoria que distribuye como $N(\mu_0, \sigma_0^2)$ donde el estadístico construido sigue una distribución t de Student bajo la hipótesis nula. Particularmente el test es utilizado para hipótesis del estilo, para dos poblaciones distintas con medias iguales o para poblaciones en donde la media es igual algún valor específico.

Por ejemplo, se plantea como hipótesis que una población tiene esperanza de vida (X v.a. $N(\mu_0, \sigma_0^2)$) μ, entonces formalmente se plantea el test

$$H_0 : \mu_0 = \mu$$
$$H_1 : \mu_0 \neq \mu$$

mientras que el estadístico construido es $t = \dfrac{\overline{X} - \mu}{s/\sqrt{n}}$, con s desviación estándar muestral, n la cantidad de muestras.

1.2.3. Prueba de Wilcoxon

La prueba *t de Student* asume normalidad en la v.a., sin embargo no se puede asumir siempre esa hipótesis en un conjunto de v.a.

La *prueba de Wilcoxon* es una prueba paramétrica utilizado en el contraste de medianas en dos conjuntos, $(x_1, \ldots, x_n)$ y $(y_1, \ldots, y_n)$. Las suposiciones sobre esta prueba, más débiles que las anteriores, son

1. $z_i = x_i - y_i$ son independientes.
2. z_i tienen las misma distribución continua y simétrica con respecto a una mediana θ.

Luego el contraste de hipótesis plantea que

$$H_0 : \theta = 0$$
$$H_1 : \theta \neq 0$$

Esta prueba se utiliza como alternativa a la prueba *t de student* en el caso que no se pueda asumir normalidad en el conjunto $(x_1, \ldots, x_n)$ [7].

1.2.4. Análisis de varianza

El análisis de varianza (ANOVA por sus siglas en ingles) permite determinar si diferentes tratamientos, muestran diferencias significativas o por el contrario, puede suponerse que sus medias poblacionales no difieren significativamente.

Supongase la variable y_{ij} que indica el valor j–ésimo del tratamiento i–ésimo que puede ser explicada según la ecuación

$$y_{ij} = \underbrace{\mu + x_i}_{y_i} + \varepsilon_{ij} \tag{1.4}$$

donde x_i es la variable explicativa del i–ésimo tratamiento, ε_{ij} corresponde a un error aleatorio, es decir, la variable y_{ij} se puede explicar según una componente esperada del tratamiento i, y_i más un error aleatorio. Entonces, se considera el siguiente desarrollo sobre la ecuación (1.4).

$$\begin{aligned} y_{ij} &= y_i + \varepsilon_{ij} \\ \sum_{ij} \{y_{ij} - \overline{y}\}^2 &= n \sum_{i} \{y_i - \overline{y_i}\}^2 + \sum_{i,j} \{\varepsilon_{ij} - \overline{y_i}\}^2 \\ s_{tot} &= s_{fact} + s_{error} \end{aligned}$$

Luego, resumiendo esto en una tabla ANOVA, se tiene que

Fuente de variación	suma de cuadrados	g.l.	Cuadrado medio	F
Intergrupo	s_{fact}	$t-1$	$T = \frac{s_{fact}}{t-1}$	$F = \frac{T}{E}$
Intragrupo	s_{error}	$N-t$	$E = \frac{s_{erros}}{N-t}$	
Total	s_{tot}	$N-1$		

Cuadro 1.1: Tabla de resumen estandar para el análisis de varianza (ANOVA)

Donde F es el estadístico, que bajo la hipótesis nula H_0 (convencionalmente H_0 = los grupos son indiferentes entre si), F sigue una distribución F–Fisher de $t-1, N-t$ grados de libertad, luego la región de rechazo será dada por $F \geq F_{gl,\alpha}$ donde $F_{gl,\alpha}$ indica el α–cuantil de la distribución.

1.3. Programación y lenguajes

Para la construcción de la plataforma GIS interactiva, se utilizó el lenguaje de programación `PHP`, ejecutando lineas de código escrito en el lenguaje de programación `R` en un servidor externo. Se mencionará en esta sección, una descripción de los lenguajes utilizados para luego, en el capitulo tres, describir la estructura y construcción de la plataforma.

1. `PHP` (Hypertext pre-processor) es un lenguaje de programación originalmente diseñado para el desarrollo web de contenido dinámico.
2. `R` es un lenguaje de programación orientado a objetos, que proporciona herramientas estadísticas y gráficas para el análisis de datos.

3. Se utilizó librerías JS de código abierto, tales como bootstrap, leaflet, jQuery. Para la confección de la plataforma en cuestión, se utilizó fuertemente la librería Leaflet JS que provee, mediante lineas de código, la visualización e interactividad de una región determinada.

Capítulo 2

Métodos y materiales

En el presente capítulo se describe el método particular *ARIMA* para el análisis y pronóstico de series temporales. Esta metodología está basada en las referencias [2],[1]. Bajo este contexto, se considera que la serie $\{Y_t\}_t$ puede ser descompuesta de la forma

$$Y_t = m_t + s_t + X_t$$

donde m_t y s_t son componentes deterministas que representan la tendencia y estacionalidad. Se introducirá el concepto de serie estacionaria para X_t, y se explicará el ajuste de parámetros de $\{Y_t\}_t$ para su aproximación según el modelo *ARIMA*.

El pronóstico y el ajuste de parámetros de la serie temporal está estrechamente relacionado con las funciones *ACV* y *ACF* que, heurísticamente, encuentra patrones cada cierto intervalo de tiempo. Se presentarán los métodos que en general, se utilizan en la práctica y en particular, el establecido por `R`.

Finalmente, se exponen los datos con los que se trabajaron los modelos mencionados en el párrafo anterior y que serán discutidos en los capítulos posteriores.

2.1. Función Auto-Covarianza y modelo *ARMA*

En lo que sigue se supondrá que la serie $(Y_t)_t$, puede descomponerse en suma de series temporales de la forma

$$Y_t = m_t + s_t + X_t \tag{2.1}$$

donde m_t es una función de *lento* cambio conocida como **tendencia**, s_t corresponde a una función con un periodo d conocido, llamada **componente estacional**, ambas deterministas y X_t un proceso estacionario residual de la serie Y_t. En la sección 2.2 se establecerá la aproximación de cada una de las componentes.

Se introducirá la definición formal de un proceso estacionario, y cómo a través de las funciones autocovarianza (ACV) y autocorrelación (ACF), se infiere información relevante de la serie estacionaria X_t.

Se mostrará cómo el modelo $ARIMA$ (Auto-regressive Integrated movil average) y sus parámetros, están estrechamente relacionados con las funciones ACV y ACF, el cual se basa en la descripción del proceso temporal $(X_t)_{t\in T}$, en los sucesos anteriores a un momento t determinado. En particular, comprende la integración de dos modelos que se introducirán, el modelo auto regresivo (AR) y el modelo de media móvil (MA) y algunas de sus propiedades.

En lo que sigue, considérese $(X_t)_{t\in\mathbb{N}}$ serie temporal y sea $(x_t)_{t\in[1,\dots n]}$ una muestra de tamaño n.

Definición 2.1 *(***Función media y covarianza)** *Sea $(X_t)_{t\in\mathbb{N}}$ serie temporal con $\mathbb{E}(X_t^2) < \infty$. La* **función media** *de $(X_t)_{t\in\mathbb{N}}$ es*

$$\mu_X(t) = \mathbb{E}(X_t)$$

La función covarianza de $(X_t)_{t\in\mathbb{N}}$ es

$$Cov_X(r,s) = Cov(X_r, X_s) = \mathbb{E}[(X_r - \mu_X(r))(X_s - \mu_X(s))]$$

para todo entero r y s

Definición 2.2 *(***Serie débilmente Estacionaria)** *$(X_t)_{t\in\mathbb{N}}$ es débilmente estacionaria si*

1. *$\mu_X(t)$ es independiente de t*
2. *$\gamma_X(t+h,t)$ es independiente de t para todo h*

Cuando se mencione el término *estacionario* se referirá a la definición de *débilmente estacionario* a menos que se mencione lo contrario.

Luego, dentro del estudio de los modelos autoregresivos, y el estudio en general de las series temporales, se introduce la función autocovarianza (ACV) y la función autocorrelación (ACF) que contribuyen como herramienta a la inferencia de propiedades de la serie temporal X_t, como sigue:

Definición 2.3 *(***ACV ACF)** *Para una serie estacionaria X_t, se define la función autocovarianza (ACV), y la función autocorrelación (*ACF*) como*

$$\gamma(k) = Cov(X_{t+h}, X_t)$$
$$\rho(k) = \frac{\gamma(k)}{\gamma(0)}$$

Acontinuación se mencionan dos procesos temporales para ejemplificar las funciones ACV, ACF, (en p.9 Brockwell P.J., 2002).

Ejemplo (*Paseo aleatorio*)

Considere el paseo aleatorio canónico $S_t = X_1 + \cdots + X_t$ donde $\{X_t\}_{t\in\mathbb{N}}$ es un conjunto de v.a. iid que distribuye como ruido blanco

Definición 2.4 (**Ruido blanco**) *Sea* $(\varepsilon_t)_{t\in\mathbb{N}}$ *secuencia no correlacionada de v.a. con* $\mathbb{E}(\varepsilon_t) = 0$ *y* $\mathbb{E}(\varepsilon_t^2) = \sigma^2 < \infty$, *entonces*

$$\gamma_X(t+h,t) = \begin{cases} \sigma^2 & \text{si} \quad h = 0 \\ 0 & \text{si} \quad h \neq 0 \end{cases}$$

se dirá que $\{\varepsilon_t\}$ *es ruido blanco y se denotará por*

$$\{\varepsilon_t\} \sim WN(0,\sigma^2)$$

Luego para $S_t = X_1 + \cdots + X_t$ paseo aleatorio, calculando ACV

$$\begin{aligned} \gamma(t,t+k) &= \mathrm{Cov}(S_{t+k}, S_t) \\ &= \mathrm{Cov}(S_t + X_{t+1} + \ldots + X_{t+k}, S_t) \\ &= \mathrm{Cov}(S_t, S_t) + \underbrace{\mathrm{Cov}(X_{t+1}, S_t) + \ldots + \mathrm{Cov}(X_{t+k}, S_t)}_{=0\text{por independencia de } X_t} \\ &= \sum_{\mathrm{i},j=1}^{t} \mathrm{Cov}(X_\mathrm{i}, X_j) \\ &= \sum_{\mathrm{i}=1}^{t} \underbrace{\mathrm{Cov}(X_\mathrm{i}, X_\mathrm{i})}_{\sigma^2} \\ &= t\sigma^2 \end{aligned}$$

Luego el paseo aleatorio S_t es un proceso no estacionario.

Ejemplo Considere el conjunto $\{X_t\}_{t\in\mathbb{N}}$ conjunto iid que sigue distribución $WN(0,\sigma^2)$, además considere $\theta \neq 1$ y S_t definido por la ecuación

$$S_t = X_t + \theta X_{t+1}$$

se tendrá que

$$\gamma(t,t+k) = \begin{cases} \sigma^2(1+\theta^2) & k = 0 \\ \sigma^\theta & |k| = 1 \\ 0 & |k| > 1 \end{cases}$$

luego S_t es un proceso estacionario. En la figura 2.1 se aprecia el comportamiento de la función ACF con respecto k, para el proceso simulado en `R`.

En la figura 2.2 se grafican ambos ejemplos descritos, simulados en `R`.

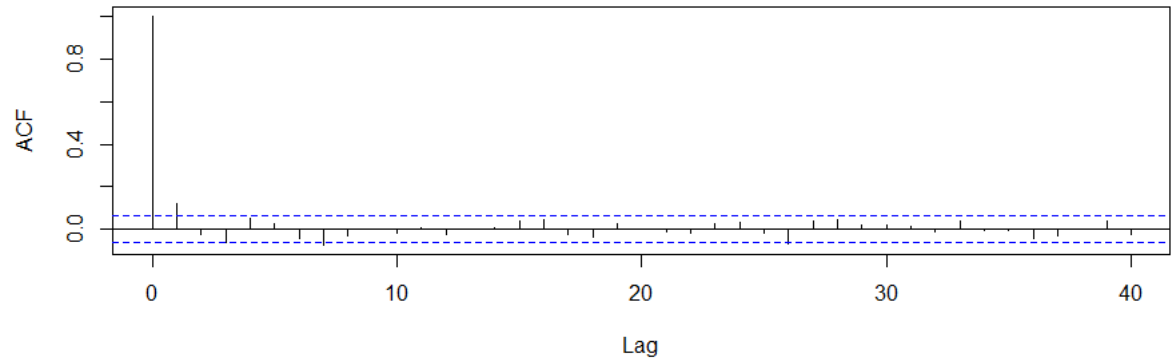

Figura 2.1: Función autocorrelación para el proceso definido por $S_t = X_t + ,7X_{t+1}$

Observación La función $\gamma(k)$ y $\rho(k)$ evidencian cuantas variables anteriores al instante t inciden en el $t-$ésimo evento, en el último ejemplo, la variable S_t depende de X_t y X_{t-1}, es decir, S_t depende (en algún sentido) de S_{t-1}, pero deja de depender de los instantes previos, en la figura 2.1 se observa que solo el primer *Lag* (retraso) $\rho \neq 0$, lo que coincide con lo expuesto en el ejemplo. Más generalmente, si se esta en presencia de un proceso estacionario, la función $\gamma(k)$ y $\rho(k)$ indican en algún grado, cual es el conjunto $\{X_{t-1}, \ldots, X_{t-k}\}$ que inciden en X_t. El teorema formaliza la intuición detrás de esta observación.

2.1.1. Estimador para $\gamma(k)$

El estimador de las funciones ACV y ACF estan definidos para una serie $\{X_t\}_t$ de tamaño n

$$\hat{\gamma}(k) = n^{-1} \sum_{t=1}^{n-|k|} (x_{t+|k|} - \overline{X})(x_t - \overline{X})$$

$$\hat{\rho}(k) = \frac{\hat{\gamma}(k)}{\hat{\gamma}(0)}$$

Luego, la matriz autocovarianza Γ definida por $\Gamma_n[\mathrm{i}, j] = (\gamma(|j - \mathrm{i}|)$, será aproximada por $\hat{\Gamma}$ definida por $\hat{\Gamma}_n[\mathrm{i}, j] = (\hat{\gamma}(|j - \mathrm{i}|)$. De igual forma la matriz autocorrelación $R_k = \gamma(0)^{-1}\Gamma_k$, será aproximada por el estimador $\hat{R}_k = \hat{\gamma}(0)^{-1}\hat{\Gamma}_k$.

Como se mostrara más adelante, en las figuras 2.3 y 2.4, una vez que k es suficientemente grande $\rho(k)$ no se aleja del cero. Formalmente se enuncia el siguiente resultado sin demostración.

Teorema 2.5 *2.1 Si $\{y_t\}_t$ es un proceso iid, de media 0 y varianza σ^2, entonces $\hat{\rho}(k)$ es*

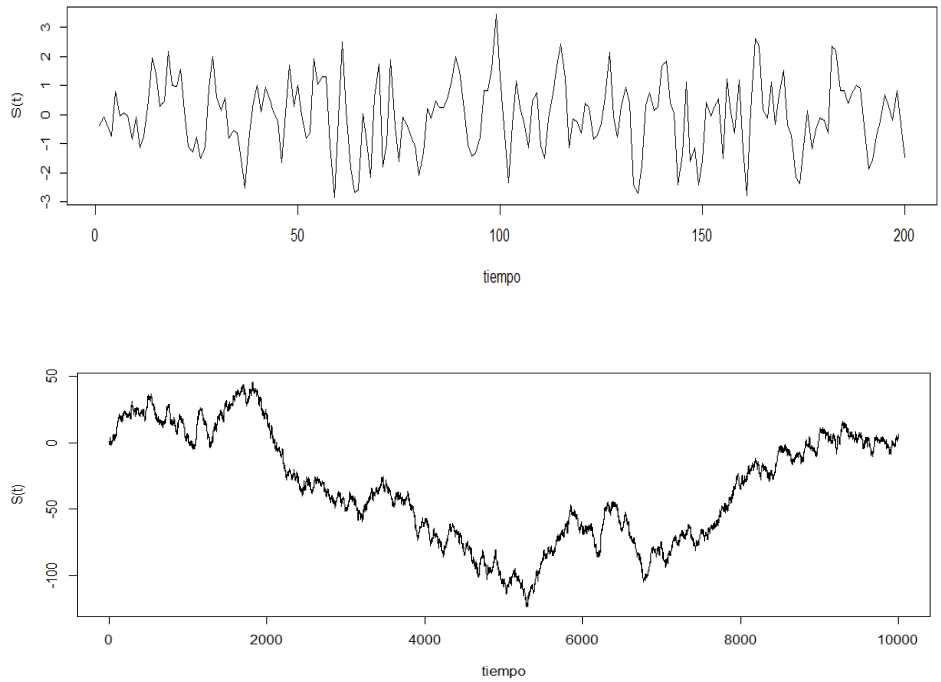

Figura 2.2: Gráficos de los procesos definidos por $S_t = \sum_{i=1}^{t} X_i$ (primera figura) y por $S_t = X_t - ,7 \cdot X_{t-1}$ (segunda figura

.

aproximadamente $N(0, \frac{1}{n})$*, de acuerdo el teorema central del límite, se tiene que*

$$\sqrt{n}\,\hat{\rho}(k) \xrightarrow{d} N(0,1)$$

Luego, con un nivel de significación del 95 % la muestra se encuentra en $[\frac{-1,96}{\sqrt{n}}, \frac{1,96}{\sqrt{n}}]$*. (en Brockwell P.J., 2002)*

Luego, a través de esta aproximación, Ljung-Box realizaron un test para probar la distribución independiente de los datos dentro de la serie temporal. La prueba se enuncia como sigue, dado un proceso $\{Y_t\}_t$ con muestra $\{y_t\}_{t\epsilon[1,\dots,n]}$

$$H_0 : \{y_t\}_t \text{ son independientes}$$
$$H_1 : \{y_t\}_t \text{ no lo son}$$

Entonces, definiendo el estadístico

$$Q(h) = n(n+2)\sum_{k=1}^{h} \frac{\hat{\rho}(k)}{n-k}$$

distribuye como χ^2_h, con h grados de libertad bajo la hipotesis nula H_0. Luego, la región de rechazo, para un nivel de significación α esta dada por

$$Q > \chi^2_{1-\alpha,h}$$

donde $\chi^2_{1-\alpha,h}$ indica el $\alpha-$cuantil de χ^2_h (en p.35-38 Brockwell P.J., 2002).

