UNIVERSIDAD CENTRAL DE VENEZUELA

FACULTAD DE CIENCIAS

ESCUELA DE MATEMÁTICAS

POSTGRADO EN MODELOS ALEATORIOS

PREDICCIÓN DE PRECIOS DEL MERCADO MEXICANO DE AUTOS SEMINUEVOS UTILIZANDO MODELOS DE SERIES TEMPORALES

Trabajo de Grado de Maestría

A ser desarrollado por el:
LIC. ADRIAN BALMORE VALENCIA AVILA
Firma: _____

Tutor:
Dr. JOSÉ BENITO HERNÁNDEZ
Firma: _____

UNIVERSIDAD CENTRAL DE VENEZUELA
FACULTAD DE CIENCIAS
COMISIÓN DE ESTUDIOS DE POSTGRADO

Comisión de Estudios de Postgrado

VEREDICTO

Quienes suscriben, miembros del jurado designado por el Consejo de la Facultad de Ciencias de la Universidad Central de Venezuela, para examinar el **Trabajo de Grado** presentado por: Adrian Balmore Valencia Avila, **Cédula de identidad** 19.965.817, bajo el título "**PREDICCIÓN DE PRECIOS DEL MERCADO MEXICANO DE AUTOS SEMINUEVOS UTILIZANDO MODELOS DE SERIES TEMPORALES**", a fin de cumplir con el requisito legal para optar al grado académico de **MAGÍSTER SCIENTIARUM, MENCIÓN MODELOS ALEATORIOS**, dejan constancia de lo siguiente:

1.- Leído como fue dicho trabajo por cada uno de los miembros del jurado, se fijó el día 21 de enero de 2019 a las 12:00 pm, para que el autor lo defendiera en forma pública, lo que éste hizo en el **Postgrado de Matemática**, mediante un resumen oral de su contenido, luego de lo cual respondió satisfactoriamente a las preguntas que le fueron formuladas por el jurado, todo ello conforme con lo dispuesto en el Reglamento de Estudios de Postgrado.

2.- Finalizada la defensa del trabajo, el jurado decidió **Aprobarlo** por considerar, sin hacerse solidario con la ideas expuestas por el autor, que se ajusta a lo dispuesto y exigido en el Reglamento de Estudios de Postgrado.

Para dar este veredicto, el jurado estimó que el trabajo proporciona un aporte al estudio de series de tiempo aplicadas a problemas reales, en particular en el área de ventas de automóviles donde fueron desarrollados, permitiendo la creación de modelos que puedan manejar grandes volúmenes de datos usando herramientas computacionales.

En fe de lo cual se levanta la presente ACTA, a los 21 días del mes de enero del año 2019, conforme a lo dispuesto en el Reglamento de Estudios de Postgrado, actuó como Coordinador del jurado el **tutor Dr. José Benito Hernández.**

Dra. Mairene Colina
C.I.12.761.954
Institución UCV

Dra. Elvia Flores
C.I. 4.128.598
Institución UCV

Dr. José Benito Hernández
C.I. 10.867.203
Institución UCV
Tutor

RESUMEN

En latinoamérica, la industria automotriz juega un papel importante; específicamente en México, es un mercado de gran relevancia para el desarrollo del país, ya que genera el 3.6% del PIB mexicano y es un importante factor dentro de su economía.

Dentro del sector automotriz, se encuentra el mercado de seminuevos, el cual se encarga de la comercialización de vehículos usados y ha presentado un fuerte crecimiento en los últimos cinco años en la economía mexicana. En general, en dicho proyecto se propuso obtener predicciones de precios de mercado para las diferentes marcas y modelos de autos mediante la utilización de modelos de series temporales con el fin de poder tomar desiciones de negocios acertadas que contribuyan al crecimiento de la empresa.

La segmentación de los vehículos para la posterior predicción de sus precios, se realizó basándose en los precios recopilados de los portales y/o anuncios clasificados y su análisis fue implementado con el software R. En relación a la segmentación obtenida de los árboles de clasificación se pudo observar que se presentó incongruencias en las agrupaciones a nivel de marca, ya que marcas consideradas competencia directa como por ejemplo, Mercedes Benz, BMW ó Audi, resultaron en grupos diferentes.

Este inconveniente se resolvió con el método de K-medias, sin embargo, luego de realizar las predicciones se encontraron varios problemas con el supuesto de estacionaridad, por lo que se tuvo que realizar ciertos tratamientos a los datos y diferenciación para lograr el cumplimiento de esta condición. Por otra parte, en la segmentación obtenida de los árboles de clasificación se pudo observar que aquellos grupos con mayor cantidad de registros, es decir, mayor cantidad de modelos de vehículos obtuvieron una mejor predicción de los precios en comparación al resto de los clusters.

Palabras clave: Arboles de Clasificación, K-medias, Cluster, Series Temporales, ARMA, ARIMA, MA, AR.

ÍNDICE GENERAL

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

INTRODUCCIÓN

El sector automotriz es una de las industrias de más impacto en la economía mundial, dada su elevada participación en la producción, en el valor agregado, en el empleo y en las exportaciones; asi mismo, se le considera como uno de los más dinámicos y modernos pues por su desempeño, es uno de los sectores económicos más importantes por los ingresos que genera. La producción automotriz constituye una industria de alta complejidad tecnológica por involucrar insumos de alta diversidad y especificidad; está integrado por una serie de actividades que van desde la fabricación de diversas partes automotrices hasta el ensamblado de automóviles en general.

En latinoamérica, la industria automotriz juega un papel importante; especificamente en México, es un mercado de gran relevancia para el desarrollo del país, ya que genera el 3.6% del PIB mexicano y es un importante factor dentro de su economía. Otra de las características importantes es su grado de sensibilidad ante el proceso de apertura que ha caracterizado a la economía mexicana en los últimos años, ya que las empresas ensambladoras han decidido invertir en este país por sus condiciones geográficas privilegiadas, mano de obra barata, bajos costos de operación y por el Tratado de Libre comercio de América del Norte.

Dentro del sector automotriz, se encuentra el mercado de seminuevos, el cual se encarga de la comercialización de vehículos usados y ha presentado un fuerte crecimiento en los últimos cinco años en la economía mexicana.

