Análisis de los datos del robo de vehículos asegurados: una aplicación de las series de tiempo

Alejandro Román Vásquez

2 de mayo del 2012

Índice general

1. Introducción **5**

1.1. Contexto, motivación y propósito del proyecto 5

1.2. Estructura de la tesis 8

1.3. Análisis del robo de vehículos asegurados en la república mexicana 9

2. Modelos ARIMA de Series de Tiempo **13**

2.1. Elementos de Series de Tiempo 13

2.1.1. Introducción 13

2.1.2. Estacionariedad y función de autocorrelación 14

2.1.3. Función de Autocorrelación Parcial 16

2.1.4. Estimación de la media y de las funciones de autocovarianza, autocorrelación y autocorrelación parcial 17

2.1.5. Procesos de ruido blanco 19

2.2. Modelos de series de tiempo estacionarios 20

2.2.1. Principales representaciones de las series de tiempo 20

2.2.2. Procesos de promedios Móviles 22

2.2.3. Procesos autoregresivos 24

2.2.4. Procesos autoregresivos de promedios móviles 26

2.3. Modelos de series de tiempo no estacionarios 28

2.3.1. No estacionariedad en media 28

2.3.2. Procesos autoregresivos integrados de promedios móviles 29

2.3.3. Términos constantes en los modelos ARIMA 30

2.3.4. Estacionariedad respecto a una tendencia determinista 31
2.3.5. No estacionaridad en varianza y covarianza 31
2.4. Modelos de series de tiempo estacionales 32
2.4.1. Modelos MA(Q) y AR(P) estacionales 33
2.4.2. Modelos ARMA estacionales multiplicativos 33
2.4.3. Modelos no estacionarios con estacionalidad: ARIMA estacionales multiplicativos 34

3. Construcción de los modelos ARIMA 35
3.1. Identificación del modelo 36
3.1.1. Prueba de raíz unitaria 37
3.2. Ajuste del modelo: estimación de los parámetros 38
3.2.1. Estimación de máxima verosimilitud condicional 38
3.2.2. Función de máxima verosimilitud exacta 39
3.2.3. Propiedades de los estimadores de máxima verosimilitud 41
3.3. Diagnóstico del modelo 42
3.3.1. Análisis Residual 42
3.3.2. Análisis de Sobreajuste 44
3.3.3. Criterio de selección del modelo 45
3.4. Análisis de intervención y análisis de datos atípicos 46
3.4.1. Análisis de intervención 47
3.4.2. Análisis de datos atípicos 49
3.5. Pronósticos del modelo 52
3.5.1. Pronósticos con mínimo error cuadrático medio 52
3.5.2. Pronósticos de modelos estacionarios de ARMA 54
3.5.3. Pronósticos de modelos no estacionarios ARIMA 57
3.5.4. Límites de predicción 58
3.5.5. Actualización de los pronósticos 59
3.5.6. Predicciones ponderadas 60

3.5.7. Selección del modelo basado en los errores de pronóstico 61

4. Modelos ARIMA: robo de vehículos asegurados 62

4.1. Clasificación de los estados 62

4.1.1. Criterio de clasificación 62

4.1.2. Agrupamiento Jerárquico de Aglomeración 65

4.1.3. Obtención de series mensuales y clasificación final 67

4.2. Modelación de los indicadores 70

4.2.1. Indicadores por región 70

4.2.2. Modelado de algunas series representativas 73

4.2.3. Proceso estacionario 73

4.2.4. Proceso no estacionario (raíz unitaria) 78

4.2.5. Proceso estacionario respecto a una tendencia lineal 83

4.2.6. Datos atípicos 89

4.2.7. Análisis de intervención 90

4.3. Resultados: modelos ARIMA para cada indicador 93

4.3.1. Procesos generadores 93

5. Conclusiones y consideraciones finales 97

6. Apéndice A: Gráficas del proceso de modelado 100

6.1. Series de prima de riesgo mensual 100

6.1.1. Zona 1A 100

6.1.2. Zona 2A 100

6.1.3. Zona 3A 104

6.1.4. Zona 3B 106

6.1.5. Zona 3C 107

6.1.6. Zona 4A 109

6.1.7. Zona 4B 111

6.1.8. Zona 5A 112
6.1.9. Zona 5B 112
6.2. Series de porcentaje de robo 115
6.2.1. Zona 1A 115
6.2.2. Zona 3A 117
6.2.3. Zona 3B 119
6.2.4. Zona 3C 121
6.2.5. Zona 4A 123
6.2.6. Zona 4B 125

Capítulo 1

Introducción

1.1. Contexto, motivación y propósito del proyecto

El seguro es una actividad económica que tiene como finalidad cubrir, mediante el concurso mutuo de todos los integrantes del mismo, la parte del costo social y financiero por la ocurrencia de siniestros individuales que son aleatorios, pero que son estadísticamente mesurables y predecibles en conjunto. La institución del seguro ha evolucionado históricamente desde un concepto rudimentario de ayuda mutual, hasta la actualidad, donde existen para cada tipo de protección una entidad especializada para ese suceso.

Dentro del este contexto de especialización, la actividad aseguradora se puede dividir en dos grandes rubros: seguros para personas y seguros de daños. Los primeros tienen por objetivo cubrir riesgos[I] que afecten la integridad física de los asegurados; destacan los seguros de vida, de accidentes, de gastos médicos. La segunda área busca cubrir riesgos que afecten los bienes materiales de las personas; se pueden mencionar seguros contra incendio, contra terremotos, contra inundaciones, entre otros.

El aseguramiento de vehículos es parte importante del área de seguro de daños. Dentro de las diversas pérdidas que puede ser sujeto el propietario de un medio de transporte automotor, el robo representa una gran amenaza que pone en riesgo su posición financiera, y más aún, su integridad humana. Además, las compañías aseguradoras se ven afectadas directamente con este fenómeno, pues el resarcimiento para cubrir una pérdida de este tipo es considerablemente alto.

La situación de inseguridad actual del país, presupone un ambiente, donde el delito de robo de automóviles se ha incrementado sustancialmente y de manera marcada en algunas regiones del país, y además del brutal impacto social que esto conlleva, está lesionando ampliamente la suficiencia de las compañías aseguradoras para hacer frente a estas obligaciones. Estas afectaciones se transmiten de forma directa a la contraparte asegurada, pues tienen que pagar un precio por adelantado por este servicio de protección (que comúnmente se conoce como prima), y dado que el aumento del robo vulnera la posición financiera de las aseguradoras, estas se ven en la necesidad

[I] Un riesgo se define como una eventualidad que de ocurrir traería como consecuencia un desequilibrio económico para el individuo que lo sufre.

de aumentar este cobro, lo que se traduce en primas más altas para este tipo de coberturas.

El sector asegurador, en particular el ramo de automóviles, tiene diversos mecanismos comerciales que durante varios años permitieron recurrir a subsidios de costos entre coberturas, de tal forma que conseguían ofrecer precios accesibles a la sociedad y con ello estimular la cultura del seguro. Sin embargo, los indicadores básicos que se utilizan para analizar el comportamiento de los siniestros han reflejado impactos en coberturas que durante mucho tiempo mantenían un comportamiento muy estable, lo que altera y vulnera la posibilidad de mantener precios nivelados.

Si bien la tendencia nacional del robo de autos es creciente, existen entidades de la república mexicana donde el robo de vehículos refleja otros comportamientos, de ahí que es pertinente analizar el problema en diferentes regiones que permitan clasificar el comportamiento del delito, en función de la homogeneidad de los indicadores de ciertos estados.

Resulta evidente la importancia de analizar de forma más precisa el crecimiento actual de este riesgo. Por ello, este trabajo busca generar indicadores regionalizados que proporciones herramientas para el análisis de dicha situación y funcionen como un complemento de las técnicas actuales del cálculo de primas.

Dado el carácter grupal, se decidió escoger cantidades relativas que permitan realizar una comparación entre las diversas zonas. De esta manera, se determinó analizar dos indicadores: la prima de riesgo[II] y la proporción de robo. De acuerdo a Molinaro[1] y Osorio González[2] se definen de la siguiente forma:

$$\text{Prima de Riesgo} = \frac{\text{Monto de las afectaciones a la cobertura de robo}}{\text{Unidades expuestas}} \tag{1.1}$$

$$\text{Proporción de robo} = \frac{\text{Afectaciones a las coberturas de robo}}{\text{Unidades expuestas}} \tag{1.2}$$

El monto se refiere al capital necesario para resarcir el daño. Las afectaciones son básicamente el número de siniestros del delito de robo. En general, cuando se asigna una unidad de riesgo, se asocia a la unidad de seguro un determinado periodo de tiempo; con estos fines, las unidades expuestas representan la proporción del tiempo que los vehículos estuvieron expuestos durante ese periodo. Generalmente, se toma un año como medida temporal, por lo que se habla de prima de riesgo y proporción de robo anual, aunque se pueden escoger otros lapsos de tiempo.

Dado del carácter heterogéneo del delito en la república mexicana, se decidió que las cantidades fueran analizadas de forma regionalizada. Se determinó que la clasificación se realizara con base en la tasa de crecimiento de la proporción de robo anual. El tomar una cantidad relativa, determina de forma más minuciosa el aumento puro del delito, dejando de lado el factor del crecimiento de robo debido al ligero incremento de las unidades expuestas que se da año con año.

Resumiendo, y con fines de claridad, se establece a continuación el propósito general del proyecto, y los objetivos particulares que permiten alcanzar esta meta.

Propósito general:

[II] A la prima de riesgo también se le conoce como prima pura o cuota de repartición.

- Analizar el robo de vehículos en la república mexicana a través de los indicadores relativos regionalizados de prima de riesgo y proporción de robo.

Objetivo 1:

- La regionalización se implementará con base en la tasa de crecimiento de la proporción de robo de vehículos de los años 2008, 2009 y 2010.

Objetivo 2:

- El análisis de los indicadores se realizará empleado series temporales mensuales de cada una de las regiones durante el periodo del 2008 al 2010 para generar pronósticos a un año que permitan estimar y proyectar cuál será su comportamiento a mediano plazo, para cuantificar el riesgo.

Para abordar y lograr el propósito general, y en específico los objetivos particulares, se propone aplicar las siguientes técnicas estadísticas:

1. Para realizar la clasificación se usaran herramientas estadísticas multivariadas del área de análisis de conglomerados. Dentro de los diversos algoritmos heurísticos se selecciona el *Agrupamiento Jerárquico de Aglomeración*, pues su desarrollo teórico es muy natural, y tiene una implementación rápida, como se puede constatar en Everitt[3].
2. El estudio de los indicadores se realizará empleando modelos paramétricos ARIMA[III] , que son bastante flexibles y convenientes para modelar series temporales. Sirven para modelar comportamientos estacionarios y no estacionarios; permiten la incorporación de cambios estructurales en los niveles de la serie, y el manejo de datos aberrantes. Además, los pronósticos generados por estos modelos son aceptables a corto y mediano plazo. Se seguirá de cerca la metodología propuesta por Box y Jenkins[4] para la aplicación de los modelos ARIMA.

Para la implementación de estas técnicas estadísticas, se decidió usar el software estadístico R[6] que es de licencia libre. La aplicación del algoritmo de Agrupamiento Jerárquico de Aglomeración es muy accesible, y el modelado de las series temporales es muy práctico y versátil por la cantidad de librerías, paqueterías y funciones que se han desarrollado

En este punto, resulta conveniente mencionar ciertos aspectos referentes a los objetivos. El primero de ellos es alusivo al periodo del análisis, el cual abarca los años 2008, 2009 y 2010. Como lo establece Jiménez[5], a partir del 2008, el sector asegurador mexicano sufrió una transición en el manejo de la información. La forma tradicional de reportar y almacenar los datos cambio de forma radical, con el objetivo de generar bases más completas y confiables, que permitan integrar de forma más rápida y práctica la información más relevante del sector asegurador. El proceso propuesto e impulsado por la AMIS[IV], se tradujo en la formación de una base de

[III] *Autoregressive Integrated Moving Average.*
[IV] Asociación Mexicana de Instituciones de Seguros, A.C.

datos llamada SESA[V], que opera de forma eficiente desde el 2008. De esta manera, al trabajar con esta base, se establece un límite natural inferior para el manejo de la información.

El análisis de datos de forma mensual se determinó por varias razones. La primera y más trascendente es que este proyecto tiene como antecedente un trabajo sobre el análisis del robo de vehículos asegurados en toda la república mexicana, que se presentó en el XXVI Foro Nacional de Estadística realizado en la ciudad de Villahermosa Tabasco. En este estudio, se decidió usar datos mensuales, porque se había identificado cierta dependencia estacional en el comportamiento del delito, es decir, se había notado que en periodos vacacionales los robos disminuyen, mientras que en lapsos normales los hurtos aumentaban. Esta hipótesis se verificó, pues se identificó un modelo ARIMA estacional con periodicidad de 12 meses como el proceso generador de los datos. Los resultados de este estudio, se muestran de forma breve y concisa en la última sección de este capítulo; se decidió incluirlos pues el proyecto sirvió como base y motivación para el desarrollo de esta tesis.

La otra razón del desglose mensual es debida al corto periodo con el que se cuenta por el uso de la base SESA. Al contar solamente con tres años, una diversificación anual, es exageradamente pobre para el uso de los modelos ARIMA por las propiedades asintóticas de los estimadores y de la posible identificación del proceso generador. Un análisis bimestral con 18 datos, pudo haberse llevado a cabo, aunque en realidad, siguen siendo pocos datos. El uso mensual, implica una segregación de 36 datos, que permite una modelación más completa y adecuada.

Por último, hay que mencionar que si bien la base de datos con la que se trabajó es muy completa y permitió obtener las afectaciones y los montos de los siniestros de forma mensual y por estados de la república mexicana, fue necesario estimar las unidades expuestas de forma mensual, pues no se tenía como una variable accesible en el SESA.

Hay que mencionar que la base de datos con la que se trabajó resultó ser muy completa, pues permitió obtener las afectaciones y los montos de los siniestros de forma mensual y por estados de la república mexicana. Sin embargo, fue necesario estimar las unidades expuestas, pues no se contaba con esta información de forma mensual en el SESA.

Con todo lo anterior, queda delimitado el contexto, motivación y propósito del proyecto. En la siguiente sección se delinea como está establecida la organización de los capítulos de esta tesis.

1.2. Estructura de la tesis

La tesis está dividida en 5 capítulos. El primero es justamente esta parte introductoria. En el capítulo 2 se presentan los elementos fundamentales de las series temporales. Se aborda primero el estudio de los procesos estocásticos discretos estacionarios en el sentido débil, y se definen las funciones de autocorrelación y autocorrelación parcial, de suma importancia, y se ilustran sus propiedades y la forma de estimarlas. Se describen los procesos de ruido blando, mostrando que son una base fundamental de los modelos ARMA[VI], que se utilizan para modelar series estacionarias. Se estudian las propiedades más importantes de los procesos autoregresivos y de

[V] Sistema Estadístico del Sector Asegurador.
[VI] *Autoregresive moving average*

promedios móviles, para después hacer una integración de ambos lo que lleva a los modelos ARMA. Se explica cuál es el análisis necesario para tratar series no estacionarias, estableciendo que en ciertas circunstancias se pueden aplicar transformaciones para convertirlas en series temporales estacionarias, lo que da pie a los modelos ARIMA. Finalmente, en la última sección, se toca el tema de la estacionalidad y como integrarla en los modelos.

El capítulo 3 está dedicado a la metodología necesaria para modelar series de tiempo a través de procesos ARIMA. Primero se estudia cómo identificar el modelo, para lo cual las funciones de autocorrelación muestral y autocorrelación parcial muestral juegan un papel transcendental. Después se estudia la estimación de parámetros en el modelo, siguiendo varias técnicas, entre las que destaca el principio de máxima verosimilitud. Posterior a esto se describe el proceso de diagnóstico, el cual incluye el análisis residual y el análisis de sobreajuste, lo que permite determinar si se cumplen las principales suposiciones teóricas. A continuación, se hace una reseña sobre como modelar intervenciones ajenas al proceso que cambien el nivel de la serie, así como el manejo adecuado de datos irregulares. Finalmente se introduce el concepto de pronóstico, el cual es de suma importancia pues es uno de los principales objetivos de la construcción de modelos de series de tiempo. Se explica que el criterio que minimiza la estimación de un valor futuro es el error cuadrático medio, y se describe su aplicación tanto en procesos estacionarios como modelos ARIMA.

En el cuarto capítulo se implementan en la base de datos los conceptos teóricos que se desarrollaron con anterioridad para modelar las series temporales. La primera sección habla sobre el proceso de la clasificación de los estados. Se describe el criterio de regionalización, indicando como fue la búsqueda y desglose de la información. Después se hace una pequeña reseña teórica sobre el Agrupamiento Jerárquico de Aglomeración y se aplica a los datos para generar una clasificación preliminar; se mencionan los principios y criterios que llevaron a la regionalización final de 9 zonas, lo que se traduce en 18 series temporales. La siguiente sección está dedicada a la aplicación directa de la metodología ARIMA a los indicadores. Se explica de forma exhaustiva la modelación de ciertas series características y representativas, pues la información del todo el modelado es considerablemente grande, por lo que los resultados faltantes se incluyen en los anexos. En la última parte del capítulo se incluyen los resultados los procesos generadores con las estimaciones de sus parámetros, y se muestran los pronósticos de todo el año 2011 de todas las series.

En la última parte de la tesis se dan las conclusiones y consideraciones finales. Se realizan ciertas comparaciones y se indica cuáles fueron las ventajas y desventajas de los modelos. Se prepara el terreno para un posible trabajo a futuro, con ideas de cómo se pueden mejorar los modelos.

1.3. Análisis del robo de vehículos asegurados en la república mexicana

Como se mencionó con anterioridad, parte de la motivación de este trabajo estuvo basada en un proyecto preliminar sobre el comportamiento del robo de vehículos en la república mexicana. En esta sección se hablará de forma breve y concisa de sobre este estudio.

El objetivo de ese trabajo era ver como se comporta el robo mensual de vehículos asegurados a

mediano plazo, verificar la presencia de componentes estacionales y usar el modelo para medir la calidad de los nuevos datos reportados a través de pronósticos de un periodo por delante.

El estudio se realizó con el robo total de vehículos mensual de la república mexicana desde el 2000 hasta el 2010, información que se pueden ver en la figura 1.1. Los datos fueron proporcionados por la AMIS, y provenían de una base de datos que recopila mensualmente la información de este delito.

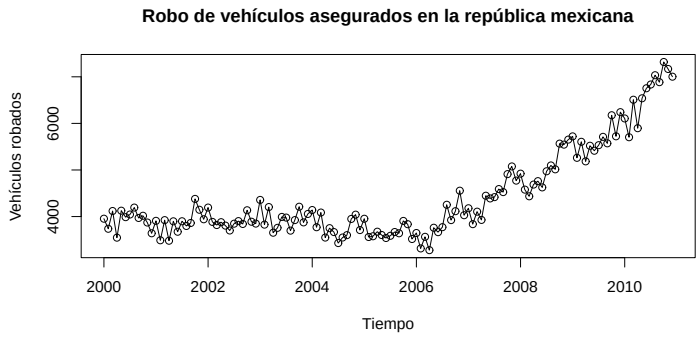


Figura 1.1: Gráfica de la serie temporal de los datos del estudio en cuestión.

El análisis de estos datos se realizó empleado la metodología específica de los modelos ARIMA, la cual se verá en detalle en los capítulos siguientes. A continuación se mencionaran los aspectos más relevantes de esta técnica de modelado.

Se identificó un proceso no estacionario con estacionalidad ARIMA$(0,1,2)x(1,0,0)_{12}$ (este tipo de modelos se estudiaran en el capítulo siguiente). Parte de la identificación de este proceso se realizó al analizar la función de autocorrelación muestral que se observa en la siguiente figura.

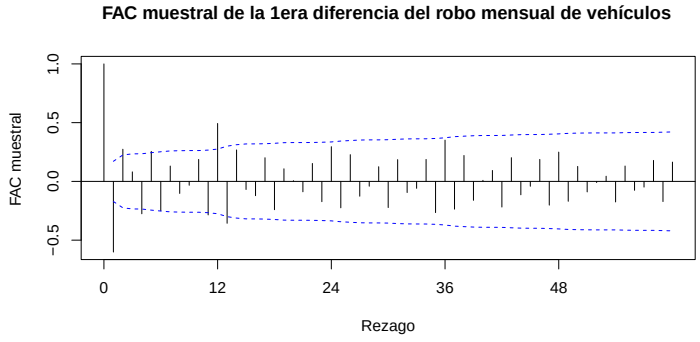


Figura 1.2: Gráfica de la función de autocorrelación muestral. Se observa un patrón estacional de 12 meses.

Las estimaciones de los parámetros se realizaron empleando máxima verosimilitud. En la siguiente tabla se muestran los resultados.

El modelo propuesto como proceso generador de los datos pasó de forma satisfactoria las pruebas de diagnóstico a las cuales fue sometido: normalidad, no correlación de los residuales de forma individual y grupal, y análisis de sobreajuste para los parámetros.

Parámetros	θ_1	θ_2	Φ_1
Estimación	-0.6185	0.2712	0.5451
Error estándar	0.0906	0.1005	0.0867

Cuadro 1.1: Parámetros estimados del modelo ARIMA$(0,1,2)x(1,0,0)_{12}$ para el robo de vehículos asegurados en México.

El modelo puede ser usado para evaluar la calidad de nuevos datos. Esto se verificó empleando los pronósticos de un periodo por delante desde enero de 2009 hasta diciembre de 2010, haciendo una comparación con los valores observados. En la gráfica de la figura 1.3, se muestran estas predicciones con su respectivo intervalo de confianza al 95 por ciento, graficando también la información que se tiene en ese periodo. Si se llega a presentar un dato fuera de estos intervalos, es posible que sea un valor irregular, que debe ser sometido a revisión.

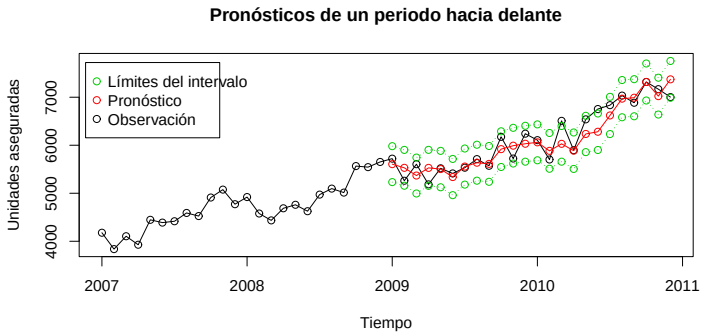


Figura 1.3: Medición de la calidad de nuevos datos empleando pronósticos de un periodo hacia delante.

Para analizar la calidad de las predicciones del modelo se realizaron pronósticos dentro de un intervalo temporal para el cual se tienen datos. Partiendo del origen de diciembre del 2008 se estimaron los pronósticos de hasta 24 periodos hacia delante (figura 1.4). Se observa que durante el primer año, las predicciones son aceptables, pero a partir del segundo año los pronósticos no son tan buenos; lo anterior se debe a que en un proceso no estacionario la varianza de las predicciones crece al aumentar la distancia al origen del pronóstico, como se observa en los intervalos de confianza.

Se obtienen las siguientes conclusiones sobre este proyecto:

- El modelo propuesto ARIMA$(0,1,2)x(1,0,0)_{12}$ describe de forma adecuada el comportamiento del proceso del robo mensual de vehículos asegurados en la República Mexicana.
- La calidad de la base de datos es aceptable y el modelo sirve para evaluarla respecto a las nuevas observaciones que se vayan agregando.
- Las predicciones dentro del intervalo de un año por delante son aceptables, por lo que la estimación del número de robos acumulada durante el año puede ser usada como complemento para el cálculo de primas de riesgo en el sector asegurador.

Otra conclusión del trabajo fue que si bien se observa que el crecimiento general del delito es

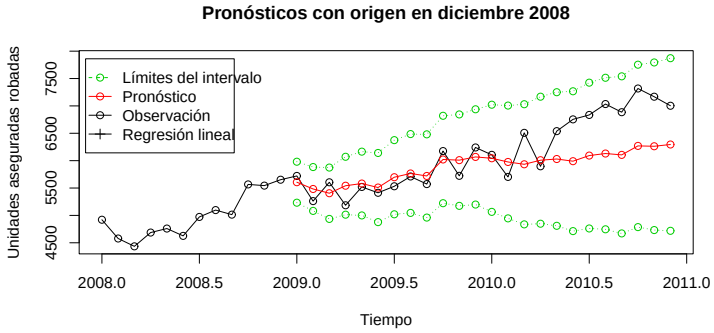


Figura 1.4: Observación de la calidad de los pronósticos dentro de un intervalo temporal para el cual se tienen datos.

a la alza, resultaba conveniente hacer un análisis de forma regionalizada para cuantificar de forma más precisa el fenómeno, pues se tenía cierto conocimiento de que en algunos estados el robo no había crecido, e incluso en otros se presentaba una disminución del delito. Esto fue una motivación y dio pie a este trabajo de tesis. Estos resultados, serán empleados posteriormente para hacer ciertas comparaciones con aquellos que resulten de este proyecto.

Capítulo 2

Modelos ARIMA de Series de Tiempo

2.1. Elementos de Series de Tiempo

2.1.1. Introducción

Un proceso estocástico es una familia de variables aleatorias $Y(\omega, t)$, donde ω pertenece a un espacio muestral Ω y t a un conjunto de índices T. Para un valor fijo de t, $Y(\omega, t)$ es una variable aleatoria; para un valor de ω dado, $Y(\omega, t)$ como función de t, se le conoce como realización del proceso estocástico. La población total que consiste de todas las realizaciones, se le conoce como ensamble del proceso estocástico. Desde este punto de vista, una serie de tiempo es sólo una realización de cierto proceso estocástico.

De igual manera que una variable aleatoria $X(\omega)$ se identifica simplemente como X, así la serie de tiempo $Y(\omega, t)$, como proceso estocástico, se identifica como $Y(t)$ o Y_t. Si el proceso toma valores en los reales se denota como proceso real-valuado; la mayoría de las series con las que se trabajan toman valores en los reales. Finalmente las realizaciones o series temporales que se pretenden analizar, pertenecen a un conjunto discreto, esto es el conjunto T al cual pertenece la variable t corresponde al conjunto de los enteros, por lo que el proceso estocástico se puede representar como $\{Y(\omega, t) = 0, \pm 1, \pm 2, \ldots\}$; por lo regular a la variable t se le identifica con el tiempo, porque la mayoría de los procesos estocásticos evolucionan temporalmente.

No es posible determinar plenamente la estructura probabilística de una serie de tiempo dado que los datos de la serie corresponden a una fracción de una realización que es infinita. Afortunadamente no se necesita trabajar con estas distribuciones, pues mucha de la información de estas distribuciones conjuntas puede ser descrita en términos de sus valores esperados, varianzas y covarianzas.

Para un proceso con valores reales de la forma $\{Y(\omega, t) = 0, \pm 1, \pm 2, \ldots\}$ su ***función valor esperado*** se define como:

$$\mu_t = E(Y_t) \tag{2.1}$$

Esto es, μ_t es el valor esperado del proceso en el tiempo t, que generalmente puede ser diferente para cada valor de t.

La ***función de varianza*** del proceso es:

$$\sigma_t^2 = E[(Y_t - \mu_t)^2] \tag{2.2}$$

La ***función de autocovarianza*** entre Y_t y Y_s está dada por:

$$\gamma_{t,s} = E[(Y_t - \mu_t)(Y_s - \mu_s)] = Cov(Y_t, Y_s) \tag{2.3}$$

donde s es también un entero $s = 0, \pm 1, \pm 2 \ldots$. Se observa de las relaciones anteriores (2.2) y (2.3) que cuando s es igual a t se tiene que $\gamma_{t,t} = \sigma_t^2$.

La ***función de autocorrelación*** (ACF por sus siglas en ingles[I]) para Y_t y Y_s se define como:

$$\rho_{t,s} = \frac{\gamma_{t,s}}{\sqrt{\sigma_t^2 \sigma_s^2}} = Corr(Y_t, Y_s) \tag{2.4}$$

Las autocovarianzas y las autocorrelaciones son medidas de la posible relación de dependencia (lineal) entre las variables aleatorias. Valores de $\rho_{t,s}$ cerca de ± 1 indican una fuerte relación de dependencia (lineal) y valores cercanos a cero muestran lo contrario. Si $\rho_{t,s} = 0$ se dice que las variables aleatorias Y_t y Y_s no están correlacionadas.

Es fácil probar de las definiciones anteriores que se cumplen las siguientes relaciones:

$$\begin{cases} \rho_{t,t} = 1 \\ \rho_{t,s} = \rho_{s,t} \\ |\rho_{t,s}| \leq 1 \end{cases} \tag{2.5}$$

2.1.2. Estacionariedad y función de autocorrelación

Para hacer inferencias estadísticas sobre la estructura de un proceso estocástico usualmente se deben hacer simplificaciones basadas en suposiciones que hasta cierto punto sean razonables. La más importante de las suposiciones es la de estacionariedad. La idea se centra en que las leyes que gobiernan el comportamiento del proceso no cambian con el tiempo; en un sentido, el proceso está estadísticamente en equilibrio. Se manejan comúnmente dos tipos de estacionariedad: una que se denomina fuerte o estricta y otra que se conoce como débil, y que se definen en seguida.

Un proceso Y_t se dice que es estrictamente estacionario si la función de distribución conjunta de Y_{t_1},Y_{t_2},$\ldots$,Y_{t_n}, es la misma que la función de distribución conjunta de $Y_{t_{1-k}}$,$Y_{t_{2-k}}$,...,$Y_{t_{n-k}}$ para todos los posibles valores de los puntos en el tiempo t_1,t_2,...,t_n y los posibles valores enteros que puede tomar k[II]; es decir, si se cumple la igualdad en distribución $\triangleq$ de:

$$F_{Y_{t_1},...,Y_{t_n}}(x_1, ..., x_n) \triangleq F_{Y_{t_{1-k}},...,Y_{t_{n-k}}}(x_1, ..., x_n) \tag{2.6}$$

[I] En inglés se escribe como autocorrelation function

[II] Dado la variable aleatoria Y_t, si $k > 0$, al valor $Y_(t-k)$ se le conoce como el valor Y_t rezagado en k unidades de tiempo. Al valor Y_{t+k}, se le denota como el valor futuro de Y_t en k unidades de tiempo.

Para cualquier n en el conjunto $(t_1, t_2, ..., t_n)$ y cualquier entero k. Las $x_i,\ i = 1, 2, ..., n$ son números reales.

Si la serie es estrictamente estacionaria se sigue que la distribución de Y_t es la misma que Y_{t-k} para todo t y k; en otras palabras las $Y's$ tienen la misma distribución debido a que es cuando $n = 1$ en la ecuación 2.6. De lo anterior se sigue que $E(Y_t) = E(Y_{t+k})$ para cualquier t y k por lo que en una serie estrictamente estacionaria la función del valor esperado es constante en el tiempo $\mu_t = \mu$. Además, se cumple que el segundo momento también es constante en el tiempo por lo que la función de varianza es constante en el tiempo $\sigma_t^2 = \sigma^2$ (siempre y cuando la varianza del proceso sea finita).

Tomando ahora $n = 2$, en la ecuación 2.6, la distribución bivariada de Y_t y Y_s es la misma que la de las variables de Y_{t-k} y Y_{s-k}, de donde se sigue que $Cov(Y_t, Y_s) = Cov(Y_{t-k}, Y_{s-k})$ para cualquier t,s y k. Tomando $s = k$ se tiene que:

$$\begin{aligned}\gamma_{t,s} &= Cov(Y_t, Y_s) = Cov(Y_{t-s}, Y_{s-s}) = Cov(Y_{t-s}, Y_0) \\ &= Cov(Y_0, Y_{t-s}) = Cov(Y_0, Y_{s-t}) = (Y_0, Y_{|t-s|}) = \gamma_{0,|t-s|}\end{aligned}$$

Entonces para un proceso estrictamente estacionario la covarianza entre Y_t y Y_s sólo depende de la diferencia en tiempo $|t - s|$.

En un proceso estacionario, se puede simplificar la notación y escribir las funciones de autocovarianza y autocorrelación como:

$$\gamma_k = E[(Y_t - \mu)(Y_{t-k} - \mu)] = Cov(Y_t, Y_{t-k}) \tag{2.7}$$

$$\rho_k = \frac{\gamma_k}{\gamma_0} = Corr(Y_t, Y_{t-k}) \tag{2.8}$$

Las relaciones de la ecuación 2.5 quedan ahora como:

$$\begin{cases}\rho_0 = 1 \\ \rho_k = \rho_{-k} \\ |\rho_k| \leq 1\end{cases} \tag{2.9}$$

Existe otra definición similar a la estacionariedad estricta, pero matemáticamente más débil, pues ya no se pide que las distribuciones conjuntas de $Y_{t_1}, Y_{t_2}, ..., Y_{t_n}$ y de $Y_{t_{1-k}}, Y_{t_{2-k}}, ..., Y_{t_{n-k}}$ sean iguales.

Un proceso Y_t es ***estacionario débil*** (o estacionario de segundo orden) si se cumplen dos cosas: que la media y la varianza existan y que sean constantes en el tiempo ($\mu_t = \mu$ y $\sigma_t^2 = \sigma^2$), y que la autocovarianza y la autocorrelación dependan sólo de la diferencia en tiempo (γ_k y ρ_k).

Otro nombre con que se le conoce a la estacionaridad débil es ***estacionaridad en covarianza***. En el análisis de las series de tiempo generalmente se toma sólo en consideración la estacionaridad débil, pues permite considerar más procesos que el sólo tomar la estacionaridad estricta. Hay que destacar que si el proceso es Gaussiano [III] se puede mostrar que las dos definiciones de

[III] Un proceso estocástico se dice que es Gaussiano o normal si la distribución de probabilidad conjunta es normal.

estacionaridad coinciden. A partir de aquí, cuando se hable de la estacionaridad de un proceso se referirá solamente a la estacionaridad en covarianza.

2.1.3. Función de Autocorrelación Parcial

La función de autocorrelación (dada por la ecuación 2.8) es muy útil para investigar algunas características de los procesos estacionarios. Pero además de esta función, existe otra relación que es también bastante útil para estudiar los procesos.

La función de autocorrelación parcial (PACF[IV] por sus siglas en inglés) se define como la correlación entre Y_t y Y_{t-k} una vez que el efecto lineal de las variables intermedias $Y_{t-1}, Y_{t-2}, ..., Y_{t-k+1}$ se han removido. Matemáticamente se expresa como:

$$\phi_{k,k} = Corr(Y_t, Y_{t-k}|Y_{t-1}, Y_{t-2}, ..., Y_{t-k+1}) \tag{2.10}$$

Se puede plantear la ecuación 2.10 desde otro punto de vista. Considere predecir Y_t basado en una función lineal que depende de las variables $Y_{t-1}, Y_{t-2}, ..., Y_{t-k+1}$, esto es, $Y_t = \beta_1 Y_{t-1} + \beta_2 Y_{t-2}, + ... + \beta_{k-1} Y_{t-k+1}$ donde las betas se seleccionan para minimizar el error cuadrático medio de la predicción. Si asumimos que las betas han sido escogidas, entonces al pensar en el pasado, se sigue por la estacionariedad que el mejor predictor de Y_{t-k} basado en las mismas variables $Y_{t-1}, Y_{t-2}, ..., Y_{t-k+1}$ es $Y_{t-k} = \beta_1 Y_{t-1} + \beta_2 Y_{t-2}, + ... + \beta_{k-1} Y_{t-k+1}$. La función de autocorrelación parcial para el rezago k, se define entonces como la correlación entre los errores de predicción de Y_t y Y_{t-k}, esto es:

$$\begin{aligned} \phi_{k,k} &= Corr(Y_t - predictor, Y_{t-k} - predictor) = Corr(Y_t - \hat{Y}_t, Y_{t-k} - \hat{Y}_{t-k}) \\ &= Corr(Y_t - (\beta_1 Y_{t-1} + ... + \beta_{k-1} Y_{t-k+1}), Y_{t-k} - (\beta_1 Y_{t-1} + ... + \beta_{k-1} Y_{t-k+1})) \end{aligned}$$

Obtener una expresión clara y general para cualquier valor k a partir de las relación anterior puede ser un poco complicado. Los interesados pueden revisar el desarrollo en Wie[7].

Finalmente existe otra forma de derivar la función de autocorrelación parcial que refleja de forma más clara como calcular explícitamente los valores de ϕ_{kk} para cualquier valor de k. Considérese un modelo de regresión donde la variable dependiente Y_t (que sigue un proceso estacionario Y_t con media cero y varianza γ_0) se pone en función de k variables de rezago $Y_{t-1}, Y_{t-2}, ..., Y_{t-k}$, es decir:

$$Y_t = \phi_{k1} Y_{t-1} + \phi_{k2} Y_{t-2} + ... + \phi_{kk} Y_{t-k} + e_t \tag{2.11}$$

donde ϕ_{ki} denota el parámetro i de la regresión y e_t es un término de error que tiene media cero y no está correlacionado con $Y_{t-1}, Y_{t-2}, ..., Y_{t-k}$. Al multiplicar ambos lados de la expresión 2.11 por Y_{t-j} y calcular el valor esperado se obtiene:

$$\gamma_j = \phi_{k1} \gamma_{j-1} + \phi_{k2} \gamma_{j-2} + ... + \phi_{kk} \gamma_{j-k} \tag{2.12}$$

[IV] Proviene de partial autocorrelation function

donde $E(Y_{t-i}Y_{t-j}) = \gamma_{j-i}$ ya que se ha supuesto que el proceso tiene media cero y $E(e_tY_{t-j}) = 0$ pues como se dijo el error no está correlacionado con el proceso. Al dividir entre la varianza del proceso γ_0 se tiene la siguiente ecuación:

$$\rho_j = \phi_{k1}\rho_{j-1} + \phi_{k2}\rho_{j-2} + ... + \phi_{kk}\rho_{j-k} \tag{2.13}$$

Aplicando esta relación a los valores de $j = 1, 2, ..., k$ se obtiene el siguiente sistema de ecuaciones:

$$\begin{aligned}
\rho_1 &= \phi_{k1}\rho_0 + \phi_{k2}\rho_1 + ... + \phi_{kk}\rho_{k-1} \\
\rho_2 &= \phi_{k1}\rho_1 + \phi_{k2}\rho_0 + ... + \phi_{kk}\rho_{k-2} \\
&\vdots \\
\rho_k &= \phi_{k1}\rho_{k-1} + \phi_{k2}\rho_{k-2} + ... + \phi_{kk}\rho_0
\end{aligned}$$

Donde se han usado las propiedades de la función de autocorrelación (ecuación 2.9). Resolviendo el sistema usando a regla de Cramer se obtiene que:

$$\phi_{kk} = \frac{\begin{vmatrix} 1 & \rho_1 & \rho_2 & ... & \rho_{k-2} & 1 \\ \rho_1 & 1 & \rho_1 & ... & \rho_{k-2} & \rho_{k-3} \\ ... & ... & ... & ... & ... & ... \\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} & ... & \rho_1 & 1 \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 & \rho_2 & ... & \rho_{k-2} & \rho_{k-1} \\ \rho_1 & 1 & \rho_1 & ... & \rho_{k-2} & \rho_{k-2} \\ ... & ... & ... & ... & ... & ... \\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} & ... & \rho_1 & 1 \end{vmatrix}} \tag{2.14}$$

2.1.4. Estimación de la media y de las funciones de autocovarianza, autocorrelación y autocorrelación parcial

Una serie de tiempo estacionaria se puede caracterizar por su media μ, su varianza σ^2, sus autocorrelaciones γ_k, y sus autocorrelaciones parciales ϕ_{kk}. Los valores exactos del proceso se podrían calcular si se conocieran todas las realizaciones posibles del ensamble; o podrían ser estimados si varias realizaciones independientes estuvieran a la mano. En la mayoría de las aplicaciones no es fácil tienes varias realizaciones del proceso, generalmente se cuenta con una solo realización. Sin embargo, existe una alternativa para los procesos estacionarios: remplazar los promedios del ensamble por promedios en el tiempo. Bajo ciertas condiciones, y con buenas propiedades estadísticas, se pueden estimar la media, la varianza, y la funciones de autocorrelación empleando promedios temporales.

Estimación de la media

Al contar con una sola realización $Y_1, Y_2, ..., Y_n$el estimador que aparece de forma natural para estimar la media de un proceso estacionario es la media muestral definida como:

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \tag{2.15}$$

Definido de esta forma el estimador es insesgado, pues se cumple que $E(\bar{Y}) = \mu$. Además $\bar{Y}$ es un estimador consistente para la media μ pues al calcular la varianza:

$$\begin{aligned} Var(\bar{Y}) &= \frac{1}{n^2}\sum_{t=1}^{n}\sum_{s=1}^{n} Cov(Y_t, Y_s) = \frac{\gamma_0}{n^2}\sum_{t=1}^{n}\sum_{s=1}^{n}\rho_{(t-s)} \\ &= \frac{\gamma_0}{n^2}\sum_{k=-(n-1)}^{n-1}(1-|k|)\rho_k = \frac{\gamma_0}{n}\sum_{k=-(n-1)}^{n-1}(1-\frac{|k|}{n})\rho_k \end{aligned}$$

se ve que cuando $n \to \infty$, la $Var(\bar{Y}) \to 0$. De esta forma se ve que la media muestral $\bar{Y}$ posee buenas propiedades estadísticas.