Observación Aún más fuerte será este test, cuando se realice el mismo análisis pero para una serie bivariada. La generalización de este estadístico se utiliza para establecer independencia entre dos series temporales y será mencionada en la sección 4.1.

2.1.2. Modelo de media móvil

El modelo de media móvil es una aproximación para modelar series temporales univariadas, que se basa en la explicación de la variable aleatoria por los ruidos blancos previos al instante t. Formalmente se define:

Definición 2.6 *(***Proceso MA de orden q)** *X_t es un proceso de media móvil de orden q si*

$$X_t = c + \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} \tag{2.2}$$

donde $\theta_1, \dots \theta_q$ constantes y $\varepsilon_j \sim WN(0, \sigma^2)$

Considérese el operador backshift B tal que

$$B^j X_t = X_{t-j}$$

entonces de manera compacta, se escribirá (2.2) de la forma $X_t = c + \theta(B)\varepsilon_t$ donde θ corresponde al polinomio $\theta(z) = 1 + \sum_{i=1}^{q} \theta_i z^i$. Sea $q \geq 0$, si $\gamma(h) = 0 \qquad \forall\, |h| > q$ X_t se dirá $q-$correlacionada.

Proposición 2.7 *Si $\{X_t\}_t$ es un proceso $q-$ relacionado, entonces puede ser representado por un proceso $MA(q)$ dado por la ecuación* (2.2)

Observación Cabe destacar que dentro de una muestra, los datos observados son $\{x_t\}_t$, por lo que la ecuación (2.2) no es una regresión en el sentido usual.

2.1.3. Modelo Auto regresivo

El modelo AR de orden p sobre $(X_t)_{t\in\mathbb{N}}$ plantea X_t en función de los valores previos al instante t. Más específicamente, este modelo establece la dependencia lineal de X_t con

respecto a los p valores aleatorios previos al ocurrido en t, es decir, X_t es combinación lineal de $\{X_{t-1}, X_{t-2}, \ldots, X_{t-p}\}$.

$$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \varepsilon_t$$

con error $\varepsilon \sim WN(0, \sigma^2)$ asociado al instante t, luego, formalmente se define el proceso AR de orden p como sigue.

Definición 2.8 *(***Modelo *AR* de orden p)** *X_t es un proceso auto-regresivo de orden p si*

$$X_t - c - \sum_{i=1}^{p} \phi_i X_{t-i} = \varepsilon_t \tag{2.3}$$

donde $\phi_1, \ldots, \phi_p$ son los parámetros del modelo, y ε_t ruido blanco asociado al instante t

Al igual que la ecuación (2.2), el modelo (2.3) se escribirá de la forma

$$\phi(B)X_t - c = \varepsilon_t$$

con ϕ el polinomio $\phi(z) = 1 - \sum_{i=1}^{p} \phi_i z^i$

Ejemplo Considerar a $(X_t)_t$ que se describe como $AR(1)$ con $\phi_1 < 1$. Se obtiene un paseo aleatorio ponderado, estacionario

$$X_t = \phi_1 X_{t-1} + \varepsilon_t$$

```
ts.sim <- arima.sim(list(order = c(1,0,0), ar = 0.2), n = 200)
ts.plot(ts.sim)
r.hat = acf(ts.sim,lag.max=20,type="correlation",plot=F)
plot(r.hat)
```

Cuadro 2.1: Cuadro mostrando simulación código en R

Una vez descrito los modelos AR y MA, se generalizará la descripción de X_t en términos de la forma establecida en las ecuaciones (2.3) y en (2.2).

Definición 2.9 *(***Modelo *ARMA (p,q)*)** *X_t es un proceso ARMA de orden (p, q) si*

$$\Phi(B)X_t = \theta(B)\varepsilon_t \tag{2.4}$$

En general se considera $\mathbb{E}(X_t) = \mu \equiv 0$ para todo t, de manera contraria basta reemplazar el proceso $\{X_t\}_t$ por $\{X_t - \mu\}$.

Simulación proceso $AR(1)$

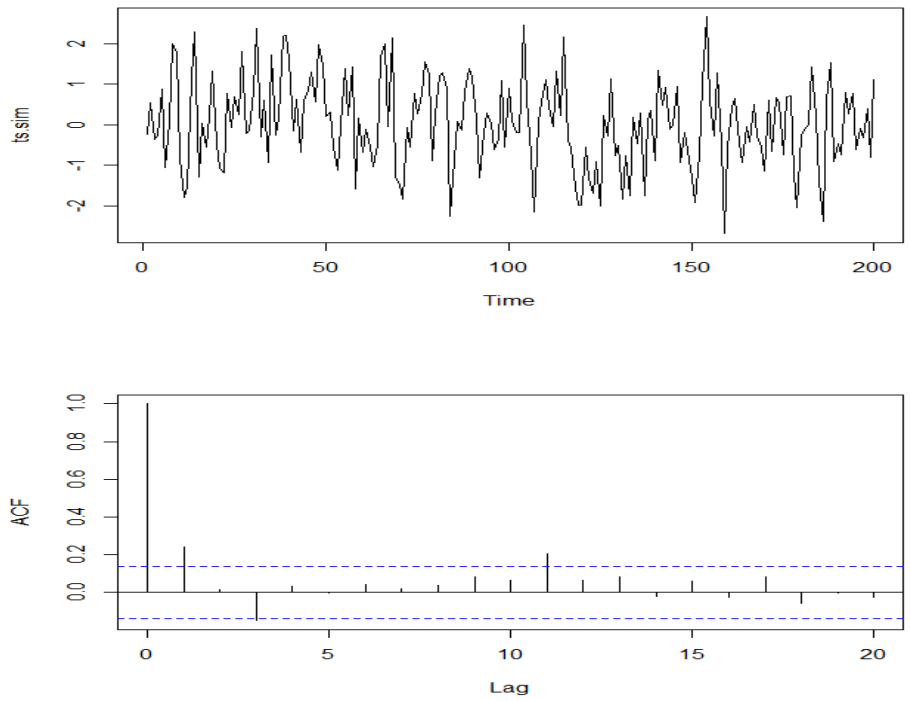

Figura 2.3: Simulación de proceso estacionario $AR(1)$ con 200 eventos. El primer gráfico muestra la serie temporal del proceso dado por la ecuación $y_t - ,2y_{t-1} = \varepsilon_t$. Mientras que el segundo gráfico muestra la función $\rho(k)$ para el proceso $\{y_t\}_t$. Esta simulación fue ejecutada según la secuencia de código que se muestra más abajo.

ARMA(1,1)

Considérese el proceso estacionario $(X_t)_t$ tal que puede ser descrito por el modelo ARMA(1,1), esto quiere decir que

$$X_t - \phi X_{t-1} = \varepsilon_t + \theta \varepsilon_{t-1} \tag{2.5}$$

De esta forma, considérese el polinomio $\Phi(z) = 1 - \phi z$, si $|\phi| < 1$ entonces podemos escribir Φ^{-1} en su serie de potencias

$$\frac{1}{1-\phi} = 1 + \phi + \phi^2 + \phi^3 + \cdots = \sum_{k=0}^{\infty} \phi^k$$

y considerando el polinomio $\Psi(z) = \sum_{k=0}^{\infty} \phi^k z^k$ aplicando Ψ a (2.5) se obtiene que

$$\begin{aligned}\Psi(B)\Phi(B)X_t = X_t &= \Psi(B)\Theta(B)\varepsilon_t \\ &= (1 + \phi B + \phi^2 B^2 + \phi^3 B^3 + \cdots)(1 + \theta B)\varepsilon_t \\ &= \left(1 + (\phi + \theta)B + \phi(\phi + \theta)B^2 + \cdots\right)\varepsilon_t \\ &= \{1 + \sum_{j=1}^{\infty}(\phi + \theta)\phi^{j-1}B^j\}\varepsilon_t\end{aligned}$$

es decir, $X_t = \{1 + \sum_{j=1}^{\infty}(\phi+\theta)\phi^{j-1}B^j\}\varepsilon_t$ es solución estacionaria de la ecuación (2.5). Por otro lado, si $|\phi| > 1$, entonces $\Phi^{-1}(z) = -\sum_{j=1}^{\infty}\phi^{-j}z^{-j}$ es decir,

$$X_t = \{-\theta\phi^{-1} - (\phi + \theta)\sum_{j=1}^{\infty}\phi^{-j-1}B^{-j}\}\varepsilon_t = -\theta\phi^{-1}\varepsilon_t - (\phi + \theta)\sum_{j=1}^{\infty}\phi^{-j-1}\varepsilon_{t+j} \tag{2.6}$$

es solución estacionaria de (2.5)

```
>ts.sim <- arima.sim(list(order = c(1,0,1), ar = 0.7, ma =1.2), n =
   200)
>ts.plot(ts.sim)
#funciones auto covarianza y auto correlacion
#g.hat = acf(ts.sim,lag.max=20,type="covariance",plot=F)
>r.hat = acf(ts.sim,lag.max=20,type="correlation",plot=F)
>plot(r.hat)
```

Existencia

Considerando un proceso descrito por el modelo *ARMA(p,q)*, cabe preguntar respecto de la existencia y la forma de una serie temporal $(X_t)_{t\in\mathbb{N}}$ estacionaria que satisface la ecuación (2.4).

Anteriormente, se muestra la existencia de una solución estacionaria para el proceso $(X_t)_{t\in\mathbb{N}}$ descrito por ARMA(1,1), de manera similar, se enuncia el siguiente resultado.

Proposición 2.10 *Existencia La ecuación* (2.4) *tiene una solución* $(X_t)_{t\in\mathbb{N}}$ *estacionaria si y solo si*

$$\Phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p \neq 0 \qquad \forall |z| = 1$$

Bosquejo de demostración:

Considerando la ecuación (2.4) se tiene que

$$\Phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$$
$$\Theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q$$

Simulación proceso $ARMA(1,1)$

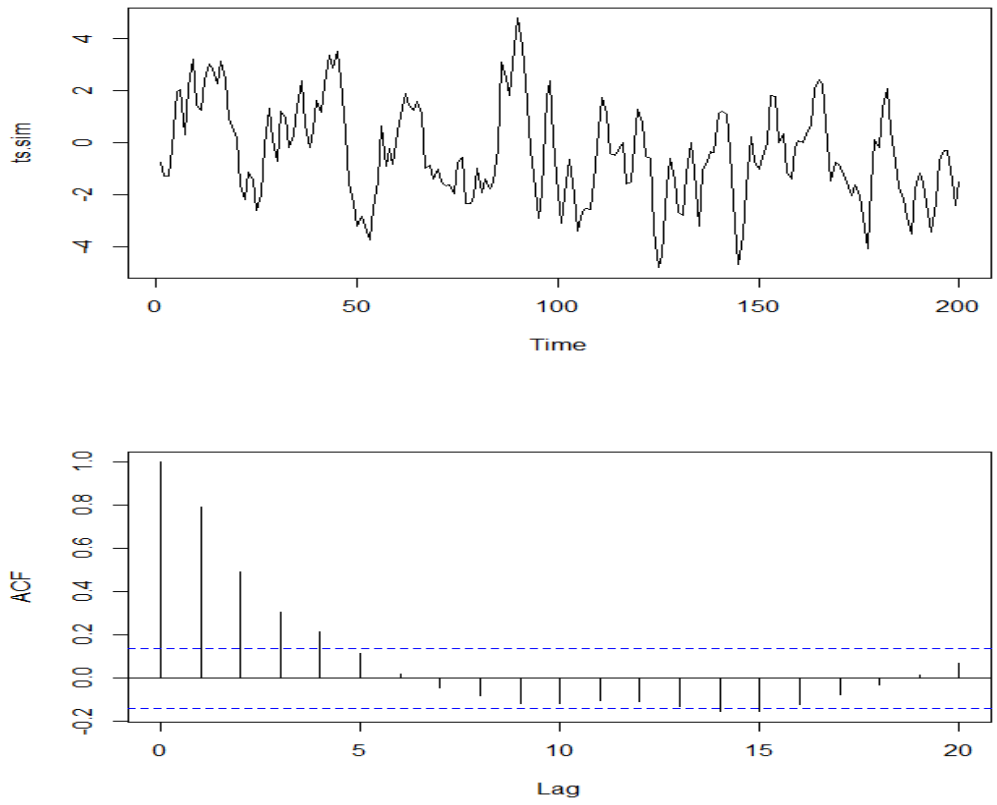

Figura 2.4: Simulación de proceso $ARMA(1,1)$ con 200 eventos. El primer gráfico muestra la serie temporal de el proceso dado por la ecuación $y_t - ,7y_{t-1} = \varepsilon_t + 1,2\varepsilon_{t-1}$. Mientras que el segundo gráfico muestra la función $\rho(k)$ para el proceso $\{y_t\}_t$. Esta simulación fue ejecutada según la secuencia de código que se muestra más abajo.

luego, si $\Phi(z) \neq 0$ para todo $|z| = 1$, entonces existe $\delta > 0$ tal que $\Phi(z) \neq 0$ para todo z tal que $1 - \delta < |z| < 1 + \delta$, luego se puede escribir

$$\Phi(z)^{-1} = \sum_{j=-\infty}^{\infty} \chi_j z^j \qquad 1 - \delta < |z| < 1 + \delta$$

con $\sum_{j=-\infty}^{\infty} |\chi_j| < \infty$. Se define entonces $\chi(B) = \Phi^{-1}(B) = \sum_{j=-\infty}^{\infty} \chi_j B^j$ y operando $\chi(B)$ en la ecuación (2.4) se obtiene que

$$\chi(B)\Phi(B)X_t = X_t = \chi(B)\Theta(B)\varepsilon_t \tag{2.7}$$

luego, $\chi(B)\Theta(B)\varepsilon_t$ es solución estacionaria de (2.4) (en p. 55-57 Brockwell P.J., 2002).

2.2. Descomposición clásica de una serie temporal y modelo *ARIMA*

Como se menciona más arriba, se trabaja sobre el supuesto que la serie temporal Y_t puede descomponerse de la forma (2.1), sin embargo, para el modelo $ARMA(p, q)$ se pretende trabajar sólo con series estacionarias. Esta sección, se centra en la remoción de la tendencia y la componente estacional de la serie original Y_t para obtener los residuales estacionarios X_t.

Más específicamente, se considera en una primera instancia $s_t \equiv 0$, y la estimación y eliminación de la tendencia m_t, y, finalmente, se muestra la estimación y eliminación de la componente estacional s_t. Cabe mencionar que para le eliminación de las componentes, se utilizan fuertemente los operadores $\nabla = (1 - B)$ y $\nabla_{\mathrm{d}} = (1 - B^{\mathrm{d}})$ para $\mathrm{d} > 1$.

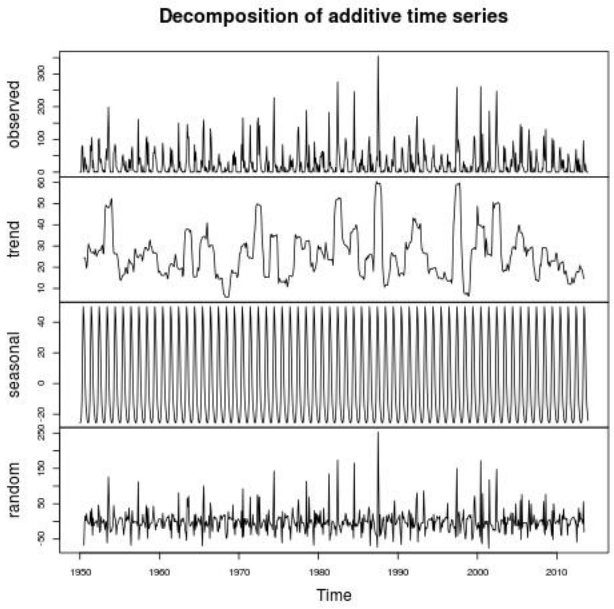

Figura 2.5: Descomposición de un proceso temporal en sus componentes. La serie de arriba corresponde a la serie original, la segunda serie corresponde a la componente T_t calculada mediante el modelo MA, la tercera corresponde a S_t a la componente estacional, y la última gráfica corresponde a los residuos ε_t

2.2.1. Estimación y eliminación de la tendencia en ausencia de la componente estacional

Para un modelo descrito según la ecuación (2.1), en ausencia del componente estacional $s_t \equiv 0$, se tendrá que éste se reduce a

$$Y_t = m_t + X_t \tag{2.8}$$

sin pérdida de generalidad, se considera que $\mathbb{E}(X_t) = 0$, de manera contraria, se puede reemplazar m_t por $m_t + \mathbb{E}(X_t)$ y X_t por $X_t - \mathbb{E}(X_t)$.En general m_t será polinomio por partes, o simplemente un polinomio a lo largo del intervalo.