El comportamiento de este mercado sirvió como base para llevar a cabo dicho trabajo de grado, en donde se desea predecir los precios de diferentes marcas de vehículos en el mercado mexicano mediante series temporales. Dicho estudio se implementará a traves del *software* libre R, en donde se pretende realizar en primera instancia una segmentación mediante árboles de clasificación para obtener así grupos de vehículos homogéneos lo que proporcionará la confianza necesaria para hallar los mejores *clusters* de autos para una correcta y prolongada validez en su estimación.

En general, en dicho proyecto se propondrá obtener predicciones de precios de mercado para las diferentes marcas y modelos de autos mediante la utilización de modelos de series temporales con el fin de poder tomar desiciones de negocios acertadas que contribuyan al crecimiento de la empresa.

Especificamente se llevarán a cabo los siguientes objetivos:

- Estudiar y adquirir las habilidades necesarias en el manejo de árboles de clasificación y series temporales.
- Determinar a jucio experto las variables que se considerarán para la clasificación de los autos.
- Obtener una agrupación de marcas y modelos usando árboles de clasificación con el fin de reducir la cantidad de grupos en los cuales se aplicarán los modelos predictivos de series temporales; este análisis se llevará a cabo con ayuda del paquete *rpart* del *software* libre R.
- Conseguir el mejor modelo de series temporales (AR, MA o ARMA) que se adapte a los grupos de autos obtenidos en la clasificación.
- Comparar los resultados obtenidos en los modelos de series temporales contra un modelo empírico desarrollado.

CAPÍTULO 1

DESCRIPCIÓN DE LA EMPRESA

(KAVAK MÉXICO)

La industria automotriz al ser un mercado de gran envergadura; la compra y venta de automóviles por internet se ha convertido en un mercado que presenta riesgos tanto para compradores como para vendedores. Los fraudes y el peligro de caer en una estafa, son problemáticas comunes que muchas veces inciden en la intención de compra.

KAVAK es una *startup* que busca cambiar el cómo se lleva a cabo la compra y venta de un auto usado en México, ya que propone para quien vende su auto, una garantía de venta; mientras que para el comprador, ofrece un vehículo en perfectas condiciones mecánicas ya que están certificados por técnicos especializados y verificados legalmente por un equipo de gestores. Además otorga la posibilidad de devolver el vehículo si no cumple con sus expectativas recibiendo así mismo un reembolso por el valor total de la compra.

En su portal se busca humanizar la compra de un auto a través de internet, agilizar cualquier trámite y garantizar también la integridad de quien necesita vender su vehículo, ya que cuenta con personas con experiencia en ventas por internet, y que además, han sufrido en carne propia los contratiempos de la compra de autos seminuevos. En general, sus principales aspectos son:

- Calidad: todo auto pasa un riguroso proceso de inspección de 240 puntos antes de convertirse en un auto Kavak.
- Transparencia: ayudan a compradores y vendedores en un ambiente confiable, sencillo y por encima de todo transparente, es decir, se elimina todo el sufrimiento del proceso de compra y venta de autos usados, ya que se encargan directamente de todos los trámites y de la entrega del auto.
- Eficiencia: por un lado, se utiliza las últimas tecnologías para comparar precios de mercado en tiempo real y garantizando así que las ofertas sean las más competitivas. Por otro lado, al ser una plataforma *on line* no tienen costosas salas de exhibición ni vendedores comisionados lo que le permite tener bajos costos y trasladar esos beneficios a sus clientes.

Entre las áreas que constituye KAVAK se encuentra el área de *Pricing*, la cual se enfoca en la búsqueda de precios de mercado y generación de ofertas para clientes interesados, apoyándose de la experticia de los especialistas que la conforman.

CAPÍTULO 2

MARCO TEÓRICO

Para poder realizar la predicción de precios deseada, se dividirá el proyecto en dos secciones. La primera sección consta de realizar una agrupación de marcas y modelos para obtener grupos homogéneos mediante árboles de clasificación. Luego de realizar la segmentación se pasará a la sección de predicción con modelos de series temporales.

2.1 ANÁLISIS DE CONGLOMERADOS

El análisis *cluster*, también conocido como análisis de conglomerados es un conjunto de técnicas estadística multivariantes utilizadas para clasificar a un conjunto de individuos en grupos homogéneos, es decir, se busca dividir un conjunto de objetos en grupos de forma que los perfiles de los objetos en un mismo grupo sean muy similares entre sí (cohesión interna del grupo) y los de los objetos de *clusters* diferentes sean distintos (aislamiento externo del grupo).

La creación de los *cluster* puede efectuarse de diversos modos en función de los algoritmos de cálculo utilizados. Entre los más usados destacan los métodos jerárquicos aglomerativos, los cuales comienza con los objetos o individuos de modo individual; de este modo, se tienen tantos *clusters* iniciales como objetos. Luego se van agrupando de modo que los primeros en hacerlo son los más similares y al final, todos los subgrupos se unen en un único *cluster*; y los métodos jerárquicos divisivos que actúan al contrario, parten de un grupo único con todas las observaciones y se van dividiendo según lo lejanos que estén.

2.1.1 ÁRBOLES DE CLASIFICACIÓN

Para la realización de este proyecto se usará en primera instancia los árboles de clasificación, lo cual es una forma gráfica y analítica de representar todos los eventos (sucesos) que pueden surgir a partir de una decisión asumida en cierto momento. Así mismo, ayudan a tomar la decisión "más acertada", desde un punto de vista probabilístico, ante un abanico de posibles decisiones. Los

métodos basados en árboles son métodos jerárquicos bastante populares en *data mining*, pudiéndose usar para clasificación y regresión. Son útiles para la exploración inicial de datos, y apropiados cuando hay un número elevado de datos y existe incertidumbre sobre la manera en que las variables explicativas deberían introducirse en el modelo.

Los modelos de árbol, donde la variable de destino puede tomar un conjunto finito de valores se denominan árboles de clasificación. En estas estructuras de árbol, las hojas representan etiquetas de clase y las ramas representan las conjunciones de características que conducen a esas etiquetas de clase. Por el contrario, los árboles de decisión, donde la variable de destino puede tomar valores continuos (por lo general números reales) se llaman árboles de regresión.