Estimación de la autocovarianza

Partiendo de un proceso estacionario, se busca estimar las autocovarianzas γ_k para distintos valores de k. La forma natural de hacerlo es calcular la covarianza muestral entre pares separados k unidades de tiempo, esto es, entre $(Y_1, Y_{1+k}), (Y_2, Y_{2+k}), ..., (Y_n, Y_{n+k})$. Sin embargo, tomando en cuenta que el proceso es estacionario, por lo que posee una media y varianza común, se tiene el siguiente estadístico que estima la función de autocovarianza, y que se le conoce como la función de autocovarianza muestral:

$$\hat{\gamma}_k = \frac{1}{n}\sum_{t=k+1}^{n}(Y_t - \hat{Y})(Y_{t-k} - \hat{Y})) \tag{2.16}$$

Se puede ver en Wie[7] que al calcular el valor esperado de γ_k se obtiene:

$$E(\hat{\gamma}_k) \cong \gamma_k - \frac{k}{n}\gamma_k - \frac{n-k}{n}Var(\hat{Y}) \tag{2.17}$$

De la relación anterior se observa que el estimador $\hat{\gamma}_k$ no es insesgado para cualquier valor de k. Sin embargo, cuando $n \to \infty$, el sesgo tiende a cero, por lo que $\hat{\gamma}_k$ es insesgado asintóticamente.

El cálculo de la varianza del estimador $\hat{\gamma}_k$ puede resultar complicado para un proceso estacionario en general. Sin embargo, si el proceso $\{Y_t\}$ es gaussiano se puede obtener la siguiente aproximación de acuerdo a Bartlett[8]:

$$Var(\hat{\gamma}_k) \cong \frac{1}{n}\sum_{i=-\infty}^{\infty}(\gamma_i^2 - \gamma_{i+k}\gamma_{i-k}) \tag{2.18}$$

Estimación de la autocorrelación

Una vez estimada la autocovarianza del proceso estacionario, se puede obtener un estimador de la función de autocorrelación ρ_k, al dividir el estimador $\hat{\gamma}_k$ entre el estimador de la varianza del proceso $\hat{\gamma}_0$, que se obtiene de la relación 2.16 al hacer $k = 0$:

$$\hat{\gamma}_0 = \frac{1}{n}\sum_{t=1}^{n}(Y_t - \bar{Y})^2 \tag{2.19}$$

De esta forma, el estimador de la función de autocorrelación $\hat{\rho}_k$ es:

$$\hat{\rho}_k = \frac{\sum_{t=k+1}^{n}(Y_t - \bar{Y})(Y_{t-k} - \bar{Y})}{\sum_{t=1}^{n}(Y_t - \bar{Y})^2} \tag{2.20}$$

A $\hat{\rho}_k$ se le conoce como la función de autocorrelación muestral, y también se denotará como como r_k. La gráfica de la autocorrelación muestral contra el rezago k, se le conoce como ***correlograma***.

Para un proceso en general $\{Y_t\}$, por la definición $\hat{\rho}_k$ las propiedades de este estadístico no son fáciles de obtener, ni siquiera su valor esperado. Se tiene de nueva cuenta que considerar una aproximación cuando el proceso estacionario es Gaussiano. En este caso, Bartlett[8] mostró que la covarianza entre $\hat{\rho}_k$ y $\hat{\rho}_{k+j}$ para $k > 0$ y $k + j > 0$ está dada por:

$$Cov(\hat{\rho}_k, \hat{\rho}_{k+j}) \cong \frac{1}{n} \sum_{i=-\infty}^{\infty} (\rho_i \rho_{i+j} + \rho_{i+k+j}\rho_{i-k} - 2\rho_k \rho_i \rho_{i-k-j} - 2\rho_{k+j}\rho_i \rho_{i-k} + 2\rho_k \rho_{k+j}\rho_i^2) \quad (2.21)$$

Además se cumple que cuando n es grande, $\hat{\rho}_k$ se distribuye aproximadamente normal con media ρ_k, y varianza dada por:

$$Var(\hat{\rho}_k) \cong \frac{1}{n} \sum_{i=-\infty}^{\infty} (\rho_i^2 + \rho_{i+k}\rho_{i-k} - 4\rho_k \rho_i \rho_{i-k} + 2\rho_k^2 \rho_i^2) \quad (2.22)$$

Estimación de la autocorrelación parcial

Al obtener la función de autocorrelación muestral $\hat{\rho}_k$, la función de autocorelación parcial muestral $\hat{\phi}_{kk}$, se puede obtener al sustituir ρ_i por $\hat{\rho}_i$ en la ecuación 2.14. Sin embargo, este proceso involucra el cálculo de determinantes, lo que puede resultar un poco complicado. De forma alternativa, Durbin[9] desarrollo un método recursivo para calcular $\hat{\phi}_{kk}$ empezando con $\hat{\phi}_{kk} = \hat{\rho}_1$. Las ecuaciones recursivas son las siguientes:

$$\hat{\phi}_{k+1,k+1} = \frac{\hat{\rho}_{k+1} - \sum_{j=1}^{k} \hat{\phi}_{kj}\hat{\rho}_{k+1-j}}{1 - \hat{\phi}_{kj}\hat{\rho}_j} \quad (2.23)$$

$$\hat{\phi}_{k+1,j} = \hat{\phi}_{k,j} - \hat{\phi}_{k+1,k+1}\hat{\phi}_{k,k+1-j} \quad (2.24)$$

Para $j = 1, 2, ..., k$. El método es válido también para calcular la función de autocorrelación ϕ_{kk} con las respectivas funciones de autocorrelación ρ_i.

2.1.5. Procesos de ruido blanco

Un ejemplo importante de un proceso estacionario es el llamado proceso de ruido blanco , que es una secuencia de variables aleatorias independientes e idénticamente distribuidas e_t. Para el ruido blanco se tiene que su media y su varianza son constantes $E(e_t) = 0$ y $Var(e_t) = \sigma_e^2$. Además su función de autocovarianza es:

$$\gamma_k = \begin{cases} \sigma_e^2 & k = 0; \\ 0 & k \neq 0 \end{cases} \quad (2.25)$$

La de autocorrelacion queda:

$$\rho_k = \begin{cases} 1 & k = 0 \\ 0 & k \neq 0 \end{cases} \quad (2.26)$$

Finalmente la función de autocorrelación parcial es:

$$\phi_{kk} = \begin{cases} 1 & k = 0 \\ 0 & k \neq 0 \end{cases} \tag{2.27}$$

Aunque en la práctica estos procesos casi no ocurren, su importancia radica en el hecho de que juega un rol importante para construir otros procesos en el análisis de series de tiempo.

Un proceso de ruido blanco se dice que es Gaussiano, si la función de distribución conjunta del proceso es normal.

2.2. Modelos de series de tiempo estacionarios

Los modelos autoregresivos de promedios móviles (ARMA por sus siglas en inglés[V]) son ampliamente manejados para describir el comportamiento de las series de tiempo estacionarias; estos modelos son sumamente útiles en el modelado de procesos en el mundo real.

2.2.1. Principales representaciones de las series de tiempo

Para introducir los modelos ARMA, se comienza con el estudio de los modelos de promedios móviles (MA por sus siglas en inglés[VI]) y luego con los modelos autoregresivos (AR por el idioma inglés[VII]) que son dos representaciones muy útiles para expresar las series de tiempo.

La primera representación de Y_t es una combinación lineal ponderada de términos presentes y pasados de un proceso de ruido blanco e_t:

$$Y_t = e_t + \psi_1 e_{t-1} + \psi_2 e_{t-2} + \dots \tag{2.28}$$

La representación anterior se conoce como ***proceso lineal general*** o ***proceso de promedios móviles***.

Para este modelo, se pueden calcular la media, la varianza, la función de autocovarianza y la función de autocorrelación del proceso lineal general, considerando las propiedades del ruido blanco; al hacerlo se obtienen los siguientes resultados:

$$E(Y_t) = 0 \tag{2.29}$$

$$Var(Y_t) = \sigma_e^2 \sum_{i=0}^{\infty} \psi_i^2 \tag{2.30}$$

$$\gamma_k = \sigma_e^2 \sum_{i=0}^{\infty} \psi_i \psi_{i+k} \tag{2.31}$$

[V] *Autoregressive Moving Average*
[VI] *Proviene de moving average*
[VII] *En inglés se denota por Autoregressive*

$$\rho_k = \frac{\sum_{i=0}^{\infty} \psi_i \psi_{i+k}}{\sum_{i=0}^{\infty} \psi_i^2} \tag{2.32}$$

Para que Y_t sea una serie estacionaria se debe satisfacer que $Var(Y_t) < \infty$ y que $\gamma_k < \infty$, para ello se debe cumplir que la suma cuadrada infinita de los pesos no diverja, es decir $\sum_{i=0}^{\infty} \psi_i^2 < \infty$ (donde $\psi_0 = 1$, es el coeficiente de e_t). Si $Var(Y_t) < \infty$, se sigue que $\gamma_k < \infty$, como se observa en la siguiente relación:

$$|\gamma_k| = |E(Y_t, Y_{t+k})| \leq [Var(Y_t)Var(Y_{t+k})]^{1/2} = \sigma_e^2 \sum_{i=0}^{\infty} \psi_i^2 \tag{2.33}$$

Entonces la única consideración para que Y_t sea estacionaria es que la varianza del proceso esté acotada.

Existe una forma reducida de escribir la expresión del proceso lineal general 2.28. Esto se logra usando el operador de retraso denotado por la letra B, el cual opera sobre sobre el índice del tiempo generando el retraso de una unidad de tiempo:

$$BY_t = Y_{t-1} \tag{2.34}$$

Empleando el operador de retraso B, la forma compacta del proceso de promedios móviles es:

$$Y_t = \sum_{i=0}^{\infty} \psi_i B^i e_t = \psi(B) e_t \tag{2.35}$$

donde $\sum_{i=0}^{\infty} \psi_i B^i = \psi(B)$

La segunda representación de la serie de tiempo es una combinación lineal de sus valores pasados $Y_{t-1}, Y_{t-2}, Y_{t-3}, ...$ más un término de ruido blanco e_t:

$$Y_t = \pi_1 Y_{t-1} + \pi_2 Y_{t-2} + \cdots + e_t \tag{2.36}$$

Esta representación se conoce como ***proceso autoregresivo***, pues como se observa, es una regresión sobre la misma variable Y_t. Usando el operador de retraso, el proceso autoregresivo se puede escribir en forma compacta:

$$(1 - \sum_{i=0}^{\infty} \pi_i B^i) Y_t = \pi(B) Y_t = e_t \tag{2.37}$$

donde $\pi(B) = 1 - \sum_{i=0}^{\infty} \pi_i B^i$.

Si se cumple que la suma del valor absoluto de los pesos está acotada, es decir $1 + \sum_{i=0}^{\infty} |\pi_i| < \infty$, entonces un proceso autoregresivo se puede escribir como un proceso de promedios móviles al despejar Y_t en la relación 2.37:

$$Y_t = \frac{1}{\pi(B)} e_t = \psi(B) e_t \tag{2.38}$$

Si se cumple lo anterior, entonces se dice que el proceso autoregresivo es ***invertible***.

Para que el proceso de promedios móviles que resulta de un proceso autoregresivo invertible sea estacionario, es decir que se satisfaga que $\sum_{i=0}^{\infty} \psi_i^2 < \infty$, se debe de cumplir que las raíces del

polinomio de retraso $\pi(B)$ como función de B, estén fuera del círculo unitario. Esto es, si δ es una raíz de la expresión $\pi(B) = 0$, entonces se debe cumplir que $||\delta||_2 > 1$. Hay que tener en cuenta que las raíces de polinomio pueden ser valores reales o complejos.

Por otro lado,si se tiene un proceso de promedios móviles estacionario, se puede invertir y generar un proceso autoregresivo:

$$e_t = \frac{1}{\psi(B)} Y_t = \pi(B) Y_t \tag{2.39}$$

Para que la expresión anterior tenga sentido matemático, se debe cumplir ahora que las raíces de $\psi(B)$como función de B, estén fuera del circulo unitario. De esta forma, se observa que existe una estrecha relación entre los procesos de promedios móviles y los procesos autoregresivos.

Si bien, estas representaciones son útiles para modelar las series de tiempo, en la práctica no son empleadas pues contienen una infinidad de parámetros, que es imposible estimar de un número finito de observaciones. En vez de ello, se construyen modelos con un número finito de parámetros.

2.2.2. Procesos de promedios Móviles

Dentro de los procesos de promedios móviles, si sólo un número finito q de pesos ψ son distintos de cero se tiene un proceso de promedios móviles de orden q:

$$Y_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - ... - \theta_q e_{t-q} \tag{2.40}$$

Se observa que $\psi_1 = -\theta_1$, $\psi_2 = -\theta_2$,..., $\psi_q = -\theta_q$ y $\psi_k = 0$ para $k > q$. Este proceso se abrevia como MA(q). Usando el polinomio de retraso $\theta_q(B) = (1 - \theta_1 B - \theta_2 B^2 - ... - \theta_q B^q)$, el modelo se puede escribir como:

$$Y_t = \theta_q(B) e_t \tag{2.41}$$

Como se tiene un número finito de parámetros este proceso es estacionario. Además, si las raíces del polinomio $\theta_q(B)$ están fuera del círculo unitario, entonces el proceso es invertible. Estos modelos son útiles para describir fenómenos donde los eventos producen efectos inmediatos que sólo perduran un periodo breve de tiempo. Para especificar algunas propiedades del proceso MA(q), se verá un caso especial: el proceso MA(1).

Proceso de promedios Móviles de primer orden

El modelo MA(1) está dado por:

$$Y_t = e_t - \theta_1 e_{t-1} = (1 - \theta_1 B) e_t \tag{2.42}$$

donde e_t es un proceso de ruido blanco con varianza σ_e^2. El polinomio de retraso en este caso es $\theta_1(B) = (1 - \theta_1 B)$. Como sólo se tiene un peso θ_1, el proceso es estacionario. Para que sea invertible, las raíces de $\theta_1(B)$ deben estar fuera del circulo unitario. En este caso, la raíz es $B = 1/\theta_1$, por lo que se tiene que cumplir que $|\theta_1| < 1$ para tener la propiedad de invertibilidad.

Claramente se tiene que la media es cero $E(Y_t) = 0$ y que la varianza es $Var(Y_t) = \sigma_e^2(1 + \theta_1^2)$. Al calcular la covarianza para un rezago se tiene que:

$$Cov(Y_t, Y_{t-1}) = Cov(e_t - \theta_1 e_{t-1}, e_{t-1} - \theta_1 e_{t-2}) = Cov(-\theta_1 e_{t-1}, e_{t-1}) = -\theta_1 \sigma_e^2$$

Y para dos rezagos:

$$Cov(Y_t, Y_{t-2}) = Cov(e_t - \theta_1 e_{t-1}, e_{t-2} - \theta_1 e_{t-3}) = 0$$

En general se cumple $Cov(Y_t, Y_(t-k)) = 0$ para $k \geq 2$, por lo que el proceso no tiene correlación más allá del primer rezago.

En resumen, la función de autocovarianza para el modelo MA(1) es:

$$\gamma_k = \begin{cases} \sigma_e^2(1+\theta_1^2) & k = 0 \\ -\theta_1\sigma_e^2 & k = 1 \\ 0 & k > 2 \end{cases} \tag{2.43}$$

Por lo que la función de autocorrelación queda como:

$$\rho_k = \begin{cases} \frac{-\theta_1}{1+\theta_1^2} & k = 1 \\ 0 & k = 2 \end{cases} \tag{2.44}$$

Se tiene pues, un solo valor para $k = 1$, que puede ser positivo o negativo dependiendo del signo del parámetro θ_1.

La función de autocorrelación parcial del proceso MA(1) se encuentra sustituyendo los valores de ρ_k de la ecuación anterior en la relación que se dedujo para encontrar los valores de $\phi_k k$ en términos de ρ_k (2.14). Sustituyendo se puede ver que:

$$\phi_{11} = \rho_1 = \frac{-\theta_1}{1+\theta_1^2} = \frac{-\theta_1(1-\theta_1^2)}{1-\theta_1^4}$$

$$\phi_{22} = \frac{\rho_1^2}{1-\rho_1^2} = \frac{-\theta_1^2}{1+\theta_1^2+\theta_1^4} = \frac{-\theta_1^2(1-\theta_1^2)}{1-\theta_1^6}$$

En general se obtiene:

$$\phi_{kk} = \frac{-\theta_1^k(1-\theta_1^2)}{1-\theta_1^{2(k+1)}} \tag{2.45}$$

Para valores de $k \geq 1$. A diferencia de la función de autocorrelación que se corta después del primer rezago, la función de autocorrelación parcial tiene un decaimiento exponencial pues se cumple que $|\theta_1| < 1$ por la invertibilidad.

Proceso de promedios Móviles de orden q

Para el proceso general de orden q, $Y_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - ... - \theta_q e_{t-q}$, a través de cálculos similares a los que se realizaron para el proceso MA(1), se puede obtener la función de autocovarianza y autocorrelación:

$$\gamma_k = \begin{cases} (1+\theta_1^2+\theta_2^2+...+\theta_q^2)\sigma_e^2 & k = 0 \\ (-\theta_k + \theta_1\theta_{k+1} + ... + \theta_q\theta_{q-k})\sigma_e^2 & k = 1, 2, ..., q \\ 0 & k > q \end{cases} \tag{2.46}$$

$$\rho_k = \begin{cases} \frac{(-\theta_k+\theta_1\theta_{k+1}+...+\theta_q\theta_{q-k})}{(1+\theta_1^2+\theta_2^2+...+\theta_q^2)} & k = 1, 2, ..., q \\ 0 & k > q \end{cases} \tag{2.47}$$

Con los valores de la función de autocorrelación, se pueden encontrar los valores de la función de autocorrelación parcial a través de la ecuación 2.14. Encontrar una expresión explícita como ocurrió en el proceso MA(1) puede resultar complicado; lo importante en la práctica es conocer el comportamiento cualitativo de la función de autocorrelación parcial. En general la función de autocorrelación posee un decaimiento exponencial como ocurrió en el proceso MA(1), si las raíces del polinomio $\theta_q(B) = (1 - \theta_1 B - \theta_2 B^2 - ... - \theta_q B^q)$ son reales. En el caso de que existan raíces complejas, el comportamiento es una mezcla de decaimiento exponencial con amortiguamiento senoidal.

2.2.3. Procesos autoregresivos

Cuando en un proceso autoregresivo, sólo un número finito de pesos π son distintos de cero, se tiene un proceso autoregresivo de orden p:

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + e_t \tag{2.48}$$

donde $\pi_1 = \phi_1$, $\pi_2 = \phi_2, ...$, $\pi_p = \phi_p$ y $\pi_k = 0$ para $k > p$. Este proceso se conoce como AR(p). Empleando el polinomio de retraso $\phi_p(B) = (1 - \phi_1 B - \phi_2 B^2 - ... - \phi_p B^p)$, el proceso se puede escribir como:

$$\phi_p(B) Y_t = e_t \tag{2.49}$$

Al tener un número finito de parámetros el proceso es invertible. Para que el proceso sea estacionario, las raíces del polinomio $\phi_p(B)$ deben estar fuera del círculo unitario. Estos modelos son útiles para describir situaciones en las que el valor presente depende de los valores pasados más un término de error. Al igual que ocurrió con los procesos de promedios móviles, para tener presente algunas de las propiedades de los procesos AR(p), se verá primero el modelo autoregresivo más simple: el proceso AR(1).

Proceso Autoregresivo de primer orden

El proceso AR(1) está dado por:

$$Y_t = \phi_1 Y_{t-1} + e_t \tag{2.50}$$

que puede ser escrito como:

$$Y_t - \phi_1 Y_{t-1} = (1 - \phi_1 B) Y_t = e_t$$

donde e_t es un proceso de ruido blanco con varianza σ_e^2. El polinomio de retraso para este modelo es $\phi_1(B) = (1 - \phi_1 B)$. Al contar sólo un peso ϕ_1, el proceso es invertible. Para que sea estacionario, las raíces de $\phi(B)$ deben encontrarse fuera del círculo unitario. Este polinomio sólo cuenta con una raíz, que es $B = 1/\phi_1$; entonces se tiene que cumplir que $|\phi_1| < 1$ para que el proceso sea estacionario.

Como ocurrió en el proceso MA(1), la media del modelo AR(1) es cero $E(Y_t) = 0$. La varianza se calcula al aplicar el operador en ambos lados de la ecuación 2.50, obteniéndose:

$$Var(Y_t) = \phi_1^2 Var(Y_{t-1}) + \sigma_e^2$$

Al ser el proceso estacionario $Var(Y_t) = Var(Y_{t-1})$, por lo que se llega a:

$$Var(Y_t) = \frac{\sigma_e^2}{1 - \phi_1^2} \tag{2.51}$$

Para encontrar la función de autocovarianza se multiplica ambos lados de la ecuación 2.50 por $Y_{t-k}(k = 1, 2, 3, ...)$ y se toma el valor esperado:

$$E(Y_t Y_{t-k}) = E(\phi_1 Y_{t-1} Y_{t-k} + e_t Y_{t-k}) = \phi_1 E(Y_{t-1} Y_{t-k}) + E(e_t Y_{t-k})$$

Como se ha supuesto que el proceso es estacionario y con media cero entonces $E(Y_t Y_{t-k}) = \gamma_k$ y $E(Y_{t-1} Y_{t-k}) = \gamma_{k-1}$. Además el término de ruido blanco e_t es independiente de Y_{t-k} por lo que $E(e_t Y_{t-k}) = 0$. De esta forma se obtiene:

$$\gamma_k = \phi_1 \gamma_{k-1} \quad para\ k = 1, 2, 3, ... \tag{2.52}$$

Como $\gamma_0 = Var(Y_t)$, la función de autocovarianza queda en general:

$$\gamma_k = \phi_1^k \frac{\sigma_e^2}{1 - \phi_1^2} \quad para\ k = 1, 2, 3, ... \tag{2.53}$$

Así, la función de autocorrelación está dada por:

$$\rho_k = \phi_1^k \quad para\ k = 1, 2, 3, ... \tag{2.54}$$

Como se cumple que $|\phi_1| < 1$ para que el proceso sea estacionario, entonces la función de autocorrelación decae exponencialmente.

Como $\rho_k = \phi_1^k$, el determinante que está en el numerador en la ecuación 2.14 se hace cero para todos los valores de $k \geq 2$. El único valor que sobrevive es $\phi_{11} = \rho_1 = \phi_1$. En resumen, la función de autocorrelación parcial es:

$$\phi_{kk} = \begin{cases} \rho_1 = \phi_1 & k = 1 \\ 0 & k > 2 \end{cases} \tag{2.55}$$

Al igual que la función de autocorrelación para el proceso MA(1), la función de autocorrelación parcial del proceso AR(1) tiene sólo un valor para $k = 1$, que es positivo o negativo dependiendo del signo del parámetro ϕ_1.

Es importante resaltar que existe cierta dualidad en el comportamiento de la funciones de autocorrelación y autocorrelación parcial de los procesos MA(1) y AR(1). Como se mencionó, la función de autocorrelación del proceso MA(1) se corta después del rezago 1, comportamiento que comparte con la función de autocorrelación parcial del modelo AR(1). Por otro lado, la función de autocorrelación parcial del proceso MA(1) tiene un decaimiento exponencial, lo que ocurre también con la función de autocorrelación del modelo AR(1). Este comportamiento se extiende para modelos más generales, como se verá a continuación.

Proceso Autoregresivo de orden p

Para calcular la función de autocovarianza en el proceso autoregresivo general de orden p, $Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + ... + \phi_p Y_{t-p} + e_t$, se realiza un proceso similar al que se hizo en el modelo AR(1): se multiplica ambos lados de ecuación 2.48 por Y_{t-k} y se toma el valor esperado. Se obtiene la siguiente relación:

$$\gamma_k = \phi_1 \gamma_{k-1} + \phi_2 \gamma_{k-2} + \cdots + \phi_p \gamma_{k-p} \quad k > 0 \tag{2.56}$$

Por lo tanto, se tiene la siguiente relación recursiva para la función de autocorrelación:

$$\rho_k = \phi_1 \rho_{k-1} + \phi_2 \rho_{k-2} + \cdots + \phi_p \rho_{k-p} \quad k > 0 \tag{2.57}$$

La ecuación anterior se puede ver como una ecuación en ***diferencias homogénea***:

$$\rho_k - \phi_1\rho_{k-1} - \cdots - \phi_p\rho_{k-p} = (1 - \phi_1 B - ... - \phi_p B^{k-p})\rho_k = \phi_p(B)\rho_k \tag{2.58}$$

El comportamiento de la solución de una ecuación en diferencias homogénea está determinado por las raíces del polinomio de retraso $\phi_p(B)$. Si las raíces del polinomio son reales, se tiene ρ_k exhibe un decaimiento exponencial. Si existen raíces complejas, además del decaimiento exponencial, la solución presenta un amortiguamiento senoidal. La función de autocorrelación parcial de los modelos MA tiene un comportamiento muy parecido, por lo que se observa la dualidad mencionada entre la función de autocorrelación de los modelos AR y la función de autocorrelación parcial de los modelos MA.

De nueva cuenta, para encontrar los valores de la función de autocorrelación parcial, se emplea la ecuación 2.14, sustituyendo los valores de la función de autocorrelación. Al tener una ecuación recursiva para ρ_k, se cumple que cuando $k > p$, la última columna del determinante que está en el numerador de ϕ_{kk}, puede ser escrita como una combinación lineal de columnas previas, por lo que el determinante es igual a cero, y la función de autocorrelación parcial se corta después del rezago p. Este comportamiento es similar a la función de autocorrelación de los modelo MA, observándose de nueva cuenta la dualidad existente entre ϕ_{kk} de los modelos AR y ρ_k de los modelos MA.

2.2.4. Procesos autoregresivos de promedios móviles

Se pueden combinar los modelos de promedios móviles MA con los modelos autoregresivos AR como una extensión natural para formar modelos más generales de series de tiempo: los llamados procesos autoregresivos de promedios móviles ARMA. Si la serie Y_t sigue un modelo ARMA(p, q), entonces cumple la siguiente relación:

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - ... - \theta_q e_{t-q} \tag{2.59}$$

Usando los polinomios de retraso se tiene de forma equivalente la ecuación:

$$\phi_p(B)Y_t = \theta_q(B)e_t \tag{2.60}$$

Para que el proceso sea estacionario se requiere que las raíces del polinomio $\phi_p(B) = 0$ estén fuera del círculo unitario. Si esto se cumple, el proceso se puede escribir únicamente con términos de promedios móviles:

$$Y_t = \psi(B)e_t \tag{2.61}$$

donde:

$$\psi(B) = \frac{\theta_q(B)}{\phi_p(B)} \tag{2.62}$$

La invertibilidad se logra si las raíces del polinomio $\theta_q(B) = 0$ residen en el exterior del círculo unitario. En este caso el ruido blanco se puede representar solamente con términos autoregresivos:

$$e_t = \pi(B)Y_t \tag{2.63}$$

donde:

$$\pi(B) = \frac{\phi_q(B)}{\theta_p(B)} \tag{2.64}$$

Como ocurrió con los modelos MA y AR, se introduce primero el modelo más sencillo ARMA$(1,1)$, para después generalizar sus propiedades para procesos de mayor orden.

Proceso autoregresivo de promedios móviles de primer orden

El proceso ARMA$(1,1)$ se expresa como:

$$Y_t = \phi_1 Y_{t-1} + e_t - \theta_1 e_{t-1} \tag{2.65}$$

el cual puede ser escrito como:

$$(1 - \phi_1 B) Y_t = (1 - \theta_1 B) e_t \tag{2.66}$$

Para que el proceso sea estacionario se debe cumplir que $|\phi_1| < 1$, y para que sea invertible se necesita que $|\theta_1| < 1$.

Para obtener la función de autocorrelación se debe notar primero que:

$$E(e_t Y_t) = E(e_t(\phi_1 Y_{t-1} + e_t - \theta_1 e_{t-1})) = \sigma_e^2 \tag{2.67}$$

y que:

$$E(e_{t-1} Y_t) = E(e_{t-1}(\phi_1 Y_{t-1} + e_t - \theta_1 e_{t-1})) = \phi_1 E(e_{t-1} Y_{t-1}) - \theta_1 E(e_{t-1} e_{t-1}) \tag{2.68}$$

$$= \phi_1 \sigma_e^2 - \theta_1 \sigma_e^2 = (\phi_1 - \theta_1)\sigma_e^2 \tag{2.69}$$

COMO LE HAGO PARA NUMERAR SOLO UNA RELACIÓN

Ahora bien, al multiplicar la ecuación 2.65 por Y_{t-k} y calcular el valor esperado se obtiene:

$$\gamma_k = \phi_1 \gamma_{k-1} + E(e_t Y_{t-k}) - \theta_1 E(e_{t-1} Y_{t-k}) \tag{2.70}$$

Al usar las relaciones 2.67 y 2.68 se obtiene la función de autocovarianza:

$$\gamma_k = \begin{cases} \phi_1 \gamma_1 + (1 - \theta_1(\phi_1 - \theta_1))\sigma_e^2 & k = 0 \\ \phi_1 \gamma_0 - \theta_1 \sigma_e^2 & k = 1 \\ \phi_1 \gamma_{k-1} & k > 1 \end{cases} \tag{2.71}$$

Al sustituir el valor de γ_1 en la expresión de γ_0 se pueden simplificar las relaciones para sólo tener los parámetros ϕ_1, θ_1 y σ_e^2. Finalmente, al dividir entre γ_0, se obtiene la función de autocorrelación:

$$\rho_k = \begin{cases} \frac{(\phi_1 - \theta_1)(1 - \phi_1 \theta_1)}{1 + \theta_1^2 - 2\phi_1 \theta_1} & k = 1 \\ \phi_1 \rho_{k-1} & k > 1 \end{cases} \tag{2.72}$$

De la ecuación anterior se observa que la función de autocorrelación combina características de los procesos MA(1) y AR(1), pues en $k = 1$ el parámetro θ_1 es necesario para realizar los cálculos (cómo ocurre en el modelo MA(1)), pero para $k > 1$ la función de autocorrelación decae exponencialmente (característica del proceso AR(1)).

La función de autocorrelación parcial resulta en una expresión muy complicada, y es preferible no hacer todo el desarrollo (por cuestiones de espacio). Lo importante es conocer su comportamiento cualitativo, el cual es similar al de la función de autocorrelación. De igual manera, existe una

mezcla de los procesos AR(1) y MA(1), sólo que ahora el decaimiento para $k > 1$ se debe a las características del modelo MA(1).

Proceso autoregresivo de promedios móviles de orden (p,q)

Para el proceso general de orden (p, q), cálculos similares a los que se realizaron para el modelo ARMA(1,1) se aplican para calcular la función de autocovarianza; se obtiene la siguiente relación:

$$\gamma_k = \phi_1\gamma_{k-1} + \cdots + \phi_p\gamma_{k-p} \quad para \;\; k \geq (q+1) \tag{2.73}$$

Al dividir entre γ_0 se obtiene la expresión para ACF:

$$\rho_k = \phi_1\rho_{k-1} + \cdots + \phi_p\rho_{k-p} \quad para \;\; k \geq (q+1) \tag{2.74}$$

La ecuación anterior se puede ver como una ecuación homogénea en diferencias (como ocurrió en el proceso AR(p), ecuación 2.58), por lo que después del rezago q la función de autocorrelación tiene un comportamiento similar al del modelo AR(p). Las primeras q correlaciones tienen una dependencia de los parámetros autoregresivos y de promedios móviles, como ocurrió en la ACF del proceso ARMA(1,1).

En analogía con la función de autocorrelación parcial del proceso ARMA(1,1), los primeros p valores dependen de los parámetros de los procesos MA(q) y AR(p), pero después del rezago p la PACF tiene un comportamiento como en el proceso MA(q).

2.3. Modelos de series de tiempo no estacionarios

Hasta el momento, sólo se han analizado modelos de series de tiempo con estacionariedad en covarianza. Sin embargo, muchos procesos generan series que no son estacionarias. La no-estacionariedad se puede deber a que la media μ_t o la varianza σ_t^2 de la serie cambian con el tiempo.

Los modelos que se discutirán en esta sección, son aquellos que provienen de procesos que no son estacionarios en media, los cuales a través de una transformación se vuelven estacionarios en el sentido débil.

2.3.1. No estacionariedad en media

Cuando la media del proceso cambia con el tiempo, se pueden generar dos tipos de modelos para describir esta dependencia temporal.

Si el proceso posee una tendencia que se cree que no cambiará en el tiempo, es decir, que es intrínseca del fenómeno y que perdurará en el futuro, se pueden considerar funciones deterministas para generar un modelo. Este enfoque consiste básicamente en generar modelos de regresión lineal. El tipo de función que se propone depende de la tendencia del fenómeno; se puede proponer dependencia lineal, polinomial, senoidal, entre otras. Los parámetros de las funciones se estiman a través de mínimos cuadrados ordinarios (sin duda una de las técnicas estadísticas más usadas).

Existen muchas series temporales donde no es evidente que el proceso seguirá una tendencia determinista que no cambiará en el tiempo. Por ejemplo puede darse el caso de que una serie presente una tendencia creciente en ciertos intervalos temporales pero una tendencia decreciente en otros, por lo que ajustar una función determinista no es posible. En estas situaciones se dice que el proceso posee una ***tendencia estocástica***. Como ejemplo se tiene una caminata aleatoria dada por la siguiente ecuación:

$$Y_t = Y_{t-1} + e_t \tag{2.75}$$

Se pueden hacer simulaciones de este proceso y se observaría la tendencia estocástica, pues el nivel de la media a cada tiempo está dado por:

$$\mu_t = Y_{t-1} \tag{2.76}$$

Es posible modelar la tendencia estocástica, siempre y cuando el proceso posea una no estacionariedad homogénea la cual de acuerdo a Box y Jenkins [4] es aquella que se presenta cuando las diferentes partes de las series se comportan de forma muy similar excepto por su diferencia en los niveles locales de la media. El término homogéneo hace referencia a que el comportamiento local es igual sin importar el nivel, por lo que si $\Psi(B)$ representa el operador que describe este comportamiento homogéneo, quiere decir que se puede aplicar en la serie y debe obtenerse el mismo valor sin importar el nivel de la media, esto es:

$$\Psi(B)(Y_t + C) = \Psi(B)(Y_t) \tag{2.77}$$

Para cualquier constante C. La ecuación anterior implica que el operador $\Psi(B)$ debe ser de la forma:

$$\Psi(B) = \phi(B)(1 - B)^d \tag{2.78}$$

para algún d entero mayor a cero, donde $\phi(B)$ es el operador autoregresivo. De esta forma, un proceso no estacionario homogéneo puede ser reducido a una serie estacionaria aplicando la respectiva diferencia en la serie; es decir el proceso Y_t es no estacionario homogéneo, pero la serie con diferencia $W_t = (1 - B)^d Y_t$ para alguna $d \geq 1$ es estacionaria.

2.3.2. Procesos autoregresivos integrados de promedios móviles

Una serie de tiempo que no es estacionaria en el sentido homogéneo puede ser reducida a una serie estacionaria a través de una transformación de diferencia. Una serie Y_t sigue un proceso autoregresivo integrado de promedios móviles (ARIMA[VIII]por sus siglas en inglés) si la diferencia d de la serie $\{W_t = (1 - B)^d Y_t\}$ es un proceso estacionario ARMA. Si la serie estacionaria W_t está descrita por un modelo ARMA(p, q), entonces se dice que Y_t es un proceso ARIMA(p, d, q). En notación de polinomios de retraso el modelo ARIMA(p, d, q) es:

$$\phi_p(1 - B)^d Y_t = \theta_q(B)e_t \tag{2.79}$$

donde $\phi_p(B)$ es el operador AR de la estacionalidad y $\theta_q(B)$ el operador MA de la invertibilidad, con raíces diferentes. Si el proceso no contiene términos autoregresivos ($p = 0$), el modelo se conoce como integrado de promedios móviles y se denota como IMA(d, q). Por otro lado, sino

[VIII] *Autoregressive Integrated Moving Average Models*

contiene términos de promedios móviles ($q = 0$) el modelo se llama autoregresivo integrado y se abrevia como ARI(p, d).

Existe una forma alternativa de escribir el modelo ARIMA(p, d, q) que es bastante útil en cuestiones de predicción, ya que facilita los cálculos de los pronósticos[IX]. Se ejemplificará con el caso más sencillo $d = 1$, es decir, con un modelo ARIMA$(p, 1, q)$. En este caso $\{W_t = (1 - B)Y_t = Y_t - Y_{t-1}\}$ sigue un proceso ARMA(p, q), por lo que se puede escribir como:

$$W_t = \phi_1 W_{t-1} + \cdots + \phi_p W_{t-p} + e_t - \theta_1 e_{t-1} - ... - \theta_q e_{t-q} \tag{2.80}$$

Al sustituir los valores de W_t con los datos de la serie Y_t:

$$Y_t - Y_{t-1} = \phi_1(Y_{t-1} - Y_{t-2}) + \cdots + \phi_p(Y_{t-p} - Y_{t-p-1}) + e_t - \theta_1 e_{t-1} - ... - \theta_q e_{t-q} \tag{2.81}$$

Al reescribir la serie se obtiene:

$$Y_t = (1-\phi_1)Y_{t-1} + (\phi_2 - \phi_1)Y_{t-2} + \cdots + (\phi_p - \phi_{p-1})Y_{t-p} - Y_{t-p-1} + e_t - \theta_1 e_{t-1} - ... - \theta_q e_{t-q} \tag{2.82}$$

La última expresión se conoce como la ecuación en diferencias del modelo, y aunque aparenta ser un modelo ARMA$(p + 1, q)$, el polinomio característico tiene una raíz igual a uno, por lo que el modelo no es estacionario, como efectivamente se sabía de antemano.

2.3.3. Términos constantes en los modelos ARIMA

Se ha manejado hasta el momento que el modelo estacionario ARMA oscila alrededor del cero, es decir, que tiene media igual a cero; lo anterior se debe a la representación del modelo lineal general (ecuación 2.28). Sin embargo, se podría dar el caso que el proceso tenga una media $\mu \neq 0$, y que la representación del proceso lineal general sea:

$$Y_t = \mu + e_t + \psi_1 e_{t-1} + \psi_2 e_{t-2} + \ldots \tag{2.83}$$

Si se tiene un proceso ARIMA(p, d, q), se sabe que la serie $\{W_t = (1 - B)^d Y_t\}$ sigue un proceso estacionario ARMA(p, q). Incorporar la media distinta de cero a la serie estacionaria $\{W_t\}$, se puede hacer de dos formas, que al final resultan equivalentes. La primera es restar la media en todos los términos autoregresivos:

$$W_t - \mu = \phi_1(W_{t-1} - \mu) + \cdots + \phi_p(W_{t-p} - \mu) + e_t - \theta_1 e_{t-1} - ... - \theta_q e_{t-q} \tag{2.84}$$

La otra forma es considerar un término constante θ_0 al modelo:

$$W_t = \theta_0 + \phi_1 W_{t-1} + \cdots + \phi_p W_{t-p} + e_t - \theta_1 e_{t-1} - ... - \theta_q e_{t-q} \tag{2.85}$$

La relación entre las dos expresiones está dada por:

$$\theta_0 = \mu(1 - \phi_1 - ... - \phi_p) \tag{2.86}$$

Se puede escoger entre una u otra representación dependiendo del tipo de parametrización que se desee implementar.

[IX] La parte predictiva de los modelos ARIMA se verá en el siguiente capítulo.

El término constante θ_0 en la serie estacionaria $\{W_t\}$, se traduce en un polinomio determinista de grado d, y el proceso se puede escribir como $Y_t = Y_t' + \mu_t$, donde Y_t' es el modelo ARIMA(p, d, q) con media cero, y μ_t es el término determinista. Para ejemplificar esto considérese el proceso IMA(1,1) con término constante:

$$Y_t = \theta_0 + Y_{t-1} + e_t - \theta_1 e_{t-1} \tag{2.87}$$

Iterando hacia el pasado, hasta el principio de la serie se llega a:

$$Y_t = t\theta_0 + e_t - (1-\theta_1)e_{t-1} - (1-\theta_1)e_{t-2} - ... - (1-\theta_1)e_1 - \theta_1 e_0 \tag{2.88}$$

donde se observa el término $t\theta_0$ determinista lineal respecto a t, donde θ_0 es la pendiente.

2.3.4. Estacionariedad respecto a una tendencia determinista

Como se establece en la sección 2.3.1, la no estacionariedad en media también se puede modelar a través de funciones deterministas que presuponen en esencia un análisis de regresión lineal. Sin embargo, es pertinente señalar que en algunas ocasiones los términos de error en la regresión pueden estar correlacionados. En este tipo de situaciones es conveniente restar la tendencia ajustada de la serie original, para modelar los residuales correlacionados a través de procesos estacionarios ARMA.

En la mayoría de los casos, la dependencia que se presupone seguirá el fenómeno en cuestión es lineal, lo que lleva al siguiente tipo de modelos:

$$Y_t = \beta_0 + \beta_1 t + r_t \tag{2.89}$$

donde β_0 y β_1 son los parámetros de la tendencia lineal y r_t es un proceso estacionario ARMA.