Suavizamiento con filtro finito en torno al promedio

Sea $q \in \mathbb{N} - \{0\}$, considere

$$\begin{aligned} W_t &= \frac{1}{2q+1} \sum_{j=-q}^{q} Y_{t+j} \\ &= \frac{1}{2q+1} \sum_{j=-q}^{q} m_{t+j} + X_{t+j} \\ &= \frac{1}{2q+1} \sum_{j=-q}^{q} m_{t+j} + \underbrace{\frac{1}{2q+1} \sum_{j=-q}^{q} X_{t+j}}_{\sim 0} \end{aligned}$$

asumiendo que m_t se aproxima de forma lineal en el intervalo $[t-q, t+q]$, entonces se aproxima m_t por

$$\hat{m}_t = \frac{1}{2q+1} \sum_{j=-q}^{q} m_{t+j} = W_t$$

Ajuste polinomial

El ajuste polinomial por trozos considera $m_t = \sum_{i=0}^{k} a_i t^i$ en cada intervalo establecido $\{[t_0, t_1], [t_1, t_2], \ldots, [t_{n-1}, t_n]\}$. En la práctica, el largo de cada partición es *'pequeña'*, generalmente $\max_i |t_i - t_{i-1}| \leq 3$, en consecuencia $k \leq 3$. El ajuste para cada intervalo, suele ser mínimos cuadrados.

Eliminación de tendencia por diferencia

Considerando m_t polinomio de grado 1, y aplicando el operador $\nabla = (1 - B)$ a la ecuación

$$Y_t = m_t + X_t = c_0 + c_1 t + X_t$$

se tiene que

$$\begin{aligned}(1-B)Y_t &= (1-B)(m_t - X_t)\\ &= m_t - X_t - m_{t-1} + X_{t-1}\\ &= c_0 + c_1 t - c_0 - c_1(t-1) + \nabla X_t\\ &= c_1 + \nabla X_t\end{aligned}$$

más general, para un polinomio de grado k, $m_t = \sum_{i=0}^{k} a_i t^i$, aplicando el operador ∇^k, definido por $\nabla^k X_t = \nabla^{k-1}(X_t - X_{t-1})$, a m_t se tiene que $\nabla^k m_t = k! a_k$

Luego, para la ecuación (2.8),

$$\begin{aligned}\nabla^k Y_t &= \nabla^k(\sum_{i=0}^{k} a_i t^i) + \nabla^k(X_t)\\ &= k! a_k + \nabla^k(X_t)\end{aligned}$$

En la práctica, en búsqueda de un ajuste poco costoso a nivel computacional, k es un número natural pequeño, $k < 4$.

2.2.2. Estimación de la tendencia y la estacionalidad

Método por diferencia

De igual forma que antes, se puede eliminar la componente periódica de la serie original mediante diferencia. Considere que la serie Y_t tiene una componente estacional s_t de periodo d, es decir, para todo t se tiene la igualdad $s_t = s_{t+k\cdot d}$ para todo $k \in \mathbb{Z}$. Luego, aplicando el operador $\nabla_d = (1 - B^d)$ a la ecuación (2.1), se tiene que

$$\begin{aligned}\nabla_d Y_t &= \nabla_d m_t + \underbrace{\nabla_d s_t}_{s_t - s_{t-d} = 0} + \nabla_d X_t\\ &= m_t - m_{t-d} + X_t - X_{t-d}\end{aligned}$$

que entrega una descomposición para $\nabla_d Y_t$ de la forma (2.8) con tendencia $m_t - m_{t-d}$ y con un término de ruido $X_t - X_{t-d}$. Luego, la tendencia $m_t - m_{t-d}$ puede ser removida con los métodos explicados anteriormente, en particular, aplicando ∇^k.

Es decir, los polinomios involucrados permiten la generalización del modelo $ARMA(p, q)$.

2.2.3. Modelo *ARIMA* para series con tendencia y estacionalidad

El modelo *ARIMA* busca generalizar el modelo *ARMA* expuesto anteriormente para serie con tendencia (que se pueda escribir de la forma 2.8). Como se describe más arriba, el modelo

se generaliza eliminando la componente determinista bajoel método de eliminación, a través del operador ∇.

En rigor, dada una serie Y_t en presencia de tendencia, dado $\mathrm{d} > 0$ apropiado, se tendrá que el proceso $X_t = \nabla^{\mathrm{d}} Y_t$ es un proceso estacionario. Luego este proceso, podrá ser ajustado por el modelo $ARMA(p,q)$, lo que permite extender el modelo $ARMA$ para series no estacionarias. Formalmente, el método $ARIMA(p,\mathrm{d},q)$ se define de la forma

Definición 2.11 *(***Modelo ARIMA (p,d,q)***) Sea* $\mathrm{d} \in \mathbb{N}$*, si* $(Y_t)_t$ *es serie temporal tal que* $Y_t = (1-B)^{\mathrm{d}} X_t$ *admite un proceso ARMA* (p,q)

Ahora, considere Y_t de la forma (2.1), con $m_t = c_0 + c_1 t$, y con periodo s, aplicando el operador $\nabla^s \nabla$ a la serie, se tiene que

$$\begin{aligned}
\nabla^s \nabla Y_t &= \nabla^s \nabla (m_t + s_t + X_t) \\
&= \nabla^s (\underbrace{m_t - m_{t-1}}_{=c_1} + s_t - s_{t-1} + X_t - X_{t-1}) \\
&= (c_1 - c_1 + \underbrace{s_t - s_{t-s}}_{=0} - \underbrace{s_{t-1} + s_{t-1-s}}_{=0} + X_t - X_{t-s} - X_t - 1 + X_{t-1-s}) \\
&= \nabla^s \nabla X_t
\end{aligned}$$

Luego se generaliza el modelo para series de la forma 2.1, con $s_t \not\equiv 0$, eliminando s_t a través del método de eliminación por diferencia. Entonces, para Y_t de periodo s, el proceso $\nabla_s Y_t = \nabla_s m_t + \nabla_s X_t$ es un proceso de la forma (2.8). Formalmente, se define el modelo $EstacionalARIMA(p,\mathrm{d},q) \times (P,D,Q)_s$ (o $SARIMA$) de la forma

Definición 2.12 *(***Modelo estacional ARIMA** $(p,\mathrm{d},q)\times(P,D,Q)_s$**)** *Sean* d, D *dos enteros positivos, entonces* $(X_t)_t$ *es un proceso estacional ARIMA* $(p,\mathrm{d},q) \times (P,D,Q)$ *con periodo* s*, si* $Y_t = (1-B)^{\mathrm{d}}(1-B^s)^D X_t$ *es un proceso causal ARMA(p,q) definido por*

$$\phi(B)\Phi(B^s)Y_t = \theta(B)\Theta(B^s)\varepsilon_t \qquad \varepsilon_t \sim WN(0,\sigma^2) \tag{2.9}$$

con $\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$*,* $\Phi(z) = 1 - \Phi_1 z - \cdots - \Phi_P z^P$*,* $\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q$*,* $\Theta(z) = 1 + \Theta_1 z + \cdots + \Theta_Q z^Q$

2.3. Ajuste de parámetros

La búsqueda de los parámetros para modelar una serie Y_t según $ARMA(p,q)$, envuelve varios problemas interrelacionados. Desde la estimación de p, q hasta los parámetros $\{\phi_1,\ldots,\phi_p\}, \{\theta_1,\ldots,\theta_q\}$ y la varianza del ruido blanco σ^2

`R` utiliza la función `arima` que tiene como datos de ingreso (o `input`) los parámetros (p, d, q), $(P, D, Q), s$, que corresponde a los grados de los polinomios $\phi, \theta, \Phi, \Theta$ y el periodo de la estacionalidad s. El ajuste de los parámetros $\{\phi_1, \ldots, \phi_p\}$, $\{\theta_1, \ldots, \theta_q\}$, se realiza mediante el estimador de máxima verosimilitud o minimización condicional de suma de cuadrados. A continuación se enuncia el calculo de ambos estimadores.

2.3.1. Estimador de máxima verosimilitud

Para encontrar el estimador de máxima verosimilitud, existen dos pasos fundamentales. Primero, se debe encontrar la función de verosimilitud, y para esto, se mostrará como obtener la función de verosimilitud para un proceso parametrizado por $AR(1)$ y $MA(1)$. Con esto, se extenderá sin mostrar, para la parametrización $ARMA(p, q)$. Segundo, se debe encontrar el parámetro que maximice $L(\psi|\{x_1, \ldots, x_n\})$ función de verosimilitud, $\hat{\psi} = \arg\max_{\psi} L(\psi)$.

Para esta sección, se restringe al caso en que las variables del proceso $\varepsilon_t \sim WN(0, \sigma^2)$, se distribuyan de forma $\varepsilon_t \sim N(0, \sigma^2)$.

Función de verosimilitud $AR(1)$

Considere un proceso $\{X_t\}_t$ parametrizado por $AR(1)$, se tendrá que

$$X_t = c + \phi X_{t-1} + \varepsilon_t \tag{2.10}$$

entonces, los parámetros a estimar son c, ϕ, σ^2, dado que ε_t es un proceso $N(0, \sigma^2)$. Los X_t también seguirán la misma distribución. Además, cabe notar para $t = 1$

$$\mathbb{E}(X_1) = \mu = \frac{c}{1-\phi}$$
$$\mathbb{E}[(X_1 - \mu)^2] = \frac{\sigma^2}{1-\phi^2}$$

Entonces, la densidad para la primera observación

$$f_{X_1}(x_1|\psi) = \frac{1}{\sqrt{2\pi}\sqrt{\sigma^2/(1-\phi^2)}} \exp\left[\frac{-\{x_1 - \mu\}^2}{2\sigma^2/(1-\phi^2)}\right]$$

Luego, la distribución condicional para X_2 dado la observación $X_1 = x_1$, según (2.10), $X_2 ? c + \phi X_1 + \varepsilon_2$, en este caso, se tiene que

$$(X_2|X_1 = x_1) \sim N((c + \phi x_1), \sigma^2)$$

lo que implica que

$$f_{X_2|X_1=x_1}(x_2|x_1;\psi) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[\frac{-\{x_2 - c - \phi x_1\}^2}{2\sigma^2}\right]$$

Inductivamente, se tiene que

$$f_{X_t|X_{t-1},X_{t-2},\ldots,X_1}(x_t|x_{t-1},\ldots,x_1;\psi) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[\frac{-\{x_t - c - \phi x_{t-1}\}^2}{2\sigma^2}\right]$$

Luego, la distribución conjunta dada por

$$f_{X_t,\ldots,X_1}(x_t,\ldots,x_1|\psi) = f_{X_t|X_{t-1}}(x_t|x_{t-1};\psi)\cdot f_{X_{t-1},X_{t-2},\ldots,X_t}(x_{t-1},\ldots,x_1|\psi)$$

Finalmente, la función de verosimilitud L y su logaritmo $(\psi|\{x_t\}) = \log(L(\psi|\{x_t\}))$ (que en general es más fácil de maximizar), están dados, respectivamente, por

$$L(\psi|\{x_t\}) = f_{X_1}(x_1|\psi)\cdot\prod_{\mathrm{i}=2}^{T} f_{X_\mathrm{i}|X_{\mathrm{i}-1}}(X_\mathrm{i}|X_{\mathrm{i}-1};\psi) \tag{2.11}$$

$$\mathcal{L}(\psi|\{x_t\}) = \log(f_{X_1}(x_1|\psi)) + \sum_{\mathrm{i}=2}^{T}\log[f_{X_\mathrm{i}|X_{\mathrm{i}-1}}(X_\mathrm{i}|X_{\mathrm{i}-1};\psi)] \tag{2.12}$$

Función de verosimilitud *MA(1)*

Considérese un proceso $\{X_t\}_t$ parametrizado por $MA(1)$, de la forma

$$X_t = \mu + \varepsilon_t + \theta\varepsilon_{t-1}$$

con $\varepsilon_t \sim N(0,\sigma^2)$, sea $\psi = (\mu,\theta,\sigma^2$ los parámetros a estimar. Si el valor ε_{t-1} se conoce con certeza, entonces

$$X_t|\varepsilon_{t-1} \sim N(\mu + \theta\varepsilon_{t-1},\sigma^2)$$

para $t = 1$, $X_1|\varepsilon_0 \sim N(\mu,\sigma^2)$, y de forma similar al desarrollo en la sección anterior, se tendrá que la función de densidad condicional es

$$f_{X_t|X_t,\ldots,X_1,\varepsilon=0}(x_t|x_t,\ldots,x_1,\varepsilon=0;\psi) = f_{X_t|\varepsilon_{t-1}}(x_t|\varepsilon_{t-1};\psi) \tag{2.13}$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[\frac{-\varepsilon_t^2}{2\sigma^2}\right] \tag{2.14}$$

Luego $\mathcal{L}(\psi|\{x_t\}) = \frac{-T}{2}\log[2\pi] - \frac{-T}{2}\log[\sigma^2] - \sum_{\mathrm{i}=1}^{T}\frac{\varepsilon_\mathrm{i}^2}{2\sigma^2}$

Función de verosimilitud *ARMA(p,q)*

Para un proceso *ARMA(p,q)* de la forma

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = c + \varepsilon_t + \theta_1\theta_{t-1} + \theta_q\varepsilon_{t-q}$$
$$X_t = c + \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + \varepsilon_t + \theta_1\theta_{t-1} + \theta_q\varepsilon_{t-q}$$

Luego, los parámetros a estimar son $\psi = (c, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q, \sigma^2)$. De manera similar a las secciones anteriores, se calculará la función de verosimilitud. Para esto, primero es necesario plantear los datos iniciales de $\{X_t\}_t, \{\varepsilon_t\}_t$. Una opción para lo anterior, es definir $\mathbf{x_0} = (x_0, x_{-1}, \ldots, x_{-p}), \varepsilon_\mathbf{0} = (\varepsilon_0, \ldots, \varepsilon_{-q})$. Luego, la secuencia $(\varepsilon_1, \ldots, \varepsilon_T)$ puede ser calculada con los datos $(x_1, \ldots, x_T)$ iterando sobre

$$\varepsilon_t = x_t - c - \phi_1 x_{t-1} - \cdots \phi_p x_{t-p} - \theta_1\varepsilon_{t-1} - \cdots - \theta_q\varepsilon_{t-q}$$

para $t = 1, \cdots, T$, luego el logaritmo de la función de verosimilitud

$$\mathcal{L}(\psi|\mathbf{x}_0, \varepsilon_0) = -\frac{T}{2}\log[2\pi] - \frac{T}{2}\log[\sigma^2] - \sum_{i=1}^{T}\frac{\varepsilon_t^2}{2\sigma^2}$$

2.3.2. Minimización condicional de suma de cuadrados *(CSS)*

El método de minimización presentado acá, como uno de los métodos utilizados en `R`, basa la estimación de los parámetros en la minimización de los residuos cuadrados ε_t^2. Entonces, la construcción de la función objetivo a partir de ε_t^2, para un proceso $\{X_t\}_t$, parametrizado por el modelo *ARMA(p,q)* de la forma

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = c + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_q\varepsilon_{t-q}$$
$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} - c - (\theta_1\varepsilon_{t-1} + \theta_q\varepsilon_{t-q}) = \varepsilon_t$$

De esta forma, se considera ε_t, dado $(x_t, \ldots, x_{t-p}, \varepsilon_{t-1}, \ldots, \varepsilon_{t-q})$ se considera

$$\varepsilon_t = X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} - c - (\theta_1\theta_{t-1} + \theta_q\varepsilon_{t-q})$$
$$\varepsilon_t^2 = \underbrace{\{x_t - \phi_1 x_{t-1} - \cdots - \phi_p x_{t-p} - c - (\theta_1\theta_{t-1} + \theta_q\varepsilon_{t-q})\}^2}_{\tilde{\phi}(B)x_t - c - \tilde{\theta}(B)\varepsilon_t}$$
$$S(\psi) = \sum_{t=p+1}^{T}\varepsilon_t^2 = \sum_{t=p+1}^{T}\{\tilde{\phi}(B)x_t - c - \tilde{\theta}(B)\varepsilon_t\}^2$$

Luego, la función objetivo a minimizar es $S(\psi)$, donde $\psi = \{\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q, c\}$ (en Hamilton, James D. 1994)

2.4. Predicción

El problema presentado en esta sección es, dado una serie $\{X_t\}_{t\in\{1,\dots,n\}}$. predecir los eventos X_{n+h} para $h > 0$. Un primer objetivo es aproximar X_{n+h} por una combinación lineal de los elementos de la serie temporal, entonces, el primer predictor lineal es de la forma

$$P_n X_{n+h} = a_0 + a_1 X_n + a_2 X_{n-1} + \cdots + a_n X_1$$

Si el problema es abordado desde el punto de vista de mínimos cuadrados, se considera que la función objetivo a minimizar es

$$S(a_0, \dots, a_n) = \mathbb{E}[(X_{n+h} - P_n X_{n+h})^2]$$

Luego, para cada $h > 0$, aplicando el operador $\frac{\partial}{\partial a_j}$ para $j \in \{1, \dots n\}$, se tienen n ecuaciones

$$\frac{\partial S(a_0, \dots, a_n)}{\partial a_j} = 0 \qquad \forall\, j\{1, \dots n\}$$

lo que computacionalmente es muy costoso. Sin embargo, se han desarrollado algoritmos más eficientes que el descrito recientemente. Se mencionará el filtro de Kalman, método utilizado por `R`. Entonces, para un proceso estacionario general, el predictor $P_n X_{n+1}$ está basado en los n valores previos, mientras que $P_{n+1} X_{n+2}$ está basado en los $(n+1)$ valores previos.