Entre los principales usos del análisis basado en árboles se encuentra:

- Segmentación: identifica personas que probablemente sean miembros de una clase concreta.
- Estratificación: asigna los casos a una categoría de entre varias, por ejemplo, grupos de alto riesgo, bajo riesgo y riesgo intermedio.
- Predicción: crea reglas y las utiliza para predecir eventos futuros. Las predicciones también pueden significar intentos de relacionar atributos predictivos con valores de una variable continua.
- Reducción de datos y clasificación de variables: selecciona un subconjunto útil de predictores a partir de un gran conjunto de variables para utilizarlo en la creación de un modelo paramétrico formal.
- Identificación de interacción: identifica las relaciones que pertenecen sólo a subgrupos específicos y las especifica en un modelo paramétrico formal.
- Fusión de categorías y creación de tramos de variables continuas: codifica las variables continuas y las categorías de los predictores del grupo, con una pérdida mínima de información.

2.1.2 ANÁLISIS DE K - MEDIAS

K-*means* o K-medias es uno de los algoritmos de aprendizaje no supervisado más simples para resolver el problema de la clusterización o agrupamiento, el cual tiene como objetivo la partición

de un conjunto de n observaciones en k grupos (prefijado) en el que cada observación pertenece al grupo cuyo valor medio es más cercano, es decir se aproxima por etapas sucesivas un cierto número de *clusters* haciendo uso de los centroides de los puntos que deben representar. .

El algoritmo se compone de los siguientes pasos:

1. Sitúa K puntos en el espacio en el que "viven" los objetos que se quieren clasificar. Estos puntos representan los centroides iniciales de los grupos, es decir, se calcula la media por variable de los registros que pertenecen al mismo grupo.
2. Asigna cada objeto al grupo que tiene el centroide más cercano.
3. Tras haber asignado todos los objetos, recalcula las posiciones de los K centroides.
4. Repite los pasos 2 y 3 hasta que los centroides se mantengan estables. Esto produce una clasificación de los objetos en grupos que permite dar una métrica entre ellos.

Aunque se puede probar que este algoritmo siempre termina, no siempre la distribución que se alcanza es la más óptima, ya que es muy sensible a las condiciones iniciales.

2.2 SERIES TEMPORALES - MODELO EMPÍRICO

Una serie temporal se define como una colección de observaciones de una variable reunida secuencialmente en el tiempo. Estas observaciones se suelen recoger en instantes de tiempo equiespaciados. Si los datos se toman en instantes temporales de forma continua, se debe o bien digitalizar la serie, es decir, tomar sólo los valores en instantes de tiempo equiespaciados, o bien acumular los valores sobre intervalos de tiempo.

La característica fundamental de las series temporales es que las observaciones sucesivas no son independientes entre sí, y el análisis debe llevarse a cabo teniendo en cuenta el orden temporal de las observaciones. Los métodos estadísticos basados en la independencia de las observaciones no son válidos para el análisis de series temporales porque las observaciones en un instante de tiempo dependen de los valores de la serie en el pasado.

Se pueden considerar varios posibles objetivos:

- **Descripción:** se estudia una serie temporal considerando:

  - Tendencia: si los datos presentan forma creciente.
  - Estacionalidad: si existe influencia de ciertos periodos de cualquier unidad de tiempo.
  - Observaciones extrañas o discordantes: si aparecen *outliers*.

- **Predicción:** se observan los valores de una serie, se pretende normalmente no sólo explicar el pasado, sino también predecir el futuro.

2.2.1 COMPONENTES DE UNA SERIE TEMPORAL

- **Tendencia:** refleja el movimiento de la serie a largo plazo, es decir, la evolución promedio de la serie a largo plazo. La cual se notará por T_t.

- **Estacionalidad:** se notará por E_t y representa fluctuaciones de la serie en periodos de tiempo inferiores a un año que se repiten con periodicidad conocida, es decir, pretende recoger los crecimientos o decrecimientos en los valores de la serie que se producen por el hecho de encontrarnos en una determinada época, en general una estación del año. Las razones de la estacionalidad pueden ser de dos tipos:
  - Físico-naturales: tiempo meteorológico, ciclos biológicos.
  - Institucionales: vacaciones escolares, fiestas horarios comerciales, etc.

2.2.2 ESQUEMAS PARA SERIES TEMPORALES

Habitualmente se reconocen dos tipos de esquema:

- **Esquema (modelo) aditivo:** supone que las observaciones se generan como suma de las dos componentes, es decir,

$$Y_t = T_t + E_t$$

  en este caso cada componente se expresa en el mismo tipo de unidad que las observaciones. La estacionalidad, en este modelo, es independiente de las demás componentes, es decir, no depende del valor que tome el otro componente de la serie.

- **Esquema (modelo) multiplicativo:** supone que las observaciones se generan como producto de las dos componentes, es decir,

$$Y_t = T_t \times E_t$$

en este modelo no se cumple la hipótesis de independencia del esquema aditivo.

2.2.3 PROCEDIMIENTOS PARA DETERMINAR EL TIPO DE MODELO

La idea de los métodos presentados consiste en poner de manifiesto si las fluctuaciones de la serie son aproximadamente constantes o si se modifican en función de la tendencia:

- **Representación grafica de la serie:** realizada la representación grafica de la serie, se observara si las fluctuaciones de esta varían con la tendencia o bien si permanecen constantes. A pesar que este procedimiento no es totalmente fiable, puede ser de utilidad en aquellas series temporales con tendencia muy marcada.

- **Grafico de la desviación típica-media:** Este método se aplica llevando sobre unos ejes cartesianos puntos de la forma (media, desviación típica). Es decir se calculará previamente las medias y las desviaciones típicas para cada instante o periodo y se llevará sobre los ejes cartesianos. Concretamente, se supone un conjunto de observaciones que se extiende a n años y que en cada año se dispone de s observaciones referidas a s estaciones (meses, trimestres,...) Se calculan las medias y desviaciones típicas de cada año y se llevan sobre los ejes cartesianos (las medias sobre el eje de abscisas y las desviaciones típicas sobre el eje de ordenadas). Si la nube de puntos se distribuye aproximadamente en torno a una recta paralela al eje de abscisas, estamos ante un modelo aditivo. Si las desviaciones típicas crecen al aumentar las medias anuales, es razonable pensar que la tendencia aparece multiplicada por las demás componentes y el esquema adecuado será el multiplicativo.