Es importante determinar si un proceso posee una tendencia determinista o si se tiene un proceso con tendencia estocástica que sea no estacionario homogéneo. Existen diversas técnicas para lograr identificar estas diferencias en los procesos generadores; en el siguiente capítulo se mencionan unas pruebas estadísticas que se usan con este fin.

2.3.5. No estacionaridad en varianza y covarianza

Como se describió con anterioridad, un proceso que es no estacionario homogéneo, se puede reducir a un proceso estacionario aplicando las respectivas diferencias. Sin embargo, muchas series no estacionarias, no son homogéneas. La no homogeneidad se debe a que la varianza y covarianza pueden depender del tiempo. Para reducir esta dependencia, se pueden usar transformaciones que estabilizan la varianza.

Es común que la varianza de un proceso no estacionario cambie cuando el nivel de la serie se modifique, es decir, que la varianza siga la siguiente relación funcional:

$$Var(Y_t) = cf(\mu_t) \tag{2.90}$$

Para encontrar una función H que estabiliza la varianza, se realiza lo siguiente. Primero, se aproxima H usando Taylor, alrededor del punto μ_t:

$$H(Y_t) \cong H(\mu_t) + H'(\mu_t)(Y_t - \mu_t) \tag{2.91}$$

Al tomar la varianza de la expresión anterior se obtiene:

$$Var[H(Y_t)] \cong [H'(\mu_t)]^2 Var(Y_t) = c[H'(\mu_t)]^2 f(\mu_t) \tag{2.92}$$

donde se ha usado la relación 2.90 y el hecho de que $H(\mu_t)$ y $H'(\mu_t)$ son términos constantes pues se han evaluado en el punto μ_t.

De esta forma, para que la transformación estabilice la varianza, es decir, que se cumpla que $Var[H(Y_t)] = c$, la derivada de $H(Y_t)$ se escoge como:

$$H'(\mu_t) = \frac{1}{\sqrt{f(\mu_t)}} \tag{2.93}$$

Lo que implica que la transformación debe ser:

$$H(\mu_t) = \int \frac{1}{\sqrt{f(\mu_t)}} d\mu_t \tag{2.94}$$

Dependiendo dye cómo sea la forma funcional respecto a la media $f(\mu_t)$, la transformación se obtiene empleado la ecuación anterior. Si por ejemplo, la varianza es proporcional al cuadrado del nivel entonces la transformación necesaria es logarítmica:

$$H(\mu_t) = log(Y_t) \quad si \quad Var(Y_t) = c\mu_t^2 \tag{2.95}$$

Se podría dar el caso también que la varianza pueda ser proporcional al nivel de la serie, por lo que la función que estabiliza la varianza es la raíz cuadrada:

$$H(\mu_t) = \sqrt{Y_t} \quad si \quad Var(Y_t) = c\mu_t \tag{2.96}$$

De forma más general, se puede incluir una familia de funciones conocidas como transformaciones de potencia, introducidas por Box y Cox[**?**] [W. Esta clase de funciones, contienen a las transformaciones anteriores como casos particulares. Para un valor dado del parámetro λ, la trasformación se define como:

$$g(Y_t) = \begin{cases} \frac{Y_t^\lambda - 1}{\lambda} & para\ \lambda \neq 0 \\ Log(Y_t) & para\ \lambda = 0 \end{cases} \tag{2.97}$$

Un ventaja de las transformaciones de potencia, es que se puede tratar a λ como un parámetro y estimar su valor con los datos. Por ejemplo, se puede incluir a λ como un parámetro, y se escoge como aquel valor que minimiza el error cuadrático medio de los residuales.

2.4. Modelos de series de tiempo estacionales

Existen series temporales que presentan un comportamiento que se repite después de un periodo regular de tiempo; reciben el nombre de series de tiempo estacionales y se pueden encontrar en el ámbito científico, económico, por mencionar algunos. El comportamiento estacional en este tipo de procesos se puede deber a diversos factores; un ejemplo en particular es el clima que favorece el turismo en ciertas épocas del año lo cual repercute directamente en las áreas de ventas y negocios.

2.4.1. Modelos MA(Q) y AR(P) estacionales

La componente estacional se puede incluir en los procesos de promedios móviles. Sea s el periodo en el cual se repite el comportamiento. Se define el modelo de promedios móviles estacional MA(Q) de orden Q con periodo s como:

$$Y_t = e_t - \Theta_1 e_{t-s} - \Theta_2 e_{t-2s} - ... - \Theta_Q e_{t-Qs} \tag{2.98}$$

En este caso el polinomio característico estacional de promedios móviles está dado por:

$$\Theta_Q(B) = (1 - \Theta_1 B^s - \Theta_2 B^{2s} - ... - \Theta_Q B^{Qs}) \tag{2.99}$$

Por su construcción la serie es estacionaria y para que sea invertible las raíces del polinomio $\Theta_Q(B)$ deben estar fuera del círculo unitario.

Es útil notar que el modelo estacional MA(Q) se puede ver como un caso especial de un modelo no estacional MA(q) cuyo orden q es igual a Qs, pero con los parámetros $\theta's$ igual a cero excepto en los rezagos estacionales $s, 2s, 3s, ..., Qs$. De esta forma, MA(Q) tiene un comportamiento que es similar al modelo de promedios móviles no estacional MA(q), sólo que la función de autocorrelación es distinta de cero en los rezagos $s, 2s, 3s..., Qs$. En analogía con la ecuación 2.47, la ACF es:

$$\rho_k = \frac{(-\Theta_k + \Theta_1\Theta_{k+1} + ... + \Theta_Q\Theta_{Q-k})}{(1 + \Theta_1^2 + \Theta_2^2 + ... + \Theta_Q^2)} \quad k = 1, 2, ..., Q \tag{2.100}$$

La estacionalidad también se puede incluir en los procesos autoregresivos. El proceso de autoregresivo estacional AR(P) de orden P con periodo s está dado por:

$$Y_t = \Phi_1 Y_{t-s} + \Phi_2 Y_{t-2s} + \cdots + \Phi_P Y_{t-Ps} + e_t \tag{2.101}$$

donde el polinomio de retrasos estacional autoregresivo es:

$$\Phi_P(B) = (1 - \Phi_1 B^s - \Phi_2 B^{2s} - ... - \Phi_P B^{Ps}) \tag{2.102}$$

Por el número finito de términos, el proceso autoregresivo estacional es invertible; para que sea estacionario, las raíces del polinomio $\Phi_P(B)$ deben estar fuera del círculo unitario. Como ocurrio con el proceso de promedios móviles estacional, AR(P) se puede ver como un caso especial del proceso no estacional autoregresivo AR(p) con onden $p = Ps$ y con coeficientes $\phi's$ no cero sólo en los rezagos $s, 2s, 3s..., Ps$. Por lo anterior, la función de autocorrelación tiene un decaimiento exponencial, con una posible mezcla de amortiguamiento senoidal, pero sólo en los rezagos estacionales.

2.4.2. Modelos ARMA estacionales multiplicativos

Raramente se necesitan modelos que incorporan solamente una dependencia estacional. Combinando los procesos estacionales con los no estacionales, se pueden generar modelos que incluyan una dependencia estacional, pero también una relación con los vecinos cercanos de la serie temporal.

Se define el modelo ARMA$(p, q)x(P, Q)_s$ estacional multiplicativo con periodo estacional s como:

$$\Phi_P(B)\phi_p(B)Y_t = \Theta_Q(B)\theta_q(B)e_t \tag{2.103}$$

donde $\Theta_Q(B)$ y $\Phi_P(B)$ son los polinomios estacionales (ecuaciones 2.99 y 2.102, respectivamente); $\theta_q(B)$ y $\phi_p(B)$ son los polinomios del proceso ARMA (sección 2.2). La ecuación anterior puede incluir al término θ_0 en el caso de que la media del proceso estacionario sea distinta de cero.

De igual forma que ocurrió con los procesos estacionales de promedios móviles y autoregresivos, el modelo ARMA$(p,q)x(P,Q)_s$ estacional multiplicativo se puede ver como un caso particular del modelo ARMA con orden MA igual a $q+Qs$ y orden AR igual a $p+Ps$.

2.4.3. Modelos no estacionarios con estacionalidad: ARIMA estacionales multiplicativos

Puede ocurrir el caso de que un modelo con estacionalidad no sea estacionario, por ejemplo cuando el proceso es casi periódico en el periodo estacional s. Un caso particular seria el promedio mensual de temperatura a través de los años, donde cada enero es aproximadamente el mismo, y así ocurre con los demás meses.

Como sucedió con los modelos ARIMA, una herramienta que ayuda para transformar la serie estacional no estacionaria a un proceso estacional estacionario es la diferencia estacional de orden D con periodo s, la cual se define como:

$$\nabla_s^D Y_t = (1-B^s)^D \tag{2.104}$$

Incorporando estas ideas, se dice que la serie Y_t sigue un modelo ARIMA$(p,d,q)x(P,D,Q)_s$ estacional multiplicativo con órdenes regulares p,d y q, y órdenes estacionales P,D y Q, y con periodo estacional s, si la serie con diferencia $\{W_t = \nabla^d \nabla_s^D Y_t\}$ se comporta como un proceso ARMA$(p,q)x(P,Q)_s$ estacional multiplicativo con periodo estacional s. Usando la notación de polinomios de retraso se tiene el modelo:

$$\Phi_P(B)\phi_p(B)\nabla^d\nabla_s^D Y_t = \Theta_Q(B)\theta_q(B)e_t \tag{2.105}$$

Al igual que ocurrió con el modelo ARMA estacional multiplicativo (ecuación 2.103) se puede incluir un término θ_0 en caso de que la media del proceso que resulta de las diferencias regular y estacional sea distinta de cero.

Los modelos estacionales representan una clase amplia y flexible para describir fenómenos que poseen un comportamiento repetitivo después de cierto intervalo de tiempo. En la práctica se ha encontrado que muchas series pueden ser modeladas con estos procesos, y se ha encontrado empíricamente que los órdenes estacionales P,D y Q tienen buenos resultados con valores de a lo más dos unidades [11].

Capítulo 3

Construcción de los modelos ARIMA

Construir un modelo ARIMA para describir el comportamiento de un proceso temporal no es un trabajo fácil. Se requiere seguir una serie de pasos para obtener el modelo más adecuado; la metodología implementada sigue en términos generales aquella propuesta por P. Box y M. Jenkins en 1976, usada ampliamente y reconocida por muchos autores como el método de Box y Jenkins".

Lo primero que se tiene que hacer es la ***especificación o identificación*** preliminar de un modelo que pueda ser adecuado para explicar los datos observados. El modelo seleccionado en este punto es tentativo, pues será sometido a revisión y análisis, pudiendo cambiar durante el proceso de modelado. Es importante recalcar que el modelo debe cumplir con el principio de parsimonía: debe contener el mínimo número de parámetros que permitan describir el proceso temporal.

Una vez identificado, se tendrán que estimar sus parámetros a partir de los datos observados. El ***ajuste del modelo*** consiste en encontrar los mejores estimadores de los parámetros desconocidos. Las técnicas estadísticas de estimación pueden ser mínimos cuadrados o máxima verosimilitud.

Ya con el modelo ajustado, se procede a realizar un ***diagnóstico o chequeo*** el cual sirve para determinar su adecuación; se debe revisar que ajuste bien los datos y que cumpla todas las suposiciones.

Durante la etapa de diagnóstico o al final de ella, es posible que se identifiquen algunos puntos de la serie temporal que posean valores atípicos y que no sigan el comportamiento intrínseco del proceso. El origen de estos valores aberrantes puede ser diverso. Una posible fuente podría ser que ciertos eventos como por ejemplo factores climáticos o económicos que tengan una influencia directa en algunos valores de la serie de tiempo. Otra posibilidad sería el desacierto humano que se ve reflejado en errores de medición o copiado de la información. El ***análisis de intervención*** y el ***análisis de valores atípicos*** son técnicas que se implementan para poder solventar estas dificultades. Es plausible que los datos aberrantes enmascaren el comportamiento general del proceso, por lo que una vez identificados, es necesario regresar al principio del modelado para ratificar el modelo o identificar uno nuevo.

Una vez que ya no se aprecian deficiencias en el modelo y se han identificado todos los datos atípicos (si es que los hubiera), se procede a hacer el ***pronóstico o predicción*** del modelo. En el análisis de series temporales uno de los objetivos primordiales es poder pronosticar valores futuros; debido a esta importancia, las predicciones deben ser sometidas a un análisis, pues un modelo con predicciones pobres no sería adecuado. En dado caso que el poder predictivo del modelo sea bajo, se necesita regresar sobre los pasos del modelado para obtener un mejor modelo; si los pronósticos son apropiados, se ha completado el todo el ciclo, y se cuenta con una expresión matemática que sirve para describir y pronosticar un fenómeno temporal.

En este capítulo se describirá detalladamente cuál es la metodología de cada uno de los pasos del modelado para construir modelos ARIMA.

3.1. Identificación del modelo

La especificación del modelo es un punto crucial en el desarrollo de los modelos paramétricos autoregresivos integrados de promedios móviles. Como se vio en el capítulo anterior, las características y el comportamiento general de estos procesos están en términos de sus funciones de autocorrelación ρ_k y autocorrelación parcial ϕ_{kk}. Como en la práctica estas funciones se desconocen, se deben estimar de la serie temporal $Y_1, Y_2 \ldots, Y_n$ con las funciones muestrales de autocorrelación $\hat{\rho_k}$ y autocorrelación parcial $\hat{\phi}_{kk}$. Entonces, para lograr la identificación del modelo, el principal objetivo es ver los patrones de las funciones muestrales e identificarlos con los correspondientes comportamientos teóricos de los modelos ARIMA, y de esta forma determinar el orden y el número de parámetros del modelo. Además de lo anterior, la especificación del modelo también requiere saber si la serie temporal necesita algún tipo de trasformación, ya sea para estabilizar la varianza o calcular la diferencia, y determinar si se debe incluir el parámetro determinístico θ_0 cuando $d \geq 1$.

A continuación se describirán una serie de pasos que constituyen una metodología adecuada y útil para la identificación tentativa del modelo que describe la serie.

Paso 1.- *Graficar los datos y seleccionar (si es necesario) las transformaciones adecuadas.*

En el análisis de series temporales, graficar los datos siempre es el primer paso. De la gráfica se pueden apreciar muchos aspectos importantes como saber si tiene un comportamiento estacionario o no estacionario, ver si existe alguna tendencia, identificar si el proceso presenta estacionalidad, detectar datos atípicos o aberrantes, entre otras cosas.

Además, al graficar los datos se puede ver si la varianza del proceso cambia respecto al nivel de la media, por lo que puede ser un indicativo de una posible transformación de varianza, pudiéndose usar alguna de las transformaciones de potencia (sección 2.3.5).

Paso 2.- *Calcular y examinar las funciones muestrales de autocorrelación y autocorrelación parcial de los datos transformados para determinar si es que se necesita una transformación de diferencia para hacer que la serie sea estacionaria.*

Como se mencionó anteriormente, al graficar los datos se puede ver si tienen un comportamiento no estacionario. Sin embargo, otro indicativo es que $\hat{\rho}_k$ tenga un decaimiento lineal muy lento y

que $\hat{\phi}_{kk}$ tenga sólo un rezago con un valor grande.

Para remover el comportamiento no estacionario, se debe tomar la respectiva diferencia de las series $(1-B)^d Y_t$, con $d \geq 1$. En la mayoría de los casos, la serie se vuelve estacionaria con d igual a 1, o 2.

Paso 3.- *Computar y analizar las funciones muestrales de autocorrelación y autocorrelación parcial de la serie estacionaria para identificar el orden de y de los polinomios autoregresivos y de promedios móviles.*

Una vez en este paso, la serie es estacionaria débil, aunque es posible que no se haya requerido hacer ninguna transformación para que lo sea. Al examinar el comportamiento de $\hat{\rho}_k$ y de $\hat{\phi}_{kk}$ se pueden identificar patrones que son similares al comportamiento teórico de las funciones ACF y PACF de los modelos estacionarios ARMA.

La variación muestral y la correlación existente de $\hat{\rho}_k$ y $\hat{\phi}_{kk}$ puede disfrazar un poco el comportamiento teórico de ACF y PACF, por lo que en estos primeros pasos de especificación, se debe observar solamente las características primordiales de las funciones de correlación muestral, y dejar de un lado los detalles. En las siguientes etapas, el modelo tentativo se va enriqueciendo y puede ir mejorando.

3.1.1. Prueba de raíz unitaria

Aunque un decaimiento lineal de la función de autocorrelación muestral es un indicio fuerte de la serie posee un comportamiento no estacionario, es útil medir cuantitativamente la evidencia de esta conducta en el proceso generador de los datos. Además, resulta importante determinar si esta no estacionariedad es debía a una tendencia determinista o si el proceso cuenta con una raíz unitaria. En este sentido Dickey y Fuller[12] propusieron una prueba para identificar si hay un raíz unitaria en un proceso estacionario. Para verificar la existencia de la raíz en un modelo $AR(p)$, se puede realizar la siguiente prueba de hipótesis: $H_0 : \pi = 0$ contra $H_a : \pi < 0$ usando la regresión:

$$\Delta Y_t = c_t + \pi Y_{t-1} + \sum_{i=1}^{p-1} \phi_i \Delta Y_{t-i} + e_t \tag{3.1}$$

Donde c_t es una función determinista, que por lo general se toma como cero, constante o lineal $\beta_0 + \beta 1t$, y e_t es un proceso de ruido blanco. El estadístico Aumentado de Dickey y Fuller (ADFI) [I] es el estadístico t del coeficiente estimado $\hat{\pi}$ empleado mínimos cuadrados:

$$\text{estadístico } - \ ADF = \frac{\hat{\pi}}{\text{error estándar } (\hat{\pi})} \tag{3.2}$$

Sin embargo, la distribución del estadístico-ADF, no es la distribución t bajo la hipótesis nula, sino que sigue una distribución asintótica no convencional que está en función de procesos de

[I] Augmented Dickey Fuller

Wiener y depende de cómo está conformado el término determinista, ver Wie[7]. Porcentajes de esta distribución límite han sido tabulados, y se pueden encontrar en Fuller[13].

En la práctica, aun después de la primera diferencia, puede ser que el proceso autoregresivo sea infinito, pero puede ser aproximado por un AR finito al aumentar el orden de las diferencias conforme se va incrementado el tamaño de la muestra. Said y Dickey[14] mostraron que al acrecentar el orden del proceso AR con el tamaño de muestra, el estadístico-ADF posee la misma distribución de muestra grande que en el caso que de la primera diferencia sea efectivamente un proceso autoregresivo finito.

3.2. Ajuste del modelo: estimación de los parámetros

Una vez que se ha identificado un modelo tentativo, el siguiente paso es estimar los parámetros que están involucrados en la descripción matemática del proceso usando los valores observados de la serie de tiempo $Y_1, Y_2, \ldots, Y_n$. Para realizar la estimación de los parámetros, se necesita que la serie sea estacionaria, por lo que si no lo es, se tiene que trabajar con la nueva serie que resulta de las respectivas transformaciones de diferencia que generan procesos estacionarios. Existen diferentes procedimientos de estimación, pero sólo se estudiaran los que resultan de emplean la metodología de máxima verosimilitud, pues sus estimadores poseen buenas propiedades de muestras grandes.

3.2.1. Estimación de máxima verosimilitud condicional

El modelo estacionario general ARMA(p, q) con media distinta de cero, está dado por la ecuación 2.84; tomando los términos $W_t - \mu$ como Y_t, el modelo se puede representar como:

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q} \tag{3.3}$$

Los términos de error $\{e_t\}$ son variables aleatorias independientes e idénticamente distribuidas $N(0, \sigma_e^2)$ por lo que la función de densidad conjunta está dada por:

$$P(\mathbf{e}|\phi, \theta, \mu, \sigma_e^2) = (2\pi\sigma_e^2)^{-n/2} exp[-\frac{1}{2\sigma_e^2}\sum_{i=1}^{n} e_t^2] \tag{3.4}$$

donde los términos en negritas significan cantidades vectoriales del ruido blanco y los parámetros: $\mathbf{e} = (e_1, e_2, \ldots, e_n)$, $\phi = (\phi_1, \phi_2, \ldots, \phi_n)$ y $\theta = (\theta_1, \theta_2, \ldots, \theta_n)$. Despejando de la ecuación 3.3 el término e_t, se puede escribir la función de máxima verosimilitud en función del conjunto de parámetros $L(\phi, \theta, \mu, \sigma_e^2)$. Como usualmente ocurre, se toma logaritmo de la función de máxima verosimilitud para que los cálculos sean más sencillos (lo cual no afecta la búsqueda del máximo pues el logaritmo es una función creciente):

$$ln[L(\phi, \theta, \mu, \sigma_e^2)] = -\frac{n}{2} ln(2\pi\sigma_e^2) - \frac{1}{2\sigma_e^2}\sum_{t=1}^{n} e_t^2(\phi, \theta, \mu) \tag{3.5}$$

Para poder evaluar la función anterior, a parte de los valores del proceso $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$, se necesitan una serie de condiciones iniciales $\mathbf{Y}_* = (Y_{1-p}, \ldots, Y_{-1}, Y_0)$ y $\mathbf{a}_* = (a_{1-q}, \ldots, a_{-1}, a_0)$; por lo anterior, la ecuación 3.5 se renombra como logaritmo de máxima verosimilitud condicional, y se reescribe como:

$$ln[L_*(\phi, \theta, \mu, \sigma_e^2)] = -\frac{n}{2} ln(2\pi\sigma_e^2) - \frac{1}{2\sigma_e^2} S_*(\phi, \theta, \mu) \tag{3.6}$$

donde el término $S_*(\phi, \theta, \mu)$, es la suma de mínimos cuadrados condicionales:

$$S_*(\phi, \theta, \mu) = \sum_{t=1}^{n} e_t^2(\phi, \theta, \mu | \mathbf{Y}, \mathbf{Y}_*, \mathbf{a}_*) \tag{3.7}$$

Las cantidades $\hat{\phi}, \hat{\theta}$ y $\hat{\mu}$ que maximizan la ecuación 3.6, se llaman ***estimadores de máxima verosimilitud condicionales***.

Para especificar las condiciones iniciales del vector $\mathbf{Y}_*$, se puede usar el hecho de que la serie es estacionaria y usar la media muestra $\hat{Y}$ para estimar esos valores. Cuando se tiene ruido blanco, el vector de condiciones iniciales se estima con la media del proceso que es igual a cero.

Como el logaritmo de la función de máxima verosimilitud condicional incluye a los datos sólo a través del término $S_*(\phi, \theta, \mu)$, estos estimadores son los mismos que los que se obtienen de minimizar la suma de mínimos cuadrados condicionales, que como se observa, no involucra la estimación de σ_e^2. La optimización de $S_*(\phi, \theta, \mu)$ generalmente involucra técnicas numéricas, pues se obtienen ecuaciones no lineales, aunque cuando no hay términos de promedios móviles, la minimización se puede hacer analíticamente.

Una vez obtenido los estimadores de los parámetros $\hat{\phi}, \hat{\theta}$ y $\hat{\mu}$, el estimador de σ_e^2 se calcula usando la siguiente relación:

$$\sigma_e^2 = \frac{S_*(\phi, \theta, \mu)}{d.f.} \tag{3.8}$$

Donde $d.f.$ es el número de grados de libertad que es igual al número de términos que intervienen en la suma de $S_*(\phi, \theta, \mu)$, menos el número de parámetros estimados.

3.2.2. Función de máxima verosimilitud exacta

La función de máxima verosimilitud derivada en la sección anterior, es una aproximación, pues los valores de las condiciones iniciales se tienen que aproximar. Se puede derivar la función de máxima verosimilitud exacta, aunque el procedimiento en términos generales es más complejo. A continuación se ilustrará el desarrollo de esta función para el modelo general AR(1) con media distinta de cero, el cual está dado por:

$$Y_t - \mu = \phi_1(Y_{t-1} - \mu) + e_t \tag{3.9}$$

donde como se sabe $\{e_t\}$ es un proceso de ruido blanco.

Se considera primero la función de densidad conjunta dada por la ecuación 3.4, adaptada al modelo AR(1), pero el vector $\boldsymbol{e}$ se toma desde $e_2, \ldots, e_n$:

$$P(\mathbf{e}|\phi_{\mathbf{1}}, \mu, \sigma_e^2) = (2\pi\sigma_e^2)^{-(n-1)/2} exp[-\frac{1}{2\sigma_e^2}\sum_{i=1}^{n} e_t^2] \tag{3.10}$$

Por otro lado se tiene la siguiente ecuación:

$$\begin{cases} Y_2 - \mu &= \phi_1(Y_1 - \mu) + e_2 \\ Y_3 - \mu &= \phi_1(Y_2 - \mu) + e_3 \\ & \vdots \\ Y_n - \mu &= \phi_1(Y_{n-1} - \mu) + e_n \end{cases} \tag{3.11}$$

Condicionando $Y_1 = y_1$, la relación anterior, define una transformación lineal entre $e_2, \ldots, e_n$ y $Y_2, \ldots, Y_n$. Entonces, la función de densidad conjunta de $Y_2, \ldots, Y_n$ dado $Y_1 = y_1$ se puede obtener al sustituir en la ecuación 3.10 los términos de los e's en función de las Y's dados por la transformación lineal de la ecuación 3.11:

$$P(y_2, \ldots, y_n|y_1) = (2\pi\sigma_e^2)^{-(n-1)/2} exp\Big[-\frac{1}{2\sigma_e^2}\sum_{i=2}^{n}((Y_t - \mu) - \phi_1(Y_{t-1} - \mu))^2\Big] \tag{3.12}$$

Ahora considérese la distribución de Y_1.La representación lineal general del proceso AR(1) con media distinta de cero es:

$$Y_t = \mu + e_t + \phi_1 e_{t-1} + \phi_1^2 e_{t-2} + \phi_1^3 e_{t-3} + \ldots \tag{3.13}$$

De la relación anterior se sigue que Y_1 tiene una distribución normal, con media μ y varianza $\sigma_e^2/(1-\phi^2)$. De esta forma, al multiplicar la función de densidad conjunta de $Y_2, \ldots, Y_n$ condicionada a $Y_1 = y_1$, por la función de Y_1, se obtiene la función de densidad conjunta de $Y_1, Y_2, \ldots, Y_n$. Como función de los parámetros ϕ_1, μ y σ_e^2 la función de máxima verosimilitud buscada es:

$$L(\phi_1, \mu, \sigma_e^2) = (2\pi\sigma_e^2)^{-\frac{n}{2}}(1-\phi_1)^{1/2} exp\Big[-\frac{1}{2\sigma_e^2} S(\phi_1, \mu)\Big] \tag{3.14}$$

donde el término $S(\phi_1, \mu)$ se conoce como la suma de mínimos cuadrados no condicionales:

$$S(\phi_1, \mu) = \sum_{i=2}^{n}((Y_t - \mu) - \phi_1(Y_{t-1} - \mu))^2 + (1-\phi_1)(Y_1 - \mu)^2 \tag{3.15}$$

Como se mencionó en la sección anterior, se emplea por lo regular el logaritmo de la función de máxima verosimilitud para encontrar el máximo. Al aplicarlo a la ecuación 3.14 se obtiene la función $l(\phi_1, \mu, \sigma_e^2)$:

$$l(\phi_1, \mu, \sigma_e^2) = ln(L(\phi_1, \mu, \sigma_e^2)) = -\frac{1}{n} ln(2\pi\sigma_e^2) + \frac{1}{2} ln(1 - \phi_1) - \frac{1}{2\sigma_e^2} S(\phi, \mu) \tag{3.16}$$

Como ocurrió con la función de máxima verosimilitud condicional, una vez obtenidos los estimadores $\hat{\phi_1}, \hat{\mu}$ se puede estimar σ_e^2 usando una relación parecida a la ecuación 3.8, sólo que usando la suma de mínimos cuadrados no condicionales:

$$\hat{\sigma}_e^2 = \frac{S(\phi_1, \mu)}{d.f.} = \frac{S(\phi_1, \mu)}{n - 2} \tag{3.17}$$

En este caso los grados de libertad son $n - 2$, porque se estimaron dos parámetros.

La estimación de ϕ_1 y μ se puede realizar minimizando el término $S(\phi_1, \mu)$. Esta optimización se debe llevar a cabo numéricamente, pues el término $(1 - \phi_1)(Y_1 - \mu)^2$ hace que las ecuaciones resultantes de $\partial S/\partial\phi_1 = 0$ y $\partial S/\partial\mu = 0$ sean no lineales. Los estimadores resultantes se conocen como estimadores de mínimos cuadrados no condicionales.

La forma exacta de la función de máxima verosimilitud para el modelo general ARMA es complicada. Como referencia, Newbold[15] la derivó para el modelo general ARMA(p, q).

3.2.3. Propiedades de los estimadores de máxima verosimilitud

Las propiedades de muestra grande de los estimadores de máxima verosimilitud coinciden con las de los mínimos cuadrados (condicionados y no condicionados); en general estas propiedades se pueden derivar modificando un poco la teoría estándar de máxima verosimilitud. Detalles se pueden encontrar en Shumway y Stoffer[16]. Sea $\alpha = (\phi, \theta)$ el vector de los parámetros del modelo ARMA(p, q). En general se cumple que $\widehat{\alpha}$ (el estimador de α) tiene una distribución asintótica multivariada normal NM$\alpha, V(\hat{\alpha})$ donde $V(\hat{\alpha})$ es la matriz de varianza-covarianza de $\hat{\alpha}$. La estimación de la matriz $V(\hat{\alpha})$ está dada por:

$$\hat{V}(\hat{\alpha}) = \hat{\sigma}_{\hat{\alpha}_i \hat{\alpha}_j} \tag{3.18}$$

donde $\hat{\sigma}_{\hat{\alpha}_i \hat{\alpha}_j}$ es la covarianza muestra entre $\hat{\alpha}_i$ y $\hat{\alpha}_j$.

A continuación se mostrarán algunos resultados para los modelos ARMA más sencillos: AR(1), MA(1) y ARMA(1,1).

Conforme a lo anterior, para n grande, los estimadores son insesgados y tienen una distribución normal. Sus varianzas y covarianzas son aproximadas por:

$$AR(1) : Var(\hat{\phi_1}) \approx \frac{1 - \phi_1^2}{n} \tag{3.19}$$

$$MA(1) : Var(\hat{\theta}_1) \approx \frac{1 - \theta_1^2}{n} \tag{3.20}$$

$$ARMA(1,1): \begin{cases} Var(\hat{\phi}_1) & \approx [\frac{1-\phi_1^2}{n}][\frac{1-\phi_1\theta_1}{\phi_1-\theta_1}]^2 \\ Var(\hat{\theta}_1) & \approx [\frac{1-\theta_1^2}{n}][\frac{1-\phi_1\theta_1}{\phi_1-\theta_1}]^2 \\ Corr(\hat{\phi}_1, \hat{\theta}_1) & = \frac{\sqrt{(1-\phi_1^2)(1-\theta_1^2)}}{1-\phi_1\theta_1} \end{cases} \tag{3.21}$$

3.3. Diagnóstico del modelo

Una vez que se han estimado los parámetros del modelo, se necesita evaluar para determinar si es adecuado, y en dado caso que no lo sea, sugerir modificaciones apropiadas para que lo sea. Se presentarán dos enfoques complementarios: el análisis residual y el análisis de sobreajuste.

3.3.1. Análisis Residual

Una de las suposiciones básicas de los modelos estacionarios ARMA es que el proceso estocástico $\{e_t\}$ es ruido blanco: son variables aleatorias no correlacionadas, con media cero y varianza constante. Para cualquier modelo, los residuales $\hat{e}_t$'s son estimaciones de ese proceso de ruido blanco, por lo tanto se debe hacer un análisis detallado de esta serie de residuales para ver que se cumple la suposición. Para los modelos autoregresivos AR(p) los residuales se definen como:

$$\hat{e}_t = Y_t - \hat{\phi}_1 Y_{t-1} + \hat{\phi}_2 Y_{t-2} + \cdots + \hat{\phi}_p Y_{t-p} \tag{3.22}$$

En el caso general de los modelos estacionarios ARMA que contienen términos de promedios móviles, se invierte el proceso generándose una serie autoregresiva infinita (ecuación 2.63)):

$$\hat{e}_t = Y_t - \hat{\pi}_1 Y_{t-1} + \hat{\pi}_2 Y_{t-2} + \hat{\pi}_3 Y_{t-3} + \ldots \tag{3.23}$$

Las π's que aparecen en la ecuación anterior se estiman de forma indirecta con las estimaciones de las ϕ's y θ's que intervienen en el modelo.

En las definiciones anteriores (ecuaciones 3.22 y 3.23) se han considerado modelos donde la media del proceso estacionario es igual a cero. En el caso de que exista una $\mu \neq 0$, se incluye en las definiciones de los residuales sustituyendo los términos Y_t con $Y_t - \mu$ y realizando los respectivos desarrollos.

Normalidad de los residuales

Para verificar la normalidad de los residuales se realizan varias pruebas. La primera es graficar los ***residuales contra el tiempo***. Si el modelo es adecuado, se esperaría que los residuales estén distribuidos de forma aleatoria alrededor del cero, sin mostrar ningún patrón o tendencia; en dado caso que existiera algún tipo de comportamiento regular, esto sería un motivo para suponer que los residuales no tienen una distribución normal.

Para indagar más sobre si la distribución de los residuales es normal, se pueden hacer ***histogramas*** que permitan dar una idea de la función de densidad. Como complemento, se pueden

generar gráficas de los ***cuantiles de los residuales*** contra los cuantiles teóricos de una normal, y ver si los puntos siguen una línea recta, lo que hablaría de un comportamiento gaussiano por parte de los residuales. Además de lo anterior, se pueden realizar algunas ***pruebas estadísticas*** formales para verificar la normalidad; destacan las prueba de Shapiro-Wilk y la de Jarque-Bera. En estas pruebas de bondad de ajuste, parten de la hipótesis nula de la que distribución de los residuales sigue un comportamiento normal.

Correlación de los residuales

Una forma de verificar la correlación de los residuales es calcular su función de autocorrelación muestral $\hat{\rho}_{\hat{e}_k}$ [II]. Si efectivamente los residuales fueran un proceso de ruido blanco, para n grande $\hat{\rho}_{\hat{e}_k}$ se distribuiría de forma normal con media cero. Al usar la ecuación 2.22, su varianza se aproxima como

$$Var(\hat{\rho}_{\hat{e}_k}) \approx \frac{1}{n} \tag{3.24}$$

Además por la ecuación 2.21, las autocorrelaciones muestrales son aproximadamente cero:

$$Corr(\hat{\rho}_{\hat{e}_k}, \hat{\rho}_{\hat{e}_j}) \approx 0 \tag{3.25}$$

Desafortunadamente, los residuales de modelos correctamente especificados se comportan de forma un poco distinta. En general, para n grande, las correlaciones muestrales $\hat{\rho}_{\hat{e}_k}$ tienen una distribución normal con media cero; sin embargo para valores de k pequeños, la varianza generalmente es menor a $1/n$ y para valores de j cercanos a estas k, los valores de $\hat{\rho}_{\hat{e}_k}$ y $\hat{\rho}_{\hat{e}_j}$ están altamente correlacionados. Cuando el valor de k es grande, se cumplen las aproximaciones de las ecuaciones 3.24 y 3.25. Como un ejemplo de estos resultados, se tienen las aproximaciones de las varianzas de $\hat{\rho}_{\hat{e}_k}$ para un modelo AR(1) correctamente especificado:

$$Var(\hat{\rho}_{\hat{e}_k}) \approx \frac{\phi_1^2}{n} \tag{3.26}$$

$$Var(\hat{\rho}_{\hat{e}_k}) \approx \frac{1-(1-\phi_1^2)\phi_1^{2(k-1)}}{n} \quad para \quad k>1 \tag{3.27}$$

Resultados para modelos estacionarios más generales ARMA se pueden encontrar en Box y Pierce[17].

Correlación de los residuales en conjunto

Además de analizar la correlación de los residuales en rezagos individuales, es conveniente tener una prueba que tome en cuenta su relación pero como conjunto. Por ejemplo se puede dar el caso de que varios de los valores de la función de autocorrelación muestral de los residuales estén por debajo pero muy cercanos al límite, pero al tomarlos todos en conjunto puede ser excesivo. Para tomar en cuenta esta posibilidad Box y Pierce[17] propusieron el siguiente estadístico:

$$Q = n(\hat{\rho}_{\hat{e}_1}^2 + \hat{\rho}_{\hat{e}_2}^2 + \cdots + \hat{\rho}_{\hat{e}_k}^2) \tag{3.28}$$

Mostraron que si el modelo ARMA(p,q) estaba correctamente especificado, entonces para n grande, el estadístico Q debería tener una distribución muestral aproximada chi-cuadrada con

[II] El subíndice $\hat{e}_k$ se emplea para identificar que es la correlación del rezago k pero de los residuales $\hat{e}$.

$K - p - q$ grados de libertad. El valor de K se selecciona en cierto sentido de forma un poco arbitraria, pero tan grande como para que los valores de los pesos Ψ's (escrito el modelo como un proceso lineal general) sean despreciables para valores de $j > K$. El ajustar un modelo erróneo tiende a inflar el valor de Q, por lo que se puede hacer una prueba estadística que rechace el modelo especificado ARMA(p, q) si el valor del estadístico supera cierto valor crítico de la distribución chi-cuadrada.

La distribución muestra de Q está basada en un resultado asintótico. Ljung y Box[18] se dieron cuenta que para valores de $n = 100$ la aproximación de la distribución chi- cuadrada no resulta satisfactoria. Por esto, modificaron el estadístico ligeramente para que su aproximación por una distribución chi-cuadrada fuera plausible con valores típicos de distribuciones muestrales. Propusieron de esta forma, en 1978, el estadístico de Ljung-Box:

$$Q_* = n(n+2)\left(\frac{\hat{\rho}^2_{\hat{e}_1}}{n-1} + \frac{\hat{\rho}^2_{\hat{e}_2}}{n-2} + \cdots + \frac{\hat{\rho}^2_{\hat{e}_K}}{n-K}\right) \tag{3.29}$$

Hay que notar que $(n+2)/(n-k) > 0$,para cada $k \geq 1$,lo que hace que $Q_* > Q$,lo que explica en cierto modo porque el estadístico tendía a pasar por alto modelos inadecuados.

3.3.2. Análisis de Sobreajuste

El sobreajuste es una herramienta complementaria al análisis residual que proporciona más elementos para discernir si el modelo es adecuado. Una vez que el modelo especificado ha pasado las pruebas del análisis residual, la técnica consiste en ajustar un modelo ligeramente más general, es decir, uno que contenga al original como caso particular. Se confirma la adecuación del modelo inicial si ocurre que:

- La estimación del **parámetro adicional** no es significativamente diferente del cero
- Las nuevas estimaciones de los **parámetros originales**, no cambian significativamente de su estimación inicial.

Al generalizar los modelos ARMA para hacer el sobreajuste, se debe evitar la posible **redundancia de los parámetros**. Esto significa que si el modelo especificado es un ARMA(p, q), entonces un modelo que también podría ser adecuado sería un ARMA$(p+1, q+1)$, siempre y cuando los polinomios de retrasos, tengan un factor en común. Esto es, si $\phi_p(B)Y_t = \theta_q(B)e_t$ es un modelo adecuado, entonces el modelo $(1-cB)\phi_p(B)Y_t = (1-cB)\theta_q(B)Y_t$ podría también serlo para cualquier constante c. Para tener una sola parametrización, se deben cancelar los factores comunes en los polinomios característicos AR y MA.

$$(1-cB)\phi_p(B)Y_t = \phi_{p+1}(B)Y_t = (1-cB)\theta_q(B)e_t = \theta_{q+1}(B)e_t \tag{3.30}$$

Para evitar la redundancia de los parámetros se sugiere lo siguiente:

- No incrementar los órdenes de las partes autoregresivas y de promedios móviles de forma simultánea.

- Extender el modelo en las direcciones sugeridas por el análisis residual.

3.3.3. Criterio de selección del modelo

Puede ser que al final de la etapa de diagnóstico existan dos o más modelos que sean buenas aproximaciones para describir el comportamiento del fenómeno estocástico. En estas instancias, se puede hacer uso del principio de parsimonia para elegir un modelo adecuado. Sin embargo, en algunas ocasiones no será suficiente esta regla y será necesario el uso de algún criterio para seleccionar el modelo más adecuado. En esta sección se introduce el ***criterio de información de Akaike***, conocido como AIC por sus siglas en inglés[III]. Este criterio establece que se debe seleccionar el modelo que minimiza:

$$AIC(k) = -2ln(\text{función de máxima verosimilitud}) + 2k$$

donde en el contexto de los modelos ARIMA, $k = p+q$ o $p+q+1$ si existe un término constante θ_0. El término k, sirve como una función de penalización para asegurarse que se selecciona el modelo con menor número de parámetros.

El valor AIC es en el fondo un estimador de la divergencia promedio de Kullback-Leibler del modelo estimado respecto al verdadero. Esta divergencia se interpreta como una medida (no simétrica) de la diferencia de dos distribuciones de probabilidad.