2.4.1. Filtro de Kalman

En esta sección, se introducirá la idea de expresar un sistema dinámico en una forma particular llamada *representación estado-espacio*. El filtro de Kalman, es un algoritmo utilizado por `R` para actualizar secuencialmente la proyección lineal. Este algoritmo, provee una vía para pronosticar una muestra finita y la función de máxima verosimilitud para un proceso Gaussiano *ARMA*. El filtro de Kalman, puede verse con mayor profundidad en [2]. En esta sección, se introducirán las ecuaciones y se restringirá a la representación del proceso *ARMA*.

Sea $\mathbf{y}_t \in \mathbb{R}^n$ variables observadas al instante t, un modelo dinámico para $\mathbf{y}_t$ puede ser descrito en términos de posibles datos no observados $\xi \in \mathbb{R}^r$, llamado *vector estado*. La *representación estado-espacio* de $\mathbf{y}$, está dada por el siguiente sistema de ecuaciones

$$\xi_{t+1} = \mathbf{F}\xi_t + \mathbf{v}_{t+1} \tag{2.15}$$

$$\mathbf{y}_t = \mathbf{A}'\mathbf{x}_t + \mathbf{H}\xi_t + \mathbf{w}_t \tag{2.16}$$

$$\tag{2.17}$$

donde $\mathbf{F}, \mathbf{A}', \mathbf{H}'$ son matrices de parámetros de tamaño $(r \times x), (n \times k)$ y $(n \times r)$ respectivamente, donde $\mathbf{x}_t$ es un vector exógeno. La ecuación (2.15) se conoce como la *ecuación de*

estado, mientras que la ecuación (2.16) es conocida por *ecuación de observación*. Los vectores $\mathbf{v}_t \in \mathbb{R}^r, \quad \mathbf{w}_t \in \mathbb{R}^n$, son vectores de ruido blanco con matrices de covarianza definida por

$$\mathbb{E}(\mathbf{v}_t \mathbf{v}_\tau') = \mathbf{Q} \cdot \mathbb{1}_{\{t=\tau\}} \tag{2.18}$$

$$\mathbb{E}(\mathbf{w}_t \mathbf{w}_\tau') = \mathbf{R} \cdot \mathbb{1}_{\{t=\tau\}} \tag{2.19}$$

Además se asume que $\mathbf{v}_t$ y $\mathbf{w}_t$ son no correlacionadas, es decir, $\mathbb{E}(\mathbf{v}_t \mathbf{w}_\tau') \equiv 0$ para todo t, τ.

Considerando así un proceso univariado $\{Y_t\}_t$ parametrizado por*ARMA(p,q)*, según la ecuación (2.4).

$$y_t = \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} \tag{2.20}$$

Se considera entonces la *representación estado-espacio*, la *ecuación de estado* con $r = \max\{p, q+1\}$:

$$\xi_{t+1} = \begin{bmatrix} \phi_1 & \phi_2 & \cdots & \phi_{r-1} & \phi_r \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix} \xi_t + \begin{bmatrix} \varepsilon_{t+1} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{2.21}$$

mientras que la ecuación (2.16), con $n = 1$ se tiene que

$$y_t = [1, \theta_1, \cdots, \theta_{r-1}]\xi_t \tag{2.22}$$

Se asumirá, sin demostración, que el conjunto de ecuaciones (2.21) y (2.22) describen el mismo proceso que (2.20).

Pronóstico para $\mathbf{y}_t$

Sea $\hat{\xi}_{t+1|t} = \hat{\mathbb{E}}(\xi_{t+1}|\mathcal{Y}_t)$ donde $\mathcal{Y}_t = (\mathbf{y}_{t-1}, \ldots, \mathbf{y}_1, \mathbf{x}_{t-1}, \ldots, \mathbf{x}_1)$ denota toda la información previa al instante t, y sea la matriz de $r \times r$

$$\mathbf{P}_{t+1|t} = \mathbb{E}[(\xi_{t+1} - \hat{\xi}_{t+1})(\xi_t - \hat{\xi}_t)]$$

Entonces, dado los valores iniciales de la forma

$$\hat{\xi}_{1|0} = \mathbb{E}(\xi_1)$$
$$\mathbf{P}_{1|0} = \mathbb{E}[(\xi - \mathbb{E}\xi_1)(\xi - \mathbb{E}\xi_1)']$$

Luego, de forma iterativa se pueden calcular

$$\hat{\xi}_{t+1|t} = \mathbf{F}\hat{\varepsilon}_{t|t-|} + \mathbf{F}\mathbf{P}_{t|t-1}\mathbf{H}(\mathbf{H}'\mathbf{P}_{t|t-1}\mathbf{H} + \mathbf{R})^{-1}(\mathbf{y}_t - \mathbf{A}'\mathbf{A}'\mathbf{x}_t - \mathbf{H}'\hat{\xi}_{t|t-1})$$

y $\mathbf{P}_{t|t-1}$ por definición para $t = 1, ..., T$ con T la cantidad de muestras de Y_t. El término $\hat{\xi}_{t+1|t}$ es el mejor predictor para $\xi_{t+1|t}$ basado en función lineal de $(\mathbf{y}_t, \ldots, \mathbf{y}_1, \mathbf{x}_t, \ldots, \mathbf{x}_1)$ Entonces, la proyección de $\hat{y}_{t+1|t}$ de acuerdo el método iterativo de proyecciones, se tiene que

$$\hat{\mathbf{y}}_{t+1|t} = \mathbf{A}'\mathbf{x}_{t+1} + \mathbf{H}' \cdot \hat{\mathbb{E}}(\xi_{t+1}|\mathbf{x}_t, \mathcal{Y}_t) = \mathbf{A}'\mathbf{x}_t + \mathbf{H}' \cdot \hat{\xi}_{t+1|t}$$

(en p. 372- 381 Hamilton, James D. 1994).

2.5. Implementación en R de *ARIMA*

Dentro de `R` existen dos grandes paquetes que permiten el análisis de series temporales a través del modelo *ARIMA*: `stats` y `forecast`. A continuación, en esta sección, se mencionará cómo `R` se realiza la descomposición clásica de series temporales según la ecuación (2.1), el ajuste de parámetros y la predicción según el ajuste del modelo *ARIMA*.

`auto.arima`

La función `auto.arima` del paquete `forecast`, es una función que recibe como parámetro una serie temporal, y retorna el mejor ajuste *ARIMA*. Además, esta función recibe como entrada las variables $(p, \mathrm{d}, q) \times (P, D, Q)$, que en el caso de no ingresarlos, la función los ajusta utilizando pruebas de raíz de la unidad

Antes de explicitar el algoritmo, se describen las funciones de información que son utilizadas dentro del algoritmo, que mediante la minimización de ellos, se encuentra el mejor ajuste de los parámetros p, q. La primera función de información (por defecto en `auto.arima`) es el criterio de información de Akaike (AIC por sus siglas en inglés)

$$AIC = -2\log(L) + 2\{p + q + 1 + \mathbf{1}_{c\neq 0}\}$$

La segunda, y más utilizada en este contexto, es la función AIC corregida

$$AIC_c = AIC + \frac{2(p + q + \mathbf{1}_{c\neq 0} + 1)(p + q + \mathbf{1}_{c\neq 0} + 2)}{n - p - q - k - 2}$$

y el último criterio, es criterio de información Bayesiano (BIC por su sigla en inglés)

$$BIC = AIC + (\log(n) - 2)(p + q + \mathbf{1}_{c\neq 0} + 1)$$

Luego, el algoritmo realizado en `R`, utilizando alguno de los criterios, sigue los siguientes pasos:

1. Encuentra los parámetros d, D según el test KPSS y el test OCSB, respectivamente(insertar referencia). Si la $\mathrm{d} = D = 0$, entonces c es incluida en el modelo, de manera contraria, se fija $c \equiv 0$.

2. Luego de diferenciar la serie temporal $\nabla^{\mathrm{d}}(\nabla_s)^D y_t$, los parámetros p, q son escogidos minimizando la función objetivo *AIC*. Selecciona el mejor modelo (el que minimiza AIC) entre los siguientes cuatro *ARIMA(2,d,2), ARIMA(0,d,0), ARIMA(1,d,0), ARIMA(0,d,1).*

3. Varía p, q ajustados en el modelo fijado en el paso 2, en ± 1

4. Luego, los parámetros $\{\phi_1, \ldots, \phi_p, \theta_1, \cdots \theta_q, \Phi_1, \ldots, \Phi_P, \Theta_1, \cdots \Theta_Q, \sigma^2\}$ son ajustados mediante el estimador de máxima verosimilitud, o *CSS* (variables de entrada en la función)

`arima, Arima`

Las funciones `arima, Arima`, de los paquetes `stat y forecast` respectivamente, reciben como parámetros a $(p, \mathrm{d}, q) \times (P, D, Q)$.

Por defecto, `arima` fija $c = 0$ cuando $\mathrm{d} > 0$ y provee una estimación $\hat{\mu}$ de $\mu = \mathbb{E}(Y_t)$ cuando $\mathrm{d} = 0$. Mientras que los parámetros $(\phi_1, \ldots, \phi_p, \theta_1, \cdots \theta_q, \Phi_1, \ldots, \Phi_P, \Theta_1, \cdots \Theta_Q, \sigma^2)$ son estimados por el estimador de máxima verosimilitud o por suma condicional de cuadrados, métodos descritos en la sección 2.3.

La función `Arima`, por otro lado, es un poco más flexible en términos de la constante c. Cuando $\mathrm{d} = 0, 1$, la función provee una estimación $\hat{\mu}$, pero fija a $c = 0$ cuando $\mathrm{d} > 1$.

`forecast`

Para el pronóstico de un instante posterior al instante final de la muestra, `R` utiliza el filtro de Kalmann descrito en la sección 2.4.1.

`decompose`

Esta función en `R`, realiza la descomposición de una serie temporal $\{y_t\}_t$, de forma aditiva (como la descrita en la sección 2.2) o multiplicativa. Según el parámetro entregado, la función `decompose` retorna las series m_t, s_t, ε_t tales que $y_t = m_t + s_t + \varepsilon_t$, si la variable de entrada es aditiva o $y_t = m_t \cdot s_t \cdot \varepsilon_t$ si es multiplicativa.

La descomposición aditiva para una serie $\{y_t\}_t$ de periodo m (como parámetro de entrada) se realiza según el siguiente algoritmo.

1. Se determina la componente de tendencia m_t utilizando un modelo me *MA*, con intervalos simétricos con igual peso.

2. Luego, para calcular la componente periódica s_t se obtiene promediando la serie calculada por $\{y'_t\}_t = \{y_t - m_t\}_t$

3. Mediante el un metodo regresivo se estima $\hat{m}_t$ para finalmente obtener $\hat{\varepsilon}_t = y_t - \hat{m}_t - \hat{s}_t$.

2.6. Comparación en series temporales

2.6.1. Independencia de series temporales

Hasta ahora, se ha supuesto que el sistema en estudio puede ser modelado por un proceso $\{X_t\}_{t\in T}$, donde $X_t \in \mathbb{R},\ \forall\, t \in T$. Sin embargo, de manera más general, un sistema puede ser representado por más de una variable, de la forma

$$\{\mathbf{X}_t\}_{t\in T} = \{(X_t^1, \ldots, X_t^n)\}_{t\in T}$$

bajo este contexto y en esta sección, se aborda la independencia entre $\{X_t^{\mathrm{i}}\}_{t\in T}$ y $\{X_t^j\}_{t\in T}$ con $0 < \mathrm{i} < j \leq n$ estableciendo una distancia entre ambos proceso, para cuantificar esta distancia, se definen los siguientes operadores.

1. Para medir la variación conjunta entre los dos procesos, se define el producto interno entre dos series temporales con diferencia k, con $k \in \mathbb{Z}$

   $$\rho_{X_t,Y_t}(k) = \langle \{X_t\}_{t\in\mathbb{Z}}, \{Y_t\}_{t\in\mathbb{Z}} \rangle_k = \frac{cov(X_{t+k}, Y_t)}{\sigma_X \cdot \sigma_Y}$$

   donde su estimador esta dado por

   $$\hat{\rho}_{x_t y_t}(k) = \frac{1}{\hat{\sigma}_x \hat{\sigma}_y} \sum_{t=k}^{n} (x_{t-k} - \overline{x_t})(y_t - \overline{y_t}) \qquad k \in \{-M, \ldots, 0, \ldots M\}$$

   Se muestra en la figura 2.6, el retraso en k unidades de una serie con respecto a la otra, y como esto implica un cambio en coeficiente $\hat{\rho_{x_t y_t}}$

2. Para establecer cercanía $k \in \mathbb{Z}$ se define el operador $\mathrm{d}_k(\cdot,\cdot)$entre dos series temporales como

   $$\mathrm{d}_k(\{X_t\}_{t\in\mathbb{Z}}, \{Y_t\}_{t\in\mathbb{Z}}) = \sum_{t\in\mathbb{Z}} (X_{t+k} - Y_t)^2$$

   donde su estimador esta dado por

   $$\hat{\mathrm{d}}_k(\{x_t\}, \{y_t\}) = \sum_{t=k}^{n} (x_{t-k} - y_t)^2 \qquad k \in \{-M, \ldots, 0, \ldots M\}$$

   cabe destacar que solo para $k = 0$ se tiene que d_k es métrica.

2.6.2. Estimación y ajuste *ARIMA*

Para establecer cuan bien ajusta el modelo *ARIMA*, se plantea realizar la comparación con dos pronósticos distintos,

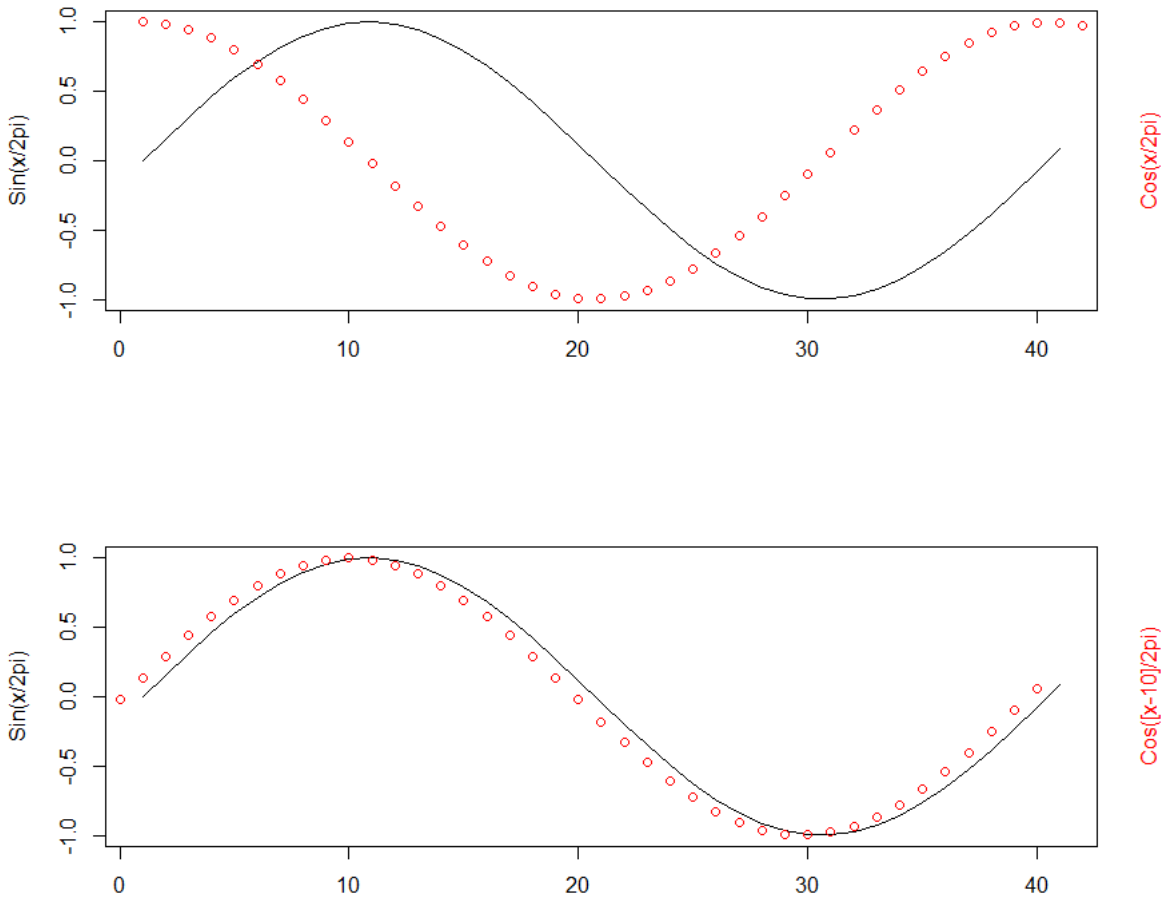

Figura 2.6: En la figura de arriba se muestra las series $\sin(x/[2\pi])$ y $\cos(x/[2\pi])$, con un coeficiente $\hat{\rho}_{x_t y_t}(0)$ =.003 entre ambas series, mientras que para $\hat{\rho}_{x_t y_t}(10) \sim 1$, se muestra en la segunda figura, como coinciden cuando se desplaza.