- **Análisis de la variabilidad de las diferencias y cocientes estacionales:** llamaremos diferencia estacional d_{ij} a la diferencia entre dos datos de la misma estación j correspondiente a dos años consecutivos $i-1$ e i, es decir,

  $$d_{ij} = y_{ij} - y_{i-1j}$$

  donde y_{ij} es el valor de la serie en el año i y estación j. De manera similar, definimos los cocientes estacionales k_{ij} como los cocientes entre dos datos de la misma estación correspondientes a dos años consecutivos:

  $$k_{ij} = y_{ij}/y_{i-1j}.$$

una vez calculados los valores d_{ij} y k_{ij}, calculamos la media y la varianza para cada distancia ($\bar{d}$, s_d y $\bar{k}$, s_k) y los coeficientes de variación sobre cada uno de ellos, $cv_d = s_d/\bar{d}$ y $cv_k = s_k/\bar{k}$. Si $cv_d < cv_k$ se elegirá esquema aditivo, en caso contrario se optará por el multiplicativo.

2.2.4 ANÁLISIS DE TENDENCIA

El propósito del ajuste de la tendencia es describir el patrón histórico de los datos. Se consideran dos procedimientos para la determinación de la tendencia en una serie cronológica: el enfoque global y el enfoque local.

- **Enfoque global:** consiste en el ajuste de las observaciones a una función matemática (usualmente por el método de mínimos cuadrados) para la obtención de la tendencia. Esto es posible ya que una serie temporal puede ser considerada como una distribución bidimensional y su nube de puntos se puede ajustar a una función. Los modelos más habituales son:

$$Lineal: \quad T_t = a + b \times t,$$

$$Parabólico: \quad T_t = a + b \times t + c \times t^2,$$

$$Exponencial: \quad T_t = a + \exp^{b \times t}.$$

  donde a, b y c son valores que se estiman a partir de los datos con los que se estudiaran. Este procedimiento además de ser el más exacto, tiene la ventaja de disponer de una medida de la bondad del ajuste, es decir, se puede obtener el valor del R^2, el cual es una medida entre 0 y 1, donde los valores próximos a 1 indican que se reproduce adecuadamente la tendencia. Desde el punto de vista práctico es conveniente tomar una función que sea sencilla, de tal forma que la estimación de los parámetros no sea muy compleja y que por otro lado intente explicar lo esencial del movimiento de la serie.

- **Enfoque local (Método de medias móviles):** este método es más flexible y no exige la suposición de una forma funcional para la tendencia al contrario que el anterior. Este método se usa bien para la obtención de la tendencia o bien como una técnica de transformar las observaciones en otras mas suavizadas para posteriormente ajustar una curva a estos valores. Generalmente entenderemos por suavizamiento de la serie la obtención de unos valores transformados con menos fluctuación. Dado un conjunto de observaciones correspondientes a una serie cronológica, la determinación de las medias móviles de orden p $(MM_{p,t})$ consiste

en aproximar el valor de la tendencia de la variable de interés en el instante t (T_t) mediante el promedio de observaciones cercanas.

Por ejemplo, la media móvil de orden 3 y 5 vienen dadas por:

$$T_t = MM_{3,t} = \frac{Y_{t-1} + Y_t + Y_{t+1}}{3} \qquad T_t = MM_{5,t} = \frac{Y_{t-2} + Y_{t-1} + Y_t + Y_{t+1} + Y_{t+2}}{5}.$$

Las medias obtenidas del proceso anterior se denominan medias móviles de orden o amplitud p. Cada media móvil se asocia al punto medio del intervalo sobre el que ha sido calculada. Este método presenta varias características:

- La elección de la amplitud "p" no tiene por que ser clara; esta elección debe estar ligada a la periodicidad de las fluctuaciones que se desean suavizar, pero este no es un problema fácil que admita una solución general.
- Hay observaciones (las primeras y las últimas) para las que no se dispone de la media móvil por lo que se produce una pérdida de información.
- Cuanto mayor sea la amplitud mejor se eliminaran las irregularidades de la serie, ya que en su cálculo se compensaran las fluctuaciones de un mayor número de observaciones. Por el contrario, cuanto menor sea el orden de las medias móviles, estas reflejaran con mayor rapidez los cambios que puedan producirse en la evolución de la serie. Obsérvese que cuanto mayor sea el orden o amplitud de las medias, mayor será el efecto de suavizamiento, pero también será menor el número de datos para cálculos posteriores. Si el orden o amplitud de las medias móviles es impar, la media móvil se asignara al instante central de la amplitud considerada, pero si el orden o amplitud es par se plantea la necesidad de centrarlas en los mismos instantes de tiempo donde se sitúan las observaciones iniciales.

2.2.5 ANÁLISIS DE ESTACIONARIDAD

La variación estacional nos indica el incremento o disminución que se ha experimentado en un período estacional dado respecto del valor medio referido a todo el año.

En el modelo multiplicativo, la componente estacional de una serie temporal, se mide con un índice denominado índice de variación estacional, expresado en porcentaje y que significa la

fluctuación del valor de la serie respecto al valor de la tendencia media del año. Por ejemplo un índice de variación estacional del 86% significa una disminución del 14% respecto del valor de la tendencia. En el modelo aditivo, para cada estación, la componente estacional indica en términos absolutos la cantidad en que se ha superado o no se ha alcanzado el valor de la tendencia media anual. En este caso la componente estacional se expresa en las mismas unidades que las observaciones.

La determinación de la variación estacional, es un aspecto fundamental en el análisis de una serie, principalmente para estudios comparativos y efectuar predicciones. En ocasiones interesa conocer las variaciones estacionales y eliminarlas del comportamiento global de la serie para poder observar mejor el comportamiento de esta ajeno a causas estacionales. La eliminación de la componente estacional de las observaciones hace comparables cantidades observadas en estaciones distintas y que se pueden presuponer influidas por este hecho. Para la obtención de la variación estacional se utilizan dos procedimientos, en ambos un aspecto común es la eliminación de la tendencia, para ello es necesario disponer de una estimación de la misma.