Sea $p(y_1, y_2, \ldots, y_n)$ la función de densidad de $Y_1, Y_2, \ldots, Y_n$ y $q_\theta(y_1, y_2, \ldots, y_n)$ la función de densidad para el modelo con parámetro θ. La divergencia de Kullback-Leibler de q_θ respecto de p, está dada por:

$$D(p, q_{\theta_0}) = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} p(y_1, y_2, \ldots, y_n) ln(\frac{p(y_1, y_2, \ldots, y_n)}{q_\theta(y_1, y_2, \ldots, y_n)}) dy_1 \ldots dy_n \tag{3.31}$$

La cantidad AIC estima $E[D(p, q_\theta)]$, donde $\hat{\theta}$ es el estimador de máxima verosimilitud del vector de parámetros θ. AIC es un estimador sesgado, pero Hurvich y Tsai [19] mostraron que el sesgo se puede remover agregando un término no estocástico al cálculo del AIC:

$$AIC_c = AIC + \frac{2(k+1)(k+2)}{n-k-2} \tag{3.32}$$

El subíndice c indica que es el criterio Akaike corregido. En la fórmula anterior, n es el tamaño de la muestra y k es el número de parámetros estimados. Hurvich y Tsai sugirieron que cuando $k/n > 10\,\%$, la cantidad AIC_c supera al valor sin la corrección AIC como método de selección. Existe otro criterio de elección parecido al AIC, el cual fue propuesto por Schwartz[20], y ha sido llamado ***criterio de información Bayesiana de Schwartz***, SBIC por sus siglas en inglés [IV], el cual se determina que se debe seleccionar el modelo que minimiza la siguiente cantidad:

[III] *Akaike Information Criteria.*

[IV] *Schwartz Bayesian Information Criterion.* Comúnmente en la literatura, se denota a este criterio simplemente como BIC.

$$SBIC(k) = -2ln(\text{función de máxima verosimilitud}) + kln(n) \tag{3.33}$$

Si el proceso sigue un modelo ARMA(p, q) el orden de los parámetros especificados al minimizar SBIC es consistente, es decir, se aproxima a los valores verdaderos al incrementarse el tamaño de la muestra.

Para poder aplicar los criterios de selección antes mencionados, es un hecho que se debe calcular la función de máxima verosimilitud; sin embargo la estimación de los modelos ARMA empleando este enfoque es propensa a errores debido a multimodalidad de la función de verosimilitud y a los problemas de sobreajuste cuando los verdaderos órdenes AR o MA son sobrepasados. Hannan y Rissanen[21] propusieron un método para evitar estos problemas, el cual genera una aproximación del SBIC (de aquel que se puede obtener empleando máxima verosimilitud), demostrando que la minimización de esta aproximación también lleva a una identificación consistente de los órdenes del proceso ARMA. Su método consiste en ajustar primero un proceso autoregresivo, cuyo orden es determinado minimizando el criterio AIC. Después se usan los residuales como aproximaciones de los términos de error. De esta manera, el modelo ARMA puede ser estimado a través de una regresión lineal de los rezagos del proceso AR ajustado y los términos de error, generando de esta manera una aproximación del SBIC.

No solamente es importante determinar el orden del proceso estacionario ARMA, sino en algunas ocasiones es trascendente saber el subconjunto de parámetros del modelo establecido que son relevantes y significativamente distintos de cero. En este sentido, se puede emplear el método propuesto por Hannan y Rissanen combinado con la regresión de pasos agigantados [V] de Furnival y Wilson[22] para hallar el subconjunto óptimo.

3.4. Análisis de intervención y análisis de datos atípicos

Como se mencionó al principio del capítulo, la serie temporal puede contener algunos datos que se salen del comportamiento general que se deben a eventos externos al proceso. Estos acontecimientos pueden ser originados por factores humanos como por ejemplo huelgas, promociones de ventas, vacaciones, entre otros, o factores naturales como cambios climáticos.

Cuando se conoce el tiempo en el que se dan estos eventos, se cuantifica su efecto sobre la serie a través de lo que se conoce como análisis de intervención. Cuando se encuentran datos aberrantes y no está claro que evento o suceso externo lo originó, se debe implementar el análisis de datos atípicos. Hay que destacar que en algunas ocasiones estos valores aberrantes pueden ser producto de desaciertos humanos como podrían ser errores en el copiado, transcripción o cálculo de cierta información.

En el modelado es necesario combinar las dos técnicas, pues a veces no se conoce de antemano el evento externo que modificó la serie por lo que primero se implementa el análisis de datos atípicos; después se puede investigar y si se encuentra la naturaleza de la perturbación, se puede

[V] La regresión de pasos agigantados (regresión by leaps and bounds) es un algoritmo que busca encontrar el subconjunto óptimo en una regresión lineal.

aplicar el análisis de intervención lo que propicia un mejor entendimiento de los factores que afectan el proceso, lo que se traduce en un modelo más completo.

3.4.1. Análisis de intervención

Los sucesos externos que pueden afectar el comportamiento general del proceso reciben el nombre de ***intervenciones***. Al hablar de intervenciones, se sabe de antemano cuál es la naturaleza del fenómeno que modifica la serie, por lo que también se conoce el tiempo de ocurrencia, el cual se denota con la variable T. Los eventos externos que se van a considerar son aquellos que modifican sólo el comportamiento de la media del proceso. También existen diferentes técnicas que se pueden aplicar cuando la varianza del proceso cambia pero no se discutirán aquí; lectores interesados podrían consultar Abraham y Wei[23].

La importancia del análisis de intervención es cuantificar qué tanto el evento externo modifica el comportamiento medio de la serie. Una prueba de hipótesis usando el estadístico t, podría parecer natural, sin embargo la correlación de los datos impide su aplicabilidad. Es por ello que Box y Tiao[24] desarrollaron la técnica para medir los cambios estructurales del proceso debido a las intervenciones.

Sea m_t la función que representa el cambio en la función valor esperado debido a una intervención, que se conoce como ***respuesta de la intervención***. Una serie $\{Y_t\}$ modificada por un evento externo se puede representar como:

$$Y_t = m_t + X_t \tag{3.34}$$

donde X_t representa una serie donde no hay intervención. Por lo regular X_t es referido como proceso subyacente natural o no perturbado. El proceso sin intervención se debe de identificar, estimar y diagnosticar antes de generar el modelo de intervención, empleando los valores de la serie antes de la fecha del suceso externo $\{Y_t : t < T\}$, los cuales se refieren como ***datos pre-intervenidos***. Una vez identificado el modelo de intervención se puede escribir tomando en cuenta el proceso de ruido blanco $\{e_t\}$ y los correspondientes polinomios autoregresivos $\phi_p(B)$ y de promedios móviles $\theta_q(B)$:

$$Y_t = m_t + \frac{\theta_q(B)}{phi_p(B)(1-B)^d} e_t \tag{3.35}$$

Para cuantificar m_t generalmente se emplean dos variables que se conocen como ***variables de intervención***. Una de ellas se debe a una intervención que ocurre en el tiempo T con un efecto que prevalece tiempos posteriores. Esto se representa a través de una ***función escalón***:

$$S_t^{(T)} = \begin{cases} 0 & t < T \\ 1 & t > T \end{cases} \tag{3.36}$$

La otra variable está relacionada con un evento externo que ocurre sólo en el periodo de tiempo T, lo cual se puede simbolizar empleando la ***función pulso***:

$$P_t^{(T)} = \begin{cases} 1 & t = T \\ 0 & t \neq T \end{cases} \tag{3.37}$$

Hay que notar que la función pulso se puede generar al diferenciar la función escalón, esto es:

$$P_t^{(T)} = S_t^{(T)} - S_{t-1}^{(T)} = (1 - B)S_t^{(T)} \tag{3.38}$$

Un modelo de intervención se puede generar usando estas dos funciones dependiendo de la forma de la intervención. Existen muchas formas de representar m_t, a continuación se ilustrarán las más comunes.

1. Un efecto producido por una intervención que se manifiesta b periodos de tiempo después del evento (donde b puede ser igual a cero).
   - Si el efecto es permanente:
   $$m_t = \omega B^b S_t^{(T)} \tag{3.39}$$
   - Si el impacto es sólo en un periodo:
   $$m_t = \omega B^b P_t^{(T)} \tag{3.40}$$
2. Un impacto generado por un evento externo en b periodos de tiempo después con una respuesta gradual.
   - Si el valor total del efecto se alcanza paulatinamente:
   $$m_t = \frac{\omega B^b}{(1 - \delta B)} S_t^{(T)} \tag{3.41}$$
   - Si el valor del efecto decrece progresivamente a sus niveles normales:
   $$m_t = \frac{\omega B^b}{(1 - \delta B)} P_t^{(T)} \tag{3.42}$$

   en ambos casos el valor de δ está entre cero y uno: $0 < \delta < 1$.

Hay que destacar que la respuesta de intervención también se puede escribir como una combinación de efectos escalón y efectos pulso como se ilustra a continuación:

$$m_t = \frac{\omega_0 B^b}{(1 - \delta B)} P_t^{(T)} + \omega_1 B S_t^{(T)} \tag{3.43}$$

Por la ecuación 3.38 la relación anterior se reescribe como:

$$m_t = \left[\frac{\omega_0 B^b}{(1 - \delta B)} + \frac{\omega_1 B^b}{(1 - B)} \right] P_t^{(T)} \tag{3.44}$$

La respuesta anterior puede representar un fenómeno de intervención que produce una respuesta inmediata $\omega_0 + \omega_1$ que decrece gradualmente hasta dejar un efecto permanente ω_1 en el sistema.

De forma más general, la respuesta de intervención se puede representar como una función racional:

$$m_t = \frac{\omega(B)B^b}{\delta(B)} I_t \tag{3.45}$$

donde I_t es la variable de intervención que puede ser la función escalón $S_t^{(T)}$ o la función pulso $P_t^{(T)}$, $\omega(B) = \omega_0 - \omega_1 B - \cdots - \omega_s B^s$ y $\delta(B) = \delta_0 - \delta_1 B - \cdots - \delta_r B^r$. El término B^b simboliza el retraso del efecto de intervención, los pesos ω_j's representan las alteraciones iniciales del evento y los coeficientes δ_j's cuantifican los efectos permanentes en el sistema. Las raíces del polinomio $\delta(B) = 0$ se toman fuera del círculo unitario.

Cuando varias intervenciones afectan el proceso, se pueden incorporar al modelo lo que resulta en una expresión más general:

$$Y_t = \sum_{j=1}^{k} \frac{\omega_j(B)B^{b_j}}{\delta_j(B)} I_{j,t} + \frac{\theta_q(B)}{\phi_p(B)(1-B)^d} e_t \tag{3.46}$$

donde $I_{j,t}$ representa las k variables de intervención que pueden ser funciones escalón o pulso dependiendo de la naturaleza del evento externo.

3.4.2. Análisis de datos atípicos

Aquellas observaciones que no se relacionan con alguna intervención y que son inconsistentes con el resto de las observaciones se conocen como datos atípicos o aberrantes. Al generar un modelo, es importante detectar y remover los efectos de los datos atípicos, ya que generalmente conllevan a una identificación errónea del proceso lo que se traduce en inferencias inválidas, que no describen la naturaleza intrínseca de la serie.

La detección de datos aberrantes fue introducida por primera vez en 1972 por Fox[25], el cual propuso dos modelos estadísticos que clasificaban las observaciones atípicas como ***aditivas*** o ***innovadoras***.

Sea $\{X_t\}$ el proceso no perturbado que sigue un proceso general estacionario e invertible ARMA(p, q). Una observación atípica aditiva (denotado como AO por sus siglas en inglés [VI]) se define como:

$$Y_t = X_t + \omega I_t^{(T)} = \frac{\theta_q(B)}{\phi_p(B)} e_t + \omega I_t^{(T)} \tag{3.47}$$

donde:

[VI] Proviene de *additive outlier*

$$I_t^{(T)} = \begin{cases} 1 & t = T \\ 0 & t \neq T \end{cases} \tag{3.48}$$

es una variable indicadora que indica la presencia o ausencia de un dato aberrante en el tiempo T.

Una observación atípica innovadora (conocida como IO por sus siglas en el idioma inglés[VII]) se define como:

$$Y_t = X_t + \frac{\theta_q(B)}{\phi_p(B)}\omega I_t^{(T)} = \frac{\theta_q(B)}{\phi_p(B)}\left(e_t + \omega I_t^{(T)}\right) \tag{3.49}$$

donde $I_t^{(T)}$ es la misma variable indicadora dada por la ecuación 3.48. A diferencia de una observación atípica aditiva, la cual afecta solamente en el tiempo T, una observación aberrante innovadora tiene un efecto que perdura más allá del tiempo T, influyendo en las observaciones $Y_T, Y_{T+1}, \ldots, Y_n$ a través de la memoria del sistema descrita por el término $\theta_q(b)/\phi_p(B)$.

De forma más general, si la serie manifiesta k datos atípicos se pueden incorporar en el siguiente esquema:

$$Y_t = X_t + \sum_{j=1}^{k} \omega_j \nu_j(B) I_t^{(T)} = \frac{\theta_q(B)}{\phi_p(B)} e_t + \sum_{j=1}^{k} \omega_j \nu_j(B) I_t^{(T)} \tag{3.50}$$

Cuando se presenta un dato aberrante AO se tiene que $\nu_j(B) = 1$, y en el caso de observaciones IO el término queda como $\nu_j(B) = \theta_q(b)/\phi_p(B)$.

Para detectar la presencia de datos atípicos se emplea la representación AR(∞) del proceso perturbado $\{Y_t\}$, es decir, expresado únicamente en términos autoregresivos:

$$a_t = \pi(B)Y_t = \frac{\theta_q(B)}{\phi_p(B)} Y_t \tag{3.51}$$

De las ecuaciones 3.47 y 3.49 se tiene que para:

$$AO : a_t = \omega\pi(B) I_t^{(T)} + e_t \tag{3.52}$$

$$IO : a_t = \omega I_t^{(T)} + e_t \tag{3.53}$$

De la ecuación 3.53 se observa que $a_t = e_t$ para $t < T$; para valores de t iguales o mayores que T se tienen las siguientes relaciones : $a_T = \omega + e_T$, $a_{T+1} = -\omega\pi_1 + e_{T+1}$, $a_{T+2} = -\omega\pi_2 + e_{T+2}, \ldots, a_n = -\omega\pi_{n-T} + e_n$. Se puede estimar ω a través de mínimos cuadrados obteniéndose:

[VII] En inglés se escribe como *innovative outlier*

$$\widehat{\omega}_{AT} = \frac{-\sum_{j=0}^{n-T} \pi_j a_{T+j}}{\sum_{j=0}^{n-T} \pi_j^2} = \frac{-\sum_{j=0}^{n-T} \pi_j a_{T+j}}{\rho^2} \tag{3.54}$$

Donde $\pi_0 = -1$ y $\rho^2 = \sum_{j=0}^{n-T} \pi_j^2$. El subíndice del estimador $\widehat{\omega}_{AT}$ sirve para indicar que es un dato atípico aditivo en el tiempo T. La varianza del estimador está dada por:

$$Var(\widehat{\omega}_{AT}) = Var\left(\frac{-\sum_{j=0}^{n-T} \pi_j a_{T+j}}{\rho^2}\right) = \frac{1}{\rho^4} Var(-\sum_{j=0}^{n-T} \pi_j a_{T+j} = \frac{1}{\rho^4}\rho^2\sigma_e^2 = \frac{\sigma_e^2}{\rho^2} \tag{3.55}$$

Similarmente, para observaciones IO, se puede estimar ω empleando la ecuación 3.53:

$$\widehat{\omega}_{IT} = a_T \tag{3.56}$$

Ahora el subíndice es un indicativo de un dato atípico IO en el tiempo T. La varianza para este estimador es:

$$Var(\widehat{\omega}_{IT}) = \sigma_e^2 \tag{3.57}$$

Se pueden formular pruebas de hipótesis para determinar si Y_T es una observación irregular. Para probar si es del tipo AO se emplea el siguiente estadístico que tiene una distribución $N(0,1)$:

$$\lambda_{1,T} = \frac{\rho\widehat{\omega}_{AT}}{\sigma_e^2} \tag{3.58}$$

Cuando se quiere probar si el dato es IO se emplea ahora la siguiente estadística que al igual que la anterior se distribuye de forma normal estándar:

$$\lambda_{2,T} = \frac{\rho\widehat{\omega}_{IT}}{\sigma_e^2} \tag{3.59}$$

Las pruebas de hipótesis anteriores se aplican cuando se conoce el tiempo de ocurrencia del dato aberrante T. Sin embargo, en la práctica no se conoce esta información y las pruebas se deben aplicar a todos los valores de la serie temporal. Para controlar la tasa de error global de las múltiples pruebas, se emplea el criterio de Bonferroni. Sea λ_1 el valor máximo que toma la estadística $|\lambda_{1,T}|$ en todos los datos de la serie:

$$\lambda_1 = \max_{1\leq t\leq n}|\lambda_{1,T}| \tag{3.60}$$

La observación que acontece en el tiempo T' en el que ocurre el máximo se considera como AO si λ_1 excede el percentil superior $(0{,}025 * 100)/n$de la distribución normal estándar. Como un dato atípico AO puede generalmente inflar el valor de la estimación de σ_e^2, se puede emplear un

estimador robusto de la varianza para aumentar la potencia de la prueba. Se puede por ejemplo emplear la media residual absoluta para estimar σ_e.

El mismo criterio se puede usar para probar que existe un valor IO en la serie. En este caso se utiliza λ_2 el cual se define como:

$$\lambda_2 = \max_{1 \leq t \leq n} |\lambda_{2,T}| \tag{3.61}$$

Si λ_2 supera el percentil superior $(0{,}025 * 100)/n$de la distribución $N(0, 1)$, la observación en la cual ocurre el máximo se considera en valor IO.

En general, la naturaleza del dato atípico no se conoce de antemano. Se puede usar como regla que si en el tiempo T se observa un valor irregular, se clasifica como AO si $|\lambda_{1,T}| > |\lambda_{2,T}|$ y como IO si se cumple la desigualdad en el otro sentido.

Cuando un dato atípico es encontrado, se incorpora al modelo y el proceso de detección se repite nuevamente hasta que se refina el modelo y no se encuentran más valores irregulares.

3.5. Pronósticos del modelo

Como se mencionó al principio del capítulo uno de los principales objetivos de la construcción de un modelo para una serie temporal es su capacidad para predecir valores futuros de la serie con una precisión aceptable.

Se desarrollará a continuación la teoría predictiva tanto de los modelos estacionarios ARMA como los no estacionarios ARIMA, mencionando al final algunos criterios de selección basados en los pronósticos de los modelos. Durante el desarrollo se asumirá que se conoce exactamente el modelo con todos sus parámetros. En la práctica esto no es posible pues se tiene que hacer estimaciones, pero sus propiedades asintóticas hacen que los resultados no difieran mucho del modelo ideal.

3.5.1. Pronósticos con mínimo error cuadrático medio

Basándose en los datos que se tienen de la serie temporal $Y_1, Y_2, \ldots, Y_t$ generalmente se desea hacer un pronóstico de lo que ocurrirá l unidades en el futuro, es decir, se quiere estimar el valor de Y_{t+l}; el tiempo t correspondiente al último valor de la serie se le conoce como ***origen de la predicción*** y al valor l como el ***tiempo para el pronóstico***.

Uno de los principales objetivos es que la estimación de Y_{t+l}, denotada como $\hat{Y}_t(l)$ sea óptima, es decir, que el error respecto al valor real sea mínimo. El criterio que se emplea para determinar el óptimo, es que la estimación minimice el error cuadrático medio:

$$E\left(\left(Y_{t+l} - \hat{Y}_t(l)\right)^2\right) \tag{3.62}$$

La cantidad $\hat{Y}_t(l)$ que minimiza la expresión anterior es la esperanza condicional de Y_{t+l} respecto a los valores de la serie $Y_1, Y_2, \ldots, Y_t$:

$$\hat{Y}_t(l) = E(Y_{t+1}|Y_1, Y_2, \ldots, Y_t) \tag{3.63}$$

La relación anterior se obtiene del hecho de que si se quiere predecir una variable aleatoria Y respecto a una función arbitraria de variables aleatorias $h(X_1, X_2, \ldots, X_n)$, la relación que minimiza el error cuadrático medio es la esperanza condicional de Y respecto a las variables aletorias $X_1, X_2, \ldots, X_n$:

$$h(X_1, X_2, \ldots, X_n) = E(Y|X_1, X_2, \ldots, X_n) \tag{3.64}$$

Para ver de donde proviene el resultado anterior, se hará el desarrollo en el caso en que se quiere predecir una variable aleatoria Y por una función de una sola variable aleatoria $h(X)$, donde la predicción que minimiza el error cuadrático medio es:

$$h(X) = E(Y|X) \tag{3.65}$$

Para derivar lo expuesto con anterioridad, se necesita primero determinar cuál es la predicción que minimiza el error cuadrático medio si se quiere pronosticar una variable aleatoria Y respecto a una constante $c \in \Re$. Para hallar el óptimo, se establece una función que depende de la constante:

$$g(c) = E[(Y - c)^2] \tag{3.66}$$

Se deriva la función anterior, se iguala a cero $g'(c) = 0$, obteniéndose la siguiente relación para el mínimo:

$$c = E(Y) = \mu_Y \tag{3.67}$$

Regresando al problema, se desea encontrar una función $h(X)$ que minimice:

$$E[(Y - h(X))^2] \tag{3.68}$$

Ahora bien, usando la propiedad del valor esperado condicional $E\big(E(Y|X)\big) = E(Y)$, se puede reescribir la expresión anterior como:

$$E\big[(Y - h(X))^2\big] = E\big(E\big[(Y - h(X))^2|X\big]\big) \tag{3.69}$$

Dado el valor de $X = x$ la función $h(X - x) = h(x)$ se puede tomar como constante:

$$E\Big(E\Big[\big(Y-h(X)\big)^2|X=x\Big]\Big)=E\Big(E\Big[\big(Y-h(x)\big)^2|X=x\Big]\Big) \tag{3.70}$$

Para cada valor de $X=x$, al ser $h(x)$ constante por la ecuación 3.67 se tiene que para cada x la función que minimiza el error cuadrático medio es:

$$h(x)=E(Y|X=x) \tag{3.71}$$

Esto se extiende automáticamente a todos los valores de x por lo que se obtiene el resultado buscado de que la función que optimiza es $h(x)=E(Y|X)$.

En las siguientes secciones se aplicarán estos resultados, en específico la ecuación 3.63, para generar los pronósticos de los modelos estacionarios ARMA y no estacionarios ARIMA.

3.5.2. Pronósticos de modelos estacionarios de ARMA

En el proceso general ARMA(p,q) con término constante θ_0, el valor en el tiempo $t+l$ está dado por la siguiente relación:

$$Y_{t+l}=\theta_0+\phi_1Y_{t+l-1}+\cdots+\phi_pY_{t+l-p}+e_{t+l}-\theta_1e_{t+l-1}-\cdots-\theta_qe_{t+l-q} \tag{3.72}$$

Usando la expresión 3.63, se aplica la esperanza condicional respecto a los valores de la serie $Y_1,Y_2,\ldots,Y_t$ en ambos lados de la relación anterior:

$$\begin{aligned}E(Y_{t+l}|Y_1,\ldots,Y_t)&\\ &=E(\theta_0|Y_1,\ldots,Y_t)+\phi_1E(Y_{t+l-1}|Y_1,\ldots,Y_t)+\cdots+\phi_pE(Y_{t+l-p}|Y_1,\ldots,Y_t)\\ &+E(e_{t+l}|Y_1,\ldots,Y_t)-\theta_1E(e_{t+l-1}|Y_1,\ldots,Y_t)-\ldots\\ &-\theta_qE(e_{t+l-1}|Y_1,\ldots,Y_t)\end{aligned} \tag{3.73}$$

Donde se ha usado el hecho de que la esperanza condicional es un operador lineal. Se puede reescribir la ecuación anterior en términos de $\widehat{Y}_t(l)$, quedando como:

$$\begin{aligned}\widehat{Y}_t(l)=\theta_0+\phi_1\widehat{Y}_t(l-1)+\cdots+\phi_p\widehat{Y}_t(l-p)+E(e_{t+l}|Y_1,\ldots,Y_t)\\ -\theta_1E(e_{t+l-1}|Y_1,\ldots,Y_t)-\cdots--\theta_qE(e_{t+l-q}|Y_1,\ldots,Y_t)\end{aligned} \tag{3.74}$$

Adicionalmente se ha empleado la propiedad de que la esperanza condicional de una constante es la constante para tener que $E(\theta_0|Y_1,Y_2,\ldots,Y_t)=\theta_0$. La expresión precedente es una ecuación recursiva pues para obtener las predicciones en un tiempo en el futuro se necesita obtener las predicciones en tiempos iniciales. Para evaluar la relación recursiva se deben usar las siguientes relaciones:

$$E(e_{t+j}|Y_1,\ldots,Y_t) = \begin{cases} 0 & para\ j > 0 \\ e_{t+j} & para\ j < 0 \end{cases} \tag{3.75}$$

$$E(Y_{t+j}|Y_1,\ldots,Y_t) = \begin{cases} \widehat{Y}_t(j) & para\ j > 0 \\ Y_{t+j} & para\ j < 0 \end{cases} \tag{3.76}$$

Cuando $j > 0$ se tiene que $E(e_{t+j}|Y_1, Y_2, \ldots, Y_t) = E(e_{t+j}) = 0$, pues e_{t+j} es independiente de $Y_1, Y_2, \ldots, Y_t$. Si $j < 0$, entonces $E(e_{t+j}|Y_1, Y_2, \ldots, Y_t) = e_{t+j}$ como resultado de otras de las propiedades de la esperanza condicional. Un argumento similar al anterior se emplea para verificar el resultado de la ecuación 3.76.

Como se observa en la ecuación 3.74, los pronósticos en $l = 1, 2, \ldots, q$ involucran algunos términos de ruido blanco. Sin embargo, para $l > q$, la parte autoregresiva desaparece completamente en las predicciones obteniéndose:

$$\widehat{Y}_t(l) = \theta_0 + \phi_1\widehat{Y}_t(l-1) + \cdots + \phi_p\widehat{Y}_t(l-p) \quad \text{para } l > q \tag{3.77}$$

De esta forma, para valores muy lejanos en el futuro, el comportamiento de los pronósticos está determinado completamente por la parte autoregresiva del modelo. Tomando en cuenta que el término constante θ_0 se puede vincular con la media del proceso μ y los parámetros autoregresivos como lo establece la ecuación 2.86, se puede reescribir la relación anterior de la siguiente forma:

$$\widehat{Y}_t(l) - \mu = \phi_1\big[\widehat{Y}_t(l-1) - \mu\big] + \cdots + \phi_p\big[\widehat{Y}_t(l-p) - \mu\big] \quad \text{para } l > q \tag{3.78}$$

En función de l, la ecuación anterior se puede ver como una ecuación homogénea en diferencias:

$$\begin{aligned} \big[\widehat{Y}_t(l) - \mu\big] - \phi_1\big[\widehat{Y}_t(l-1) - \mu\big] - \cdots - \phi_p\big[\widehat{Y}_t(l-p) - \mu\big] \\ = \big(1 - \phi_1 B - \cdots - \phi_p B^{k-p}\big)\big[\widehat{Y}_t(l-\mu\big] = \phi_p(B)\big[\widehat{Y}_t(l-\mu\big] = 0 \end{aligned} \tag{3.79}$$

Donde B opera sobre el índice l dejando inalterado los valores constantes: $B[\widehat{Y}_t(l) - \mu]$. Así, como ocurrió con la función de autocorrelación ρ_k para modelos AR de orden p, el comportamiento del término $\widehat{Y}_t(l) - \mu$ depende de las raíces del polinomio $\phi_p(B)$; si existen solamente raíces reales se presenta solamente un decaimiento exponencial, mientras que si existen raíces complejas, a parte del decaimiento exponencial se presenta un amortiguamiento senoidal. De esta manera, para modelos estacionarios ARMA, el término $\widehat{Y}_t(l) - \mu$ tiende a cero, por lo que para valores de l lejanos al origen t el pronóstico es simplemente la media del proceso μ. Esto está acorde con el comportamiento de los modelos ARMA, donde la dependencia de los valores de la serie disminuye a medida que el tiempo entre observaciones se incrementa.

Además de hacer las predicciones, es importante y necesario tener una estimación del error que se comete en cada una de ellas. Por ello se introduce el ***error del pronóstico*** en el tiempo l:

$$e_t(l) = Y_{t+l} - \widehat{Y}_t(l) \tag{3.80}$$

Para deducir una expresión general para el error del pronóstico en términos de los coeficientes de los modelos estacionarios ARMA, es necesario introducir una nueva representación que es general para los modelos no estacionarios ARIMA (de los cuales los ARMA son un caso particular). Esta forma equivalente ayuda a que los cálculos sean más sencillos.

Un proceso ARIMA se puede representar como un ***modelo lineal truncado*** cuya forma es:

$$Y_{t+l} = C_t(l) + I_t(l) \quad \text{para } l > 1 \tag{3.81}$$

donde

$$C_t(l) = \sum_{i=0}^{d} A_i l^i + \sum_{i=1}^{r} \sum_{j=0}^{p_i-1} B_{ij} l^j (G_i)^l \tag{3.82}$$

$$I_t(l) = \sum_{j=0}^{l-1} \Psi_j e_{t+l-j} \tag{3.83}$$

En $C_t(l)$ los términos A_i y B_{ij} son independientes de l y sólo dependen de los valores $Y_t, Y_{t-1}, \ldots$. La representación como modelo lineal truncado es válida, pues para un tiempo fijo t, $C_t(l)$ es la función complementaria de la siguiente ecuación en diferencias:

$$C_t(l) - \varphi_1 C_t(l-1) - \cdots - \varphi_{p+d} C_t(l-p-d) = \phi_p(B)(1-B)^d C_t(l) = \theta_0 \tag{3.84}$$

Los términos G_i de la ecuación 3.82 son los inversos de las raíces de la ecuación homogénea en diferencias $\Phi_p(B)(1-B)^d C_t(l) = 0$; el índice hace referencia a las multiplicidad de dichas raíces. La prueba de que $C_t(l)$ es la función complementaria de la ecuación 3.84, es un resultado de la teoría de ecuaciones en diferencias (ver Goldberg[26]).

El término $I_t(l)$ es una solución particular de la ecuación en diferencias homogénea:

$$\phi_p(B)(1-B)^d I_t(l) = 0 \tag{3.85}$$

Los términos ψ's están definidos a partir de la siguiente relación:

$$\phi_p(B)(1-B)^d(1 - \psi_1 B + \psi_2 B^2 + \ldots) = \theta_q(B) \tag{3.86}$$

Como $C_t(l)$ contiene $p+d$ constantes (las A's y los B's), la suma de $C_t(l)$ y $I_t(l)$ es una solución general de la ecuación que define los modelos ARIMA. Los valores específicos de A_i y B_{ij} están determinados por las condiciones iniciales del proceso $\{Y_t\}$.

Con la representación de un modelo lineal truncado se va a poder establecer una expresión general del error de pronóstico en términos de los parámetros del modelo; lo que hay que tener en cuenta es la representación explícita de $I_t(l)$ dada por la ecuación 3.83 y que para modelos invertibles (que es el caso de los modelos ARMA), $C_t(l)$ es una función de los valores de la serie $Y_1, Y_2, \ldots, Y_t$.

Para hacer el cálculo del error, se usa la ecuación 3.81 para establecer el pronóstico en un tiempo hacia delante l en función de un modelo lineal truncado:

$$\begin{aligned}\widehat{Y}_t(l) =& E(Y_{t+l}|Y_1, \ldots, Y_t) = E(C_t(l) + I_t(l)|Y_1, \ldots, Y_t) = \\ & E(C_t(l)|Y_1, \ldots, Y_t) + E(I_t(l)|Y_1, \ldots, Y_t)\end{aligned} \tag{3.87}$$

En el desarrollo anterior se ha empleado la linealidad de la esperanza condicional. Como se tiene un proceso invertible, $C_t(l)$ depende de $Y_1, \ldots, Y_t$ por lo que $E(C_t(l)|Y_1, \ldots, Y_t) = C_t(l)$. Por otro lado, $I_t(l)$ es independiente de $Y_1, \ldots, Y_t$ lo que lleva a que $E(l_t(l)|Y_1, \ldots, Y_t) = E(I_t(l))$ por las propiedades del valor esperado. Finalmente, como $E(I_t(l)) = 0$, se tiene que:

$$\widehat{Y}_t(l) = C_t(l) \tag{3.88}$$

De esta forma, para modelos estacionarios ARMA que sean invertibles, el error del pronóstico en el tiempo l se establece como:

$$e_t(l) = Y_{t+l} - \widehat{Y}_t(l) = (C_t(l) + I_t(l)) - C_t(l) = I_t(l) = \sum_{j=0}^{l-1} \psi_j e_{t+l-j} \tag{3.89}$$

Se pueden calcular la media y la varianza del error

$$E(e_t(l)) = 0 \tag{3.90}$$

$$Var(e_t(l)) = \sigma_e^2 \sum_{j=0}^{l-1} \psi_j^2 \tag{3.91}$$

3.5.3. Pronósticos de modelos no estacionarios ARIMA

Las predicciones de los modelos no estacionarios ARIMA, se hacen de forma similar a la de los modelos ARMA. El asunto es escribir el modelo ARIMA(p, d, q) en su forma de ecuación en diferencias, como se vio en la sección 2.3.2, para que aparente ser un modelo ARMA$(p + d, q)$:

$$Y_t = \varphi_1 Y_{t-1} + \cdots + \varphi_{p+d} Y_{t-p-d} + e_{t+l} - \theta_1 e_{t+l-1} - \cdots - \theta_q e_{t+l-q} \tag{3.92}$$

Los coeficientes φ's están dados por la siguiente relación de polinomios de retraso:

$$\varphi_p(B) = \phi_p(B)(1-B)^d \tag{3.93}$$

Usando esta representación, se pueden emplear las ecuaciones 3.74, 3.75 y 3.76 remplazando p con $p+d$ y ϕ_i con φ_i para generar la respectiva relación recursiva de los pronósticos.

Como forma alternativa al proceso anterior, se podría trabajar con la serie estacionaria que resulta de la transformación $W_t = (1-B)^d Y_t$ para calcular las predicciones. Una vez obtenidos los pronósticos, se "deshace"la diferencia y se suman los términos para generar predicciones en la serie original. Este procedimiento alternativo es válido básicamente porque la transformación de diferencia es un operador lineal.

En el caso que se tenga un modelo ARIMA invertible el error de pronóstico está determinado por la misma relación que en el caso de modelos estacionarios:

$$e_t(l) = I_t(l) = \sum_{j=0}^{l-1} \psi_j e_{t+l-j} \tag{3.94}$$

Las ecuaciones para la media y la varianza de los procesos ARIMA invertibles son exactamente las mismas que para los modelos estacionarios invertibles: ecuaciones 3.90 y 3.91. A diferencia de lo que ocurre en los procesos ARMA, los pesos ψ's no decaen a cero a medida que se incrementa el valor de j.

3.5.4. Límites de predicción

Además de realizar las predicciones es importante validar la precisión de las mismas. Esto se puede hacer generando un intervalo de confianza y dependiendo de su tamaño será la exactitud del pronóstico. El intervalo se genera empleando el error de pronóstico $e_t(l)$.

Como en los modelos ARIMA los términos de ruido blanco $\{e_t\}$ son variables independientes e idénticamente distribuidas de forma normal, entonces el error de predicción tendrá también una distribución normal con media cero y varianza determinada por la ecuación 3.91. El estadístico definido como:

$$\frac{e_t(l)}{\sqrt{Var(e_t(l))}} = \frac{Y_{t+l} - \hat{Y}_t(l)}{\sqrt{Var(e_t(l))}} \tag{3.95}$$

tendrá una distribución normal estándar y se puede emplear para generar un intervalo de confianza para respecto al error del pronóstico. Hay que destacar que σ_e^2 y los coeficientes ψ's necesitan ser estimados[VIII] para calcular la varianza del error, pero en el caso de muestras grandes las estimaciones modifican muy poco la distribución del estadístico. Para un nivel de confianza dado

[VIII] en realidad las ψ's se calculan a través de las estimaciones de los coeficientes ϕ's y θ's del respectivo modelo ARMA.

$1-\alpha$ se pueden usar los percentiles $-z_{1-\alpha}$ y $z_{1-\alpha}$ de la distribución normal estándar para establecer la probabilidad de encontrar estadístico entre esos valores:

$$P\left[-z_{1-\alpha} \leq \frac{Y_{t+l} - \hat{Y}_t(l)}{\sqrt{Var(e_t(l))}} \leq z_{1-\alpha}\right] = 1-\alpha \tag{3.96}$$

Lo que lleva a:

$$P\left[\widehat{Y}_t(l) - z_{1-\alpha}\sqrt{Var(e_t(l))} \leq Y_{t+l} \leq \widehat{Y}_t(l) + z_{1-\alpha}\sqrt{Var(e_t(l))}\right] = 1-\alpha \tag{3.97}$$

De esta forma el intervalo de confianza del $(1-\alpha)100\,\%$ respecto a la observación futura Y_{t+l} es:

$$\widehat{Y}_t(l) \pm z_{1-\alpha}\sqrt{Var(e_t(l))} \tag{3.98}$$

3.5.5. Actualización de los pronósticos

Cuando nuevas observaciones del proceso están disponibles es posible emplearlas para mejorar los pronósticos hechos con anterioridad. Partiendo de que el origen de predicción es el tiempo t, el pronóstico en el tiempo $l+1$ se denota como $\widehat{Y}_t(l+1)$. Una vez que se conoce el proceso en el tiempo $t+1$, se desea actualizar la predicción, que partiendo de este nuevo origen se escribe como $\widehat{Y}_{t+1}(l)$. Para evitar hacer todo el procedimiento que se describió en las secciones 3.5.2 y 3.5.3, es posible deducir una expresión que facilita los cálculos.

Partiendo de la representación del proceso lineal truncado se escribe el valor Y_{t+l+1} como:

$$Y_{t+l+1} = C_t(l+1) + I_t(l+1) = C_t(l+1) + \sum_{j=0}^{l} \psi_j e_{t+l+1-j} \tag{3.99}$$

Como $C_t(l+1)$ y e_{t+1} dependen de $Y_1, Y_2, \ldots, Y_t, Y_{t+1}$ mientras que $e_{t+2}, e_{t+3}, \ldots, e_{t+l}, e_{t+l+1}$ son independientes de $Y_1, Y_2, \ldots, Y_t, Y_{t+1}$ al aplicar $\widehat{Y}_{t+1}(l) = E(Y_{t+l+1}|Y_1, Y_2, \ldots, Y_t, Y_{t+1})$ se obtiene:

$$\widehat{Y}_{t+1}(l) = C_t(l+1) + \psi_l e_{t+1} \tag{3.100}$$

Por otro lado se tiene por la ecuación 3.88 que $C_t(l+1) = \widehat{Y}_t(l+1)$. Además como $e_{t+1} = Y_{t+1} - \widehat{Y}_t(1)$, por la ecuación 3.80, $\widehat{Y}_{t+1}(l)$ se puede escribir como:

$$\widehat{Y}_{t+1}(l) = \widehat{Y}_t(l+1) + \psi_1(Y_{t+1} - \widehat{Y}_t(1)) \tag{3.101}$$

La relación anterior se conoce como ***ecuación general de actualización***. Hay que notar que el término $(Y_{t+1} - \widehat{Y}_t(1))$ es el error de pronóstico en el tiempo una vez que la observación Y_{t+1} se hace presente.

3.5.6. Predicciones ponderadas

Para modelos ARIMA donde no hay términos de promedios móviles es claro como las predicciones están determinadas explícitamente por la serie $Y_1, Y_2, \ldots, Y_t$. Sin embargo, cuando existe la parte correspondiente MA, los pronósticos involucran términos de ruido blanco (siempre que $l < q$), por lo que la naturaleza de las predicciones en términos de los valores de la serie está oculta. Es posible desenmascarar este comportamiento, invirtiendo el modelo ARIMA en términos de $Y_{t-1}, Y_{t-2}, \ldots$ como lo establece la ecuación 2.63:

$$Y_t = \pi_1 Y_{t-1} + \pi_2 Y_{t-2} + \pi_3 Y_{t-3} + \cdots + e_t \tag{3.102}$$

Por lo anterior, se tiene que Y_{t+1} queda como:

$$Y_{t+1} = \pi_1 Y_t + \pi_2 Y_{t-2} + \pi_3 Y_{t-2} + \cdots + e_{t+1} \tag{3.103}$$

Al calcular el valor esperado condicionado a los valores de la serie $Y_1, Y_2, \ldots, Y_t$, se obtiene el pronóstico en un tiempo futuro:

$$\widehat{Y}_t(1) = \pi_1 Y_t + \pi_2 Y_{t-1} + \pi_3 Y_{t-2} + \ldots \tag{3.104}$$

En la situación de que se desee obtener la predicción a un tiempo futuro l, se tiene que calcular la respectiva ecuación recursiva:

$$\widehat{Y}_t(1) = \pi_1 \widehat{Y}_t(l-1) + \pi_2 \widehat{Y}_t(l-2) + \pi_3 \widehat{Y}_t(l-3) + \ldots \tag{3.105}$$

Las ecuaciones anteriores involucran sumas infinitas pero para cuestiones prácticas se toman los pesos hasta el índice $t-1$ considerando que los parámetros $\pi_t, \pi_{t+1}, \ldots$ se vuelven despreciables. Para un modelo ARIMA invertible los pesos π's se pueden obtener de la siguiente relación:

$$e_t = \pi(B) Y_t = \frac{\phi_p(B)(1-B)^d}{\theta_q(B)} Y_t = \frac{\varphi_p(B)}{\theta_q(B)} Y_t \tag{3.106}$$

La ecuación anterior genera las siguientes fórmulas recursivas para obtener las π's:

$$\pi_j = \begin{cases} \sum_{i=1}^{min(j,q)} \theta_i \pi_{j-i} + \varphi_j & \text{para } 1 \leq j \leq p+d \\ \sum_{i=1}^{min(j,q)} \theta_i \pi_{j-i} & \text{para } j > p+d \end{cases} \tag{3.107}$$

3.5.7. Selección del modelo basado en los errores de pronóstico

En muchas circunstancias es común encontrar varios modelos que son adecuados para representar el proceso temporal que se está estudiando. Como se mencionó en la sección 3.3.3 existen algunas técnicas que permiten elegir el mejor modelo de acuerdo a cierto criterio. Si la parte predictiva del modelo es de suma importancia, entonces se puede tener un criterio de selección basado en los errores de pronóstico $e_t(l) = Y_{t+l} - \widehat{Y}_t(l)$. Se introducen a continuación cuatro cantidades que sirven para comparar las predicciones de los modelos, teniendo en cuenta que entre menor es el valor mejor es la predicción del modelo. En términos generales, la selección se hace considerando el desempeño de los cuatro criterios.