- *Paseo Aleatorio:* Estimar X_{n+h} de la forma

$$X_{n+h} = X_n + \varepsilon_1 + \cdots + \varepsilon_h$$

  con $\varepsilon_i \sim N(0, \sigma^2)$ para algún σ^2 adecuado.
- *Por promedio:* Estimar X_{n+h} de acuerdo al valor esperado del ciclo correspondiente. Es decir $X_{n+h} = \frac{1}{S}\sum_{k=1}^{S} X_{n+h-k\cdot s}$ donde s es la periodicidad de la serie $\{X_t\}_t$ y S la cantidad de ciclos.

Para comparar cuan bien ajustan los modelos predictivos, se plantea el análisis sobre el error definido por

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

Y sobre este conjunto, se utiliza primero la *prueba de Shapiro-Wilk* para contrastar normalidad en las v.a. En el caso de asegurar normalidad en el conjunto $(e_t)_t$ se utiliza la prueba *t de Student* para establecer si la media μ de e_t es igual a cero como se describe en la sección 1.2.1 y 1.2.2.

En caso de no poder establecer normalidad en $(e_t)_t$ se utiliza la *prueba de Wilcoxon* para establecer si su mediana es identicamente nula como se describe en la sección 1.2.3.

2.7. Base de datos

Los datos recaudados para construir la plataforma GIS y para la investigación del modelo *ARIMA*,

1. *Concentración de Ozono, Concentración de Material Particulado (PM), Neumonía, Asma* Tiene fecha de inicio enero del 2007, y fecha de termino diciembre del 2012. Por una parte, los ingresos hospitalarios por Neumonía y Asma son variables mensuales. Por otro lodo, Las variables Material particulado y Ozono, en ppm (*partes por millón*) calculada por la media móvil diaria.

   Cabe mencionar que existe un sesgo en los ingresos hospitalarios por asma y neumonía, donde esta involucrada la decisión de las personas por atenderse en una u otra comuna. Se considera así las variables $AsmaSantiago$ y $NeumoniaSantiago$ como los ingresos hospitalarios de todas las comunas, de la forma

$$AsmaSantiago_t = \sum_{i \in comunas} Asma_t^i$$
$$NeumoniaSantiago_t = \sum_{i \in comunas} Neumonia_t^i$$

2. *Precipitación*

   Datos adquiridos desde CR2, y la base de datos contiene la secuencia temporal de precipitación en [mm] de estaciones de la DGA (Dirección General de Aguas) y estaciones de la DMC (Dirección Meteorológica de Chile), que contienen la distribución geográfica relevante de cada una de las estaciones (Latitud, Longitud, Altura m.s.n.m, nombre). Esta base de datos, mensual , tiene como fecha de inicio enero de 1940, y fecha de termino diciembre 2013[1] .

[1] Algunas estaciones tienen su base actualizada hasta septiembre del 2015

Capítulo 3

Resultados

En el siguiente capítulo, se exponen, finalmente, los resultados de la investigación realizada. Como primer elemento, se muestra el análisis sobre series temporales, con mayor énfasis en los ajustes del modelo *ARIMA* en las distintas variables de los datos expuestos en 2.7. Como segundo elemento, se muestra la estructura de la plataforma interactiva a través de extractos de código ejecutados e imágenes de la plataforma.

3.1. Análisis series temporales

Dentro de la sección, se muestran los análisis planteados en la sección 2 para procesos temporales, el ajuste y pronóstico del modelo *ARIMA* y, la razón de verosimilitud entre *ARIMA* y los descritos en 2.6.1, para compararlos.

3.1.1. Proyecciones modelo *ARIMA*

Se muestran a continuación, los resultados del ajuste de los parámetros bajo la función *ARIMA*, simulado con `R` para las variables Neumonía en Santiago, Concentración de Ozono, Concentración de Material particulado y precipitación. En los cuadros, se indican los parámetros calculados, junto con el error cuadrático asociado. El objetivo de esta sección es observar el comportamiento de los pronósticos realizados por el ajuste `arima`.

Neumonía Santiago

Para los ingresos por neumonía en Santiago, se ajusta el modelo para los primeros 4 años (desde enero 2007, hasta diciembre 2010), para luego comparar el comportamiento de la predicción con el último año (enero 2011, hasta diciembre 2011). La serie temporal en este caso, es aproximada por `arima` con los parámetros $(p, \mathrm{d}, q) \times (P, D, Q)_s = (2, 0, 1) \times (2, 0, 1)_{12}$ como d = 0, `R` ajusta la constante c (`intercept` en los resultados). Una vez realizado el ajuste del modelo, se realiza la predicción y se muestra en la figura 3.1 la curva pronosticada por el modelo.

```
Call:
arima(x = ts(neumoniasantiago, start = c(2007, 1), freq = 12), order
   = c(2,
    0, 1), seasonal = list(order = c(1, 0, 1), period = 12))

Coefficients:
         ar1      ar2     ma1    sar1     sma1  intercept
      0.4277  -0.1051  0.4697  0.9998  -0.9676   269.2395
s.e.  0.2976   0.2325  0.2701  0.0012   0.0944    54.0099

sigma^2 estimated as 2570:  log likelihood = -271.08,  aic = 556.17
```

Cuadro 3.1: Resultado modelo $\texttt{arima}(2, 0, 1) \times (2, 0, 1)_{12}$, para el ajuste de ingresos por neumonía de Santiago

Figura 3.1: Pronóstico (curva azul) bajo el ajuste $\texttt{arima}(1, 0, 1) \times (1, 0, 1)_6$ para la variable neumonía en Santiago. El intervalo corresponde al 80 % y 90 % respectivamente

Material particulado Las Condes

Los contaminantes de la base de datos adquirida, originalmente corresponden a procesos temporales diarios, donde cada x_t es la media móvil diaria. Se estimó considerar el dato mensual como el promedio de los días del mes. De esta forma, se aplica el modelo *ARIMA* con los parámetros $(p, \mathrm{d}, q) \times (P, D, Q)_s = (1, 0, 1) \times (1, 0, 1)_{12}$ a la serie mensual de material particulado en la comuna de Las Condes.

```
Call:
arima(x = ts(y, start = c(2007, 1), freq = 12), order = c(1, 0, 1),
   seasonal = list(order = c(1,  0, 1), period = 12))

Coefficients:
         ar1      ma1    sar1     sma1  intercept
      0.5599  -0.0165  0.7958  -0.5037    22.5770
s.e.  0.2016   0.2543  0.3295   0.4763     1.9354

sigma^2 estimated as 14.89:  log likelihood = -134.72,  aic = 281.45
```

Cuadro 3.2: Resultado modelo `arima` para material particulado comuna de Las Condes

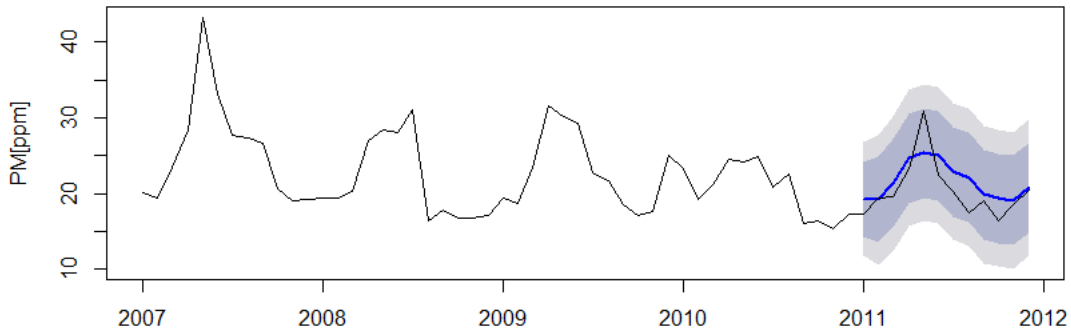

Figura 3.2: Pronóstico (azul) gráfico bajo el ajuste `arima`$(1, 0, 1) \times (1, 0, 1)_6$ para la variable mensual material particulado, comuna de Las Condes

Ozono comuna Cerillos

Para ozono, al igual que para material particulado, se consideran datos mensuales como el promedio de las medias móviles de los datos diarios.

A pesar que la serie tenga un estacional anual, se observa en la figura 3.3 que contiene un periodo cada 4 meses. Es por esto, que se fija en este caso particular $s = 4$, fijando así los parámetros $(p, \mathrm{d}, q) \times (P, D, Q)_s = (1, 0, 1) \times (2, 1, 1)_4$ a la serie mensual de Ozono, comuna de Cerrillos. Los parámetros ajustados se observan en 3.3.

```
Call:
arima(x = ts(y, freq = 12, start = c(2007, 1)), order = c(1, 0, 1),
   seasonal = list(order = c(2,
    1, 1), period = 4))

Coefficients:
          ar1     ma1     sar1     sar2    sma1
      -0.4165  0.4010  -1.4081  -0.6179  0.8974
s.e.   1.9431  1.9328   0.1864   0.1447  0.3132

sigma^2 estimated as 31.48:  log likelihood = -140.43,  aic = 292.86
```

Cuadro 3.3: Resultado modelo `arima` para media móvil de ozono (promedio mensual) para la comuna de Cerrillos

Figura 3.3

Precipitación

Originalmente, los datos de precipitación datan desde enero de 1940 hasta diciembre del 2013, sin embargo, los pronósticos se pudieron comparar con una base de datos actualizada que contiene la serie de precipitación desde enero de 2014 hasta septiembre de 2015.

El pronóstico de la estación Quinta Normal, Santiago, es comparado con la serie desde enero 2014 hasta septiembre 2015, mientras que la estación Valparaíso gobernación marítima, puede ser comparada con los datos mensuales del año 2014.

En la figura 3.4, se muestra el pronóstico realizado, mientras que los resultados del ajuste se muestran en el cuadro 3.4.

```
#########################################################
######Resultados Ajuste para Quinta normal############
#########################################################
Call:
arima(x = ts(w, start = c(2000, 1), freq = 12), order = c(1, 1, 2),
    seasonal = list(order = c(2, 1, 2), period = 12))

Coefficients:
          ar1      ma1     ma2     sar1     sar2    sma1     sma2
      -0.3711  -0.6249  -0.375  -0.9449  -0.1121  0.2423  -0.7573
s.e.     NaN      NaN     NaN   0.1241   0.1416  0.1541   0.1397

sigma^2 estimated as 234.4:  log likelihood = -657.04,  aic = 1330.08

#########################################################
########Resultados Ajuste para Valparaiso ############
#########################################################
Call:
arima(x = ts(w, start = c(2000, 1), freq = 12), order = c(1, 1, 1),
   seasonal = list(order = c(0, 2, 2), period = 12))

Coefficients:
         ar1      ma1     sma1    sma2
      0.0254  -0.8519  -1.9066  0.9927
s.e.  0.1128   0.0813   0.1179  0.1175

sigma^2 estimated as 25218:  log likelihood = -949.89,  aic = 1909.78
```

Cuadro 3.4

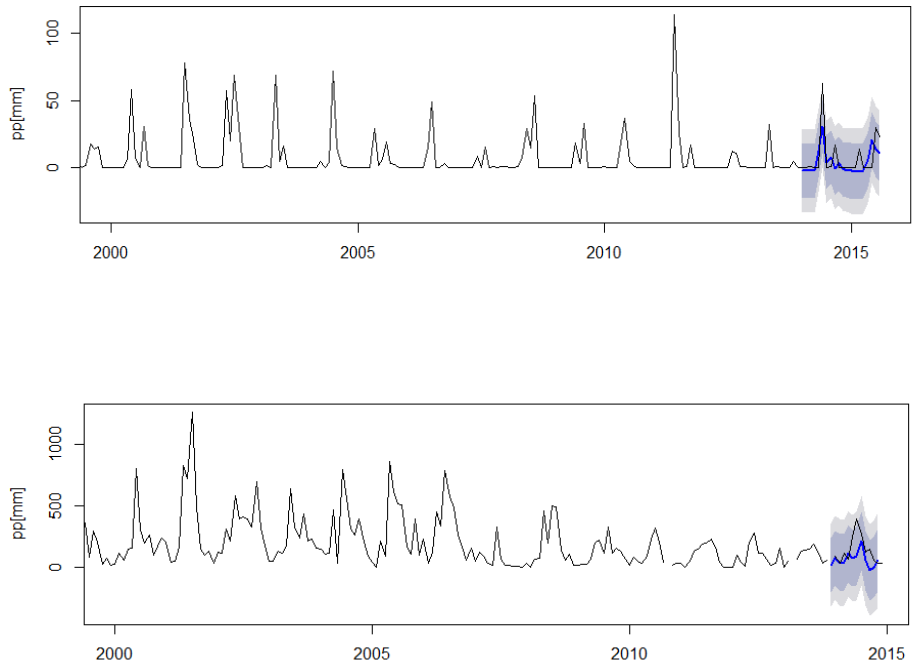

Figura 3.4: Pronóstico gráfico bajo el ajuste $\texttt{arima}(1,0,1)\times(1,0,1)_6$ para la variable material particulado, comuna de La Florida

3.1.2. Ajuste y comparación modelo *ARIMA*

La presente sección se plantean distintos métodos para establecer cuan bien ajusta la predicción bajo $ARIMA$. En conjunto con esto, se contrasta $ARIMA$ con las predicciones estándar establecidos en 2.6.1.

Estimación del error

Para establecer la comparación, se muestra el comportamiento del error definido por

$$\mathrm{e}_t = \hat{y}_t - y_y$$

donde $\hat{y}_t$ corresponde a las predicciones realizados por los tres modelos en el instante t. Para las estaciones meteorológicas de Quinta Normal, Santiago y Riecillos, con los datos desde enero 2002 hasta diciembre del 2011. Cada modelo genera un pronostico para el periodo enero-diciembre 2012 y se muestra en las figuras 3.5, 3.6, 3.7 las predicciones realizadas por los tres modelos predictivos para la estación Quinta Normal, Santiago.

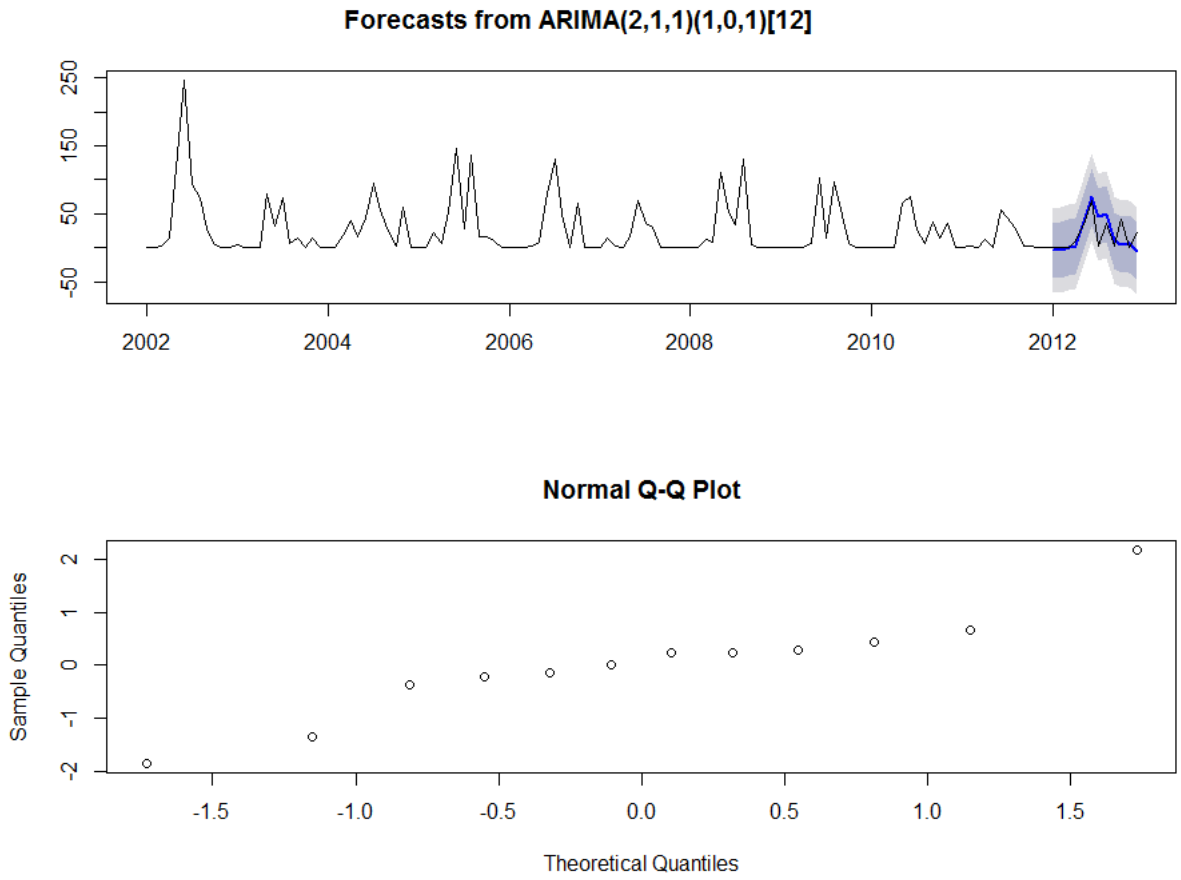

Figura 3.5: Normalidad en el error generado por el pronostico para pp en la estación meteorológica Quinta Normal, Santiago, años 2002-2011 pronosticando 2012

Como se establece en 1.2.1, se realiza primero una prueba para establecer si el error asociado a la predicción se comporta de manera normal, mediante la prueba de *Shapiro-Wilk*, y si esta tiene media nula o no a través de la prueba *t de Student*. Por otro lado de manera más genérica, se utiliza la prueba de *Wilcoxon* para decidir si el error tiene mediana nula (independiente de la distribución que el error tenga).

el cuadro 3.5 indica que el $p-$valor es .3155 para el error asociado a la predicción por $ARIMA$, lo que implica que no se rechaza la hipótesis nula, es decir que el error se comporta como una distribución normal.