- **Método de la razón (o diferencia) a la media móvil:** para las series que siguen un esquema multiplicativo, se estudia el método de la razón a la media móvil, por el contrario, si el esquema fuera aditivo se operaría de modo análogo pero en lugar de cocientes se utilizarían diferencias. Si se parte de un esquema multiplicativo:

  $$Y_t = T_t \times E_t$$

  el cociente que aisla el componente estacional de la serie es:

  $$E_t = \frac{Y_t}{T_t}.$$

Luego se procede a calcular el índice de estacionalidad utilizando medias móviles, para esto se siguen los siguientes pasos:

1. Se determina la tendencia por el método de las medias móviles.

2. Se elimina la componente tendencial dividiendo la serie original por la tendencia calculada en el paso anterior si la hipótesis es multiplicativa ó restándola si la hipótesis es aditiva.

3. Se calculan las medias aritméticas para cada estación. Si las observaciones son mensuales tendremos 12 medias: $M_1, M_2, .., M_{11}, M_{12}$; donde M_1 será la media aritmética de los valores

correspondientes a todos los meses de enero, M_2 la media aritmética de los valores correspondientes a todos los meses de febrero. Estas medias nos representan de forma aislada la importancia de la componente estacional, para ello se supone que los errores tienen un comportamiento aleatorio, lo cual implica que en unas observaciones tomará valor negativo y en otras positivo, con lo cual cuando calculamos la media de todos los meses de enero, lo que estamos haciendo es quedarnos únicamente con la estacional.

4. Obtención de la componente estacional. Para ello se calcula la media aritmética anual MA de las medias estacionales $M_1, M_2, M_3...$

   - Si el esquema que sigue la serie es multiplicativo la componente estacional se calcula mediante la obtención de los siguientes índices, conocidos como "índices de variación estacional":

   $$IVE_1 = \frac{M_1}{MA} \times 100, \quad IVE_2 = \frac{M_2}{MA} \times 100, \ldots$$

   habrá tantos índices como período o medias estacionales tengan las observaciones y nos indicarán la importancia de la variación estacional al pasar de un período a otro. Si un índice expresado en tantos por ciento nos da 80 quiere decir que en el período en donde nos encontremos, la magnitud en estudio es un 20% más baja de su tendencia media. La suma de los índices de variación estacional para esta hipótesis será siempre igual al número de índices ó, períodos si están expresados los índices en tantos por uno ó, al número de índices multiplicados por 100 si están expresados en tantos por ciento. Para datos mensuales la suma sería 12 (ó 1200), para datos trimestrales 4 (ó 400), etc.

   - Si el esquema que sigue la serie es aditivo la componente estacional se calcula restando a cada media estacional $M_1, M_2, M_3...$ la media anual MA:

   $$IVE_1 = M_1 - MA$$

   $$IVE_2 = M_2 - MA, \ldots$$

   bajo esta hipótesis, la suma de las componentes estacionales es igual a cero.

2.2.6 PREDICCIÓN

Luego de conocer los componentes de una serie, así como el modelo según el cual se relacionan, se puede conocer el valor de la serie en cualquier momento. La componente irregular (residuos)

es desconocida por naturaleza, del resto de las componentes tenemos a lo sumo una estimación, y el modelo es un tipo de esquema que imponemos artificialmente para facilitar la aproximación al estudio de la serie. Por lo tanto no se podrá conocer exactamente el valor de la serie en un momento futuro y nos conformaremos con poder hacer una predicción lo mejor posible.

La predicción se hará en base al tipo de modelo (aditivo o multiplicativo) y a las componentes tendencia y variación estacional. En un esquema aditivo, la estimación de valor de la serie en un momento futuro estará dada por:

$$Y_t = T_t + IVE_t/100$$

mientras que en un esquema multiplicativo, la estimación de valor de la serie en un momento futuro estará dada por:

$$Y_t = T_t \times IVE_t/100.$$

2.3 SERIES TEMPORALES - PROCESOS AUTOREGRESIVOS

Cuando se construye un modelo de series temporales univariante el objetivo no es conseguir el "verdadero" modelo. Es preciso ser conscientes de que estamos tratando de modelar una realidad compleja y el objetivo es lograr un modelo parsimonioso y suficientemente preciso que represente adecuadamente las características de la serie recogidas fundamentalmente en la función de autocorrelación. En este proyecto uno de los enfoques para analizar las series temporales será el dominio del tiempo, por lo que en esta sección se muestran algunas medidas teóricas que permiten hacer una descripción de las series temporales y sus estimadores, además, se introducen los modelos AR(p), MA(q) y ARMA(p,q) como aproximaciones al modelo lineal general.

2.3.1 CARACTERÍSTICAS DE UNA SERIE DE TIEMPO

Comenzaremos por caracterizar sus funciones de autocorrelación para conocer sus propiedades y posteriormente utilizarlos para modelar series y predecir.

2.3.1.1 FUNCIÓN MEDIA

Sea $\{x_t\}$ una serie de tiempo, la función media de x_t, si existe, viene dada por:

$$\mu_t=\mu_{xt}=E(x_t)=\int_{-\infty}^{\infty} x f(x)dx$$

donde E denota el operador de esperanza. Mientras que la media muestral de $\{x_t\}$ es:

$$\bar{x}=\frac{1}{n}\sum_{t=1}^{n} x_t.$$

2.3.1.2 FUNCIÓN DE AUTOCOVARIANZA (ACVF)

La autocovarianza es la covarianza entre $\{x_t\}$ y su valor en otro instante de tiempo $\{x_s\}$, es un indicador de la dependencia lineal entre dos observaciones de la serie en tiempos diferentes. Asumiendo que la varianza es finita y μ constante, la función de autocovarianza conocida como ACVF (*Autocovariance Function*) se define como:

$$\gamma(s,t)=cov(x_s,x_t)=E\left[(x_s-\mu)(x_t-\mu)\right]$$

para todo t y s $\in \mathbb{Z}$.