1. **Porcentaje de error medio (PEM)** :

$$PEM = \left(\frac{1}{M} \sum_{l=1}^{M} \frac{e_t(l)}{Y_{t+l}} \right) 100\,\% \tag{3.108}$$

2. **Error cuadrático medio (ECM)**:

$$ECM = \frac{1}{M} \sum_{l=1}^{M} (e_t(l))^2 \tag{3.109}$$

3. **Porcentaje de error absoluto medio (PEAM)** :

$$PEM = \left(\frac{1}{M} \sum_{l=1}^{M} \left| \frac{e_t(l)}{Y_{t+l}} \right| \right) 100\,\% \tag{3.110}$$

4. **Error absoluto medio (EAM)**:

$$EAM = \frac{1}{M} \sum_{l=1}^{M} |(e_t(l))| \tag{3.111}$$

Para evaluar las cantidades descritas con anterioridad, por lo regular el origen de la predicción es dentro de la misma serie $Y_1, Y_2, \ldots, Y_t$, es decir, en un cierto valor Y_r donde $r < t$, para que los pronósticos se puedan comparar con los valores restantes de la serie $Y_{r+1}, Y_{r+2}, \ldots, Y_t$. En este caso, el valor M de las ecuaciones anteriores viene siendo $M = t - r$.

Capítulo 4

Modelos ARIMA: robo de vehículos asegurados

En este capítulo se presentan los análisis que se hicieron a los datos de robo de autos. Antes de ajustar los modelos ARIMA se prepararon los datos y se clasificaron los estados de acuerdo al crecimiento del robo de autos y al costo de la prima de seguro.

4.1. Clasificación de los estados

Se explica de forma íntegra cuales fueron las técnicas involucradas en la regionalización de los estados de la república mexicana. Primero se va a describir el criterio empleado para generar la clasificación. Después se detallará la técnica de agrupamiento jerárquico de aglomeración y se explicará cómo se implementó con los datos para ordenar los estados. Finalmente se describirá la regionalización final, la cual se obtiene con un refinamiento de la clasificación anteriormente descrita.

4.1.1. Criterio de clasificación

El primer criterio para clasificar a los datos fue el crecimiento o decrecimiento del robo de autos, para ello se calcularon las proporciones de autos robados de la población asegurada para cada estado, y con estas proporciones relativas a los años de 2008, 2009 y 2010, se estima la tasa anual del crecimiento de robos para cada uno de los estados con estos datos, utilizando un modelo de regresión lineal simple; las pendientes de los ajustes lineales son las estimaciones del crecimiento o decrecimiento del robo de autos.

Para obtener las proporciones de robos de los años 2008, 2009 y 2010, se necesitan los datos tanto de las afectaciones de la cobertura de robo como de las unidades expuestas. Por esto, se procedió a trabajar con las bases de datos para conseguir esta información.

Hay que destacar cuatro aspectos importantes de esta minería de datos:

1. Una es relativa a las compañías. Para el cálculo de las proporciones se deseaba emplear todo la información del sector asegurador mexicano. Sin embargo, se encontraron dos compañías (una en el año 2008 y otra diferente en el año 2009) que presentaban inconsistencias en sus datos de afectaciones de la cobertura de robo. Como no era práctico por cuestiones de tiempo rectificar estos datos, se vio en la necesidad de remover la información de estas compañías, en estos años en específico. Esta decisión no afecta significativamente las estimaciones finales, ya que es muy poco lo que representan estos datos de la población total asegurada.

2. En el sector asegurador, existen dos tipos de aseguramiento de vehículos: contratos individuales de automóviles, y es pólizas de forma grupal. Los contratos grupales por lo regular se identifican como flotilla y generalmente son requeridos por empresas en las cuales sus vehículos circulan por diversos estados. Aquí se decidió utilizar solamente los datos de las pólizas individuales, pues los contratos grupales podrían generar un sesgo, debido a que los autos de flotilla tienen un patrón diferente en el robo de sus unidades pues existe la posibilidad de un robo fuera de la entidad de aseguramiento.

3. Otro aspecto que tuvo que ser tomando en consideración fue el coaseguro. Esta práctica consiste en que dos o más compañías aseguradoras comparten cierta parte de un riesgo asegurado(como se establece en Kass[27]); el porcentaje que asumen se determina de acuerdo a los intereses particulares de cada empresa. En el 2010, se identificó que ciertas compañías transferían el 100 % de su riesgo de robo de vehículo a otra empresa. Si dos compañías están en coaseguro, en las bases de datos aparece que cada empresa tiene ese riesgo cubierto, en este caso un vehículo asegurado. Para evitar duplicar la información, se decidió filtrar las compañías que tenían estos contratos de coaseguro del 100 %, pues en realidad, es la otra entidad aseguradora quien asume toda la responsabilidad para resarcir el daño. Esta decisión parece justificada por el hecho de que las 5 compañías que tenían estos contratos de coaseguro, son en realidad empresas pequeñas en el ámbito asegurador, que transfirieron su riesgo a compañías más grandes.

4. Hay que mencionar que en el estudio se incluyen todos los tipos vehículos asegurados con cobertura para robo de cualquier tipo, es decir, cualquier unidad automotora: vehículos de 4 puertas, camiones, tráileres, motocicletas, por mencionar algunos.

Después de aplicar las acciones antes descritas se calculan los porcentajes de robo de autos usando la fórmula siguiente:

$$Porcentaje\ de\ proporción\,\% = Proporción * 100 \tag{4.1}$$

En la tabla 4.1 se muestra las proporciones del robo de autos por estado y por año.

Los valores de la tabla 4.1 se utilizaron para ajustar las 32 ecuaciones de recta, y obtener las respectivas pendientes, que es uno de los principales objetivos. De acuerdo a Mendenhall[28] el modelo de regresión lineal simple es:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{4.2}$$

Estados	2008	2009	2010
Aguascalientes	0,29	0,33	0,38
Baja California	2,02	1,42	1,41
Baja California Sur	0,21	0,15	0,18
Campeche	0,35	0,20	0,17
Chiapas	0,65	0,46	0,52
Chihuahua	1,54	2,04	2,44
Coahuila	0,26	0,43	0,87
Colima	1,19	0,16	0,23
Distrito Federal	1,15	1,03	1,08
Durango	0,70	0,83	1,81
Estado de México	1,40	1,59	1,90
Guanajuato	0,87	0,38	0,31
Guerrero	1,61	1,49	1,63
Hidalgo	0,85	0,53	0,57
Jalisco	0,62	0,68	0,82
Michoacán	0,70	0,56	0,62
Morelos	1,07	1,13	1,33
Nayarit	0,46	0,30	0,49
Nuevo León	0,53	0,70	1,08
Oaxaca	0,96	0,74	0,60
Puebla	0,65	0,43	0,42
Querétaro	0,34	0,31	0,25
Quintana Roo	0,55	0,36	0,41
San Luis Potosí	0,30	0,30	0,54
Sinaloa	1,47	1,47	2,56
Sonora	0,56	0,43	0,55
Tabasco	0,75	0,75	0,75
Tamaulipas	0,65	0,53	1,26
Tlaxcala	0,59	0,54	0,73
Veracruz	0,54	0,39	0,55
Yucatán	0,30	0,10	0,07
Zacatecas	0,20	0,31	0,86

Cuadro 4.1: Proporciones de robo de vehículos asegurados (con cobertura para este delito) de los estados de la república mexicana para los años 2008, 2009 y 2010.

donde y_i son los valores de la variable dependiente, x_i los valores de la variable independiente o covariable, y ϵ_i son los términos de error que son variables aleatorias independientes que se distribuyen de forma normal con media cero y varianza constante σ^2. En el caso que compete, la variable depediente son los porcentajes de proporción de robo y la covariable es el tiempo dado en años.

De esta manera, empleando el paquete estadístico R, se estimaron los parámetros de las respectivas regresiones lineales simples como se establece en Faraway [29] y Verzani[32].

Los valores estimados del parámetro β_1, para cada uno de los estados se indican en la figura 4.1, en la cual se han ordenado las pendientes estimadas de menor a mayor. Se observa cantidades negativas que indican un decrecimiento, valores que son pequeños al compararlos con los demás que sugieren una estabilidad del fenómeno de robo y pendientes positivas que muestran una tendencia de crecimiento.

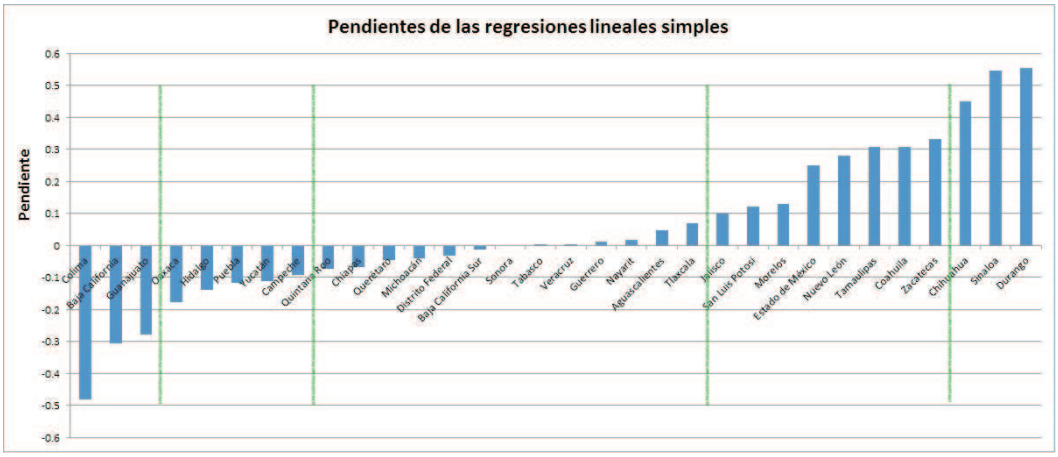


Figura 4.1: Pendientes resultantes del ajuste lineal. Se han ordenado de menor a mayor.

Al revisar la gráfica 4.1, es posible determinar 5 grupos de crecimiento homogéneo, que se clasifican como: decrecimiento alto (1), decrecimiento moderado(2), estable (3), crecimiento moderado (4), crecimiento alto (5).

Esta clasificación se obtiene con el procedimiento conocido como Agrupamiento jerárquico de aglomeración, técnica que se describe enseguida.

4.1.2. Agrupamiento Jerárquico de Aglomeración

El análisis de conglomerados (cluster analysis) es una técnica estadística utilizada para formar o discernir grupos de observaciones homogéneas, permitiendo generar una clasificación. Existen diversas técnicas y modelos para formar grupos. En el caso del presente estudio se decidió implementar la técnica de ***Agrupamiento Jerárquico de Aglomeración***; que es básicamente un algoritmo heurístico que se escogió debido a su implementación directa a través del software estadístico R, como se puede constatar en Everitt[3].

Esta técnica de clasificación inicia con n grupos de un solo elemento, y termina con un grupo

de n elementos. En cada paso, se fusionan los dos grupos más cercanos. La cercanía se puede medir con alguna de las distancias conocidas cuya definición se puede encontrar en Everitt y Landau[31]; para este trabajo se utiliza la ***distancia euclidiana*** entre dos puntos, entre un punto y un conjunto y entre dos conjuntos.

El algoritmo consiste en lo siguiente:

Inicio: Se tienen $G_1, \ldots, G_n$ grupos de un elemento:

1. Encontrar el par de grupos más cercanos G_i y G_j e integrarlos en un solo grupo G_i y borrar G_j.
2. Si el número de grupos es igual a 1 parar, sino regresar a 1)

Las distancias d_{ij} pueden formar una matriz cuya diagonal es igual a 0. En este caso, cada unidad muestral (cada estado de la república) tiene un único valor asociado que corresponde a la pendiente estimada, por lo que la medida utilizada entre dos puntos es el valor absoluto de su diferencia. En general en este trabajo se va a considerar para la distancia entre dos grupos A y B que tengan uno o más puntos, la fórmula (Everrit[3]):

$$d_{AB} = \left\{ \begin{array}{c} max(d_{ij}) \\ i \in A \\ j \in B \end{array} \right. \tag{4.3}$$

donde d_{AB} es la distancia entre los grupos A y B y d_{ij} es la distancia entre elementos de los grupos. Las clasificaciones jerárquicas se pueden representar por diagramas bidimensionales conocidos como ***dendrogramas***, los cuales van ilustrando las fusiones en cada uno de los pasos. La forma de los dendrogramas simula un árbol genealógico.

Como todos los algoritmos jerárquicos de agrupamiento reducen los datos a un solo grupo, se tiene que decidir cuál es la cantidad de grupos que mejor se ajusta a los datos. Esto se puede traducir en decidir a qué altura del dendrograma se debe "cortar"para tener un número óptimo de grupos. En realidad es una respuesta difícil. Un criterio puede ser elegir la distancia máxima de las que se generan en el dendrograma y cortar ahí; la otra es determinar los grupos de acuerdo a un número tentativo seleccionado con anterioridad.

Se aplicó este análisis de conglomerados a las pendientes que se obtuvieron de las regresiones lineales. El procedimiento, que se implementó a través de R, fue primero generar la matriz de distancias euclidianas, ecuación (4.3), para después emplearla en el algoritmo de Agrupamiento Jerárquico de Aglomeración. El dendrograma obtenido, usando un agrupamiento completo (distancia máxima), se observa en la figura 4.2.

Sobre la figura se han sobrepuesto recuadros redondeados para indicar los estados que pertenecen a cada una de las regiones.

Con esta regionalización las primas de seguros de los estados en una misma zona no tienen un comportamiento homogéneo, por lo que se consideró hacer un refinamiento de esta regionalización.

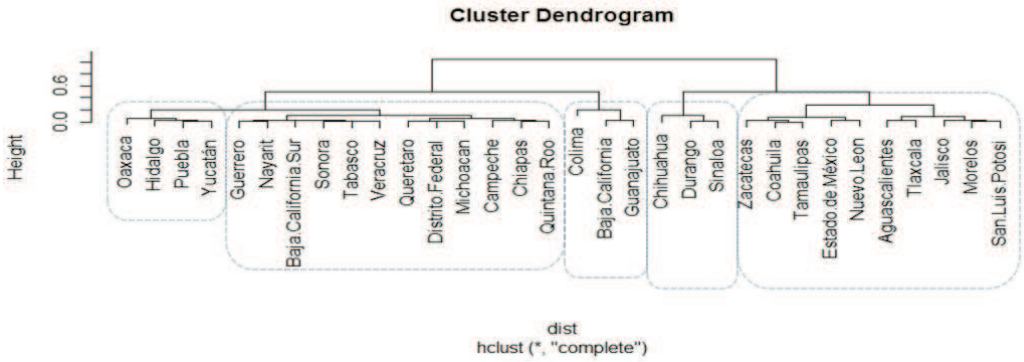


Figura 4.2: Dendrograma que resulta de aplicar el algoritmo de Agrupamiento Jerárquico de Aglomeración, donde muestra la clasificación de los estados en 5 zonas.

4.1.3. Obtención de series mensuales y clasificación final

Para calcular los porcentajes del robo de autos y posteriormente hacer la regionalización de los estados de la república se requirió tener el número total de autos asegurados. Esta información no se tenía en las bases de datos proporcionadas por la AMIS, y era necesario obtenerla; Sin embargo, como la vigencia de las pólizas no inicia a un mismo tiempo, era difícil realizar un conteo minucioso, por la cantidad de información involucrada y se buscó una manera de estimar este valor. En un principio se pensó aplicar varios cortes al mes y contar el número de vehículos asegurados en cada uno de esos momentos y con esas mediciones generar una mejor estimación del total de autos asegurados por mes, pero la cantidad de cálculos, filtros y operaciones requeridas, hizo que se desechara la idea. La estrategia seleccionada, fue hacer un solo corte al 15 de cada mes para determinar cuántas pólizas estaban vigentes en esa fecha; este número de vehículos asegurados se utilizó como una estimación de las unidades expuestas. Además, por la misma razón, se decidió emplear solamente las compañías más importantes del sector, pues la búsqueda de datos era necesaria implementarla en las bases de cada compañía, lo que resultaría en un arduo trabajo de extracción que llevaría mucho tiempo aplicar.

Una vez que se contaba con la estimación del total de vehículos asegurados, se procedió a generar las series de los indicadores mensuales del porcentaje de robo y de la prima de riesgo por estado.

Al revisar los datos de los estados que pertenecen a un mismo grupo se encontró que en términos generales el comportamiento de las gráficas que corresponden a cada estado dentro de una misma zona es el mismo (por su tipo de crecimiento); sin embargo, la cantidad de prima mensual llegaba a diferir considerablemente entre algunos estados que conformaban la misma región. Por lo tanto, se decidió reagrupar dentro de la misma zona los datos de series con primas similares, para evitar algún sesgo de la información.

Se calculó el promedio de la prima mensual de cada estado del último año. Al analizar este promedio, se puede entender mejor el fenómeno. Por ejemplo, en la zona 3, de comportamiento estable, el promedio de primas mensual para Guerrero es de $142.37, mientras que en otros estados de la misma zona, como Campeche, Querétaro y Quintana Roo, el promedio es de $14.59, $23.39 y $23.54 respectivamente. Se ve que existe una diferencia considerable, por lo que la respectiva

fusión de la información, generaba que la prima mensual de Guerrero disminuyera demasiado y la de los otros estados aumentara considerablemente.

Para hacer un refinamiento de la clasificación de los estados, se determinó que los valores mínimo y máximo de los promedios de primas mensual dentro de una misma no difirieran en más de 40 pesos. Con fines de claridad, se muestra en la tabla 4.2 como quedó la nueva clasificación.

De esta forma se pasaban de 5 zonas, a 13 zonas. Para fines prácticos este número era elevado y alguna zonas quedaban con un solo estado, por lo que se decidió agruparlas y generar menos regiones.

Se pudieron fusionar las zonas 4C y 4D, porque a pesar de que la diferencia máxima fue de $46.3 pesos, al graficar los 4 estados se observó que durante todo el periodo las gráficas no estaban tan alejadas y poseían un comportamiento similar. Se intentó aplicar la misma solución en las zonas 3C y 3D, pues la diferencia era similar, de $46.6 pesos. Sin embargo al graficar el D.F. junto con Guerrero, se apreciaron regiones en el tiempo donde las series diferían considerablemente, esto es, su comportamiento era diferente, pues si bien están en la región estable, el D.F. tiene una pendiente negativa y Guerrero positiva.

Los grupos 2A y 4A, formados solamente por los estados de Yucatán y Aguascalientes, respectivamente, se integraron al grupo 3A, porque al revisar gráficamente los datos de estos estados se encuentra que tienen un comportamiento semejante. De esta manera, se reducía el número a 10 grupos. Finalmente Guerrero, también estaba cercano a la región de crecimiento moderado, por lo que se decidió graficar la serie con los estados cuyo promedio de prima era cercano, que eran Morelos, Tamaulipas y Estado de México. Se vio que la gráfica de Guerrero estaba notablemente cercana al comportamiento de las otras gráficas, por lo que se optó por agruparlas.

Con esto, quedaron 9 zonas, de las cuales sólo dos estaban compuestas por un solo estado: la 3C por el D.F. y la 5B por Sinaloa. La clasificación final, que se puede ver gráficamente en la figura 4.3, es la siguiente:

Clasificación final:

Zona $1A$: **Baja California, Colima y Guanajuato.**
Zona $2A$: **Hidalgo, Oaxaca y Puebla.**
Zona $3A$: **Aguascalientes, Baja California Sur, Campeche, Querétaro, Quintana Roo y Yucatán.**
Zona $3B$: **Chiapas, Michoacán, Nayarit, Sonora, Tabasco y Veracruz.**
Zona $3C$: **Distrito Federal.**
Zona $4A$: **Coahuila, Jalisco, Nuevo León, San Luis Potosí, Tlaxcala y Zacatecas.**
Zona $4B$: **Estado de México, Guerrero, Morelos y Tamaulipas.**
Zona $5A$: **Chihuahua y Durango.**
Zona $5B$: **Sinaloa.**

Aunque no es un fenómeno que esté tan marcado, se puede ver que en la clasificación hay cierta relación territorial, pues se observan varios grupos de estados vecinos que poseen el mismo color.

Además, se puede apreciar que las zonas 4 y 5 que indican un mayor crecimiento de robo están en el norte de la república, mientras que las que en el centro y sur hay más estados de los sectores

Estados	Pendientes	Promedio del último año (pesos)	Zona	Nueva zona
Colima	−0,480	23,90	1	1A
Baja California	−0,306	60,58	1	1A
Guanajuato	−0,279	23,55	1	1A
Oaxaca	−0,177	37,48	2	2B
Hidalgo	−0,139	62,38	2	2B
Puebla	−0,117	40,93	2	2B
Yucatán	−0,112	5,15	2	2A
Campeche	−0,092	14,59	3	3A
Quintana Roo	−0,072	23,54	3	3A
Chiapas	−0,068	36,46	3	3B
Querétaro	−0,046	23,39	3	3A
Michoacán	−0,039	59,66	3	3B
Distrito Federal	−0,033	96,38	3	3C
Baja California Sur	−0,013	11,98	3	3A
Sonora	−0,003	46,89	3	3B
Tabasco	0,001	46,77	3	3B
Veracruz	0,004	48,23	3	3B
Guerrero	0,010	142,99	3	3D
Nayarit	0,016	39,94	3	3B
Aguascalientes	0,046	20,14	4	4A
Tlaxcala	0,070	75,94	4	4B
Jalisco	0,098	68,88	4	4B
San Luis Potosí	0,122	52,19	4	4B
Morelos	0,129	121,86	4	4C
Estado de México	0,251	168,18	4	4D
Nuevo León	0,279	89,25	4	4B
Tamaulipas	0,306	140,80	4	4C
Coahuila	0,307	70,71	4	4B
Zacatecas	0,331	83,39	4	4B
Chihuahua	0,450	158,39	5	5A
Sinaloa	0,545	343,96	5	5B
Durango	0,555	144,54	5	5A

Cuadro 4.2: En la última columna se muestra la nueva clasificación. El orden de aparición de los estados está determinado por el valor obtenido de su pendiente.

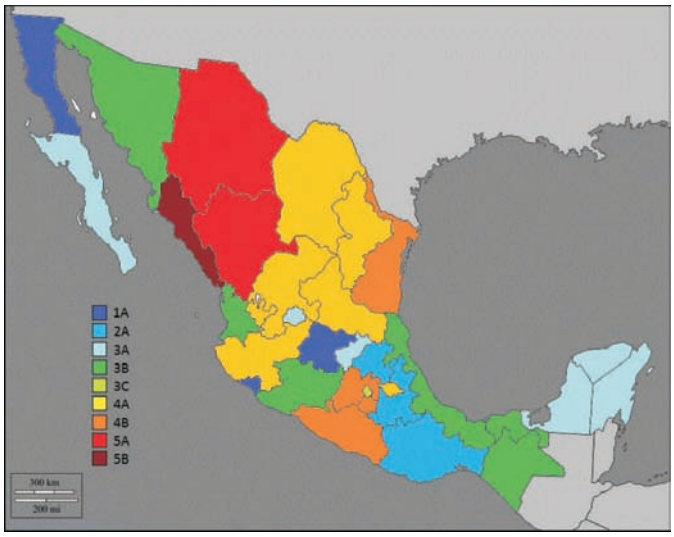


Figura 4.3: Clasificación final de los estados.

de estabilidad y disminución del delito. El contraste visual se facilita, pues se escogieron colores cálidos para indicar crecimiento de robo, y colores fríos para denotar su disminución.

4.2. Modelación de los indicadores

Una vez establecida la regionalización, se procedió al análisis de los indicadores siguiendo con detalle la metodología presentada en el capítulo 2, con el objetivo de identificar cuál de los procesos generadores de los datos introducidos en el capítulo 3 modelaba cada una de las series en cuestión.

4.2.1. Indicadores por región

Una vez formulada la clasificación, se necesitaban obtener las series tanto de prima de riesgo mensual como de porcentaje de robo; al definirse 9 grupos, se tenían entonces que calcular 18 series temporales mensuales, las cuales se muestran en las tablas 4.3 y 4.4

Para describir el comportamiento de los indicadores mensuales de todas las regiones de robo de autos, tanto del porcentaje como de la prima de seguros, se eligió el modelo ARIMA. Para algunas regiones (en particular la 2 y la 3) se encontró que con las correlaciones y las autocorrelaciones muestrales se tenía evidencia que las series eran estacionarias, en estos casos se utilizó el modelo ARMA para estimar su comportamiento. En los otros casos se observó tendencias no constantes en su comportamiento (en particular en las zonas 3, 4 y 5) por lo que se decidió conveniente estimar modelos ARIMA en estos casos. En la primera zona se observaron intervenciones, que cambiaron la evolución del nivel de la serie; se necesitó integrarlas en los modelos para estimar su efecto. Finalmente, en varias series temporales se manifestaron datos atípicos o aberrantes y su análisis y estimación fue fundamental en el proceso de modelado.

Tiempo	1A	2A	3A	3B	3C	4A	4B	5A	5B
2008-01	68.72	25.38	9.31	36.54	65.98	27.69	82.69	39.30	138.98
2008-02	47,96	38,91	10,13	23,03	67,20	25,63	63,99	59,68	94,55
2008-03	53,14	36,12	9,54	24,97	75,56	19,45	69,86	51,59	113,75
2008-04	49,92	26,84	14,32	23,80	78,61	24,04	77,14	36,72	131,08
2008-05	54,75	46,57	10,20	42,62	66,32	26,68	73,66	44,44	108,94
2008-06	47,89	47,89	17,59	31,92	62,64	29,00	78,51	59,31	103,60
2008-07	65,76	66,54	14,70	22,43	73,89	31,35	72,84	75,96	110,81
2008-08	64,23	54,63	14,20	33,59	75,07	32,43	71,09	99,25	59,34
2008-09	65,89	36,42	8,72	28,30	66,96	36,96	79,43	102,50	137,44
2008-10	58,38	61,33	8,62	49,08	86,79	34,86	80,93	114,39	214,39
2008-11	72,76	50,70	11,06	38,99	99,63	33,64	104,87	119,03	210,21
2008-12	76,04	44,89	15,63	46,06	95,19	35,61	89,46	135,57	345,35
2009-01	50,87	48,91	17,17	45,59	106,90	42,94	103,47	128,64	254,89
2009-02	55,02	40,45	11,79	38,38	90,07	42,57	83,86	112,05	221,09
2009-03	53,59	48,03	17,17	37,85	100,87	41,98	101,96	82,36	302,75
2009-04	60,58	29,54	14,84	31,93	91,42	39,72	89,23	103,83	189,64
2009-05	73,47	67,26	16,57	30,52	98,10	46,00	103,95	115,68	172,15
2009-06	69,97	42,02	12,90	30,79	96,72	36,01	112,09	92,48	150,44
2009-07	66,49	38,09	12,65	33,62	92,61	44,03	111,10	104,87	139,14
2009-08	57,89	44,44	18,88	38,25	97,66	47,67	110,61	107,34	160,99
2009-09	80,08	36,63	16,48	27,35	91,95	45,06	114,91	119,20	191,56
2009-10	70,27	37,36	17,36	34,21	103,97	54,04	124,42	126,27	208,72
2009-11	71,45	42,37	19,93	37,66	101,10	57,35	123,35	111,20	202,49
2009-12	83,57	45,30	17,63	41,96	101,37	72,78	122,71	122,43	305,45
2010-01	51,40	51,13	22,28	38,18	100,08	60,09	128,95	111,90	349,31
2010-02	30,29	29,16	17,73	39,70	91,05	51,17	123,67	112,01	260,79
2010-03	28,38	29,37	15,61	40,17	82,18	55,68	141,74	155,50	360,24
2010-04	34,32	50,92	16,13	33,85	78,72	59,78	143,31	96,05	335,65
2010-05	39,31	54,05	14,99	36,98	94,87	79,15	165,22	132,71	324,82
2010-06	29,20	47,95	14,91	41,44	101,46	82,46	150,81	114,11	334,22
2010-07	26,24	35,75	18,93	44,97	105,86	80,80	149,49	154,44	344,23
2010-08	30,36	55,59	15,52	47,91	106,08	91,85	176,93	202,76	322,52
2010-09	27,31	43,25	20,38	74,31	96,39	78,37	174,91	178,49	309,21
2010-10	39,67	42,41	20,85	53,78	102,51	91,16	186,58	221,15	346,54
2010-11	40,21	45,93	17,13	61,31	106,67	101,29	174,51	191,00	394,13
2010-12	34,47	52,46	14,16	70,46	90,74	90,53	169,27	196,32	445,80

Cuadro 4.3: Datos de las series de prima mensual por regiones; los valores son en pesos mexicanos.

Tiempo	1A	2A	3A	3B	3C	4A	4B	5A	5B
2008-01	0.0961	0.0349	0.0188	0.0379	0.0779	0.0346	0.0780	0.0636	0.0887
2008-02	0.0778	0.0381	0.0161	0.0300	0.0781	0.0350	0.0677	0.0728	0.0772
2008-03	0.0753	0.0473	0.0178	0.0318	0.0788	0.0310	0.0805	0.0760	0.0802
2008-04	0.0859	0.0380	0.0215	0.0279	0.0841	0.0319	0.0727	0.0655	0.0792
2008-05	0.0926	0.0431	0.0182	0.0390	0.0788	0.0342	0.0699	0.0714	0.0918
2008-06	0.0789	0.0388	0.0218	0.0337	0.0770	0.0370	0.0758	0.0921	0.0688
2008-07	0.0915	0.0440	0.0203	0.0284	0.0808	0.0397	0.0716	0.0872	0.0956
2008-08	0.0937	0.0471	0.0174	0.0331	0.0839	0.0389	0.0732	0.1211	0.0594
2008-09	0.0923	0.0372	0.0153	0.0273	0.0743	0.0415	0.0839	0.1127	0.1198
2008-10	0.0859	0.0385	0.0142	0.0335	0.0895	0.0419	0.0881	0.1165	0.1423
2008-11	0.0903	0.0382	0.0154	0.0338	0.1096	0.0407	0.0954	0.1189	0.1324
2008-12	0.0810	0.0374	0.0139	0.0321	0.0926	0.0400	0.0855	0.1298	0.1799
2009-01	0.1028	0.0528	0.0279	0.0476	0.1395	0.0606	0.1228	0.1446	0.1665
2009-02	0.0978	0.0495	0.0256	0.0502	0.1275	0.0560	0.1133	0.1181	0.1470
2009-03	0.1020	0.0660	0.0279	0.0405	0.1446	0.0581	0.1240	0.1070	0.1637
2009-04	0.0932	0.0426	0.0331	0.0377	0.1227	0.0571	0.1206	0.1024	0.1079
2009-05	0.1177	0.0642	0.0235	0.0372	0.1361	0.0596	0.1269	0.1218	0.1137
2009-06	0.1039	0.0542	0.0285	0.0368	0.1363	0.0565	0.1290	0.1244	0.0940
2009-07	0.1122	0.0531	0.0180	0.0433	0.1414	0.0645	0.1367	0.1397	0.1138
2009-08	0.1074	0.0610	0.0228	0.0427	0.1368	0.0659	0.1408	0.1408	0.1081
2009-09	0.1353	0.0504	0.0252	0.0361	0.1357	0.0625	0.1482	0.1227	0.1192
2009-10	0.1217	0.0487	0.0288	0.0430	0.1478	0.0747	0.1434	0.1518	0.1259
2009-11	0.1068	0.0440	0.0287	0.0376	0.1387	0.0704	0.1445	0.1339	0.1260
2009-12	0.1023	0.0357	0.0221	0.0422	0.1125	0.0702	0.1183	0.1427	0.1628
2010-01	0.0412	0.0422	0.0189	0.0366	0.0941	0.0611	0.1112	0.1425	0.1885
2010-02	0.0332	0.0326	0.0182	0.0394	0.0886	0.0484	0.1085	0.1347	0.1643
2010-03	0.0398	0.0289	0.0184	0.0382	0.0839	0.0592	0.1252	0.1461	0.2165
2010-04	0.0378	0.0395	0.0199	0.0334	0.0842	0.0553	0.1147	0.1291	0.2130
2010-05	0.0429	0.0391	0.0191	0.0365	0.0935	0.0695	0.1328	0.1443	0.1900
2010-06	0.0392	0.0341	0.0226	0.0361	0.0982	0.0770	0.1324	0.1285	0.2160
2010-07	0.0275	0.0302	0.0222	0.0399	0.0991	0.0822	0.1337	0.1816	0.2159
2010-08	0.0326	0.0402	0.0192	0.0432	0.0961	0.0864	0.1461	0.2016	0.1859
2010-09	0.0334	0.0339	0.0173	0.0588	0.0912	0.0733	0.1441	0.1869	0.1862
2010-10	0.0437	0.0329	0.0208	0.0470	0.0959	0.0865	0.1547	0.2040	0.1928
2010-11	0.0411	0.0409	0.0164	0.0557	0.0940	0.0918	0.1496	0.1930	0.2298
2010-12	0.0398	0.0427	0.0209	0.0608	0.0859	0.0818	0.1461	0.1806	0.2383

Cuadro 4.4: Datos de las series de porcentaje de robo mensual de vehículos de cada una de las zonas.

Con la finalidad de hacer ágil la presentación de los resultados, se decidió escoger una serie con cada una de las diferentes técnicas aplicadas. Los resultados para las zonas que no se presentan sus resultados de manera detallada, se incluirán en el apéndice A.

4.2.2. Modelado de algunas series representativas

La explicación minuciosa del modelado se realizará de la siguiente manera. Primero se estudiará una zona cuyos datos indican ser estacionarios, por lo que se aplicó el modelo ARMA; después se detallarán los pasos para un modelo que contiene una raíz unitaria. Posteriormente se analizará un proceso con tendencia lineal. Después se explicaran algunos de los pasos importantes para modelar datos atípicos. Al final se mencionará un caso donde se presentó una intervención y cuál fue el proceso de estimación y modelado.

4.2.3. Proceso estacionario

Se seleccionó la serie temporal de la zona $2A$ del porcentaje de robo mensual como ejemplo del procedimiento de modelación de un proceso ARMA.

1. **Gráfica de la serie Temporal**

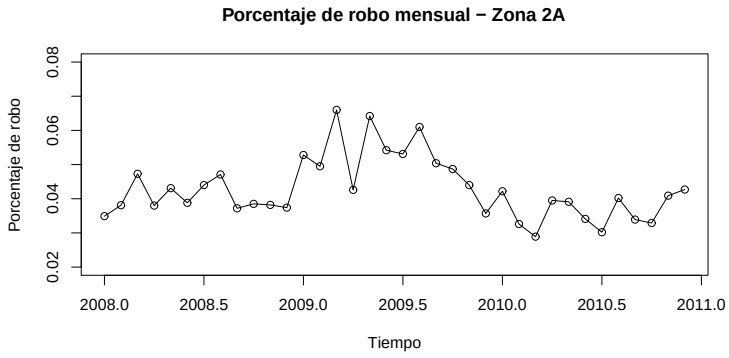
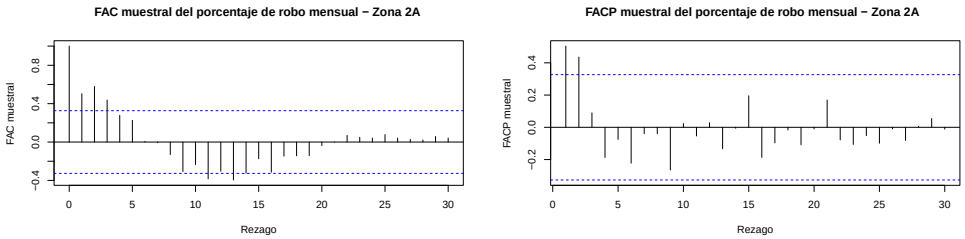


Figura 4.4: Serie temporal del porcentaje de robo mensual de la zona $2A$.

Respecto al nivel de la media, se observa que la serie de datos se comporta de forma estacionaria. Referente a la varianza, se percibe también una conducta estacionaria, lo que significa que es posible que no sea necesaria ninguna transformación para estabilizarla. En términos generales todos los datos aparentan ser regulares, aunque el del $2009 - 04$ podría ser atípico.

2. **FACM y FACPM de la serie temporal**

Las gráficas de la figura 4.5 confirman la estacionariedad del proceso. La FAC muestral exhibe un amortiguamiento senoidal, lo que indica la presencia de raíces complejas en el polinomio autoregresivo, por lo que al menos el polinomio es de orden 2. Este análisis, más



(a) FACM (b) FACPM

Figura 4.5: Gráficas de FACM y de FACPM de la serie de porcentaje de robo mensual de la zona 2*A*.

los dos rezagos significativos de la FACP muestral, sugieren un posible modelo ARMA(2,0) como el proceso generador de la serie.

3. **Prueba de raíz unitaria**

   Aparte de lo que se puede inferir de las gráficas de $\widehat{\rho}$ y $\widehat{\phi}$ respecto a la estacionariedad, es conveniente realizar una prueba estadística para verificar la hipótesis. Se implementa la prueba de ADF con hipótesis alternativa de un proceso estacionario con media distinta de cero (sección 3.1.1).

   Se emplea un valor alto para el número de rezagos (12 para ser precisos), y el algoritmo implementado en R solamente deja aquellos que son significativos mínimo al 10 %. El resultado de la prueba arroja un valor p de 0,022 por lo que la hipótesis nula de raíz unitaria se rechaza hasta el 2,5 %. Este resultado confirma la estacionariedad.

4. **Subconjunto óptimo**

   Para la búsqueda de un subcobjunto óptimo se aplica la técnica vista en la sección 3.3.3, la cual combina el método de Hannan y Rissanen [21] con la regresión de pasos agigantados[I] de Furnival y Wilson [22].

   Se busca ajustar uno proceso ligeramente mayor en la parte AR, lo que permitirá confirmar que el modelo propuesto incluye los rezagos que son realmente significativos.

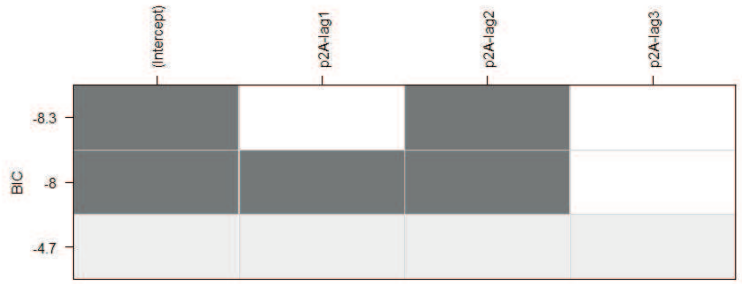


Figura 4.6: Gráficas de FACM y de FACPM de la serie de porcentaje de robo mensual de la zona 2*A*.

[I]La regresión de pasos agigantados (regresión by leaps and bounds) es un algoritmo que busca encontrar el subconjunto óptimo en una regresión lineal.

Parámetros	ϕ_1	ϕ_2	μ
Estimación	0.2815	0.4284	0.0422
Error estándar	0.1452	0.1454	0.0036

Cuadro 4.5: Parámetros estimados del modelo ARMA(2,0) para el porcentaje de robo de la zona $2A$

Parámetros	ϕ_1	ϕ_2	ϕ_3	μ
Estimación	0.2410	0.4056	0.0829	0.0421
Error estándar	0.1670	0.1525	0.1702	0.0038

Cuadro 4.6: Sobreajuste en la dirección autoregresiva: modelo ARMA(3,0).

Se valida el modelo propuesto ARMA(2,0) , ya que el rezago 2 está en los dos modelos que poseen el menor BIC, mientras que el retardo 3 no aparece en éstos. Para la estimación, se decide incorporar el rezago 1 pues está incluido en el segundo modelo, cuyo BIC es cercano al valor del primer modelo.

5. **Estimación de los parámetros**

   La estimación se implementó en el software R, empleando la técnica estadística de máxima verosimilitud. Los resultados se muestran en la tabla 4.5

   Se observa que todos los términos son significativos respecto a su error estándar, por lo que se procede al diagnóstico del modelo.