Como se observa en la figura 3.6, la predicción realizada con el modelo aleatorio no predice de manera fidedigna la serie temporal, sin embargo, la predicción es generada por variables aleatorias normales, por lo que la prueba de *Shapiro -Wilk* decide no rechazar la hipótesis nula ($(e_t)_t$ sigue una distribución normal) como se observa en el cuadro 3.6.

Por último, la prueba de *t de Student* y la prueba de *Wilcoxon* sobre la predicción realizada

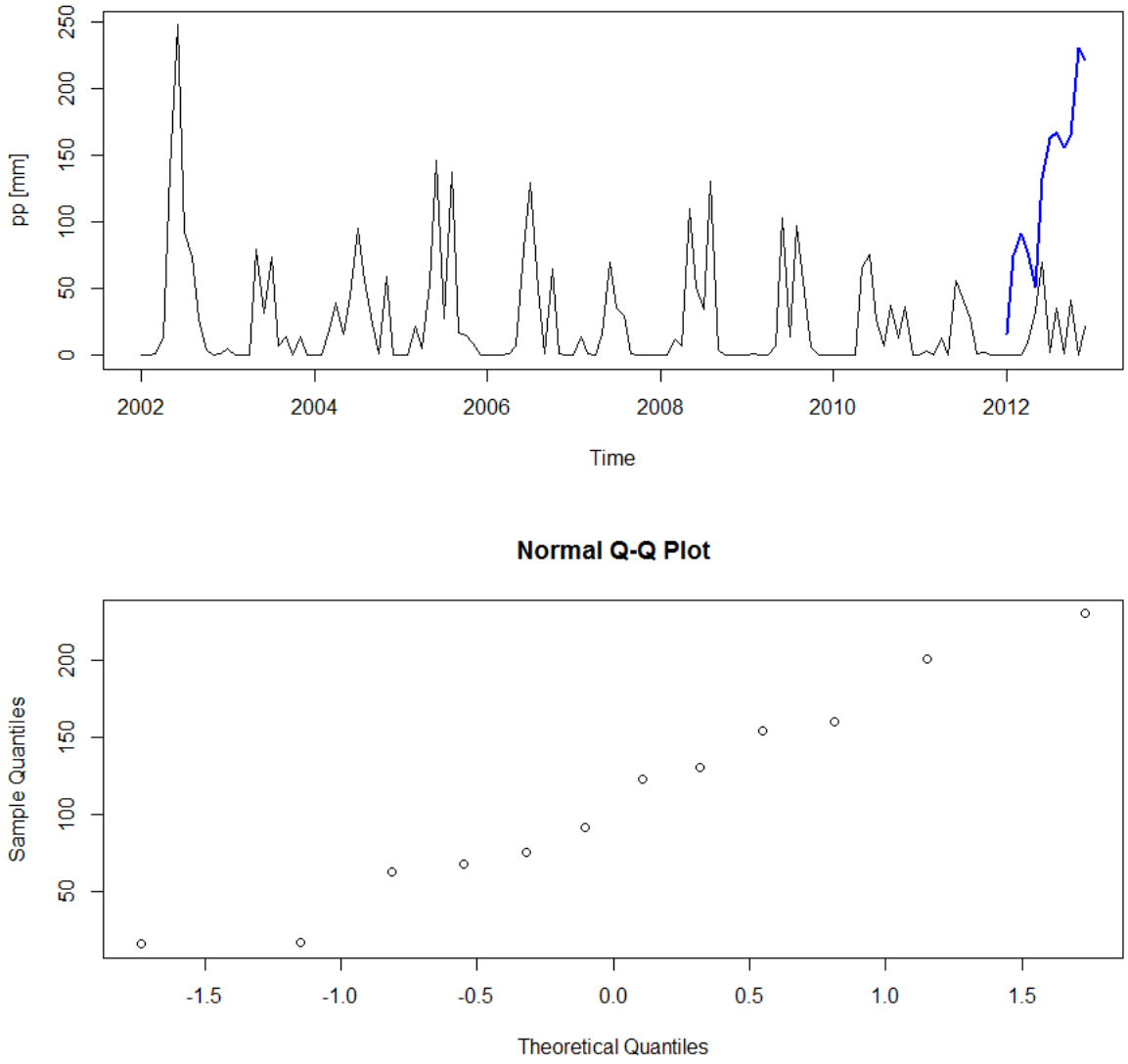

Figura 3.6: Normalidad en el error producido por el pronostico realizado de manera aleatoria, estación meteorológica Quinta Normal, Santiago, años 2002-2011 pronosticando 2012.

```
>shapiro.test(error)
Shapiro-Wilk normality test
data:  error
W = 0.92341, p-value = 0.3155

> t.test(error)

        One Sample t-test

data:  error
t = 0.033198, df = 11, p-value = 0.9741
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -12.40672  12.78672
sample estimates:
mean of x
0.1900018

> wilcox.test(error)

        Wilcoxon signed rank test

data:  error
V = 45, p-value = 0.6772
alternative hypothesis: true location is not equal to 0
```

Cuadro 3.5: Formato de resultado de las pruebas realizadas al error provocado por la predicción bajo el modelo *ARIMA*

por *ARIMA*, con un p-valor de .97 y .68 respectivamente, son los más alto de las tres predicciones como se observa en el cuadro 3.6.

Modelo predictivo	p-valor *test de Shapiro*	p-valor *t-test*	Media estimada	p-valor *test de Wilcoxon*
ARIMA	.32	.97	.19	.68
Aleatorio	.81	~ 0	159.9	~ 0
Promedio ciclico	.73	0.22	8.32	.20

Cuadro 3.6: Tabla de resultados sobre las predicciones sobre los distintos modelos predictivos para el año 2012 de la estación *Quinta Normal, Santiago*.

El mismo análisis realizado para la estación meteorológica de *Quinta Normal, Santiago* se replica para la estación meteorológica de *Riecillos*, y se resumen los resultados en el cuadro 3.7.

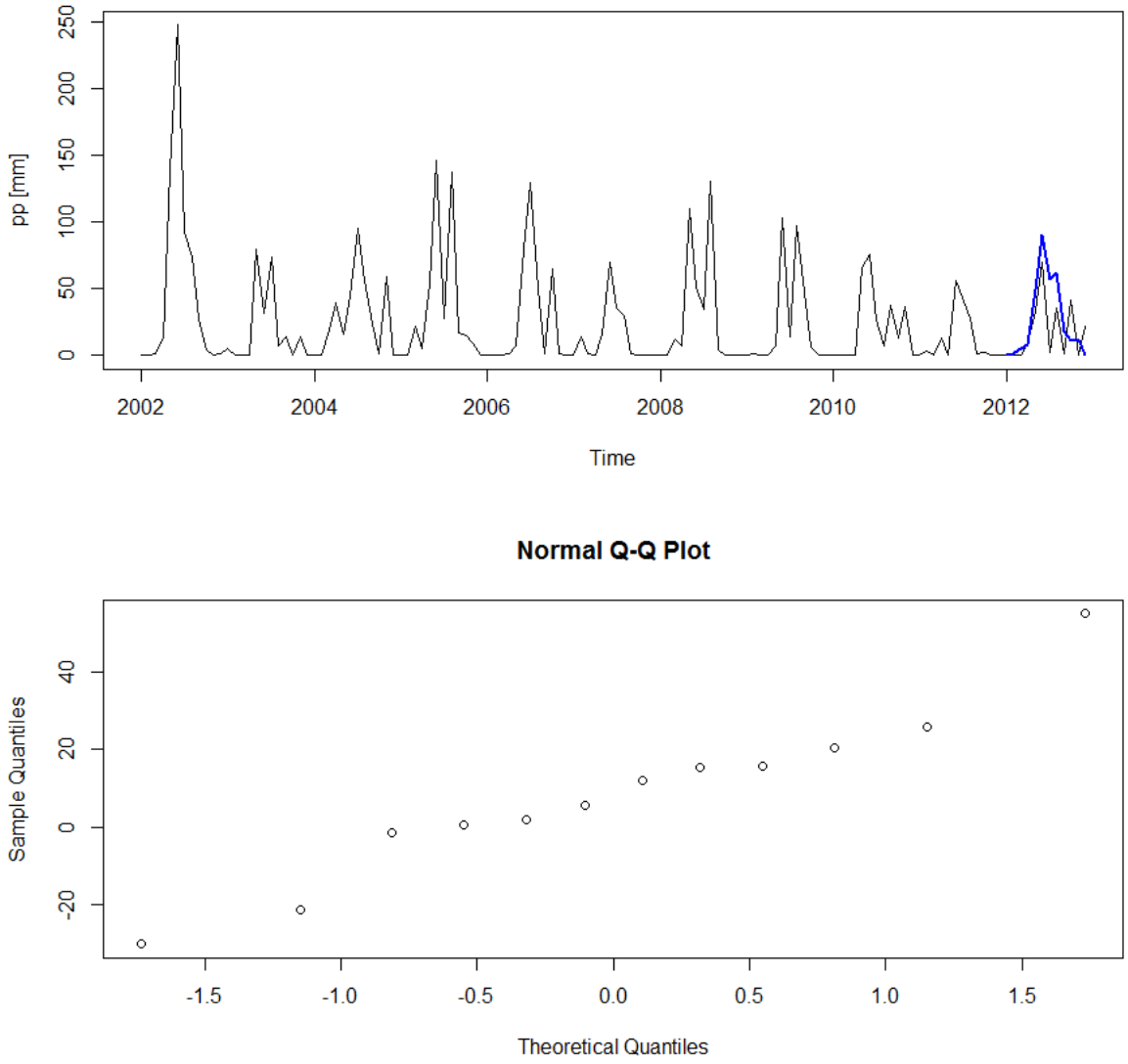

Figura 3.7: Normalidad en el error generado por pronostico de promedio ciclico, estación meteorológica Quinta Normal, Santiago, años 2002-2011 pronosticando 2012.

Modelo predictivo	p-valor *test de Shapiro*	p-valor *t-test*	Media estimada	p-valor *test de Wilcoxon*
ARIMA	.62	.41	8.66	.34
Aleatorio	.95	~ 0	67.1	~ 0
Promedio ciclico	.64	.16	17.45	.18

Cuadro 3.7: Tabla de resultados sobre las predicciones sobre los distintos modelos predictivos para el año 2012 de la estación *Riecillos*.

De igual manera que la estación Quinta Normal, Santiago, para la estación meteorológica de *Riecillos*, el pronostico según *ARIMA* es el que arroja un un P-valor más alto en las pruebas (con un P-valor igual a .41 y .34 en las pruebas *t de Student* y de *Wilcoxon* respectivamente), es decir es mejor ajustado que las otras predicciones.

3.1.3. Comparación de series temporales

Mediante la búsqueda de similitud entre series temporales, se desea encontrar qué variable explica de mejor forma a otra. Por ejemplo, se desea averiguar qué comuna explica mejor los ingresos por Neumonía. Luego, para establecer cercanía, se utilizan los coeficientes planteados en la sección 2.6.1.

Antes de establecer cercanía o similitud entre series temporales, guardan relación *Asma* y *Ozono* (ver [11]) y se evidencia en la figura 3.8 las variables que, *a priori*, estan relacionadas.

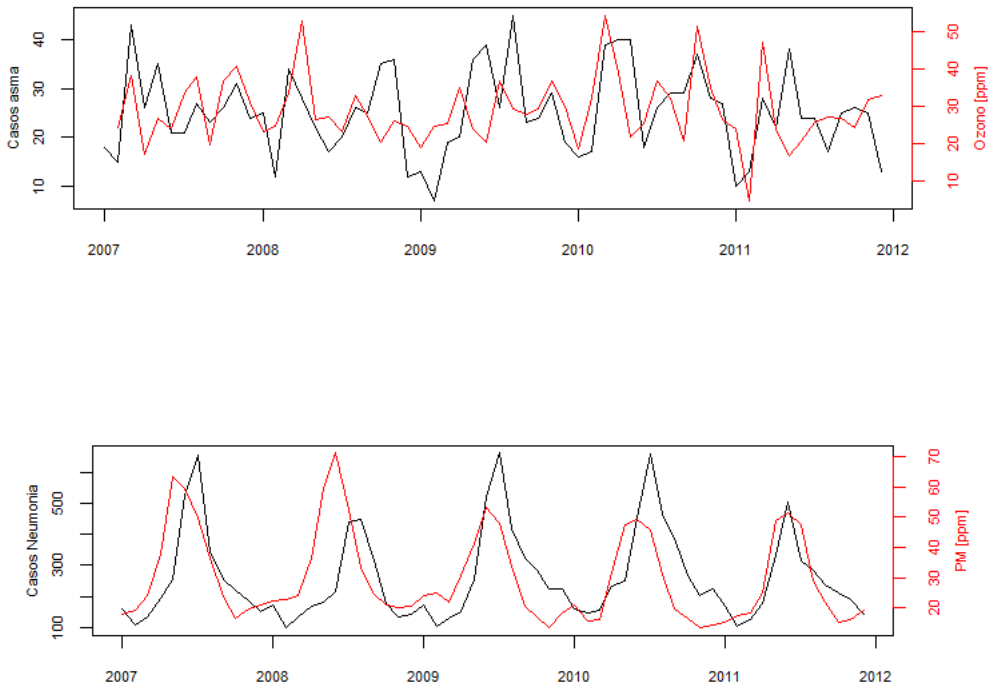

Figura 3.8: En la primera figura se muestra la serie temporal de casos de Asma junto con Ozono, comuna de Pudahuel [ppm], mientras que en la segunda, se muestran casos de Neumonía junto a Material Particulado (PM) [ppm]

Luego, una vez establecido que las variables a comparar son, *Asma Santiago con Ozono* y *Neumonía Santiago* con *Material particulado*, para cada comparación se realiza el ajuste del mejor retraso en k meses que permita encontrar la mejor serie explicativa. De esta forma, recordando que

$$\hat{\rho}_{x_t y_t}(k) = \frac{1}{\hat{\sigma}_x \hat{\sigma}_y} \sum_{t=k}^{n} (x_{t-k} - \overline{x_t})(y_t - \overline{y_t}) \qquad k \in \{-M, \ldots, 0, \ldots M\}$$

$$\hat{\mathrm{d}}_k(\{x_t\}, \{y_t\}) = \sum_{t=k}^{n} (x_{t-k} - y_t)^2 \qquad k \in \{-M, \ldots, 0, \ldots M\}$$

se muestra en el cuadro 3.8 para x_t = Casos de Neumonia y_t = Concentración de PM [ppm] para las comunas de Pudahuel, La Florida, Independencia. Dentro de esta contexto, el retraso k corresponde a meses.

En el cuadro 3.9, se muestra los coeficientes para las variables x_t = Concentración Ozono

[ppm] de las comunas de Cerrillos, Las Condes, Independencia y = Casos de Asma.