Sea $s=t+h$, donde h es el retraso o tiempo de traslación, entonces:

$$\gamma(s,t)=\gamma(t+h,t)=\gamma(h,0)=\gamma(h)$$

así

$$\gamma(h)=E\left[(x_{t+h}-\mu)(x_t-\mu)\right].$$

Finalmente la función de ACVF muestral para $h=0,1,\ldots,n-1$, se define como:

$$\hat{\gamma}(h)=\frac{1}{n}\sum_{t=1}^{n-h}(x_{t+h}-\bar{x})(x_t-\bar{x}).$$

2.3.1.3 FUNCIÓN DE AUTOCORRELACIÓN (ACF)

La función de autocorrelación ACF (*Autocorrelation Function*) para todo t y s y para $h = 0, 1, \ldots, n-1$, se define como:

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

Así, la ACF muestral es:

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

2.3.1.4 SERIES DE TIEMPOS ESTACIONARIAS

Una serie estrictamente estacionaria se define como aquella donde el comportamiento probabilístico de cada sucesión de observaciones $\{x_{t_k}\}$ es exactamente igual a la serie trasladada en el tiempo $\{x_{t_k+h}\}$ donde $h = 0, \pm 1, \pm 2, ...$ es un retraso o salto, esta versión de estacionaridad resulta ser poco útil debido a su complejidad. Una definición alternativa de estacionaridad es que una serie de tiempo es débilmente estacionaria si se cumple que la función media es constante y no depende del tiempo t, la varianza es finita y la función de covarianza $\gamma(s,t)$ solo depende de $|s-t|$, es decir, la diferencia de s y t. De aquí en adelante la estacionaridad débil se mencionará como estacionaridad, y la función media será $\mu_t = \mu$.

2.3.1.5 FUNCIÓN DE AUTOCORRELACIÓN PARCIAL (PACF)

La función de autocorrelación parcial de un proceso estacionario denotada por Φ_{hh} , para $h = 1, 2, ...$, es definida por Shumway [6] como:

$$\Phi_{11} = corr(x_{t+1}, x_t) = \rho(1),$$

$$\Phi_{hh} = corr(x_{t+h} - \hat{x}_{t+h}, x_t - \hat{x}_t),\ h \geq 2$$

donde $(x_{t+h} - \hat{x}_{t+h})$ y $(x_t - \hat{x}_t)$ no están correlacionados con $\{x_{t+1}, \ldots, x_{t+h-1}\}$.

2.3.2 PROCESOS AUTOREGRESIVOS

En la sección para el modelo empírico se consideró como únicos componentes para la elaboración de las series temporales, la tendencia y la estacionaridad. Sin embargo, en este apartado se representará la serie de tiempo como la suma de tres componentes,

$$x_t = T_t + E_t + \varepsilon_t$$

donde T_t es la tendencia, E_t representa la estacionalidad y ε_t es un ruido aleatorio, que se denominará como ruido blanco y el cual se define como una sucesión de variables aleatorias con media cero, varianza constante y que no mantienen correlación. Si presenta una distribución normal entonces es llamado ruido blanco gaussiano.

Lo anterior con lleva a un proceso autorregresivo, lo cual es un modelo de regresión en el que las variables explicativas son la misma variable dependiente retardada. Los modelos autoregresivos que se estudiarán será AR, MA y ARMA, cuya estructura es presentada a continuación.

2.3.2.1 MODELO AUTOREGRESIVO AR(p)

El modelo AR se basa en que el valor x_t puede ser expresado en función de sus p valores pasados $x_{t-1}, x_{t-2}, ..., x_{t-p}$, y se define como:

$$x_t = \nu + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \ldots + \phi_p x_{t-p} + \varepsilon_t$$

en donde ε_t es ruido blanco, ϕ_i para $i = 1, ..., p$ son parámetros con $\phi_p \neq 0$, la variable x_t es estacionaria y $\upsilon = \mu(1 - \phi_1 - \ldots - \phi_p)$.

En primer lugar es preciso comprobar si el proceso AR(p) cumple las condiciones de estacionaridad para cualquier valor de los parámetros. Esta comprobación que es relativamente sencilla para el modelo AR(1), se complica para modelos autorregresivos de orden mayor cuyos parámetros han de satisfacer restricciones complejas para ser estacionarios. Sin embargo, un proceso autorregresivo finito AR(p) se considerará estacionario si y sólo si $\mid \phi_i \mid < 1$ para todo i.

Las características del proceso AR(p) estacionario son:

1. Media

$$E(x_t) = E(\phi_1 x_{t-1} + \phi_2 x_{t-2} + \ldots + \phi_p x_{t-p} + \varepsilon_t) =$$

$$E(x_t) = \phi_1 E(x_{t-1}) + \phi_2 E(x_{t-2}) + \ldots + \phi_p E(x_{t-p}) + E(\varepsilon_t)$$

Como el proceso es estacionario, la media es constante:

$$(1 - \phi_1 - \phi_2 - \ldots - \phi_p) E(x_t) = 0 \quad \longrightarrow \quad E(x_t) = 0.$$

2. La función de autocorrelación, ρ_k, $k = 0, 1, 2, \ldots$ de un proceso AR(p) decrece exponencialmente hacia cero sin truncarse.

2.3.2.2 MODELO AUTOREGRESIVO MA(q)

El modelo de promedio móvil MA (*Moving Average*) de orden q viene dado por

$$x_t = \nu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q},$$

donde el valor x_t se puede representar como la combinación lineal del ruido blanco ε_t y su pasado y se asume sus coeficientes $\theta_1, \theta_2, ..., \theta_q (\theta_q \neq 0)$.

El modelo MA(q) es una generalización del modelo MA(1), por lo tanto sus características van a ser muy similares. Ahora bien, al introducir más retardos de la perturbación en el modelo, la memoria aumenta y la estructura dinámica representada por el modelo es más rica. Si una perturbación entra en el momento t, influye en x_t , en x_{t+1}, y en valores posteriores hasta x_{t+q}. Por lo tanto, la perturbación ε_t en un modelo MA(q) permanece q periodos en el sistema. Esta memoria más larga del MA(q) se verá reflejada en la estructura de la función de autocovarianzas y/o autocorrelación.

Ahora, para comprobar si el modelo MA(q) es estacionario para cualquier valor de los parámetros de medias móviles, se comprueba si se cumplen las condiciones de estacionaridad. Como el modelo de medias móviles finito de orden q no es más que el resultado de truncar el modelo lineal general a partir del retardo q , será estacionario bajo las mismas condiciones que el modelo general, es decir, si se cumple la condición de que la sucesión de los parámetros del modelo es convergente:

$$\sum_{i=1}^{q} \theta_i^2 < \infty.$$

Como el número de parámetros de un MA(q) es finito, esta condición siempre se cumple y, por lo tanto, todos los modelos de medias móviles finitos son siempre estacionarios para cualquier valor de sus parámetros.