6. **Sobreajuste**

   Se realizan dos sobreajustes, uno en la dirección autoregresiva y el otro en la de los promedios móviles, como se sugiere en la sección 3.3.2. Los valores obtenidos se muestran en las tablas 4.6 y 4.7.

   Los parámetros del modelo no varían de forma relevante en cada uno de los sobreajustes y los que se incorporaban no eran significativamente distintos de cero, lo que es favorable al proceso generador propuesto.

7. **Análisis residual**

   Para validar el supuesto de normalidad, se obtuvieron las gráficas de los residuales contra el tiempo y la de los cuantiles teóricos de una normal respecto a los muestrales, que se muestran en la figura 4.7.

   El comportamiento de los residuales estandarizados respecto al tiempo es de forma aleatoria, sin ningún patrón específico. En la gráfica de los cuantiles, la mayoría de los puntos

Parámetros	ϕ_1	ϕ_2	ϕ_3	μ
Estimación	0.3561	0.3919	-0.0956	0.0421
Error estándar	0.2646	0.1889	0.2722	0.0037

Cuadro 4.7: en la dirección de promedios móviles: modelo ARMA(2,1).

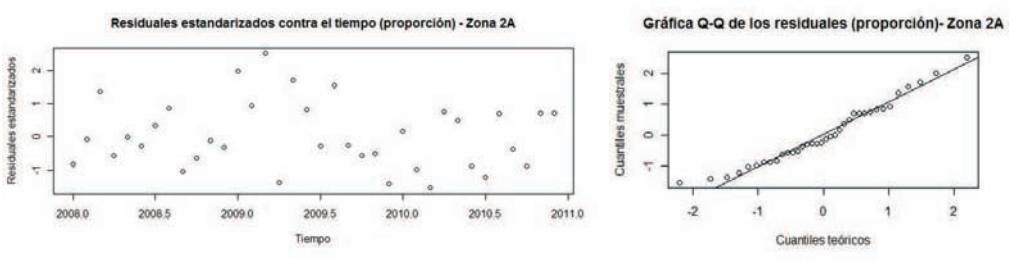


Figura 4.7: Gráficas para verificar la normalidad de los residuales.

caen sobre la línea de normalidad, aunque hay uno que se aleja un poco. Para darle formalidad al apoyo visual, se aplica la prueba de Shapiro Wilk(Verzani [32]); se obtiene un valor de p de 0.2758, por lo que no hay evidencia para rechazar la hipótesis nula de normalidad.

Para investigar la correlación de los residuales se calcula la función de autocorrelación muestral (figura 4.8). El comportamiento de la gráfica es similar al ruido blanco, lo que confirma que los residuales no están correlacionados.

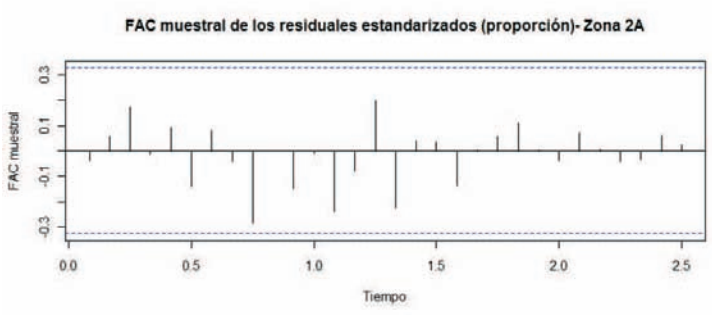


Figura 4.8: Función de autocorrelación muestral de los residuales del modelo ajustado.

8. **Búsqueda de valores atípicos**

   A pesar de que al inicio del análisis se pensó que el valor de 2009-04 podría ser un dato aberrante, al aplicar algoritmos en R para su búsqueda, se obtuvieron resultados negativos. Es posible que el valor que está un poco alejado de la línea de normalidad en la gráfica Q-Q sea este dato.

9. **Modelo ajustado y pronóstico del último trimestre**

   Para ver cómo se comporta el modelo, en la figura 4.9, se muestra una gráfica del modelo ajustado (puntos en azul) respecto a los datos originales (puntos en negro). Se quiso ver como se comportaban los pronósticos del modelo, realizando predicciones en un periodo que contiene valores observados, por lo que en la gráfica también se incluyen los pronósticos del último trimestre[II] (puntos en rojo).

   Los valores ajustados no distan mucho de los observados. El pronóstico del último trimestre es aceptable, ya las observaciones están dentro de los intervalos de confianza al 95 % (puntos en verde).

[II]Se decide solamente considerar el último trimestre, porque no se tienen muchos datos, o que modifica las estimaciones de los parámetros.

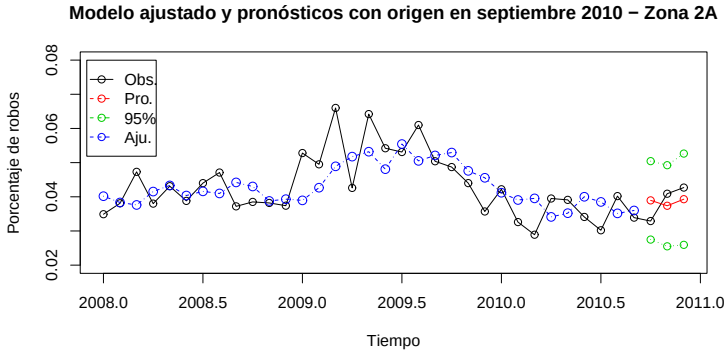


Figura 4.9: Modelo ajustado y pronósticos del último trimestre.

11-1	11-2	11-3	11-4	11-5	11-6	11-7	11-8	11-9	11-10	11-11	11-12
0.0418	0.0423	0.0420	0.0422	0.0421	0.0422	0.0421	0.0421	0.0421	0.0421	0.0421	0.0422

Cuadro 4.8: Pronósticos de cada uno de los meses del 2011 del modelo propuesto ARMA(2,0).

10. **Pronóstico de un año**

Por último se realiza el pronóstico de todo el año 2011. En la figura 4.10 se muestra la serie temporal con las predicciones de 12 meses por delante con sus respectivos intervalos de confianza al 95 %.

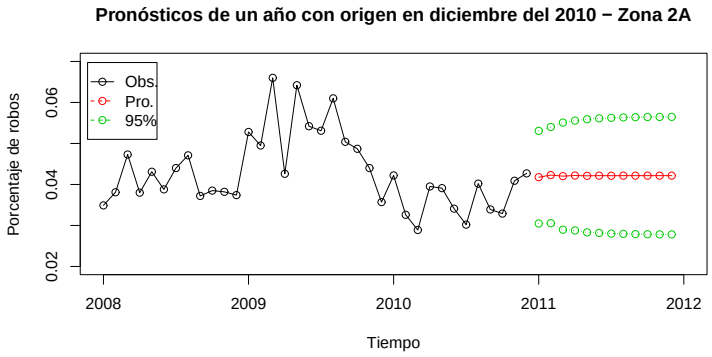


Figura 4.10: Pronósticos mensuales de un año para la serie temporal del porcentaje de robo de la zona 2A. El proceso generado de los datos se identificó como ARMA (2,0).

Los valores de las predicciones de los respectivos meses están en la tabla 4.8.

Finalmente, para obtener una estimación anual, se suman las predicciones de la tabla anterior, obteniéndose un valor del porcentaje de robo para el 2011 de ***0.5054*** con una desviación estándar de ***0.0989***, que se calcula sumando los valores de los errores estándar de los pronósticos mensuales.

4.2.4. Proceso no estacionario (raíz unitaria)

Se escogió la serie temporal de la zona $5A$ del porcentaje de robo mensual para ejemplificar la metodología de modelación de un proceso ARIMA.

Metodología:

1. **Gráfica de la serie Temporal**

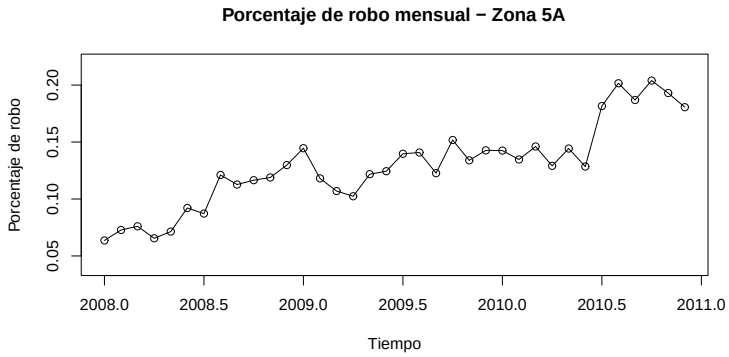
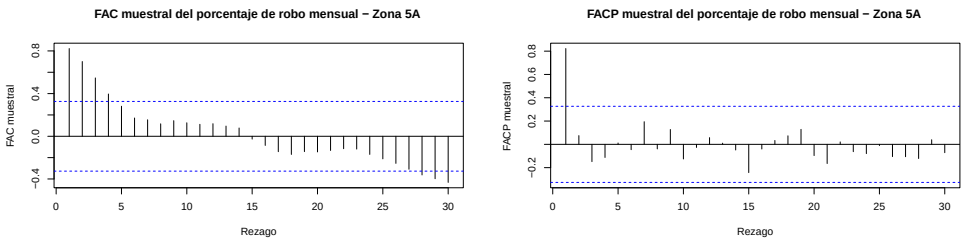


Figura 4.11: Serie temporal del porcentaje de proporción de robo mensual de la zona $5A$.

Se percibe una tendencia creciente en la serie, es decir, se tiene un proceso no estacionario en la media. Por otro lado, la varianza se comporta de forma estable, por lo que se trabajarán con los datos sin transformar. Aparentemente no hay datos aberrantes, aunque el 2010-07 podría ser irregular.

2. **FACM y FACPM de la serie temporal**



(a) FACM (b) FACPM

Figura 4.12: FACM y FACPM de la serie de porcentaje de robo mensual de la zona $5A$.

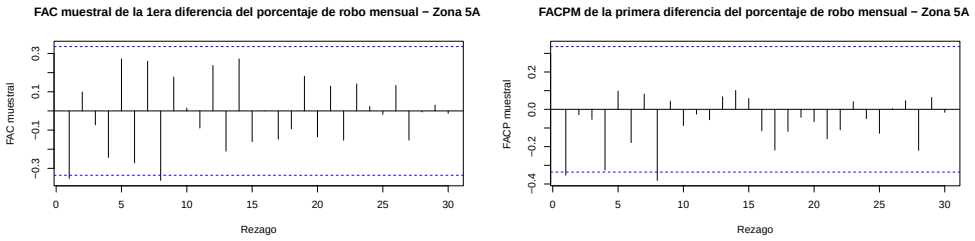
La tendencia mostrada en las gráficas de la figura 4.12, muestra un comportamiento no estacionario: decaimiento lineal de la FACM y un primer retardo grande en la FACPM. Hay que verificar si este comportamiento es debido a alguna raíz unitaria en el proceso o se presenta estacionariedad respecto a una tendencia lineal.

3. **Prueba de raíz unitaria**

Se aplica la prueba ADF para ver si el proceso tiene una raíz unitaria, con la hipótesis alternativa de un proceso estacionario respecto a una tendencia lineal.

De igual forma como ocurrió con la prueba ADF del proceso estacionario que se modeló en la sección anterior, se usó un valor grande de rezagos (10 en este caso), ya que el algoritmo en R deja aquellos que son significativos mínimo al 10 %. Se obtiene un valor p que es mayor a 0.1, por lo que no hay evidencia para rechazar la hipótesis de raíz unitaria en el proceso.

4. **FACM y FACPM de la primera diferencia de la serie temporal**



(a) FACM (b) FACPM

Figura 4.13: Gráficas de la FACM y FACPM de la primera diferencia de la serie.

Aunque no está tan marcado, en la FAC muestral se percibe un amortiguamiento senoidal; el primer rezago es significativo (-0.352), después hay un decaimiento (2,3 rezagos) y después comienza a crecer de nuevo (4,5,6,7,8) rezagos para después repetir el mismo proceso. Cabe destacar que el rezago 8 es significativo (-0.363). Por otro lado en la FACP muestral se aprecia que el retraso 1 y 8 son significativos (-0.352 y -0.381 respectivamente) y el 4 es muy alto (-0.352). Por todo lo anterior, es posible que el modelo tenga hasta el retardo 8 en la parte AR. El modelo propuesto es un ARIMA (8,1,0) o algún subconjunto de él.

5. **Subconjunto óptimo**

De nueva cuenta se aplica la técnica que combina el método de Hannan y Rissanen [21] con la regresión de pasos agigantados de Furnival y Wilson [22] para encontrar el subconjunto óptimo.

Se busca el subconjunto tomando un modelo ligeramente mayor en la parte AR, digamos un ARIMA (9,1,0), para confirmar que se incluyen los rezagos adecuados.

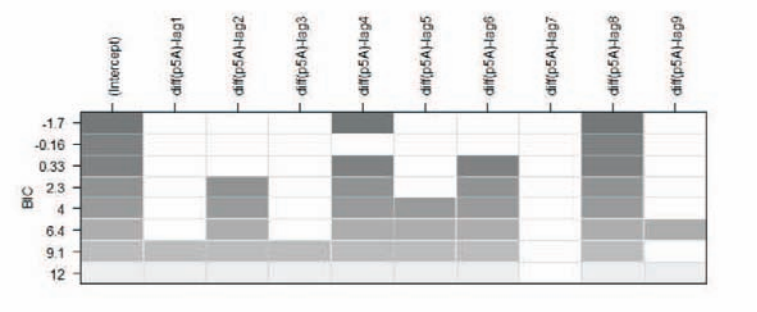


Figura 4.14: Subconjunto óptimo para el modelo ARIMA (9,1,0).

Parámetros	ϕ_4	ϕ_8	μ
Estimación	-0.3355	-0.5673	0.0031
Error estándar	0.1445	0.1556	0.0014

Cuadro 4.9: Estimaciones del subconjunto ARMA (8,1,0) para el porcentaje de robo de la zona $5A$.

Parámetros	ϕ_4	ϕ_8	ϕ_9	μ
Estimación	-0.3683	-0.5136	0.2114	0.0033
Error estándar	0.1409	0.1609	0.1648	0.0015

Cuadro 4.10: Sobreajuste en la dirección autoregresiva: subconjunto de modelo ARIMA (9,1,0).

Los parámetros que aparecen en los primeros modelos son el ϕ_4 , ϕ_8 y μ ; los demás no se consideraran en el modelado, a excepción del que aparece desde el tercer subconjunto, con un valor de BIC cercano al del segundo modelo. Con esto se valida que el proceso generador es un subconjunto del proceso ARIMA (8,1,0).

6. **Estimación de los parámetros**

   Al estimar el subconjunto propuesto, se vio que el parámetro ϕ_6 no era significativamente distinto de cero, por lo que tuvo que ser removido, y se reestimo el modelo con los términos restantes. Los resultados de la estimación de máxima verosimilitud (implementados en R) están en la tabla 4.9.

   Se observa que todos los términos son significativos respecto a su error estándar, por lo que se procede al diagnóstico del modelo a través del sobreajuste y el análisis residual.

7. **Sobreajuste**

   Se realizan dos sobreajustes, uno en la dirección AR y el otro en la MA, como se muestra en las tablas 4.10 y 4.11.

   Los parámetros del modelo no varían de forma relevante en cada uno de los sobreajustes, lo que es un buen indicador. En los dos casos se puede ver que los parámetros aumentados tienen un valor semejante a su error estándar, por lo que no existe evidencia suficiente para rechazar que estos parámetros sean diferentes de cero; en estas circunstancias se decide continuar con el primer modelo.

8. **Análisis residual**

Parámetros	ϕ_4	ϕ_8	ϕ_1	μ
Estimación	-0.3738	-0.5053	-0.3597	0.0032
Error estándar	0.1504	0.1695	0.2108	0.0009

Cuadro 4.11: Sobreajuste en la dirección de promedios móviles: subconjunto de modelo ARIMA(8,1,0).

Se elaboraron las gráficas de los residuales contra el tiempo y la de los cuantiles teóricos de una normal respecto a los muestrales, para validar la suposición de normalidad (figura 4.15).

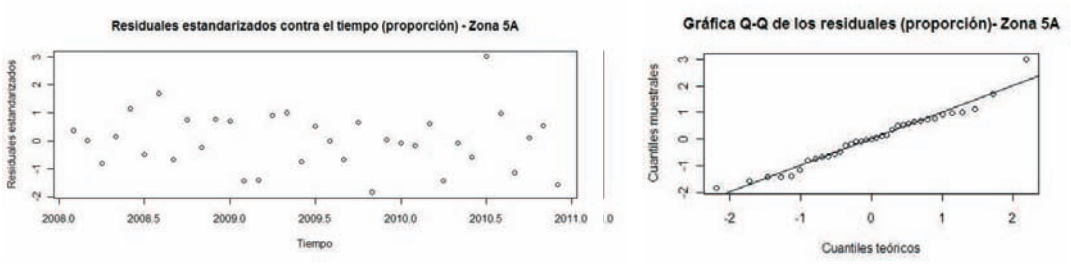


Figura 4.15: Gráficas para validar la normalidad de los residuales.

En la gráfica de los residuales estandarizados contra el tiempo no se observa ninguna tendencia marcada. En la otra, casi todos los puntos caen sobre la línea que marca la tendencia normal, aunque hay uno que está relativamente lejos. La prueba de Shapiro Wilk arroja un valor de p de 0.3698 por lo que no se puede rechazar la hipótesis de normalidad.

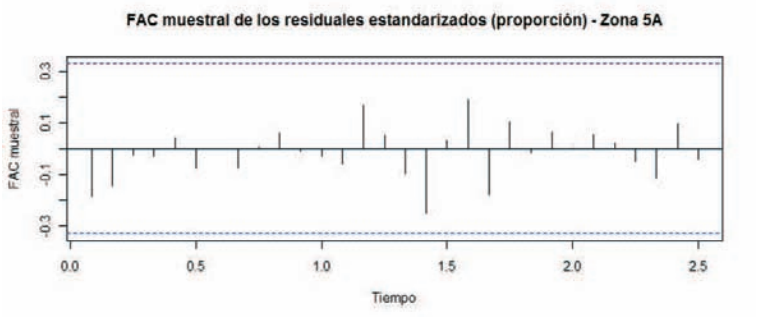


Figura 4.16: Función de autocorrelación muestral de los residuales del modelo ajustado.

Se calcula la función de autocorrelación muestral (figura 4.16). Efectivamente, el comportamiento de la gráfica es similar al ruido blanco, lo que confirma que los residuales no están correlacionados.

9. **Búsqueda de valores atípicos**

   Se implementó la búsqueda de valores irregulares del tipo AO y IO. No se encontró ninguno, a pesar de que se creía que el valor de 2010-07 podría ser un atípico.

10. **Modelo ajustado y pronósticos del último trimestre**

    Se busca comparar el modelo ajustado con la serie temporal (figura 4.17). De nueva cuenta, se realizan predicciones del último trimestre del 2010, para contrastarlas con los datos de ese periodo.

    Los valores ajustados (puntos azules) son cercanos a la serie (puntos negros). Los dos primeros pronósticos (puntos rojos) son muy, cercanos a las observaciones, mientras que el tercero no dista mucho. En general el pronóstico es aceptable, pues los tres datos están dentro del intervalo de confianza del 95 % (puntos verdes).

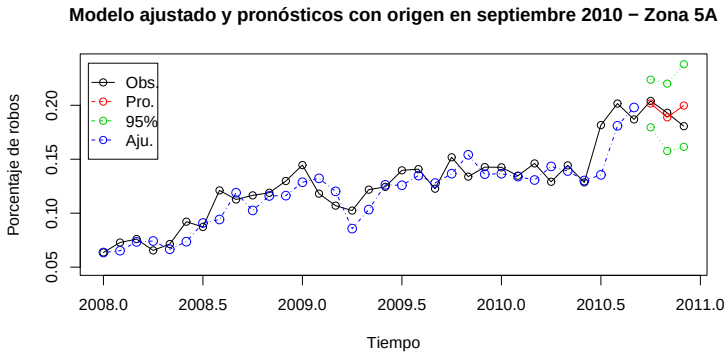


Figura 4.17: Modelo ajustado y pronósticos del último trimestre.

11-1	11-2	11-3	11-4	11-5	11-6	11-7	11-8	11-9	11-10	11-11	11-12
0.1823	0.1921	0.1735	0.1722	0.1866	0.1828	0.2004	0.2145	0.2204	0.2267	0.2423	0.2499

Cuadro 4.12: Pronósticos de cada uno de los meses del 2011 del modelo propuesto ARMA (2,0).

11. **Pronóstico de un año**

Se realizan las predicciones de 12 periodos por delante con origen en diciembre de 2010, que se muestran en la figura 4.18.

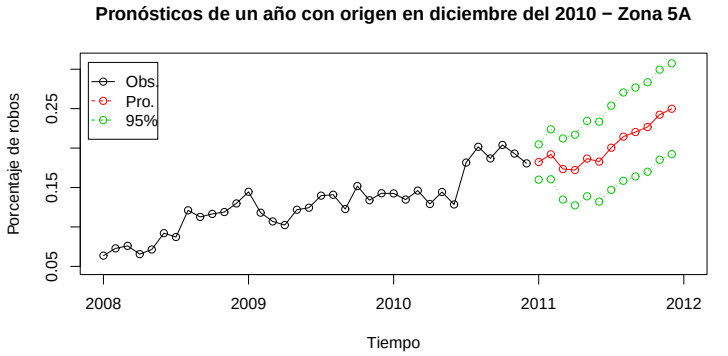


Figura 4.18: Pronósticos mensuales de un año para la serie temporal del porcentaje de robo de la zona 5A. El proceso generado se identificó como un subconjunto de un ARIMA (8,1,0).

Los pronósticos de los respectivos meses están en la tabla 4.12.

Por último, para obtener una estimación de todo el año, se suman las predicciones de la tabla anterior, calculándose un valor del porcentaje de robo del 2011 de ***2.4438*** con una desviación estándar de ***0.3486***.

4.2.5. Proceso estacionario respecto a una tendencia lineal

Para ejemplificar la modelación de un proceso estacionario respecto a una tendencia lineal, se eligió la serie del porcentaje de robo mensual de la zona $5B$.

Metodología:

1. **Gráfica de la serie Temporal**

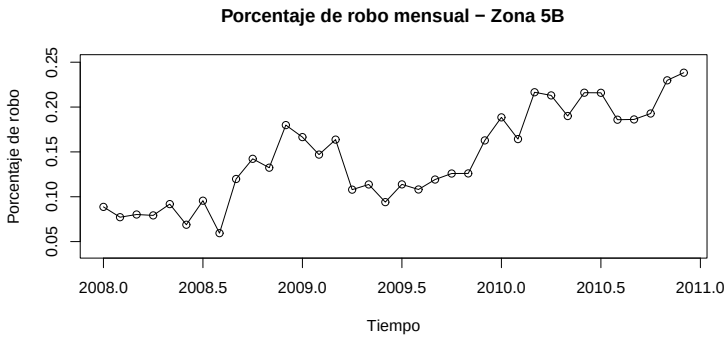
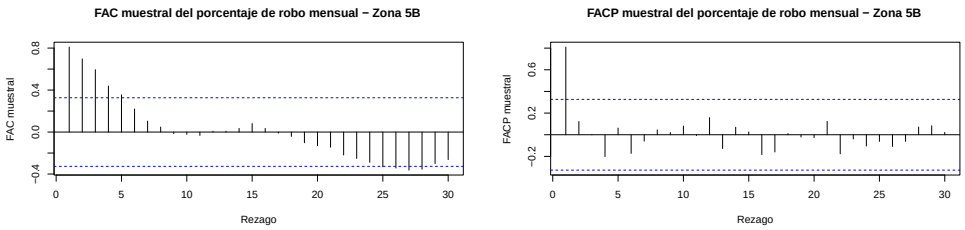


Figura 4.19: Serie temporal del porcentaje de robo de la zona $5B$.

Se observa un crecimiento en la serie, por lo que se tiene un proceso que no es estacionario(figura 4.19). No se percibe que se necesite transformar los datos para estabilizar la varianza. A simple vista no se ven datos irregulares, aunque se podría sospechar del punto del 2008-09.

2. **FACM y FACPM de la serie temporal**



(a) FACM (b) FACPM

Figura 4.20: FACM y FACPM de la serie de porcentaje de robo mensual de la zona $5B$.

El comportamiento de las gráficas de la figura 4.20 es el esperado, pues el proceso no es estacionario en media. Hay que determinar si esta conducta es debida a una tendencia estocástica o determinista.

3. **Prueba de raíz unitaria**

Se aplica la prueba de raíz unitaria de ADF tomando como proceso alternativo una tendencia lineal. Usando 6 rezagos y la misma condición de mantener aquellos que son significativos al 10 %, se obtuvo un valor p de 0.029; se establece que hay evidencia suficiente para

Parámetros	β_0	β_1
Estimación	-98.49	0.049
Error estándar	10.94	0.00542
Valor p	1.58e-10	1.53e-10

Cuadro 4.13: Parámetros estimados de la regresión lineal simple del porcentaje de robo de la zona $5B$.

rechazar la presencia de una raíz unitaria al 5 % de significancia, y considerar un proceso estacionario alrededor de una tendencia lineal.

Para eliminar la tendencia en esta serie se estima un modelo de regresión lineal simple con variable independiente, el tiempo (ecuación 2.89). A la serie original se le resta la ecuación estimada y con esto se obtiene una nueva serie que se espera sea estacionaria, y a esta serie se le modela con un proceso ARMA.

4. **Ajuste de tendencia lineal**

   El ajuste lineal, tomando como covariable el tiempo y variable dependiente el porcentaje de robo se observa en la tabla 4.13.

   Los parámetros son estadísticamente significativos, como se puede constatar por su error estándar y el valor p del estadístico t.

5. **Análisis de los residuales de la regresión lineal**

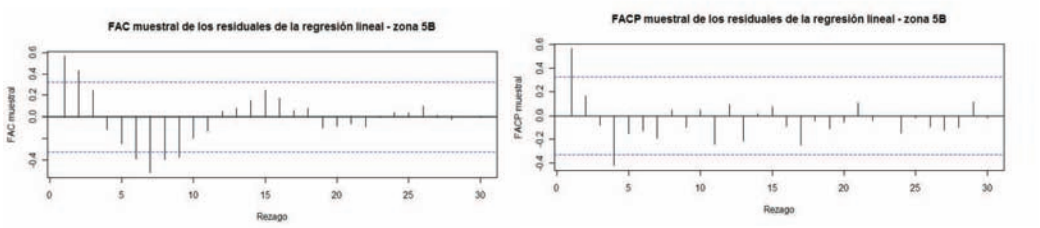


Figura 4.21: FACM y FACPM de los residuales del ajuste lineal.

Como se observa en la gráfica de la función de autocorrelación muestral (figura 4.21), los residuales están correlacionados. Se observa un comportamiento senoidal, lo que establece que existen raíces complejas, por lo que al menos se tienen dos términos autoregresivos. La FACP muestral posee dos retardos significativos, el primero (0.566) y el cuarto (-0.421). Sintetizando el análisis anterior, el proceso generador podría ser un AR con hasta 4 rezagos significativos.

6. **Subconjunto óptimo para el modelo ARMA de los residuales de la regresión lineal**

   Se busca el subconjunto óptimo para el modelo estacionario de los residuales (figura 4.22). De nueva cuenta, se realiza una búsqueda a partir de un modelo con un parámetro más, esto es un ARMA(5,0), para validar el proceso puesto.

   Efectivamente, el parámetro ϕ_5 no es significativo en los primeros cuatro modelos que son los minimizan el BIC, confirmando la hipótesis. El subconjunto propuesto incluye a

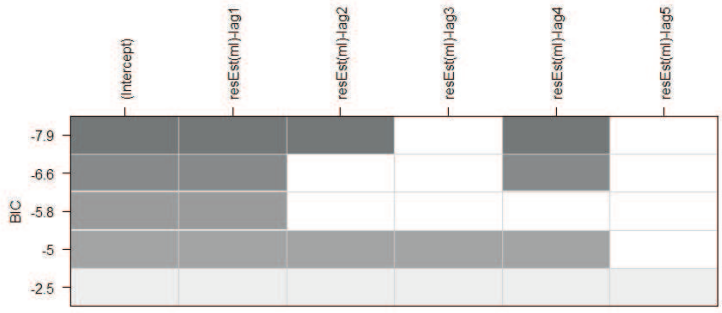


Figura 4.22: Subconjunto óptimo para el modelo ARMA(5,0) de los residuales

Parámetros	ϕ_1	ϕ_2	ϕ_4
Estimación	0.4668	0.3655	-0.3973
Error estándar	0.1421	0.1593	0.1277

Cuadro 4.14: Estimaciones del subconjunto del proceso ARMA (4,0) para los residuales de la regresión lineal de la serie del porcentaje de proporción de robo mensual de la zona $5B$.

los términos ϕ_1, ϕ_2, y ϕ_4, pues están en los primeros 4 modelos (aunque ϕ_2 no aparece en el tercer y cuarto, se decide integrarlo en la estimación, pues está en el primero). El parámetro ϕ_3 queda fuera de la estimación. Por los resultados anteriores se elige ajustar un subconjunto del proceso ARMA(4,0) a los residuales de la regresión lineal.

7. **Estimación de los parámetros del modelo ARMA a los residuales de la regresión lineal**

   En la estimación del subconjunto propuesto, el término μ no fue significativamente distinto de cero, por lo que se quitó del modelo. El nuevo subconjunto estimado, empleando máxima verosimilitud se aprecia en la tabla 4.14.

   Se ve que todos los parámetros son significativamente distintos de cero respecto a su error estándar, por lo que se procede al diagnóstico del modelo.

8. **Sobreajuste del modelo ARMA de los residuales de la regresión lineal** Se realizan los sobreajuste en la dirección autoregresiva y en la de promedios móviles. Los resultados se muestran en las tablas 4.15 y 4.16.

   El sobreajuste en la dirección de AR fue satisfactorio. En la dirección de MA se presentaron dos problemas: el término ϕ_2 no fue relevante y el parámetro θ_1 es significativo. A pesar de estos problemas, se continuará con el modelo, a reserva de lo que ocurra en el análisis residual.

Parámetros	ϕ_1	ϕ_2	ϕ_4	ϕ_5
Estimación	0.4238	0.3914	-0.3394	-0.0972
Error estándar	0.1562	0.1627	0.1568	0.1531

Cuadro 4.15: Sobreajuste en la dirección autoregresiva: modelo ARMA(5,0).

Parámetros	ϕ_1	ϕ_2	ϕ_4	θ_1
Estimación	0.7703	0.1871	-0.3763	-0.4252
Error estándar	0.2818	0.2476	0.1129	0.3229

Cuadro 4.16: Sobreajuste en la dirección de promedios móviles: modelo ARMA(4,1).

9. **Búsqueda de atípicos en el modelo ARMA para los residuales de la regresión lineal**

   No se encontró ninguno valor irregular en el proceso estacionario respecto a la tendencia lineal.

10. **Análisis residual del modelo ARMA de los residuales de la regresión lineal**

    Se elaboraron las gráficas de los residuales del ajuste del modelo ARMA del proceso estacionario alrededor de la tendencia lineal, para validar la suposición de normalidad (figura 4.23).

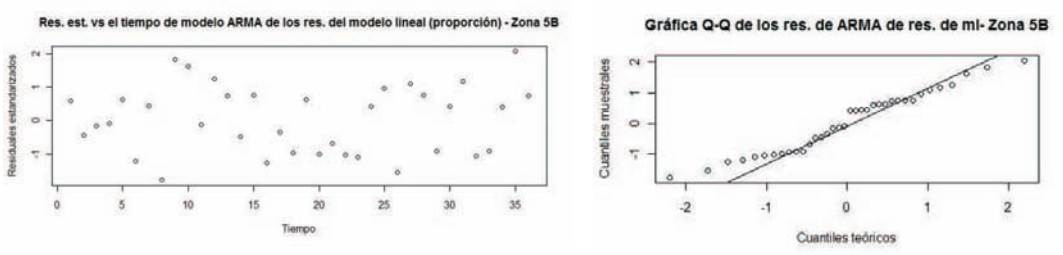


Figura 4.23: Gráficas para verificar la normalidad de los residuales del ajuste del modelo ARMA del proceso estacionario respecto a una tendencia lineal.

En la gráfica de los residuales contra el tiempo se observa un comportamiento aleatorio. La mayoría de los residuales caen sobre la línea de normalidad, aunque hay algunos que están ligeramente separados. La prueba de Shapiro-Wilk resulta en un valor p de 0.2237 por lo que no hay evidencia suficiente para rechazar la hipótesis nula de una distribución normal.

Se calcula la función de autocorrelación muestral para ver que realmente se tiene un comportamiento de ruido blanco (figura 4.24). Se observa que no hay correlación de los residuales del modelo ARMA.

11. **Estimación del modelo estacionario respecto a una tendencia lineal**

    Una vez que se tiene el proceso ARMA que modela la correlación de los residuales, hay que estimar todos los parámetros del modelo estacionario respecto a la tendencia lineal; al incluirlos todos, los valores difieren sólo un poco de aquellos que se estimaron con anterioridad, como se observa en la tabla 4.17.

12. **Búsqueda de valores atípicos y análisis residual**

    Una vez estimado el modelo en conjunto, se procedió a la búsqueda de valores atípicos, con resultados negativos. Se analizó también que el modelo cumpliera con los supuestos de normalidad y no correlación como se observa en las gráficas de las figuras 4.25 y 4.26.

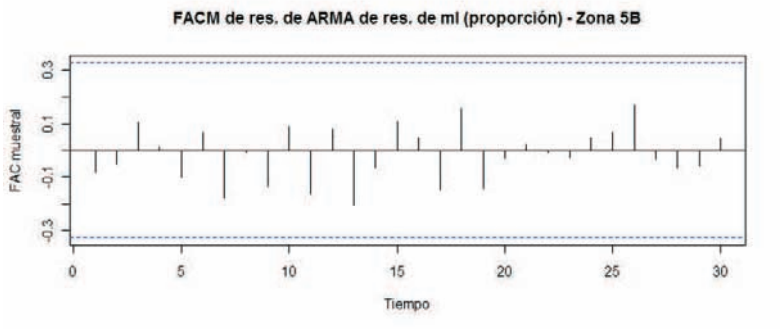


Figura 4.24: Función de autocorrelación muestral de los residuales del modelo ARMA del proceso estacionario respecto a una tendencia lineal.

Parámetros	ϕ_1	ϕ_2	ϕ_4	β_0	β_1
Estimación	0.4673	0.3738	-0.4012	-101.1348	0.0504
Error estándar	0.1325	0.1362	0.1260	1014.2269	0.0071

Cuadro 4.17: Parámetros del modelo estacionario ARMA (4,0) respecto a una tendencia lineal.

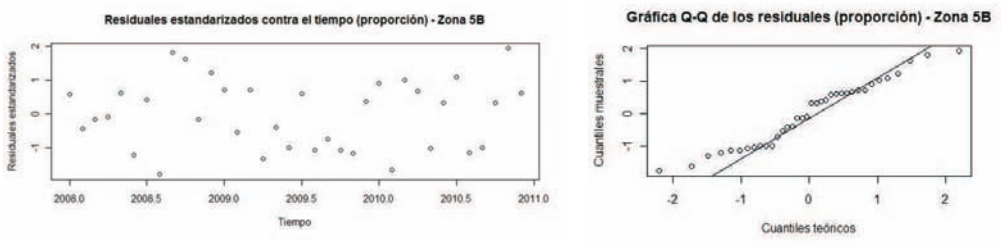


Figura 4.25: Gráficas para verificar la normalidad del modelo estacionario respecto a una tendencia lineal

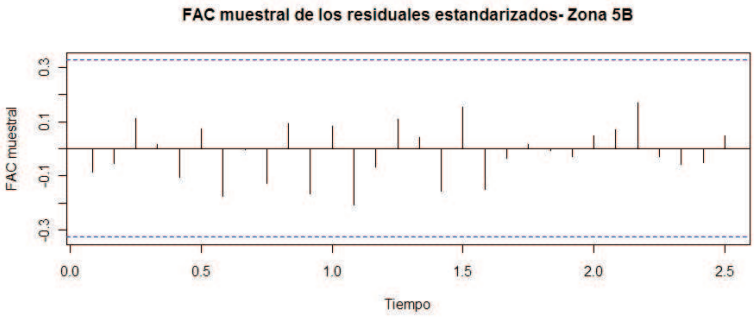


Figura 4.26: FACM de los residuales del modelo estacionario respecto a una tendencia lineal.

La normalidad se confirma con la prueba de Shapiro obteniéndose valor p de 0.1617. En la figura 4.26, se observa que no hay correlación significativa de los residuales.

13. **Modelo ajustado y pronóstico del último trimestre**

    La figura 4.27 muestra el modelo ajustado, los datos originales y las predicciones del último trimestre.

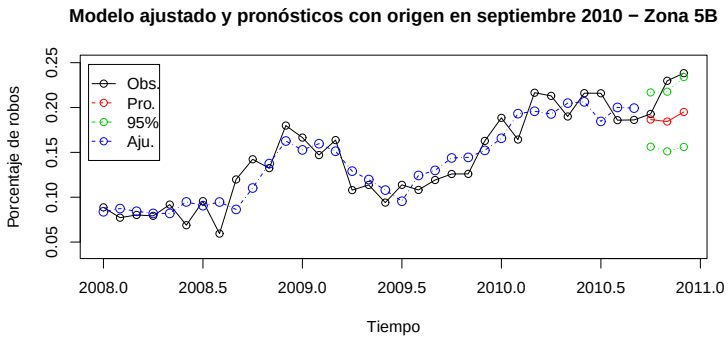


Figura 4.27: Modelo ajustado y pronósticos del último trimestre.

Los valores ajustados parecen cercanos a los datos de la serie temporal. Sólo la primera predicción es cercana al valor real; las otras dos están un poco distantes, y los intervalos de confianza no incluyen a los datos. Es posible que esto se deba al crecimiento pronunciado del valor de 2010-11 respecto al valor de 2010-10.

14. **Pronóstico de un año**

    Finalmente se calcula el pronóstico del año 2011, que se ilustra en la figura 4.28.

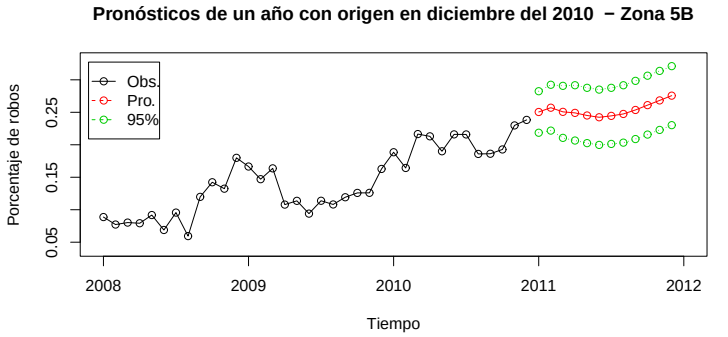


Figura 4.28: Pronósticos mensuales de un año para la serie temporal del porcentaje de robo mensual de la zona 5B. El proceso generado de los datos se identificó como estacionario ARMA(4,0) respecto a una tendencia lineal.

Los valores de los pronósticos de los respectivos meses están en la tabla 4.18.

Por último, se suman las predicciones para obtener la estimación anual del porcentaje de robo de todo el 2011 de ***3.0451*** con una desviación estándar de ***0.3061***.

11-1	11-2	11-3	11-4	11-5	11-6	11-7	11-8	11-9	11-10	11-11	11-12
0.0418	0.0423	0.0420	0.0422	0.0421	0.0422	0.0421	0.0421	0.0421	0.0421	0.0421	0.0422

Cuadro 4.18: Pronósticos de cada uno de los meses del 2011 del modelo estacionario ARMA(4,0) alrededor de una tendencia lineal.

4.2.6. Datos atípicos

La búsqueda de valores atípicos es importante, pues la inclusión de los datos irregulares puede cambiar la estimación de los parámetros, dándose el caso de que el proceso generador cambie al volverse irrelevantes algunos términos. Para ejemplificar la modelación de datos atípicos, se seleccionó la serie temporal de prima de riesgo mensual de la zona $3B$.

No se mostrará toda la metodología, pues se identificó que el proceso generador de los datos tenía una raíz unitaria, lo cual ya se explicó en la sección 4.2.4. Simplemente se presenta la gráfica de la serie (para observar el dato irregular), cuáles son sus parámetros al estimarlo sin incluir el atípico y que valores se consiguen al incorporarlo.

1. **Gráfica de la serie Temporal**

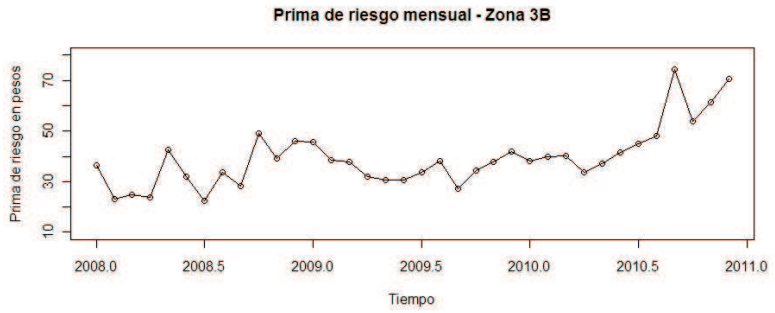


Figura 4.29: Serie temporal de prima de riesgo mensual de la zona $3B$.