	$\hat{\rho}(k)$			$\hat{\mathrm{d}}(k)$		
	$k = 0$	$k = 1$	$k = 2$	$k = 0$	$k = 1$	$k = 2$
Pudahuel	.57	.79	.66	.039	.021	.031
La Florida	.43	.70	.62	.057	.037	.042
Independencia	.41	.79	.85	.044	.027	.029

Cuadro 3.8: Tabla con los coeficientes $\hat{\rho}_{x_t y_t}(k)$ y $\hat{\mathrm{d}}_{x_t y_t}(k)$

	$\hat{\rho}(k)$				$\hat{\mathrm{d}}(k)$			
	$k = 0$	$k = 1$	$k = 2$	$k = 3$	$k = 0$	$k = 1$	$k = 2$	$k = 3$
Cerrillos	.009	.117	-.040	.197	.057	.051	.060	.048
Las Condes	.009	-.014	-.041	.167	.064	.066	.068	.056
Independencia	.032	.043	-.083	.051	.059	.057	.065	.058

Cuadro 3.9: Coeficientes calculados

Luego se muestra una mejor correlación para las variables de Concentración de Material Particulado, comuna Independencia con Neumonía en Santiago, con un coeficiente $\hat{\rho}(2) = ,85$ y con una distancia $\hat{\mathrm{d}}(2) = ,029$. Lo que se condice con un tiempo de respuesta de los ingresos hospitalarios luego del aumento en la Concentración de PM. La figura 3.9 muestra el comportamiento de las curvas de Neumonía retrasada en k con respecto a Material Particulado, comuna de Independencia.

Por otro lado, la variable que mejor explica los ingresos por Asma en Santiago, con coeficientes $\rho_{x_t y_t}(3) = ,197, \mathrm{d}_{x_t, y_t}(3) = ,048$ es la Concentración de Ozono en la comuna de Cerrillos, con un retraso de 3 meses.

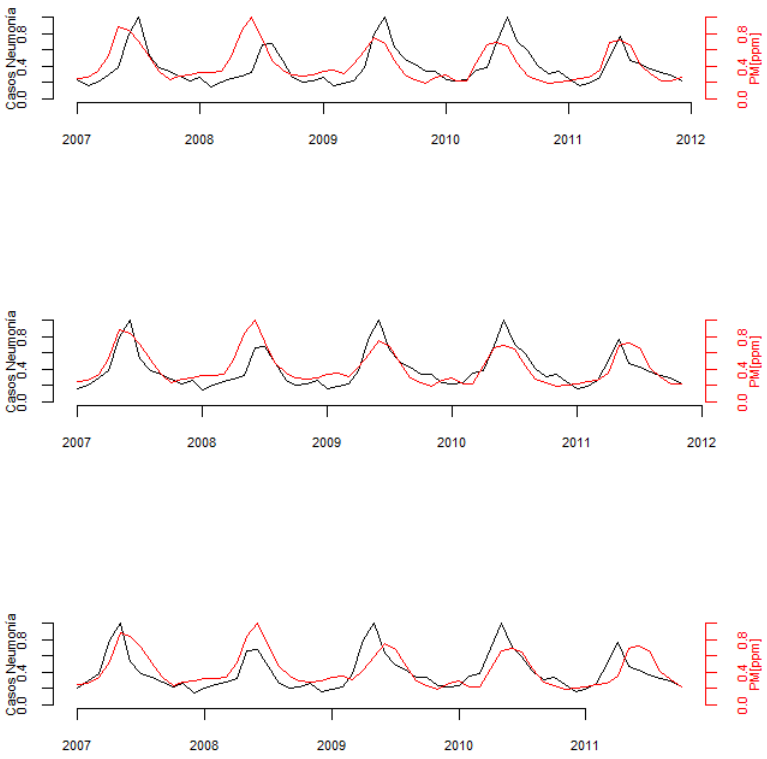

Figura 3.9: Ingresos por Neumonía retrasado en k con respecto a Material Particulado, comuna de Independencia ($k = 0, 1, 2$ de arriba a bajo).

3.2. Plataforma

En la presente sección se describe la estructura interna del GIS y los lenguajes que son utilizados en su construcción. En conjunto con esto, se explica en detalle el rol de las funciones utilizadas.

3.2.1. Plataforma interactiva

El objetivo principal de la plataforma es proveer a un usuario, información exploratoria de los datos almacenados y resultados de métodos estadísticos que relacionen o comparen variables dentro de la región.

La plataforma, específicamente, contiene la distribución geográfica de la región, división por comunas, y distribución de estaciones meteorológicas (en lo que sigue *comunas* y *estacio-*

nes, respectivamente), que pueden ser seleccionadas para obtener la información particular del objeto en cuestión. En conjunto con esto, la plataforma permite realizar interacciones usuario-plataforma, mediante el llenado de formularios o acciones (selecciones), por parte del usuario, sobre el mapa interactivo. El lenguaje `PHP`, con el cual se construye la plataforma, permite la realización de estas acciones (este lenguaje esta destinado al contenido web-dinámico). Luego, una vez establecida la acción por parte del usuario, el código `PHP` ejecuta lineas de comando escritas en lenguaje `R` en un servidor remoto, produciendo análisis y gráficas para ser mostradas en el GIS.

3.2.2. Lenguajes y librerías

Como segundo objetivo del GIS construido, se plantea que cualquier usuario de la plataforma, con o sin conocimientos estadísticos previos, pueda ser capaz de interactuar e interpretar resultados desde los análisis posibles. Es por esto, que la confección de la parte interactiva se realiza con lenguaje `PHP` en conjunto con `JavaScript`, que proveen herramientas para el contenido dinámico y amigable con el usuario.

Por otro lado, los cálculos estadísticos y producción de gráficos se realizan con el lenguaje `R`, a través de códigos ejecutados en un servidor externo.

A continuación, se detallarán las librerías (*library*) para `PHP` y `JavaScript`, y paquetes (*packages*) para `R` que son utilizados para la producción de la plataforma.

Librerías

1. *Leaflet JavaSript*

   La librería fundamental dentro de la plataforma es Leaflet JS, librería JavaScript de código libre (*open-source*), diseñada para realizar mapas interactivos y con ellos realizar aplicaciones web y móviles.

   Leaflet JS, cuenta con un diseño simple y atractivo, donde el último usuario tiene fácil acceso a información variada dentro de una región determinada. Cabe destacar, que la sintaxis de esta librería permite la fácil comunicación con los otros lenguajes de programación, necesarios para la plataforma (*PHP y R*).

2. *Paquetes en R*

   Para el análisis estadístico del modelo ARIMA, existen dos paquetes que proveen funciones en torno a este estudio: `stat y forecast`. Por un lado, `stat`, con autoría de: `R Core Team and contributors worldwide`, contiene funciones generales destinadas al análisis estadístico, uni y multivariadas. Mientras que `forecast`[1] es un paquete des-

[1]Cuya autoriaes de Rob J Hyndman, Contributors include George Athanasopoulos, Christoph Bergmeir, Carlos Cinelli, Yousaf Khan, Zach Mayer, Slava Razbash, Drew Schmidt, David Shaub, Yuan Tang, Earo

```
var map = L.map('map').setView([-33.5, -71], 8);
var basemap = L.tileLayer(
                'http://{s}.tiles.mapbox.com/v3/sarasafavi.hjegnofh/{
                   z}/{x}/{y}.png',
                      {maxZoom: 12});

var info1 = L.control({position: 'topleft'});

info1.onAdd = function (map){
                        var div = L.DomUtil.create('div', 'info');
                                {div.innerHTML = [...]} }

basemap.addTo(map);
var geojson = L.geoJson(rm,{style: style,
                                onEachFeature: onEachFeature}).addTo(
                                   map);
L.geoJson(markers,{onEachFeature: onEachFeatureMarker}).addTo(map);
```

Cuadro 3.10: Extracto de código utilizado en la plataforma interactiva

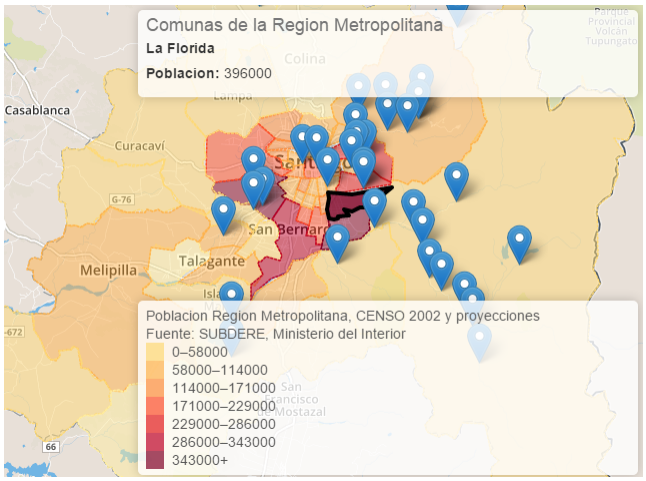

Figura 3.10: Mapa interactivo de Santiago de Chile, donde se visualiza la división por comunas de la región. Esta imagen comprende la ejecución del código mostrado en 3.10

tinado al pronóstico de series temporales, incluyendo el método *ARIMA* y funciones, que contribuyen a la búsqueda de los mejores parámetros del modelo.

Conjunto a lo anterior, se utilizó el paquete `HTML2` para generar salidas y resúmenes de los análisis, en formato HTML y que puedan ser incrustadas directamente a la plataforma.

Wang, Zhenyu Zhou.

Interacción entre los lenguajes

El esqueleto de la comunicación PHP-R establecida para el funcionamiento de la plataforma, se reduce al código en el cuadro 3.11, el cual consiste en el ingreso de un conjunto de variables y el análisis estadístico deseado (o `input`). Luego los gráficos de salida son expuestos en la plataforma mediante la secuencia descrita en el cuadro 3.12.

```
exec("Rscript ScriptR.R 'Var'");
```

Cuadro 3.11: Extracto de código para ejecutar un archivo (en particular R) utilizado en la plataforma

```
<!---- Imagen de salida ---->
<div id="grafico_de_caja" class="col-md-6"><?="<img src=$grafico1
   >"?></div>
<!---- Imagen de salida ---->
<div id="grafico_serie_descomposicion" class="col-md-6"><?="<img src=
   $grafico2>"?></div>
```

Cuadro 3.12: Código PHP para mostrar imagen

Donde `ScriptR.R`, es un conjunto de instrucciones en R que se ejecuta desde un servidor externo, y `Var` es el conjunto de variables que ingresa el usuario, para que `ScriptR.R` arroje alguna respuesta (`output`) que será entregada al usuario en forma de gráfico o tabla.

Funciones de R

Los métodos estadísticos utilizados en la plataforma fueron,

- Análisis de varianza `aov`.
- Test de independencia de medias, `t.test`.
- Test de correlación, `cor.test`.
- Del paquete `stat` se utilizó la función `decompose`, que descompone una serie temporal en la forma (2.1).
- Modelo lineal generalizado, `glm`, del paquete `stat`.
- Modelo predictivo *ARIMA*, con la función `forecast` del paquete `forecast`.

3.2.3. Funciones y cualidades

Finalmente, la plataforma construida, es una aplicación web de geo-referenciamiento que muestra la Región Metropolitana, dividida por comunas. Dentro del mapa, se identifican dos

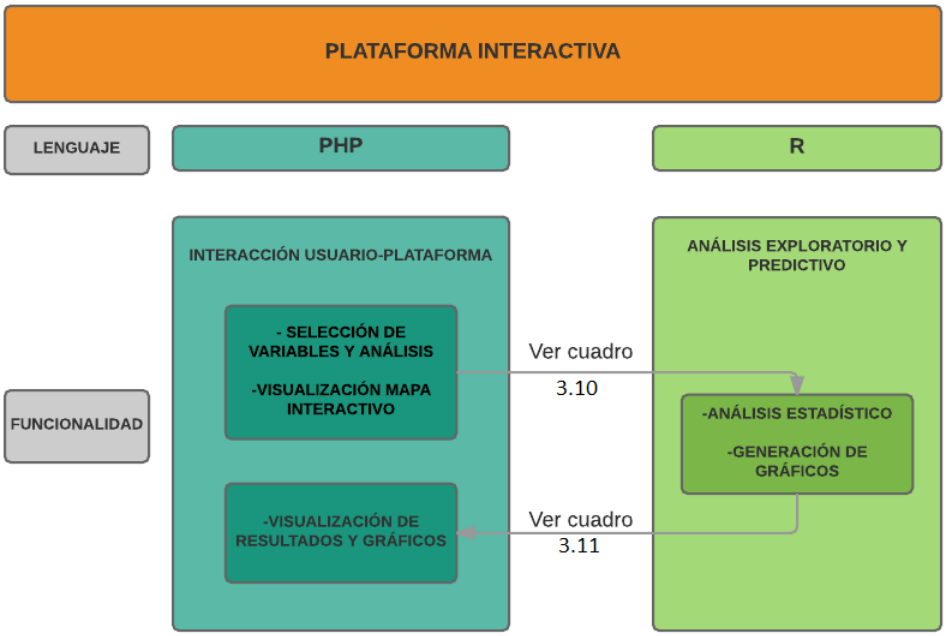

Figura 3.11: Esquema de la plataforma interactiva que resume la funcionalidad e interacción entre los lenguajes `PHP` y `R`

objetos que contienen información particular, *comunas* y *estaciones*, como se muestra en la figura 3.12.

Este prototipo contiene dos armazones (*frameworks*) principales. Primero, aquel que permite escoger variables Asma, Neumonia, Ozono, Material Particulado, sobre las comunas y da la opción de escoger test de correlación o independencia, análisis de varianza y un modelo lineal, como se muestra en la figura 3.13. El segundo armazón, permite acceder a información descriptiva acerca de los datos a través de gráficas, luego de escoger algún objeto (*comuna* o *estación*) dentro del mapa interactivo.

El primer armazón descrito anteriormente, arroja como salida un resumen de los cálculos estadísticos realizados, en el caso de `t.test` y `cor.test` (como se ve en el cuadro), mientras que en el caso de análisis anova y modelo lineal, se muestran los resultados y gráficas que contribuyen a su interpretación. A continuación, se muestra una imagen del formato de salida de estas posibilidades, y extractos de los códigos de `R` utilizados en la ejecución de ellas.

El cuadro 3.11 muestra la ejecución de un archivo `R` desde `PHP`, esto llama a ejecutar, por ejemplo, un extracto de código destinado a la parametrización de un modelo lineal, como se muestra en el cuadro 3.13 y la figura 3.11.

Una vez ejecutado el código mostrado en 3.13, arrojará como respuesta un resumen del análisis realizado desde el servidor, en conjunto de imágenes que contribuyen al entendimiento de los resultados, como se muestran en las figuras 3.14 y 3.15.

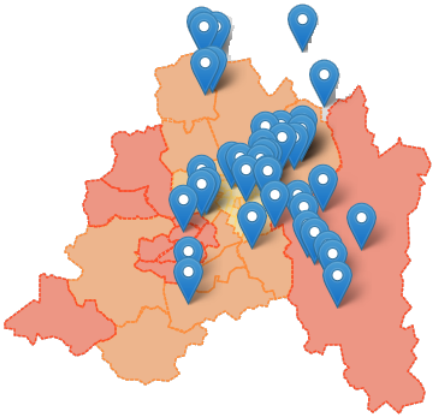

Figura 3.12: Imagen exhibiendo los objetos gráficos dispuestos dentro del mapa interactivo, *Comunas* y *Estaciones*

Figura 3.13: *Framework* destinado al ingreso de variables y análisis deseados por el usuario

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.5e+01	2.1e+00	7.3e+00	1.4e-09
ozono	-4.3e-01	1.1e-01	-3.8e+00	3.6e-04
mat	3.8e-03	4.7e-02	8.0e-02	9.4e-01
estac	-5.1e+01	1.3e+01	-3.9e+00	2.5e-04
ozono:estac	4.8e+00	1.3e+00	3.7e+00	5.2e-04
mat:estac	3.1e-01	3.3e-01	9.5e-01	3.5e-01

Figura 3.14

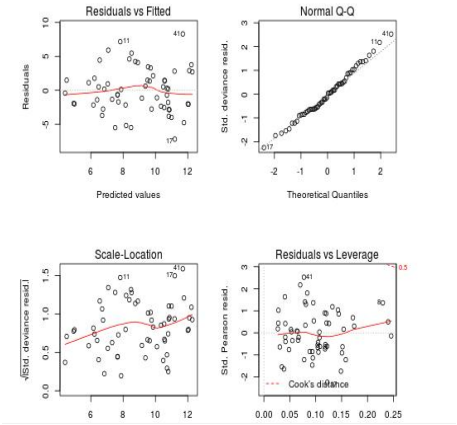

Figura 3.15

Asimismo, el mapa permite seleccionar los objetos mencionados anteriormente, lo que arrojará un resultado gráfico dependiendo del tipo de objeto que se escoja, y éste es desple-

```
args = (commandArgs(TRUE))

form<-paste0(as.character(args[1]), ' ~ ',  as.character(args[2]))
for(i 2:length(args)){
  form<-paste0(form, ' + ', as.character(args[i]))
}
modelo=glm(form)

#########################################################
####          Ejemplo
####   modelo=glm(asma.LaFlorida ~ ozono.LaFlorida+mat.LaFlorida)
#########################################################

df<-summary(modelo)\$coefficients
```

Cuadro 3.13: Extracto de código utilizado para realizar un modelo lineal. Se establece un ejemplo: asma explicado según ozono y material particulado

gado en el segundo armazón.