En relación a su función de autocovarianzas el modelo de MA(q) al tener media constante e igual a cero, varianza constante y finita, su función de autocovarianzas está truncada a partir del retardo q, es decir,

$$\rho_k = \begin{cases} \rho_k \neq 0 & k = 1, 2, \ldots, q \\ \rho_k = 0 & k > q. \end{cases}$$

2.3.2.3 MODELO AUTOREGRESIVO ARMA(p, q)

El modelo autoregresivo de promedio móvil de orden p y q , conocido como ARMA(p, q) (*Auto-regressive Moving Average*), donde p es el orden autoregresivo y q es el orden de promedio móvil, se define como:

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \ldots + \phi_p x_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q}$$

con $\phi_p \neq 0$ y $\theta_q \neq 0$,

En relación a la estacionaridad del modelo ARMA(p,q) para cualquier valor de sus parámetros, se cumple si y solo si $\mid \phi_i \mid < 1$, es decir, las condiciones de estacionaridad del modelo ARMA(p,q) vienen impuestas por el proceso autoregresivo, dado que las medias móviles finita siempre son estacionarias.

Así mismo, el modelo ARMA(p,q) va a compartir las características de los modelos AR(p) y MA(q) ya que contiene ambas estructuras a la vez, entiéndase como, media cero, varianza constante y finita y una función de autocovarianzas infinita. Por otra parte, con respecto a la función de autocorrelación es infinita decreciendo rápidamente hacia cero pero sin truncarse.

2.4 PRUEBAS DE HIPÓTESIS

En la sección anterior se describieron los modelos AR, MA y ARMA los cuales tienen como supuestos que la serie es estacionaria y los residuos son ruido blanco, para determinar la veracidad de esto se realizarán las siguientes pruebas de hipótesis:

2.4.1 TEST DE ESTACIONARIDAD DE KWIATOWSKI, PHILLIPS, SCHMIDT Y SHIN (KPSS)

Una serie de tiempo estacionaria es una serie cuyas propiedades estadísticas, tales como la media, la varianza y la autocorrelación, se mantienen constantes en el tiempo. Muchas de las metodologías estadísticas para realizar estimación de series de tiempo se basan en el supuesto de que estas series sean estacionarias (o por lo menos se conviertan a observaciones aproximadamente estacionarias).

Otra razón de interés para tratar de estacionarizar una serie de tiempo, es para obtener estadísticos muestrales de importancia que son informativos del comportamiento futuro. Por ejemplo, si una serie de tiempo aumenta consistentemente con el tiempo, la media y la varianza crecerán con el tamaño de la muestra y tenderán a subestimar estas medidas en el futuro.

Kwiatkowski, Phillips, Schmidt, y Shin [7] presentan esta prueba para probar la hipótesis nula de estacionaridad, la cual consiste en la representación de la serie como $Y_t = \alpha_0 + \mu_t + \varepsilon_t$, donde α_0 es constante, $\mu_t = \mu_{t-1} + u_t$ con $u_t \sim iid(0, \sigma_u^2)$ y ε_t un proceso estacionario, donde la hipótesis nula y la hipótesis alternativa son respectivamente:

$$\begin{cases} H_0 & \sigma_u^2 = 0 \\ H_\alpha & \sigma_u^2 > 0. \end{cases}$$

Si la hipótesis nula es aceptada entonces μ_t es una constante y Y_t es estacionaria. El estadístico de prueba KPSS está dado por:

$$KPSS = (n^{-2} \sum_{t=1}^{n} S_t^2)/\hat{\sigma}_{\hat{u}}^2$$

donde $S_t = \sum_{i=1}^{n} \varepsilon_t$ es la suma parcial de los residuales y $\hat{\sigma}_{\hat{u}}^2 = n^{-1} \sum_{i=1}^{n} \hat{\varepsilon}_t^2$ es la varianza de los residuales. La hipótesis nula es rechazada a un nivel $\alpha = 0,05$ si el estadístico de prueba KPSS es mayor que el cuantil 0,95, es decir, mayor que 0,46.

2.4.2 TEST DE KOLMOGOROV-SMIRNOV PARA NORMALIDAD

La prueba de hipótesis a realizar consiste en lo siguiente:

$$\begin{cases} H_0 & los\,datos\,siguen\,una\,distribución\,normal \\ H_\alpha & los\,datos\,no\,tienen\,una\,distribución\,normal. \end{cases}$$

Cuando la prueba de Kolmogorov-Smirnov (KS) se aplica para contrastar la hipótesis de normalidad, el estadístico de la prueba es la máxima diferencia:

$$D = \max_{1 \leq i \leq n} \mid \hat{F}_n(x_i) - F_0(x_i) \mid$$

donde es el i-esimo valor observado en la muestra (cuyos valores se han ordenado previamente), $\hat{F}_n(x_i)$ es un estimador de la probabilidad de observar valores menores o iguales que x_i, y $F_0(x)$ es la probabilidad de observar valores menores o iguales que cuando la hipótesis nula es cierta.

Si $D \leq D_\alpha$ se acepta la hipótesis nula, en caso contrario se rechaza. El valor $D_\alpha$se escoge de tal forma que $P(D > D_\alpha \mid los\,datos\,siguen\,una\,distribución\,normal) = 0.05$.

Para la revisión de esta prueba de hipótesis, en el marco del análisis de los residuales, se utiliza para garantizar que los residuales son efecto un ruido blanco, es decir, siguen una distribución normal con una media de cero o cercana, una varianza finita y estable (garantizado por la prueba de homocedasticidad) e independencia entre las observaciones del proceso de ruido blanco (prueba de Ljung-Box).

2.4.3 TEST DE HOMOCEDASTICIDAD BREUSCH-PAGAN

El fenómeno de la heterocedasticidad se da cuando se presenta una varianza no constante en los residuales o perturbaciones, en otras palabras, la varianza de los residuales cambia de manera evidente a lo largo de las observaciones. Por el contrario, homocedasticidad hace referencia a una variabilidad constante, es decir, que a lo largo del tiempo no debe observarse fluctuaciones en la dispersión de las observaciones.