A grandes rasgos se logra identificar un ligero crecimiento de la serie (figura 4.29), que al hacer las respectivas pruebas, era debido a una tendencia estocástica en el proceso. Se trabajó con los datos tal cual están, pues no se observa que la varianza sea inestable durante todo el proceso. A simple vista, se observa un valor que resalta pues posee un comportamiento irregular: 2010-09.

2. **Estimación de los parámetros**

   Durante el modelado de la serie, se identificó al modelo ARIMA (1,1,0) como el proceso generador de los datos. La estimación de los términos, sin tomar en cuenta que el dato del 2010-09 puede ser un atípico, está en la tabla 4.20.

   El término ϕ_1 es significativo, y no se puede rechazar que μ sea igual a cero. Las pruebas de normalidad arrojaron resultados que hacen rechazar la normalidad, esto podría ser causado por el dato aberrante.

Parámetros	ϕ_1	μ
Estimación	-0.4607	1.0269
Error estándar	0.1542	0.9846

Cuadro 4.19: Parámetros estimados del modelo ARIMA (1,1,0) para la prima mensual de riesgo.

Parámetros	ϕ_1	$A0$
Estimación	-0.4167	20.1549
Error estándar	0.1645	7.3447

Cuadro 4.20: Parámetros estimados del modelo ARIMA (1,1,0). AO es la estimación del valor atípico.

3. **Búsqueda de valores atípicos**

   Al implementar la búsqueda de valores aberrantes a través de las funciones implementadas en el paquete estadístico R, se encontró uno del tipo AO en 2010-09.

4. **Estimación de los parámetros considerando el valor atípico**

   Al incluir el valor irregular en la estimación, el parámetro μ se vuelve insignificante, por lo que se retira de la estimación considerándolo como cero. Los resultados finales se observan en la tabla 4.20.

   Ambas estimaciones son significativamente distintas de cero. El término ϕ_1 se reduce ligeramente respecto a la estimación anterior. Una vez calculados estos valores, la metodología de modelación continua de forma normal.

4.2.7. Análisis de intervención

Durante el análisis se notó que las series temporales de la zona $1A$ (tanto la de primas mensuales como porcentaje de robo) poseían una disminución considerable en su nivel de la media, principalmente después del 2010. Por esta razón se propuso implementar un análisis de intervención, para cuantificar y estimar esta disminución.

Como se explicó en la sección 3.4.1, es necesario identificar primero el proceso generador de los datos del periodo de pre-intervención. Una vez determinado, se analiza cual es la variable que produce el efecto de intervención, para ser estimada. Para ejemplificar esta técnica, se seleccionó la serie de porcentaje de proporción de robo mensual de la zona 1A. No se detallará sobre la modelación de los datos pre-intervenidos, pues resultó ser un proceso estacionario, que se explicó con detalle en la sección 4.2.3. Se presentará la gráfica de la serie (para observa la caída en la media) mencionando cual es el modelo antes de la intervención, para posteriormente estimar la variable que produjo esa respuesta específica.

1. **Gráfica de la serie Temporal**

   A simple vista se vuelve evidente que a inicios del 2010, ocurrió algún fenómeno que cambio de forma considerable el nivel de la media (figura 4.30). Antes de esa fecha, se observa un

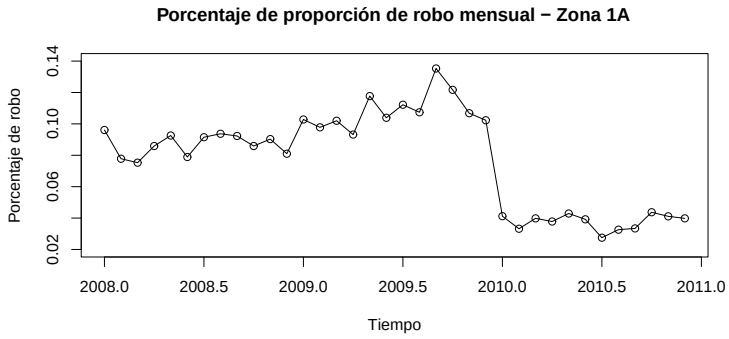


Figura 4.30: Serie temporal de porcentaje de robo mensual de la zona $1A$.

Parámetros	ϕ_1	ϕ_4	ϕ_5	μ
Estimación	0.6352	0.7033	-0.5605	0.0980
Error estándar	0.1497	0.1530	0.2006	0.0067

Cuadro 4.21: Parámetros estimados del modelo ARMA (5,0) para el periodo antes de la intervención de la serie de porcentaje de robo mensual de la zona $1A$.

comportamiento estable y estacionario tanto en media como en varianza. No se perciben datos irregulares, ni antes ni después de la intervención.

2. **Modelo antes de la intervención**

   Usando las técnicas de modelado de procesos estacionarios, se determinó que el proceso ARMA (5,0) es aquel que genera los datos antes de la intervención. A pesar de que son pocos datos ($n = 24$), la identificación fue satisfactoria. La estimación de los respectivos parámetros, empleado máxima verosimilitud, se puede ver en la tabla 4.21.

   Se tiene que todos los parámetros son significativamente distintos de cero. El modelo pasa de forma adecuada los diagnósticos, tanto de sobreajuste como el análisis residual. No se encontraron valores irregulares. Se procede a estimar el modelo de intervención.

3. **Modelo de intervención**

   Al observar detalladamente la gráfica de los datos (figura 4.30), se ve que el efecto de la intervención se mantiene durante todo el año 2010. Por esta razón, se escoge como variable de intervención aquella que produce un efecto permanente en la media, que es básicamente es una función escalón multiplicada por el parámetro ω (ecuación 3.39), el cual se estimará para medir el efecto de la intervención. Los resultados del ajuste del modelo empleando máxima verosimilitud se exhiben en la tabla 4.22.

4. **Pronóstico de un año**

   Finalmente se calcula el pronóstico de doce periodos por delante con origen en diciembre del 2010, que se ilustra en la figura 4.31.

   Los valores de las predicciones de los respectivos meses se ven en la tabla 4.23.

Parámetros	ϕ_1	ϕ_4	ϕ_5	μ	μ
Estimación	0.5937	0.5782	-0.3403	0.1013	-0.0719
Error estándar	0.1325	0.1566	0.1627	0.0074	0.0088

Cuadro 4.22: Parámetros estimados del modelo de intervención para el porcentaje de robos de la zona $1A$.

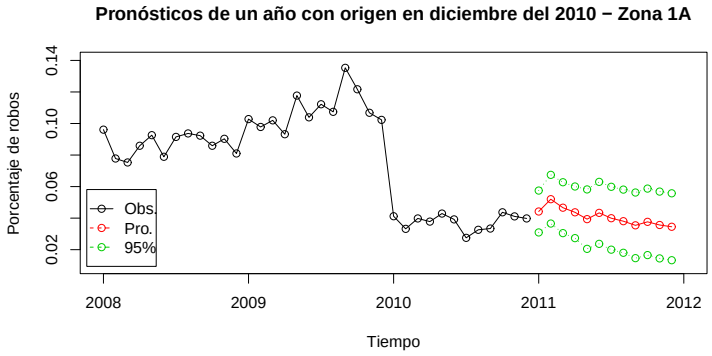


Figura 4.31: . Pronósticos mensuales de un año para la serie temporal del porcentaje de robo mensual de la zona $1A$. Se modeló la caída en la media del proceso a partir del 2010.

11-1	11-2	11-3	11-4	11-5	11-6	11-7	11-8	11-9	11-10	11-11	11-12
0.0442	0.0520	0.0466	0.0437	0.0394	0.0433	0.0399	0.0381	0.0354	0.0377	0.0356	0.0345

Cuadro 4.23: Pronósticos de cada uno de los meses del 2011 del modelo de intervención ARIMA (5,0,0) para la serie del porcentaje de robo de la zona 1A.

La estimación anual del porcentaje de robo del 2011 es de ***0.4905*** con una desviación estándar de ***0.1362***.

4.3. Resultados: modelos ARIMA para cada indicador

En la sección anterior, se explicó de forma exhaustiva cual fue la metodología de modelación para algunas series temporales representativas de los diversos procesos que se encontraron. En la primera parte de este apartado, se van a mostrar los resultados finales, es decir, cuáles fueron los procesos generadores en cada una de las zonas, tanto para las series de prima mensual como para las de porcentaje de robo. Como se mencionó con anterioridad, las gráficas y los resultados que llevaron a cada modelo, se incluyen en el apéndice A, como complemento del análisis, pero sin una explicación tan detallada. En la última parte de esta sección, se muestran finalmente, los pronósticos anualizados para cada zona y cada indicador, así como los valores anuales de los años anteriores con fines comparativos.

4.3.1. Procesos generadores

A continuación se enlistan los modelos ARIMA de cada una de las 18 series analizadas con las respectivas estimaciones de sus parámetros. Se presentaran primero las primas de riesgo mensual, para posteriormente introducir los procesos de las series de porcentaje de robo.

Modelos de para las primas de riesgo mensual:

Zona 1A **ARIMA(1,0,0) con intervención en 2010-02**

Parámetros: $\phi_1(0.2350)$ $\mu(62.74)$ $I1(-29.68)$

Zona 2A **ARIMA(9,0,0) con dato atípico IO en 2009-05**

Parámetros: $\phi_1(-0.30)$ $\phi_9(-0.51)$ $\mu(43.58)$ $I0(23.46)$

Zona 3A **ARIMA(8,0,0)**

Parámetros: $\phi_1(0.40)$ $\phi_8(0.39)$ $\mu(14.82)$

Zona 3B **ARIMA(1,1,0) con dato atípico AO en 2010-09**

Parámetros: $\phi_1(-0.42)$ $A0(20.16)$

Zona 3C **ARIMA(6,1,0)**

Parámetros: $\phi_1(-0.22)$ $\phi_6(-0.23)$

Zona 4A **ARIMA(5,1,0)**

Parámetros: $\phi_1(-0.38)$ $\phi_5(0.34)$ $\mu(1.90)$

Zona 4B **ARMA(2,0) respecto a tendencia lineal con atípico AO en 2008-01**

Parámetros: $\phi_1(0.19)$ $\phi_2(0.28)$ $\beta_0(80476.84)$ $\beta_1(40.11)$ $A0(23.85)$

Zona 5A **ARIMA(5,1,0)**

Parámetros:$\phi_1(-0.32)$ $\phi_4(-0.26)$ $\phi_5(0.29)$ $\mu(4.69)$

Zona 5B **ARIMA(8,1,0)**

Parámetros:$\phi_1(-0.43)$ $\phi_2(-0.27)$ $\phi_3(0.28)$ $\phi_6(-0.36)$ $\phi_7(-0.34)$ $\phi_8(-0.34)$

Modelos de para los porcentajes de robo mensual:

Zona 1A **ARIMA(5,0,0) con intervención en 2010-01**

Parámetros: $\phi_1(0.90)$ $\phi_4(0.25)$ $\phi_5(-0.32)$ $\mu(0.078)$ $I1(-0.033)$

Zona 2A **ARIMA(2,0,0)**

Parámetros: $\phi_1(-0.28)$ $\phi_2(0.43)$ $\mu(0.042)$

Zona 3A **ARIMA(5,0,0) con dos atípicos AO en 2009-01 y 2009-06**

Parámetros: $\phi_1(0.70)$ $\phi_4(-0.44)$ $\phi_5(0.32)$ $\mu(0.021)$ $A0(0.0069)$ $A02(0.0091)$

Zona 3B **ARIMA(1,1,0)**

Parámetros:$\phi_1(-0{,}36)$

Zona 3C **ARIMA(3,0,0)**

Parámetros: $\phi_1(0.73)$ $\phi_2(0.43)$ $\phi_3(-0.30)$ $\mu(0.099)$

Zona 4A **ARIMA(4,1,2)**

Parámetros: $\phi_2(0.37)$ $\phi_2(-0.34)$ $\theta_1(-0.48)$ $\theta_2(-0.52)$ $\mu(0.0015)$

Zona 4B **ARIMA(1,1,0) con atípicos AO en 2009-01 y 2009-12**

Parámetros: $\phi_1(-0.63)$ $\mu(0.0024)$ $A0(0.024)$ $A0(-0.036)$

Zona 5A **ARIMA(8,1,0)**

Parámetros: $\phi_4(-0.34)$ $\phi_8(-0.57)$ $\mu(0.0031)$

Zona 5B **ARMA(4,0) respecto a tendencia lineal**

Parámetros: $\phi_1(0.47)$ $\phi_2(0.37)$ $\phi_4(-0.40)$ $\mu(-101.14)$ $t(0.0504)$

A pesar que se trabajó con series temporales de pocos datos[III] , se considera que la modelación de los indicadores fue adecuada; esto se puede constatar al observar el ajuste de los modelos y otros resultados importantes en el apéndice A. Se puede apreciar que en términos generales, la parte AR domina los procesos generadores, pues hay mucho más términos autoregresivos que

[III] Algunos autores (Wie [7]) señalan que para lograr una buena identificación del modelo y estimación apta de los parámetros es conveniente tener al menos 50 observaciones.

Zona	2008	2009	2010	2011 ($\pm 95\,\%$)
$1A$	725.44	793.25	411.16	$397{,}56 \pm 175{,}64$
$2A$	536.22	520.38	537.97	$511{,}98 \pm 151{,}51$
$3A$	144.02	193.77	208.62	$190{,}48 \pm 60{,}29$
$3B$	401.33	428.11	583.06	$815{,}92 \pm 271{,}47$
$3C$	913.48	1172.74	1156.61	$1132{,}03 \pm 334{,}79$
$4A$	357.34	570.15	922.34	$1297{,}69 \pm 313{,}67$
$4B$	944.47	1301.66	1885.39	$2373{,}21 \pm 177{,}73$
$5A$	937.74	1326.35	1866.44	$2997{,}35 \pm 773{,}43$
$5B$	1768.44	2499.31	4127.46	$7055{,}52 \pm 1755{,}08$

Cuadro 4.24: Primas de riesgo anualizadas. Los pronósticos del 2011 se muestran con su respectivo intervalo al 95 %.

Zona	2008	2009	2010	2011 ($\pm 95\,\%$)
$1A$	1.041	1.303	0.452	$0{,}4905 \pm 0{,}2240$
$2A$	0.483	0.622	0.437	$0{,}505 \pm 0{,}1627$
$3A$	0.211	0.312	0.234	$0{,}248 \pm 0{,}0762$
$3B$	0.389	0.495	0.526	$0{,}7131 \pm 0{,}2285$
$3C$	1.005	1.620	1.106	$1{,}1213 \pm 0{,}4045$
$4A$	0.447	0.756	0.873	$1{,}1178 \pm 0{,}2272$
$4B$	0.942	1.569	1.599	$2{,}0645 \pm 0{,}2834$
$5A$	1.128	1.550	1.973	$2{,}4438 \pm 0{,}5735$
$5B$	1.215	1.549	2.437	$3{,}0451 \pm 0{,}5045$

Cuadro 4.25: Porcentajes de robo anualizados. Las predicciones del 2011 se muestran con su respectivo intervalo al 95 %.

de promedios móviles. Respecto al número de parámetros, la mayoría de los modelos poseen pocos parámetros, que va acode con el principio de parsimonia, que establece que un modelo más sencillo o parsimonioso que describe a los datos adecuadamente, es preferible a un modelo más complicado que solamente signifique una pequeña mejoría en la explicación de la información (Dobson [33]).

Una vez establecidos los modelos para cada indicador, se procedió a realizar pronósticos a un año para poder vislumbrar los posibles escenarios de las primas de riesgo y de los porcentajes de robo en el 2011, con base en la información de los años anteriores. Los resultados se muestran en la tablas 4.24 y 4.25, indicando cual fue la prima y la proporción de robo en años anteriores con fines comparativos.

Se puede ver que la clasificación de las zonas de acuerdo a su tasa de crecimiento, va acorde con los que se pronostica para el 2011, es decir, en zonas de disminución, las predicciones de los indicadores tienen un decrecimiento, en regiones de estabilidad los pronósticos mantienen el nivel, y en sectores de crecimiento las estimaciones aumentan.

Dentro las primeras 5 zonas ($1A$, $2A$, $3A$, $3B$ y $3C$) las primas de riesgo predichas para el año

2011 poseen resultados favorables, pues se mantienen estables, y lo que es mejor, han disminuido en algunos sectores ($1A$ y $2A$). La zona $4A$ posee un incremento que no es muy elevado en comparación con los valores de otros años. En las últimas regiones, en específico la zona $5B$, el pronóstico de la cuota de repartición se eleva considerablemente, lo cual debe ser tomado en cuenta, por si el crecimiento continua así, esto significará un problema para las aseguradoras.

Respecto al porcentaje de proporción de robo el comportamiento es similar al de las primas de riesgo. En las primeras 5 zonas, el robo relativo a las unidades expuestas se mantiene estable; las últimas 4 zonas ($4A$, $4B$, $5A$ y $5B$), la proporción de robo comienza a crecer de forma considerable.

Es importante destacar, que los intervalos de confianza al 95 %, tienen una longitud razonable dentro de las primeras zonas, mientras que va aumentando en regiones de alto crecimiento. Lo anterior se debe a que en los proceso no estacionarios el error de predicción va incrementándose conforme se van alejando las estimaciones del origen del pronóstico, como se vio en la sección 3.5.4.

Con estos resultados se tiene un panorama general del comportamiento del robo a mediano plazo, que era un objetivo primordial del proyecto. En el siguiente capítulo, se establecen las conclusiones del trabajo, mencionando aspectos relevantes, ciertos detalles, trabajo a futuro y el aprendizaje obtenido.

Capítulo 5

Conclusiones y consideraciones finales

Al realizar esta tesis se tuvo la oportunidad de utilizar los datos del sector asegurador y conocer de cerca las necesidades y problemas del mismo, esto permitió proponer modelos de series de tiempo apegados a sus necesidades.

En general, al realizar el análisis de los datos se obtuvieron los resultados esperados y se alcanzaron los objetivos planteados al inicio. El primer resultado importante fue la generación de una regionalización de la república mexicana en zonas de comportamiento homogéneo de robo de autos asegurados. El segundo resultado importante es que para cada zona de la regionalización obtenida se encontró la estimación de un modelo ARIMA ajustado a los indicadores de prima de riesgo y proporción de robo lo que permite vislumbrar su comportamiento a mediano plazo a través de los pronósticos generados.

Regionalización

La regionalización se hizo en primera instancia basados en la tasa de crecimiento o decrecimiento del robo de autos, en segunda instancia se consideró clasificar las zonas geográficas basándose en el monto de la prima cobrada. Estos dos criterios generaron 9 zonas, que cumplen las expectativas del sector asegurador. Los dos criterios de clasificación miden características diferentes, el primero agrupa estados que tienen una razón de crecimiento del robo de autos muy similar, el segundo agrupa estados con un monto de prima del seguro semejante, esto se correlaciona con la proporción de robo de autos asegurados.

Se pueden obtener conclusiones interesantes de la regionalización establecida. Como se puso de manifiesto en la sección 4.1.3, el contraste visual que se genera por las tonalidades cálidas y frías, relacionadas a aumento de robo y disminución del mismo (figura 4.3), permite identificar que en los estados del norte de la república el robo de vehículos asegurados está en aumento, mientras que en los estados del sureste el delito disminuye.

Modelos

Los procesos ARIMA propuestos para modelar los indicadores fueron satisfactorios y se considera

que describen de forma adecuada el comportamiento general de las series temporales, además las estimaciones de los parámetros y los pronósticos fueron estadísticamente apropiadas y significativas. El desglose mensual no propicio la identificación de algún componente estacional en ninguna de las series de ambos indicadores. Una posible causa de esto es que la disgregación y separación de los datos en regiones, tendió a disminuir la dependencia estacional; aunado a esto, la carencia de observaciones pudo hacer sido otro factor que propiciara la nula identificación de la dependencia de 12 meses atrás, la cual en caso de haberla encontrado, hubiera beneficiado las predicciones de los modelos.

Pronósticos

El comportamiento de crecimiento previsto en algunas regiones (zonas 4B, 5A y 5B) afectará los mecanismos comerciales de subsidios. Si la situación sigue como pronostican los modelos, a largo plazo las compañías podrán tener problemas de suficiencia para afrontar sus responsabilidades, motivo por el cual la importancia de estos resultados.

Las predicciones pueden ser empleadas como un posible panorama de lo que ocurría si el delito se sigue comportando de la misma forma. Esto puede servir como punto de partida para que se generen propuestas de cómo se puede cambiar el curso del robo de vehículos a nivel estatal, y por qué no nacional, lo que beneficiaría la parte económica de los asegurados, al disminuir las primas, y a la contraparte aseguradora, al disminuir el capital necesario para hacer frente a sus responsabilidades.

Cuando se obtenga la información de todo el 2011 se podrá contrastar los pronósticos realizados para determinar la exactitud de las estimaciones y la confiabilidad de los intervalos de confianza. Esto se podrá realizar a partir del mes de mayo del 201 2, cuando la base de SESA se actualice con toda la información del sector asegurador mexicano.

Trabajo a futuro

Después de estudiar y analizar a fondo este trabajo, surgen ideas de cómo pueden mejorarse los resultados de las predicciones. Una de ella sería buscar si datos bimestrales pueden facilitar la identificación de la estacionalidad, lo cual mejoraría las predicciones; el desarrollo de esta propuesta podría ser a mediano plazo, hasta que se contara con la información en las bases de todo el 2011, para poder tener más observaciones que permitieran una mejor modelación de las series bimestrales.

Otra idea que parece razonable es identificar algún indicador económico o social que pueda tener alguna relación directa con el robo de vehículos, y que pueda ser usado para mejorar las predicciones. Para lograr un análisis como este se tiene que hacer uso de modelos más generales que aquellos que se propusieron en esta tesis, los cuales se conocen como modelos funciones de transferencia. En éstos, la variable de respuesta y_t es una serie temporal que se relaciona con otra serie x_t que se cree tiene una relación causal, es decir, valores presentes y/o pasados de x_t influyen en la respuesta y_t. La parte estocástica, se modela a través de una serie de ruido blanco η_t. De acuerdo a Box y Jenkins [4], un modelo de función de transferencia es el siguiente:

$$y_t = \nu(B)x_t + \eta_t \tag{5.1}$$

donde el término $\nu(B) = \sum_{j=-\infty}^{\infty} \nu_j B^j$ se conoce como filtro o función de transferencia, y_t es la respuesta o variable de salida, x_t es la variable de entrada y η_t es una serie de ruido blanco.

Como se busca el caso específico de que la variable de entrada sea la *causa* de la variable de salida, se tiene que el filtro debe ser la siguiente manera:

$$y_t = \nu_0 x_t + \nu_1 x_{t-1} + \nu_2 x_{t-2} + ... + \eta_t \tag{5.2}$$

Con esto se asegura que valores pasados y a lo más presentes de x_t influyan el comportamiento presente de y_t. Un valor futuro de la variable de entrada x_{t+1} que contribuya en valores presentes o pasados de la respuesta, presuponen otro tipo de modelos que deben incluir cierta retroalimentación, lo cual presupone modelos multivariados más complejos como los VAR, VMA o VARMA (ver Wei[7] y Tsay[34]).

Una vez establecida cual es la variable de entrada y la variable de salida el objetivo es estimar la función de transferencia $\nu(B)$ y la serie del ruido η_t. Por fines de practicidad, se debe determinar un número relativamente pequeño de términos del filtro que sean significativamente distintos de cero.

Con el propósito de identificar cuantos parámetros integran el filtro, generalmente se escribe a la función de transferencia de forma racional:

$$\nu(B) = \frac{\omega(B)B^b}{\delta(B)} \tag{5.3}$$

donde $\omega(B) = \omega_0 - \omega_1 B - \cdots - \omega_s B^s$ y $\delta(B) = \delta_0 - \delta_1 B - \cdots - \delta_r B^r$, y b es un parámetro de retraso que representa el tiempo que tuvo que pasar para que la variable de entrada comienzara a afectar a la variable de salida; por la relación causal b puede tomar valores en los naturales. Para que el sistema sea estable se debe de cumplir que las raíces del polinomio $\delta(B) = 0$ estén fuera del círculo unitario. Tomando en cuenta esta representación es prudente señalar que el análisis de intervención y el manejo de datos atípicos (introducidos en el capítulo 3) son un caso particular de este tipo de modelos, donde x_t representa una intervención o un dato irregular.

Retomando la idea, se puede tratar de encontrar una variable de entrada que tenga una influencia marcada en los indicadores que se estudiaron, y generar un sistema dinámico que permita el uso de los modelos de transferencia; x_t puede ser un indicador social como el índice delictivo, o posiblemente uno económico como el desempleo; la decisión cual es la variable de entrada está en función de lo que se desea correlacionar y de lo que se sabe que puede afectar las series de respuesta. Un punto a favor de este tipo de modelado, es que se puede implementar también en el software estadístico R.

Capítulo 6

Apéndice A: Gráficas del proceso de modelado

En este anexo se muestran las gráficas y los resultados más importantes para el modelado de los indicadores, exceptuando aquellos indicadores que fueron descritos de forma exhaustiva en las secciones 4.2.3, 4.2.4 y 4.2.5. Se exhibirán primero las series de prima de riesgo, para posteriormente introducir las de porcentaje de proporción de robo.

6.1. Series de prima de riesgo mensual

6.1.1. Zona 1A

Modelo estacionario ARIMA(1,0,0) con intervención en 2010-02

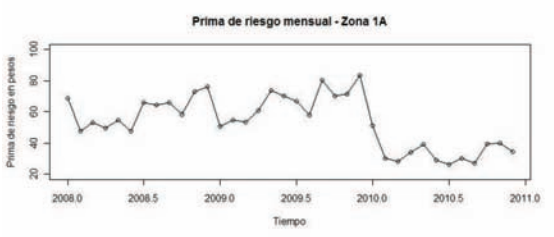
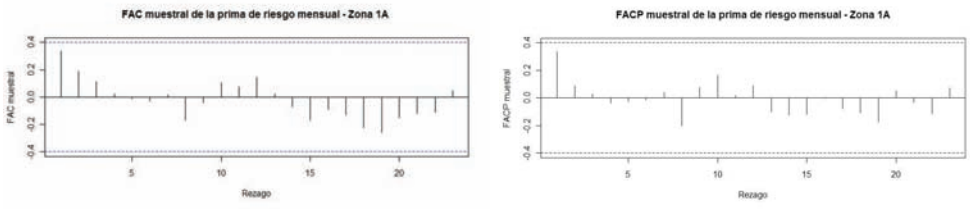


Figura 6.1: Serie temporal de la prima de riesgo mensual de la zona 1A.

6.1.2. Zona 2A

Modelo estacionario ARIMA(9,0,0) con dato atípico IO en 2009-05



(a) FACM (b) FACPM

Figura 6.2: Gráficas de FACM y de FACPM de la serie en el periodo antes de la intervención.

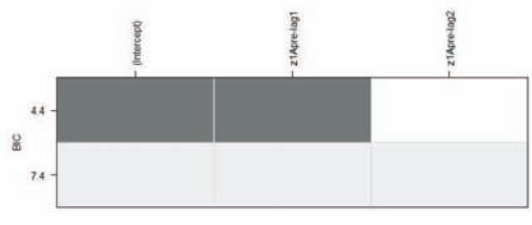
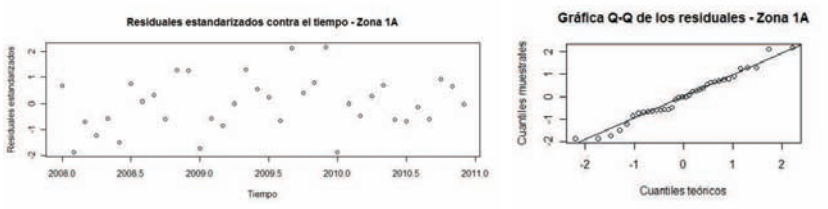


Figura 6.3: Subconjunto óptimo para el modelo antes de la intervención ARMA(1,0).



(a) Residuales contra el tiempo (b) Gráfica QQ

Figura 6.4: Gráficas para validar la normalidad de los residuales.

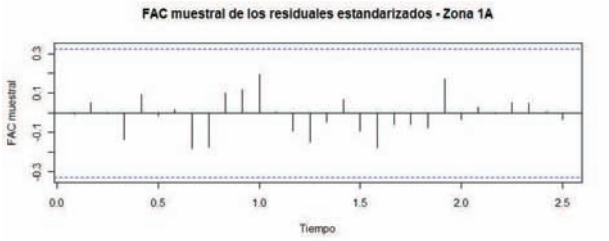


Figura 6.5: Función de autocorrelación muestral de los residuales del modelo ajustado.

Figura 6.6: Modelo ajustado y pronósticos del último trimestre.

Figura 6.7: Pronósticos mensuales del año 2011.

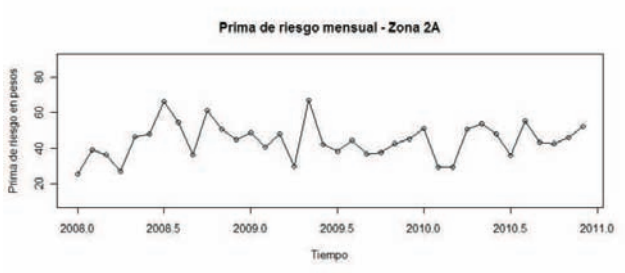
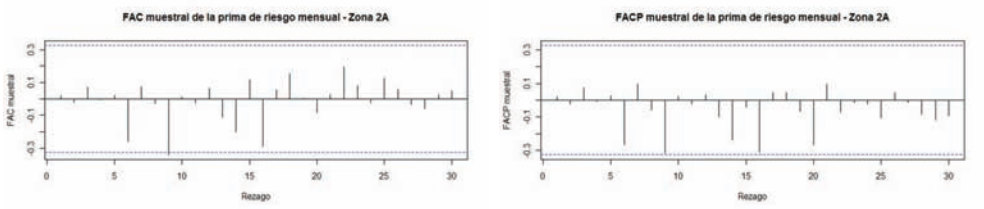


Figura 6.8: Serie temporal de la prima de riesgo mensual de la zona 2A.



(a) FACM (b) FACPM

Figura 6.9: Gráficas de FACM y de FACPM de la serie de prima de riesgo mensual de la zona 2A.

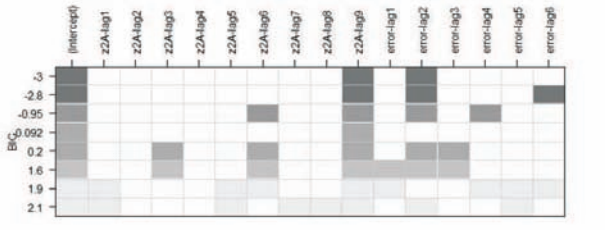


Figura 6.10: Subconjunto óptimo para el modelo de la serie de prima de riesgo mensual de la zona 2A.

(a) Residuales contra el tiempo (b) Gráfica QQ

Figura 6.11: Gráficas para validar la normalidad de los residuales.

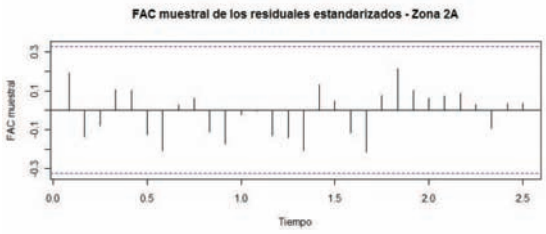


Figura 6.12: Función de autocorrelación muestral de los residuales del modelo ajustado.

Figura 6.13: Modelo ajustado y pronósticos del último trimestre.

Figura 6.14: Pronósticos mensuales del año 2011.

6.1.3. Zona 3A

Modelo estacionario ARIMA(8,0,0)

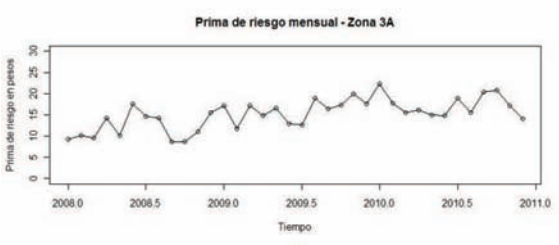
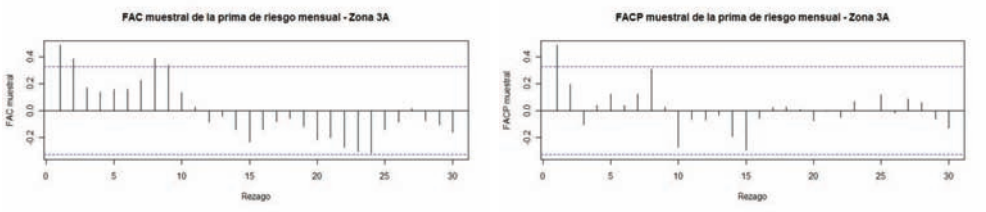


Figura 6.15: Serie temporal de la prima de riesgo mensual de la zona 3A.



(a) FACM (b) FACPM

Figura 6.16: Gráficas de FACM y de FACPM de la serie de prima de riesgo mensual de la zona 3A.

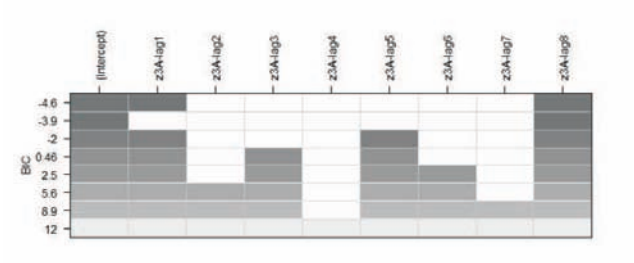
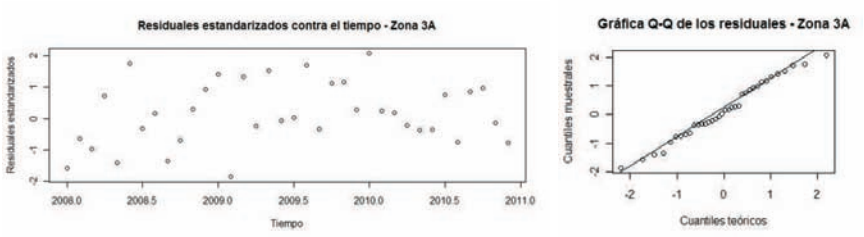


Figura 6.17: Subconjunto óptimo para el modelo de la serie de prima de riesgo mensual de la zona 3A.



(a) Residuales contra el tiempo (b) Gráfica QQ

Figura 6.18: Gráficas para validar la normalidad de los residuales.

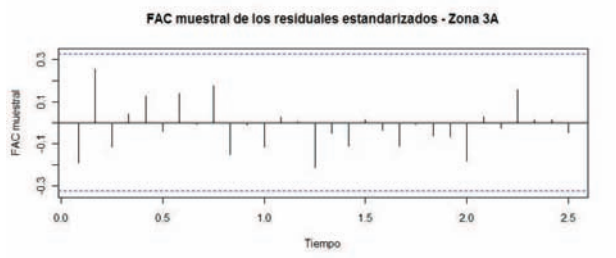


Figura 6.19: Función de autocorrelación muestral de los residuales del modelo ajustado.

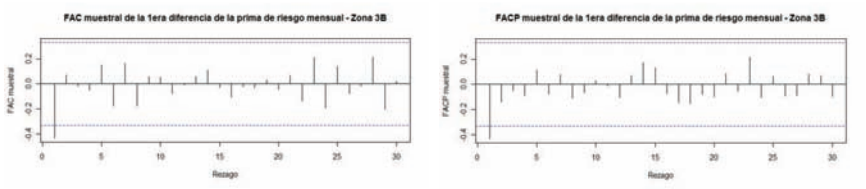
Figura 6.20: Modelo ajustado y pronósticos del último trimestre.

Figura 6.21: Pronósticos mensuales del año 2011.

6.1.4. Zona 3B

Modelo estacionario ARIMA(1,1,0) con dato atípico AO en 2010-09

Para esta serie, se incluyen los resultados más importantes de la modelación que complementan a los que se describieron en la sección 4.2.6.



(a) FACM (b) FACPM

Figura 6.22: Gráficas de FACM y de FACPM de la primera diferencia de la serie de prima de riesgo mensual de la zona 3B.

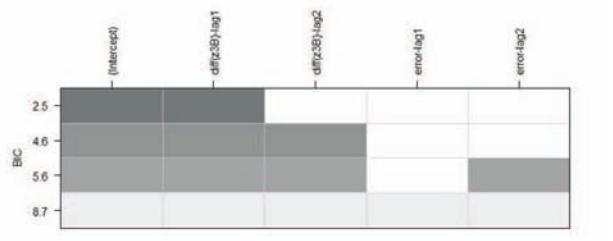
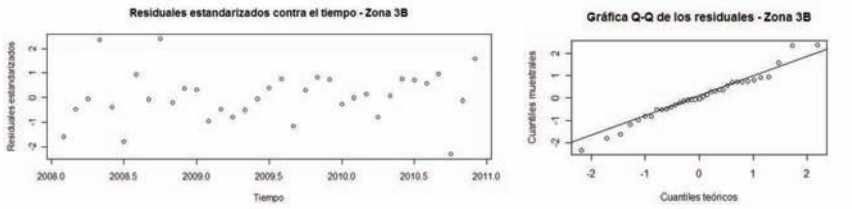


Figura 6.23: Subconjunto óptimo para el modelo de la serie de prima de riesgo mensual de la zona 3B.



(a) Residuales contra el tiempo (b) Gráfica QQ

Figura 6.24: Gráficas para validar la normalidad de los residuales.

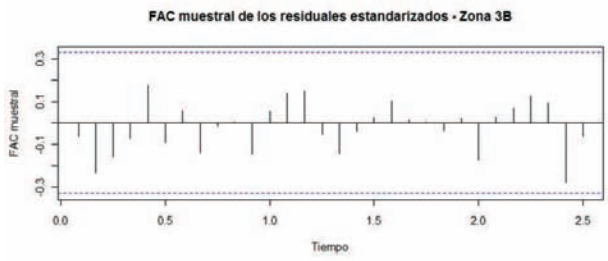


Figura 6.25: Función de autocorrelación muestral de los residuales del modelo ajustado.

Figura 6.26: Modelo ajustado y pronósticos del último trimestre.

Figura 6.27: Pronósticos mensuales del año 2011.

6.1.5. Zona 3C

Modelo con tendencia estocástica ARIMA(6,1,0)

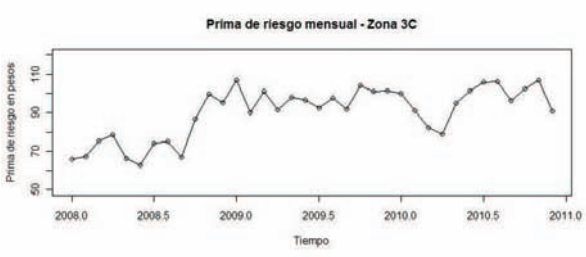
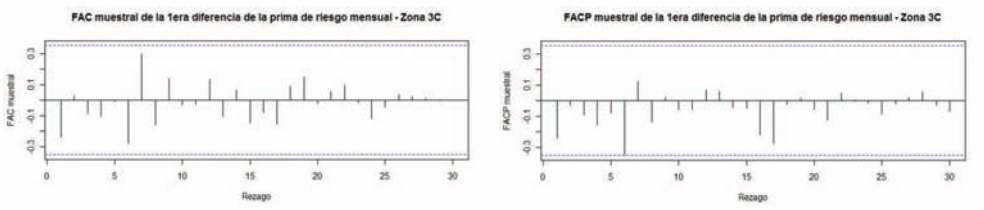


Figura 6.28: Serie temporal de la prima de riesgo mensual de la zona 3C.



(a) FACM (b) FACPM

Figura 6.29: Gráficas de FACM y de FACPM de la primera diferencia de la serie de prima de riesgo mensual de la zona 3C.

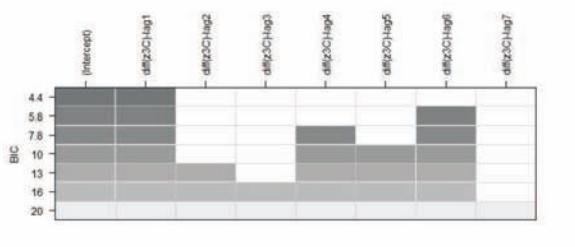


Figura 6.30: Subconjunto óptimo para el modelo de la serie de prima de riesgo mensual de la zona 3C.

(a) Residuales contra el tiempo (b) Gráfica QQ

Figura 6.31: Gráficas para validar la normalidad de los residuales.

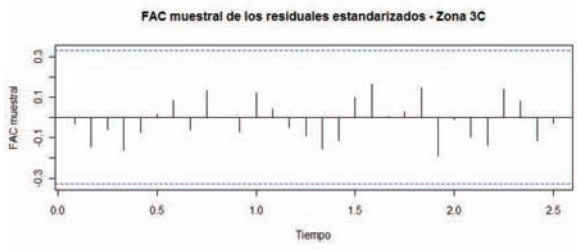


Figura 6.32: Función de autocorrelación muestral de los residuales del modelo ajustado.

Figura 6.33: Modelo ajustado y pronósticos del último trimestre.

Figura 6.34: Pronósticos mensuales del año 2011.