Análisis exploratorio de la plataforma

Por una parte, el análisis exploratorio que permite la plataforma cuenta con relaciones entre variables como se ha mencionado anteriormente (t-test, test de correlación, ANOVA, y modelo lineal). Esto se realiza seleccionando sobre el *framework* como se indica en la figura 3.13, y arroja resultados resumidos como lo indican las figuras 3.14 y 3.15

Por otro lado, se cuenta con la descripción de cada objeto (*comunas* y *estaciones*) dentro de la plataforma interactiva a través de la selección de ellos. Si el objeto seleccionado es una *comuna*, se muestran gráficas de las variables expuestas por comuna, como se ve en la figura 3.16 la selección de la comuna de *La Florida* y bajo ella, se muestra la salida gráfica que ésta tiene. En el cuadro 3.14 se muestra un extracto de la secuencia de código `R` destinada a generar la salida de dichos gráficos. Del mismo modo, si el objeto es una *estación*, se muestra la serie temporal de precipitación en conjunto con la descomposición clásica, descrita por la ecuación (2.1).

Esta posibilidad de la plataforma tiene por objetivo otorgar un primer acercamiento al usuario sobre los datos.

```
par(mar=c(5, 4, 4, 4) + 0.1)
#######################
# Grafico de neumonia
#######################
plot(neumonia)
mtext('Casos Neumonia',side=2, line=2)

#######################
# Grafico de pm sobre
# grafico anterior
#######################
plot(material.Particulado, col='red')

title('Neumonia y PM comuna La Florida')
dev.off()
```

Cuadro 3.14: Código ejecutado para la generación del gráfico expuesto en la figura 3.16

Análisis predictivo de la plataforma

Como análisis predictivo se implemento la función `arima` de `R`, que sigue la metodología del modelo *ARIMA* descrita en el capítulo 2, permitiendo al usuario visualizar una predicción de un año en las estaciones escogidas, como se ve en la figura 3.17.

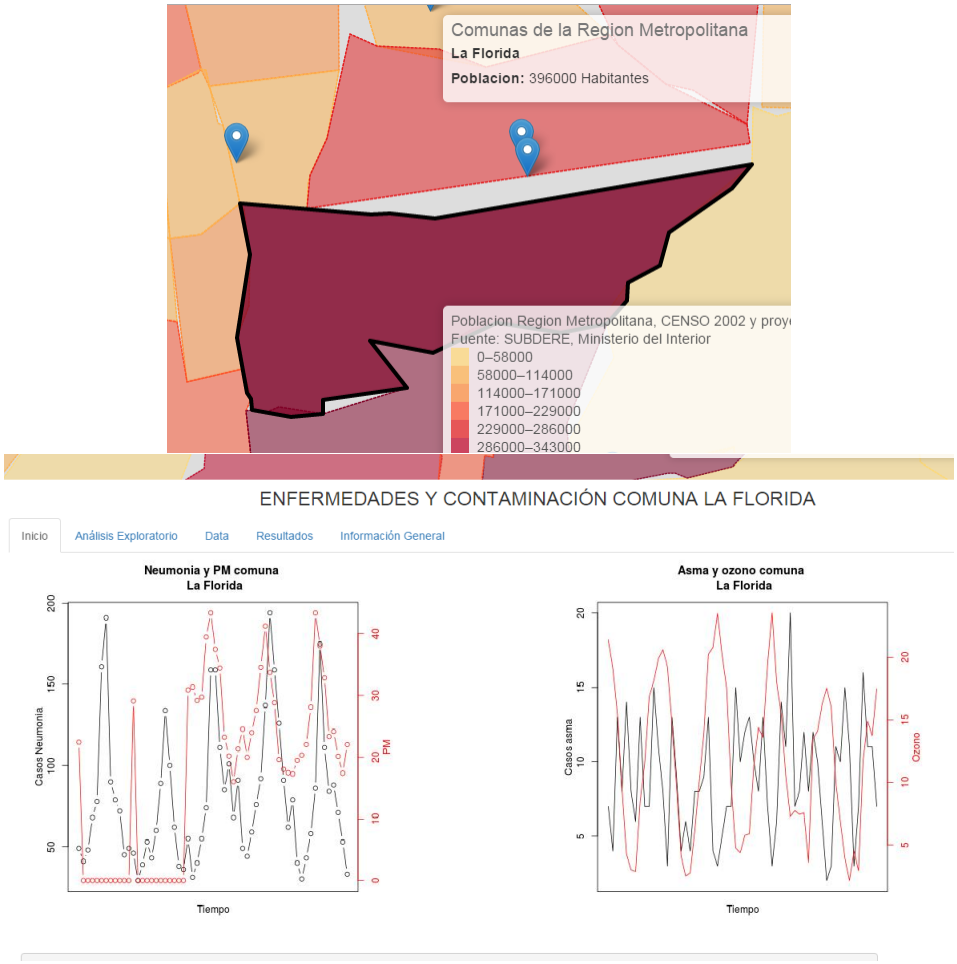

Figura 3.16: Figura donde se indica la selección de una comuna (particularmente *La Florida*) dentro del mapa interactivo (arriba) y la salida gráfica una vez seleccionada dicha comuna (abajo).

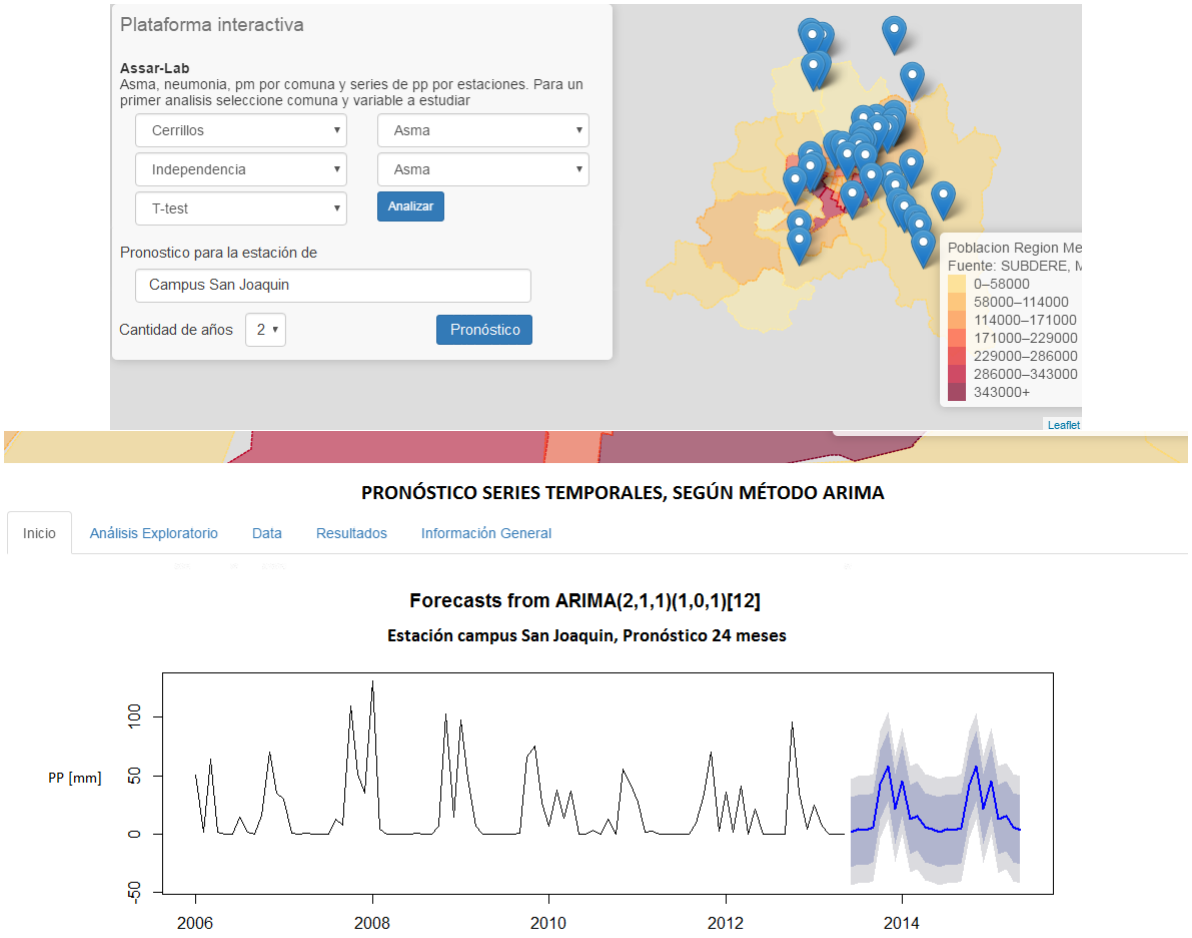

Figura 3.17: Pronostico realizado en la plataforma por el método *ARIMA*.

Capítulo 4

Discusión

4.1. Independencia de dos series temporales

Implementar un test de independencia en `R` contribuiría al estudio en series temporales dentro del desarrollo de esta herramienta estadística, dentro del contexto del modelo autoregresivo *ARIMA*. En esta sección se describe un estadístico para decidir la independencia de dos procesos que en un trabajo futuro puede ser implementado.

En lo que sigue, se considera la serie bivariada (x_t, y_t). Considérese, por ejemplo, las variables *ingresos hospitalarios por asma* y, *Concentración de Ozono*. Un problema a considerar es cómo mostrar, con algún nivel de significancia, si ambas series planteadas son o no independientes. Se describe un estadístico para probar la independencia de ambas series, el cual está en función de $\gamma_{uv}(k) = \mathbb{E}(u_{t-k}V_t)$ de manera similar al estadístico establecido por el test de Ljung-Box.

4.1.1. Test de independencia de Haugh

En esta sección, se considera, por un lado, que la serie bivariada (x_t, y_t) puede ser expresada de la forma

$$\begin{pmatrix} x_t \\ y_t \end{pmatrix} = \begin{pmatrix} F_{11}(B) & F_{12}(B) \\ F_{21}(B) & F_{22}(B) \end{pmatrix} \begin{pmatrix} \mathrm{e}_{1t} \\ \mathrm{e}_{2t}t \end{pmatrix} \tag{4.1}$$

donde F_{ij} son polinomios. Por otro lado, se puede ajustar x_t e y_t, como series univariadas, por algún buen ajuste según el modelo *MA* de tal forma que

$$\begin{pmatrix} x_t \\ y_t \end{pmatrix} = \begin{pmatrix} \psi_{11}(B) & 0 \\ 0 & \psi_{22}(B) \end{pmatrix} \begin{pmatrix} u_t \\ v_t t \end{pmatrix} \tag{4.2}$$

con U_t, V_t ruidos blancos que distribuyen como $WN(0, \sigma_1^2)$ y $WN(0, \sigma_2^2)$ respectivamente.

Si x_t, y_t son dependientes, no necesariamente e_{1t}, e_{2t} lo son. Sin embargo, la forma de la ecuación (4.2), muestra que U_t, V_t deben tener algún tipo de relación.

Para formalizar esto último, Haugh proporcionó un test formal como sigue.

Considere

$$\rho_{uv}(k) = \frac{\gamma_{uv}(k)}{\sigma_u \sigma_v}$$

donde $\gamma_{uv}(k) = \mathbb{E}(u_{t-k} v_k), \sigma_u^2 = \mathbb{E}(U_t^2), \sigma_v^2 = \mathbb{E}(V_t^2)$.

Entonces, dada n observaciones u_t, v_t ρ se aproxima por el estimador

$$\hat{\rho}_{uv}(k) = \left[\sum_{t=1}^{n} u_t^2 \sum_{t=1}^{n} v_t^2 \right]^{-1/2} \sum_{t=k}^{n} u_{t-k} v_t \quad , \qquad k \in \{-M, \ldots, 0, \ldots M\} \tag{4.3}$$

El test decide mediante el estadistico

$$s = n \sum_{k=-M}^{M} \hat{\rho}(k)^2 \qquad \text{para muestras de gran tamaño,}$$

o

$$s^* = n^2 \sum_{k=-M}^{M} \frac{1}{n - |k|} \hat{\rho}(k)^2 \qquad \text{para muestras pequeñas}$$

Dicho estadístico, bajo la hipótesis nula que x_t, y_t son independientes, es asintóticamente distribuido como χ^2_{2M+1} con $2M + 1$ grados de libertad.

4.2. Construcción plataforma

`R` cuenta con extensos paquetes destinados a la producción de mapas geográficos y a la producción de aplicaciones interactivas que son exportadas en formato HTML tales como, `maps, Rmaps, RgoogleVis, Shiny, leaflet , Rleaflet`, e incluso muchas de ellas son compatibles entre sí. Sin embargo, para poder realizar una aplicación interactiva con las características buscadas, se optó por integrar los lenguajes mencionados en la sección 2.

La plataforma fue construida en un servidor R y en el cual se han instalado los paquetes necesarios para los calculos establecidos en las secciones anteriores. Con ello, todos los datos ingresados por el usuario (`input`) son enviados a un servidor en el cual se ejecutan comandos en `R` y son devueltos al usuario como datos de salida (`output`). En este contexto, `PHP` permite justamente esta interacción dinámica entre usuario-plataforma.

Esta plataforma está disponible para su uso vía solicitud a *www.assar-lab.cl*. Si se quiere construir una plataforma del mismo tipo hay que considerar dichos requerimientos así como la arquitectura descrita en la Figura 3.12.

Capítulo 5

Conclusiones

En este trabajo de titulo se ha logrado integrar métodos estadísticos exploratorios y predictivos en una plataforma geo-referenciada. En particular, en dicha plataforma se integro el método *ARIMA*, una generalización de modelos regresivos (Modelos AR y MA), para series con tendencia y estacionalidad. Con ello, se contribuye al análisis de series temporales que puedan ser descompuestas de la forma 2.1. Este modelo asimismo, permite el pronóstico de los eventos X_{n+h} basado en los instantes previos $\{X_{n-k}, \ldots, X_n\}$.

Cabe mencionar de la sección 3, según los cuadros 3.6 y 3.7, el modelo predictivo *ARIMA* es el que mejor ajusta en términos de la distribución esperada de $(\mathrm{e}_t)_t$, la cual se distribuye de manera normal, con media cercana a cero con p-valor más alto en la prueba *t de Student* y en la prueba de *Wilcoxon*, siendo este modelo el más fidedigno entre los tres para realizar la predicción.

Por otro lado, en términos de comparación de series temporales, según la sección 3.1.3, existe un desfase de dos meses entre los casos de *Asma Santiago* y *Concentración de Ozono*, lo que se interpreta como un aumento en los casos de Asma con un tiempo de espera de dos meses luego del incremento en *Concentración de Ozono* en la comuna de *Independencia* Aplicado este desfase se logra una coincidencia $\hat{\rho}$ = ,85 y $\hat{\mathrm{d}}$ = ,029.

En relación a los impactos funcionales del estudio, la plataforma GIS construida permite el almacenamiento y manipulación de datos que están vinculados a una referenciación espacial. Su formato facilita la incorporación de variables socio-culturales, económicas y ambientales que contribuyen a la toma de decisiones de una manera más eficaz.

Como sugerencias para siguientes investigaciones, se propone abordar el problema de establecer independencia entre dos series temporales, descomponiendo la forma 4.2 e implementándola en `R`, lo cual permitiría realizar el Test de Haugh descrito en la sección 4.1. Asimismo, y en función de dotar de herramientas estadísticas a la plataforma GIS construida, El test de independencia de Haugh puede ser implementado a la plataforma, decidiendo independencia sobre las variables en cuestión.

Capítulo 6

Bibliografía

[1] Brockwell P.J. y Davis R., (2002). *Introduction to Time Series and Forecasting*, New York, Springer.

[2] Hamilton, James D. (1994) *Time Series Analysis*, Princeton University Press, Princeton.

[3] Koch Paul D. y Yang Shie-Shien. (Junio 1986). *A Method for testing the independence of two time series that accounts for a potential pattern in the cross-correlation function.* American Statiscal Association, 81, pp. 533-544.

[4] Hastie T, Tibshirani R, Friedman J. (2009) *The Elements of Statistical Learning, Data Mining, Inference and Prediction.* Springer Series in Statistics.

[5] Hothorn T. y Everitt B. S. (2009) *A Handbook of Statistical Analyses Using R. Second Edition.* Chapman and Hall.

[6] Shapiro, S. S. y Wilk, M. B. (1965). *An analysis of variance test for normality (complete samples).* Biometrika Vol. 52 No. 3/4: pp. 591–611.

[7] Wilcoxon, F. (1945) *Individual Comparisons by Ranking Methods.* Biometrics Vol. 1, No. 6, pp. 80-83.

[8] *GIS studies related to alcohol: how GIS is being used to solve alcohol - related problems.* (s.f.) `http:// www.bio.davidson.edu/ people/midorcas/gisclass/ giswebsites/martineau/GISstudies.htm`

[9] Hart A, McCulloch B, Harper C, Gardiner N, Rutherford S, Baker P, Harris P, O'Sullivan D. (2005) *Report on GIS and public health spatial applications*, Public Health Services, Queensland Health. Brisbane.

[10] Carroll L.N., Au A.P., Detwiler L.T., Fu T.C., Painter I.S., Abernethy N.F. (Abril 2014). *Visualization and analytics tools for infectious disease epidemiology: A systematic review.* Journal of Biomedical Informatics, 51, pp. 287-298.

[11] Bell, M., McDermott, A., Zeger, S., Samet, J., Dominici, F. (2004) *Ozone and Short-term Mortality in 95 US Urban Communities, 1987-2000.* ;292(19), pp. 2372–2378.