6.1.6. Zona 4A

Modelo con tendencia estocástica ARIMA(5,1,0)

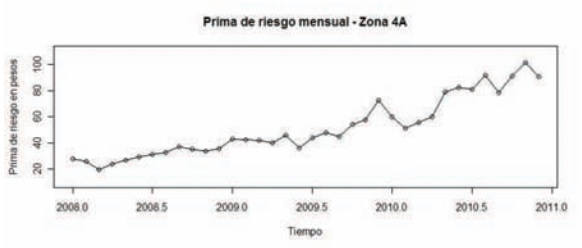
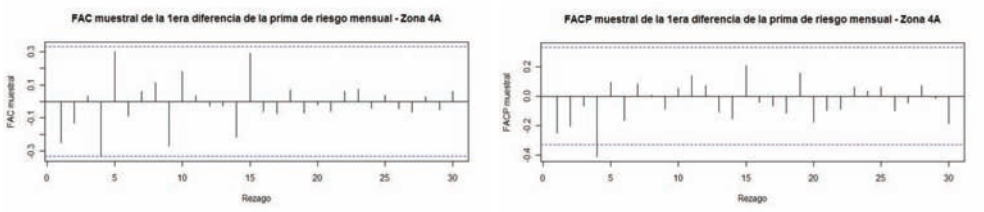


Figura 6.35: Serie temporal de la prima de riesgo mensual de la zona 4A.



(a) FACM (b) FACPM

Figura 6.36: Gráficas de FACM y de FACPM de la primera diferencia de la serie de prima de riesgo mensual de la zona 4A.

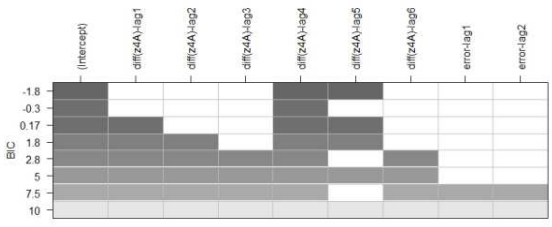


Figura 6.37: Subconjunto óptimo para el modelo de la serie de prima de riesgo mensual de la zona 4A.

(a) Residuales contra el tiempo (b) Gráfica QQ

Figura 6.38: Gráficas para validar la normalidad de los residuales.

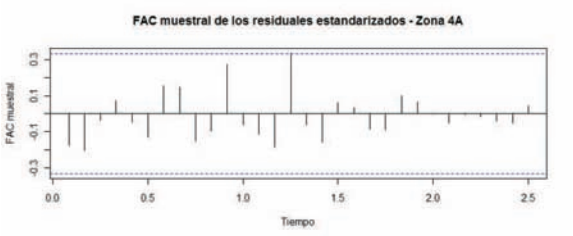


Figura 6.39: Función de autocorrelación muestral de los residuales del modelo ajustado.

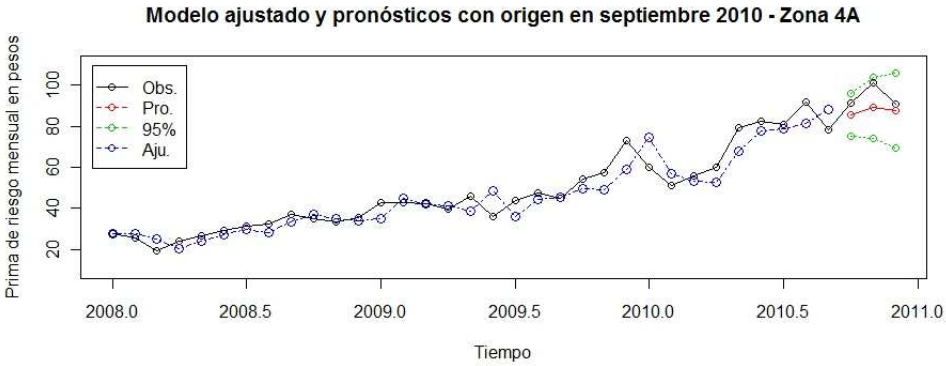


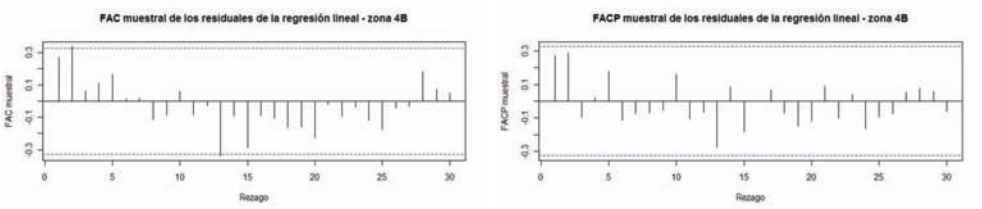
Figura 6.40: Modelo ajustado y pronósticos del último trimestre.

Figura 6.41: Pronósticos mensuales del año 2011.

6.1.7. Zona 4B

Modelo estacionario ARMA(2,0) alrededor de una tendencia lineal con atípico AO en 2008-01

Figura 6.42: Serie temporal de la prima de riesgo mensual de la zona 4B.



(a) FACM (b) FACPM

Figura 6.43: Gráficas de FACM y de FACPM de los residuales de la regresión simple.

(a) Residuales contra el tiempo (b) Gráfica QQ

Figura 6.44: Gráficas para validar la normalidad de los residuales.

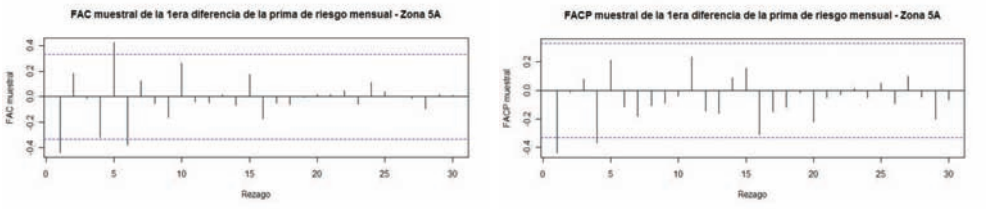
Figura 6.45: Modelo ajustado y pronósticos del último trimestre.

Figura 6.46: Pronósticos mensuales del año 2011.

6.1.8. Zona 5A

Modelo con tendencia estocástica ARIMA(5,1,0)

Figura 6.47: Serie temporal de la prima de riesgo mensual de la zona 5A.



(a) FACM (b) FACPM

Figura 6.48: Gráficas de FACM y de FACPM de la primera diferencia de la serie de prima de riesgo mensual de la zona 5A.

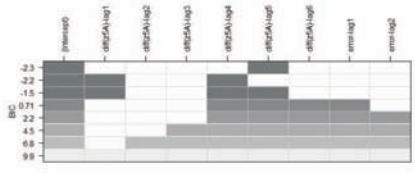


Figura 6.49: Subconjunto óptimo para el modelo de la serie de prima de riesgo mensual de la zona 5A.

6.1.9. Zona 5B

Modelo con tendencia estocástica ARIMA(8,1,0)

(a) Residuales contra el tiempo (b) Gráfica QQ

Figura 6.50: Gráficas para validar la normalidad de los residuales.

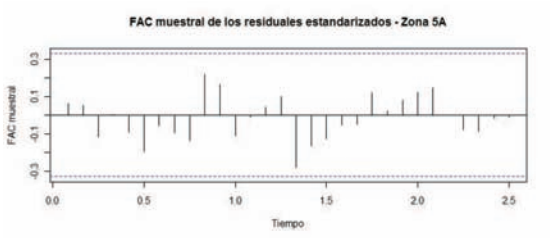


Figura 6.51: Función de autocorrelación muestral de los residuales del modelo ajustado.

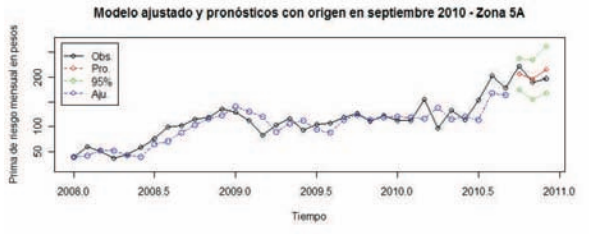
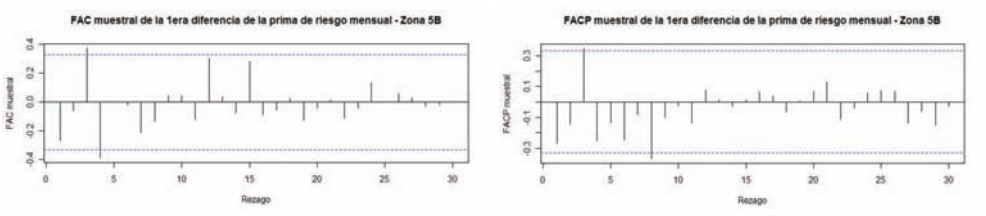


Figura 6.52: Modelo ajustado y pronósticos del último trimestre.

Figura 6.53: Pronósticos mensuales del año 2011.

Figura 6.54: Serie temporal de la prima de riesgo mensual de la zona 5B.



(a) FACM (b) FACPM

Figura 6.55: Gráficas de FACM y de FACPM de la primera diferencia de la serie de prima de riesgo mensual de la zona 5B.

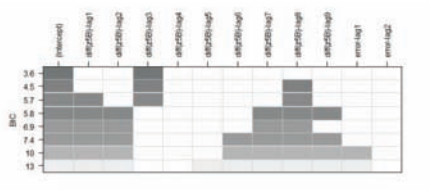
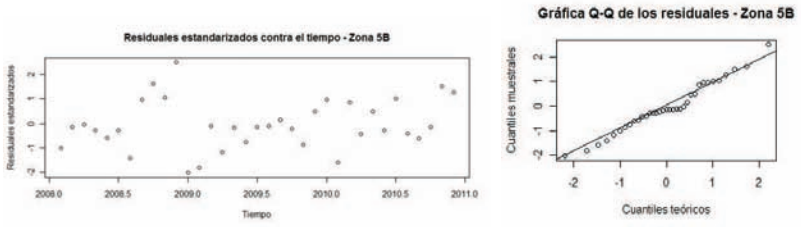


Figura 6.56: Subconjunto óptimo para el modelo de la serie de prima de riesgo mensual de la zona 5B.



(a) Residuales contra el tiempo (b) Gráfica QQ

Figura 6.57: Gráficas para validar la normalidad de los residuales.

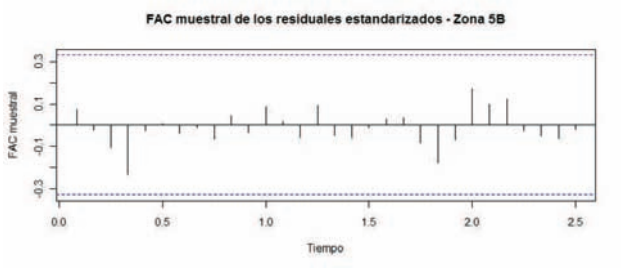


Figura 6.58: Función de autocorrelación muestral de los residuales del modelo ajustado.

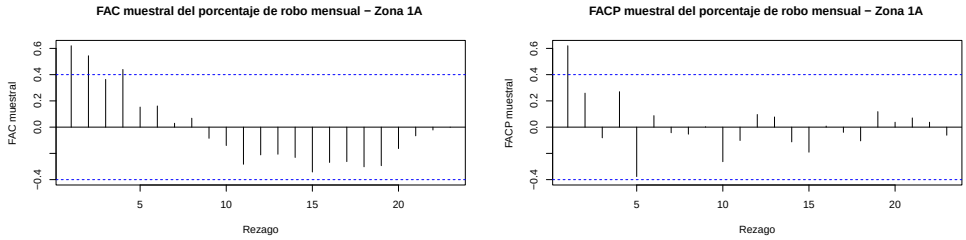
Figura 6.59: Modelo ajustado y pronósticos del último trimestre.

Figura 6.60: Pronósticos mensuales del año 2011.

6.2. Series de porcentaje de robo

6.2.1. Zona 1A

Modelo estacionario ARIMA(5,0,0) con intervención en 2010-01



(a) FACM (b) FACPM

Figura 6.61: Gráficas de FACM y de FACPM de la serie en el periodo antes de la intervención.

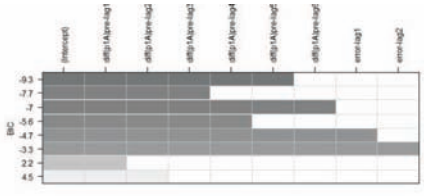
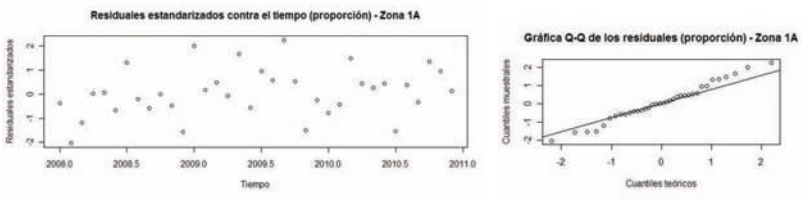


Figura 6.62: Subconjunto óptimo para el modelo antes de la intervención ARMA(1,0).



(a) Residuales contra el tiempo (b) Gráfica QQ

Figura 6.63: Gráficas para validar la normalidad de los residuales.

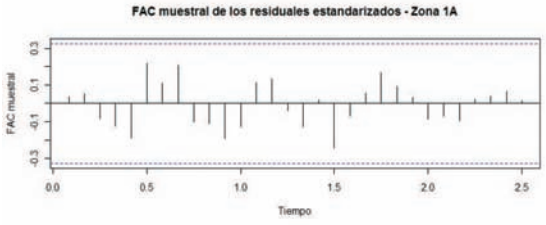


Figura 6.64: Función de autocorrelación muestral de los residuales del modelo ajustado.

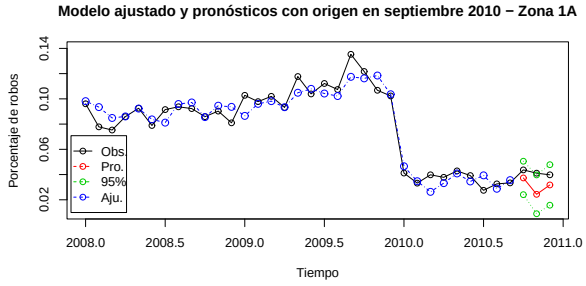


Figura 6.65: Modelo ajustado y pronósticos del último trimestre.

6.2.2. Zona 3A

Modelo estacionario ARIMA(5,0,0) con dos atípicos AO en 2009-01 y 2009-06

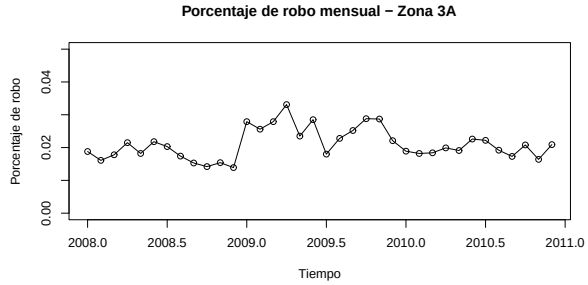
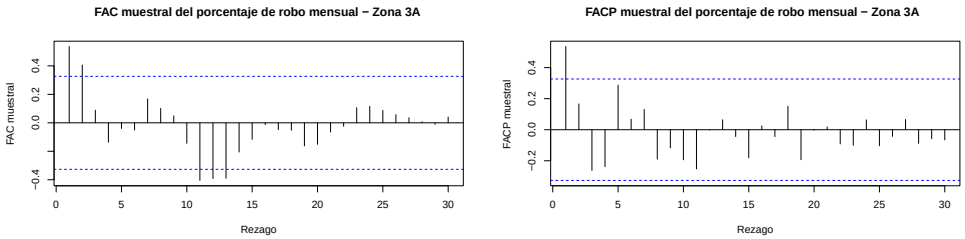


Figura 6.66: Serie temporal del porcentaje de robo mensual de la zona 3A.



(a) FACM (b) FACPM

Figura 6.67: Gráficas de FACM y de FACPM de la serie del porcentaje de robo de la zona 3A.

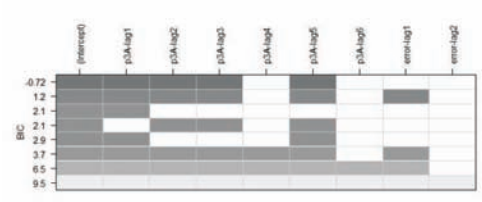
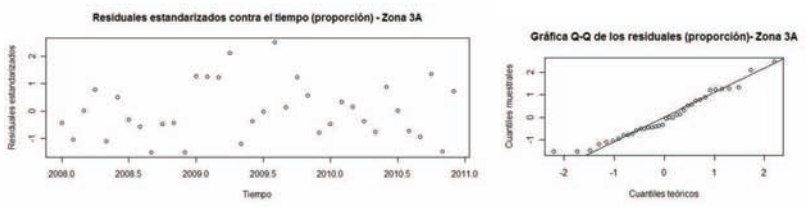


Figura 6.68: Subconjunto óptimo para el modelo de la serie de porcentaje de robo mensual de la zona 3A.



(a) Residuales contra el tiempo (b) Gráfica QQ

Figura 6.69: Gráficas para validar la normalidad de los residuales.

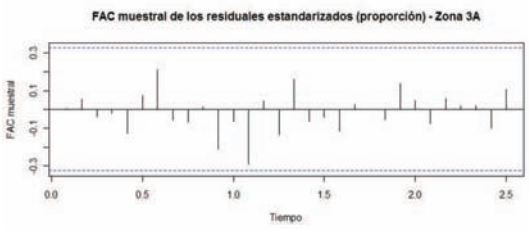


Figura 6.70: Función de autocorrelación muestral de los residuales del modelo ajustado.

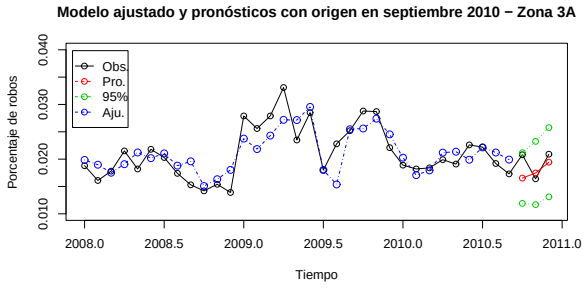


Figura 6.71: Modelo ajustado y pronósticos del último trimestre.

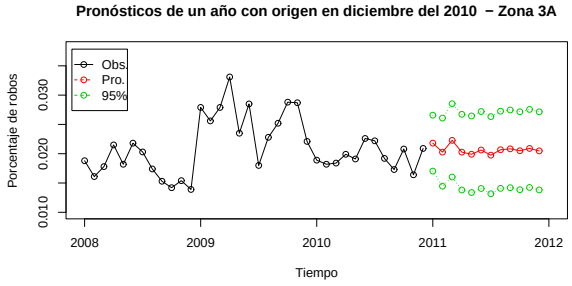


Figura 6.72: Pronósticos mensuales del año 2011.

6.2.3. Zona 3B

Modelo con tendencia estocástica ARIMA(1,1,0) con dato atípico AO en 2010-09

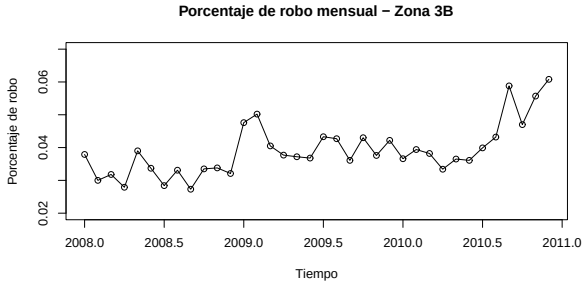
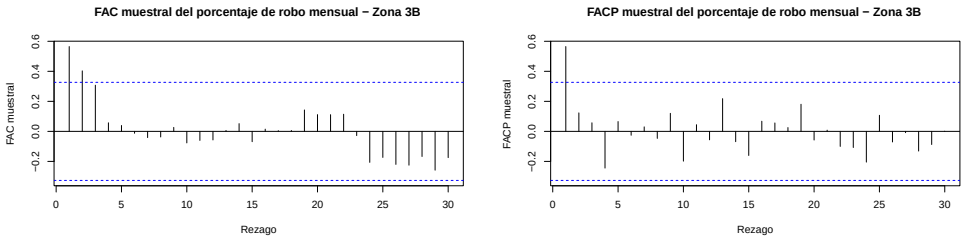


Figura 6.73: Serie temporal del porcentaje de robos mensual de la zona 3B.



(a) FACM (b) FACPM

Figura 6.74: Gráficas de FACM y de FACPM de la primera diferencia de la serie de porcentaje de robos mensual de la zona 3B.

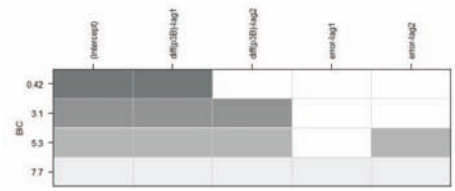
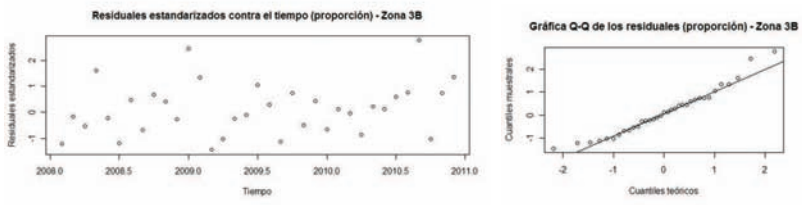


Figura 6.75: Subconjunto óptimo para el modelo de porcentaje de robo mensual de la zona 3B.



(a) Residuales contra el tiempo (b) Gráfica QQ

Figura 6.76: Gráficas para validar la normalidad de los residuales.

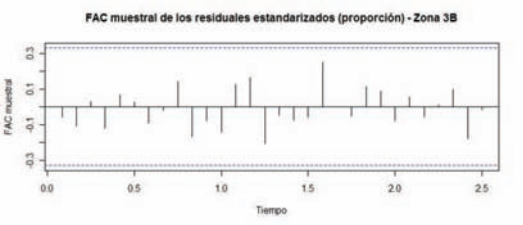


Figura 6.77: Función de autocorrelación muestral de los residuales del modelo ajustado.

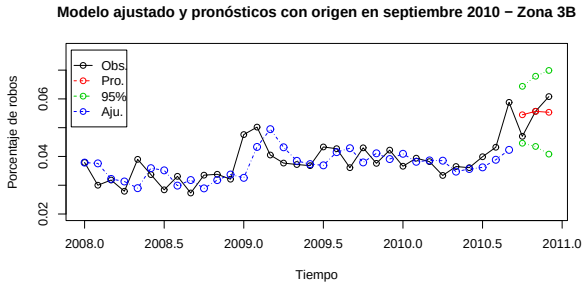


Figura 6.78: Modelo ajustado y pronósticos del último trimestre.

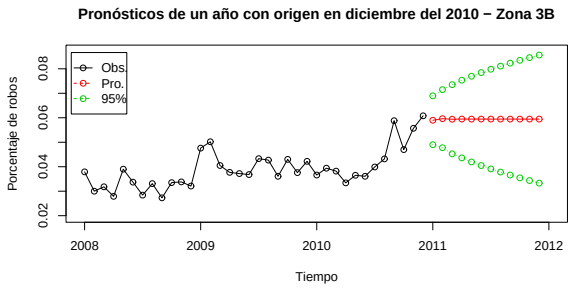


Figura 6.79: Pronósticos mensuales del año 2011.

6.2.4. Zona 3C

Modelo estacionario ARIMA(3,0,0)

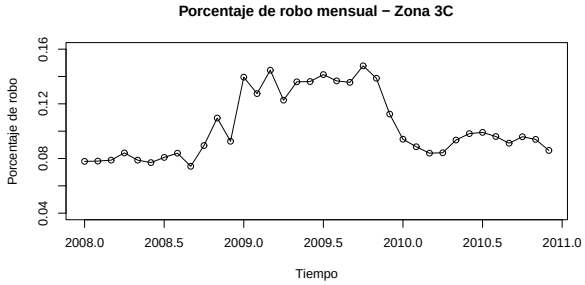
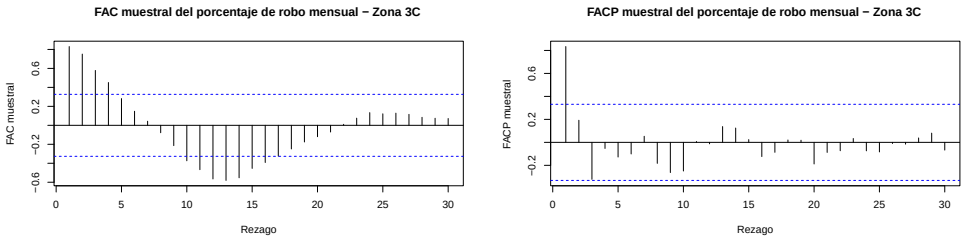


Figura 6.80: Serie temporal del porcentaje de robos mensual de la zona 3C.



(a) FACM (b) FACPM

Figura 6.81: Gráficas de FACM y de FACPM de la serie del porcentaje de robo de la zona 3C.

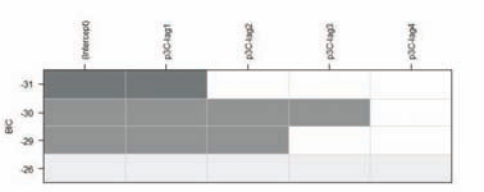
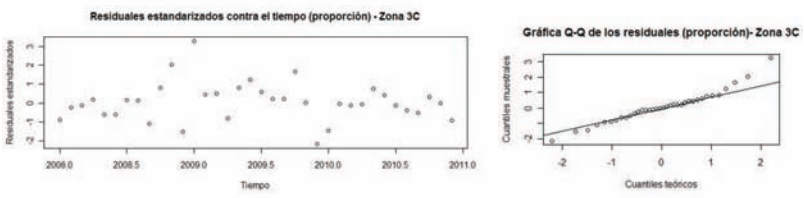


Figura 6.82: Subconjunto óptimo para el modelo de la serie de porcentaje de robo mensual de la zona 3C.



(a) Residuales contra el tiempo (b) Gráfica QQ

Figura 6.83: Gráficas para validar la normalidad de los residuales.

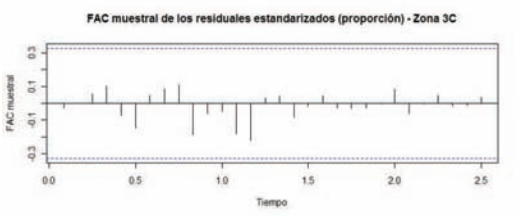


Figura 6.84: Función de autocorrelación muestral de los residuales del modelo ajustado.

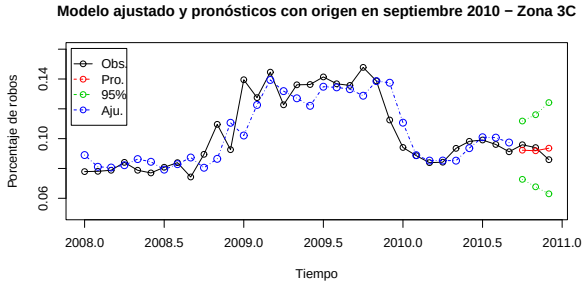


Figura 6.85: Modelo ajustado y pronósticos del último trimestre.

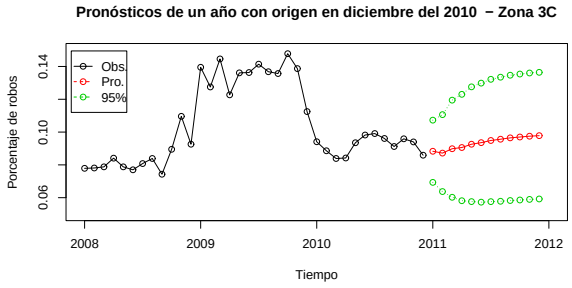


Figura 6.86: Pronósticos mensuales del año 2011.

6.2.5. Zona 4A

Modelo con tendencia estocástica ARIMA(4,1,2)

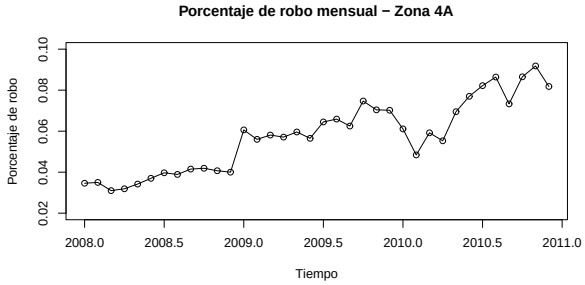
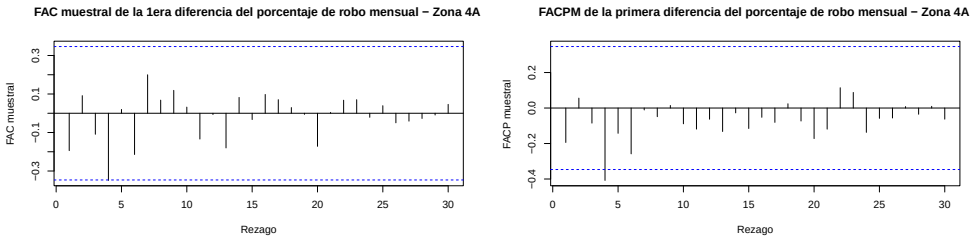


Figura 6.87: Serie temporal del porcentaje de robo mensual de la zona 4A.



(a) FACM (b) FACPM

Figura 6.88: Gráficas de FACM y de FACPM de la primera diferencia del porcentaje de robo mensual de la zona 4A.

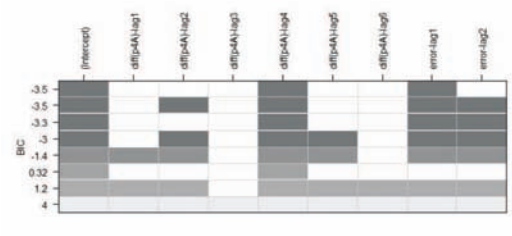
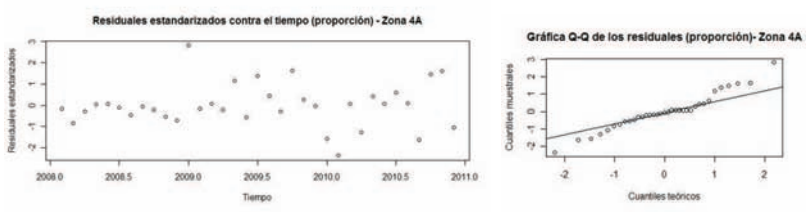


Figura 6.89: Subconjunto óptimo para el modelo de porcentaje de robo mensual de la zona 4A.



(a) Residuales contra el tiempo (b) Gráfica QQ

Figura 6.90: Gráficas para validar la normalidad de los residuales.

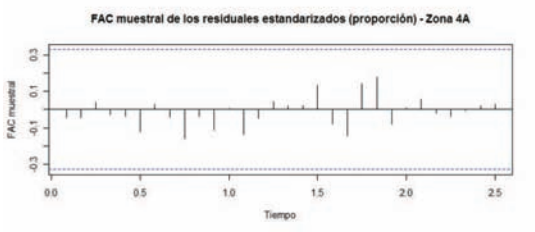


Figura 6.91: Función de autocorrelación muestral de los residuales del modelo ajustado.

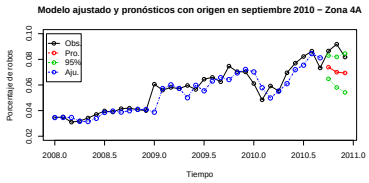


Figura 6.92: Modelo ajustado y pronósticos del último trimestre.

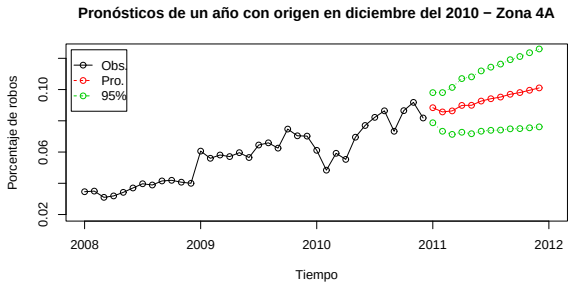


Figura 6.93: Pronósticos mensuales del año 2011.

6.2.6. Zona 4B

Modelo con tendencia estocástica ARIMA(1,1,0) con atípicos AO en 2009-01 y 2009-12

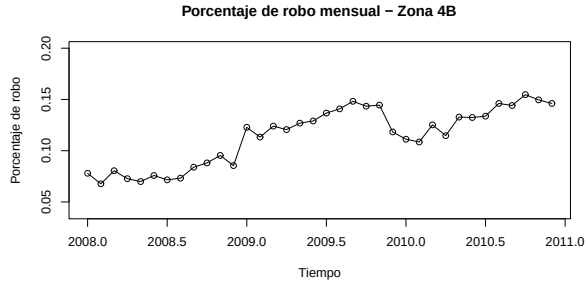
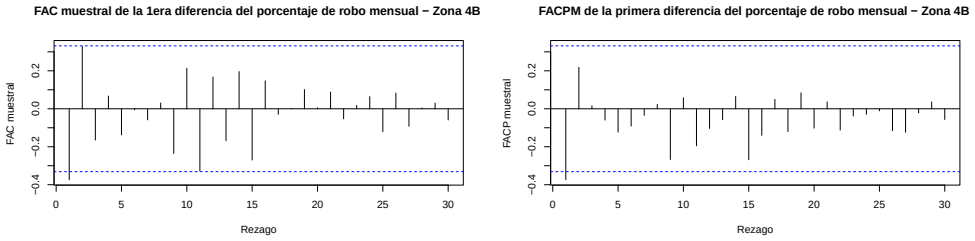


Figura 6.94: Serie temporal del porcentaje de robos mensual de la zona 4B.



(a) FACM (b) FACPM

Figura 6.95: Gráficas de FACM y de FACPM de la primera diferencia del porcentaje de robo mensual de la zona 4B.

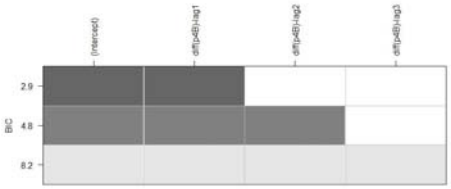
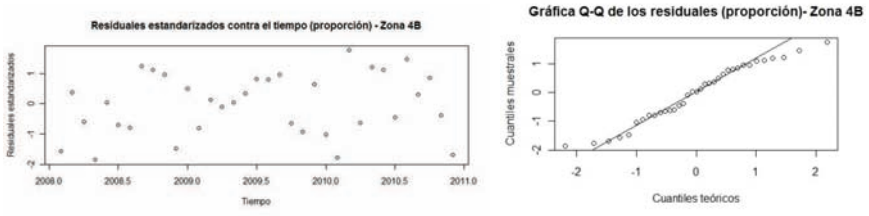


Figura 6.96: Subconjunto óptimo para el modelo de porcentaje de robo mensual de la zona 4B.



(a) Residuales contra el tiempo (b) Gráfica QQ

Figura 6.97: Gráficas para validar la normalidad de los residuales.

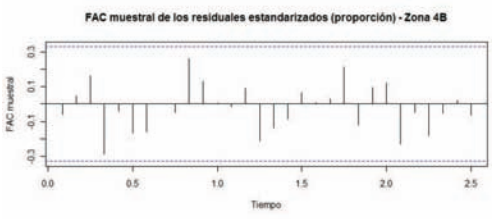


Figura 6.98: Función de autocorrelación muestral de los residuales del modelo ajustado.

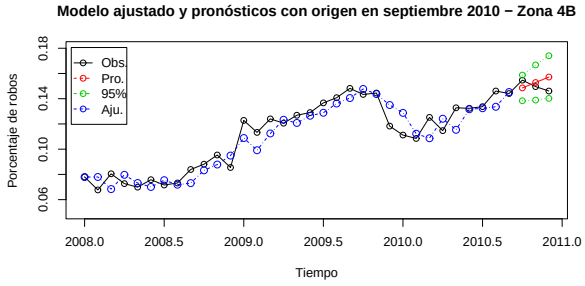


Figura 6.99: Modelo ajustado y pronósticos del último trimestre.

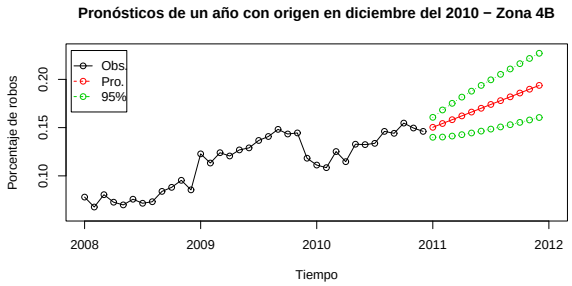


Figura 6.100: Pronósticos mensuales del año 2011.

Bibliografía

[1] Molinaro L. *Lecciones de técnica actuarial de los seguros contra los daños.* Textos universitarios, dirección general de publicaciones. 1976.

[2] Osorio González G. *Manual Básico del Seguro.* Paraguay, 2003.

[3] Everitt B., Hothorn T. *An Introduction to Applied Multivariate Analysis with R.* Springer. 2011.

[4] Box George, Jenkins Gwilym, Reinsel Gregory. *Time Series Analisys: Forecasting and Control.* John Wiley & Sons. Fourth Edition. 2008.

[5] Jiménez Palacios C. *Transición del.* 2012.

[6] The R Development Core Team. *R: A Language and Environment for Statistical Computing Reference Index*, Version 2.13.1.

[7] Wie William W.S. *Time Series Analysis : Univariate and Multivariate Methods.* Editorial Addison-Wesley , 2da edición. 2006.

[8] Bartlett, M. S. *On the Theoretical Specification of sampling properties of autocorrelated time series*, J. Royal Stat. Soc., B8,274. 1946.

[9] Durbin, J. *The fitting of time series models*, Review of the Institute of Intenratiotral Statistics, 28,233-244. 1960.

[10] Box, G. E. P., and Cox, D. R. *An analysis of transformations*, J. Roy Stat. Soc. Ser. B. 26,211-252. 1964.

[11] Cryer Jonathan, Chang Kungsik. *Time Series Analysis with Applications in R.* Springer, 2da edición, 2008.

[12] Dickey, D. A. y Fuller, W. A., *Likelihood ratio statistics for autoregressive time series with a unit root*, Econometrica 49. 1981

[13] Fuller, W. A., *Introduction to Statistical Time Series*, 2nd ed. New York: John Wiley & Sons. 1996.

[14] Said, S. E. and Dickey, D. A. (1984). *Testing for unit roots in autoregressive-moving average models of unknown order.* Biometrika, 71. 1984.

[15] Newbold, P. *The exact likelihood function for a mixed autoregressive moving average process*, Biometrika, 61,423-426. 1974.

[16] Shumway, R. H. and Stoffer, D. S., *Time Series Analysis and Its Applications (with R Examples)*, 3nd ed. New York: Springer. 2011.

[17] Box, G. E. P. and Pierce, D. A., *Distribution of residual correlations in autoregressive- integrated moving average time series models.* Journal of the American Statistical Association, 65, 1509-1526. 1970.

[18] Ljung, G. M. and Box, G. E. P., *On a measure of lack of fit in time series models.* Biometrika, 65, 553-564. 1978.

[19] Hurvich, C. M. and Tsai, C. L., *Regression and time series model selection in small samples.* Biometrika, 76, 2, 297-307.1989.

[20] Schwartz, G., *Estimating the dimension of a model.*, Ann. Statist., 6,461-464. 1978.

[21] Hannan E. y Rissanen J., *Recursive estimation of mixed autoregressive- moving average order.* Biometrika, 69. 1982.

[22] Furnival G. M. y Wilson R. W., *Regressions by leaps and bounds.* Technometrics, 16. 1974.

[23] Abraham, B., and Wei, W. W, S., *Inferences about the parameters of a time series model with changing variance.*, Metrika, 31,183-194. 1984.

[24] Box, G. E. P., and Tiao, G. C., *Intervention analysis with applications to economic and environmental problems*, J. Amer. Starist. Assoc., 70,70-79.

[25] Fox A. J., *Outliers in time series*, J. Roy. Statist. Soc., Ser. B, 43,350-363. 1972.

[26] Goldberg, S. I., *Introduction to Difference Equations.* New York: Science Editions. 1958.

[27] Kaas Rob, Goovaerts Marc, Dhaene Jan, Denuit Michel. *Modern Actuarial Risk Theory Using R.* Second edition, Springer. 2008

[28] Mendehall William, Wackerly Dennis. *Estadística matemática con aplicaciones.* 6ta edición: Thompson. 2002

[29] Faraway Julian. *Linear Models with R.* Chapman & Hall/CRC Texts in Statistical Science. 240 páginas. 2004.

[30] Verzani John. *Using R for Introductory Statistics.* Chapman & Hall. 2005.

[31] Everitt Brian S., Landau, S., Leese, M., and Stahl, D., *Cluster Analysis*, Chichester, UK: John Wiley & Sons, 5th edition. 2011

[32] Verzani John. *Using R for Introductory Statistics.* Chapman & Hall. 2005.

[33] Dobson Annette J. *An introduction to Generalized Linear Models.* Chapman & Hall. Second Edition. 2002

[34] Tsay Ruey S. *Analisys of financial time series.* Jonh Wiley & Sons, third edition. 2010.