La hipótesis nula y alternativa para el test de Breusch-Pagan son las siguientes:

$$\begin{cases} H_0 & los\, residuales\, son\, homocedasticos \\ H_\alpha & los\, residuales\, son\, heterocedasticos. \end{cases}$$

El fundamento de esta prueba consiste en suponer que los residuales son una función lineal de una o más de las variables independientes. Este supuesto se puede pensar como:

$$\varepsilon_i = \alpha_0 + \alpha_1 X_{i1} + \ldots + \alpha_p X_{ip} + u_i.$$

Una vez obtenidos los residuales, se ajusta un modelo auxiliar como sigue:

$$\hat{\varepsilon}^2 = \delta_0 + \delta_1 \hat{Y}_i$$

esto es, los residuales son una función lineal de una variable que puede ser una sucesión de números reales, de esta forma, se debe esperar que estos residuos se dispersen (o se ajusten) de forma equitativa alrededor de la recta, concluyendo así homocedasticidad.

De esta última regresión se obtiene la medida de ajuste R^2 y posteriormente se calcula el estadístico de la prueba:

$$F = \frac{\frac{R^2}{1}}{\frac{1-R^2}{n-2}}.$$

Bajo la hipótesis nula este estadístico posee una distribución $F_{1,n-2}$si este estadístico es significativo en el nivel escogido (0.05 en este caso) entonces hay evidencia de heterocedasticidad.

2.4.4 PRUEBA DE INDEPENDENCIA DE LOS RESIDUOS LJUNG-BOX

La prueba de Ljung-Box determina la independencia de los rezagos de una serie de tiempo, es decir que toma en cuenta las autocorrelaciones $\hat{p}(j)$ de orden j de los rezagos. Usualmente es usada para hacer contrastes sobre la presencia de autocorrelación entre los errores de cualquier orden y esa es su aplicación en este desarrollo.

La prueba de hipótesis es definida como:

$$\begin{cases} H_0 & los\ residuales\ se\ distribuyen\ de\ forma\ independiente \\ H_\alpha & los\ datos\ no\ se\ distribuyen\ de\ forma\ independiente. \end{cases}$$

Afirmar que los residuales se distribuyen de forma independiente es equivalente a decir que las correlaciones a través de las observaciones son cero, de modo que cualquier correlación observada en los datos es el resultado de la aleatoriedad del proceso. Un incumplimiento de esta prueba se traduce como evidencia sugestiva de que existe información útil olvidada en los residuales y que dicha información puede ser usada para mejorar la estimación.

Su estadístico:

$$Q_{LB} = n(n+2) \sum_{j=1}^{k} \frac{1}{n-j} \hat{\rho}^2(j).$$

Se distribuye aproximadamente χ^2 con k grados de libertard, donde la hipótesis nula H_0de que las primeras k autocorrelaciones son cero es rechazada si $Q_{LB} > \chi^2_{\alpha,k-p-q}$, en esta expresión p y q son los órdenes del modelo ARMA(p,q).

2.5 CRITERIOS DE INFORMACIÓN

En el análisis de series de tiempo una de las dificultades que se presenta es escoger de entre varios modelos aquel que mejor se ajuste al comportamiento de una serie, para enfrentar este inconveniente los criterios de información representan una herramienta importante ya que estiman la calidad de los modelos para una muestra y permiten escoger el mejor.

2.5.1 AKAIKE

Al ajustar un proceso autorregresivo ARMA(p,q) es necesario seleccionar un valor óptimo para p y q, lo cual es tedioso en algunas ocasiones, por esta razón es necesario introducir un factor de penalización (AIC) que ayude a determinar la mejor configuración de parámetros para un buen ajuste. Generalmente se utiliza el criterio de información de Akaike, comúnmente conocido AIC y se define como:

$$AIC = -2logL_k+2k$$

donde L_k es la función de verosimilitud maximizada y k es el número de parámetros estimados del modelo.

Este criterio cuantifica, esencialmente, la bondad de ajuste y la simplicidad o parsimonia de un modelo en un solo estadístico. Por sí solo, no tiene una interpretación practica y su signo tampoco es relevante, solo cuando se compara dos modelos que tratan de explicar las mismas observaciones es cuando el criterio es de utilidad y, en estos casos, la idea es siempre optar por el modelo con menor AIC.

Por ejemplo, si se ajusta un ARMA(p,q) y se obtiene que ARMA(1,1) minimiza el AIC, dando como resultado el siguiente valor AIC = 212 y por otra parte, se ajusta un AR(p) y se hallan que los valores que minimizan el AIC satisfacen la selección del modelo AR(2) con AIC = 213.54. Entonces como los modelos AR(2) y ARMA(1,1) muestran una pequeña diferencia entre sus valores se puede preferir el ARMA(1,1) sobre el AR(2). En dicho sentido, la utilidad del criterio yace en la diferencia entre criterios asociados a varios modelos y no en un valor del criterio por sí solo.

2.5.1.1 APROXIMACIÓN DE SEGUNDO ORDEN

El criterio de información corregido de Akaike (AICc) ha sido estudiado por Wong y Li (1998), este criterio proporciona cuando la muestra es pequeña, mejores estimaciones que el AIC. La definición del AICc es la siguiente:

$$AICc = AIC + \frac{2k(k+1)}{n-k-1}$$

cuando el cociente n/k es suficientemente grande, ambos valores (AIC y AICc) son muy similares.

Los criterios AIC y AICc son asintóticamente eficientes, pero mientras el AICc tiene propiedades similares al AIC en muestras grandes, en muestras pequeñas el funcionamiento del AICc es mucho mejor que el del AIC.

2.5.2 BAYESIANA

El criterio de información Bayesiana (BIC) propuesto por Schwarz en (1978), es una medida de bondad de ajuste de un modelo estadístico y ha sido uno de los métodos más populares usado para la selección de modelos entre un conjunto finito de modelos.

Se basa en la función de probabilidad logarítmica y está estrechamente relacionado con el criterio de información de Akaike. Se define como:

$$BIC = -2logL_k + klog(n)$$

donde k es el número de paramétros, L es el valor de máxima verosimilitud y n es el número de datos.

Dados dos modelos estimados, el modelo con el menor valor del BIC es preferido; un BIC bajo implica un número menor de variables explicativas, mejor ajuste o ambos.

CAPÍTULO 3

METODOLOGÍA

A continuación se pesentan los métodos y herramientas que se emplearon para clasificar y predecir los precios de los vehículos en el mercado mexicano. Para llevar a cabo este análisis se utilizó una base de datos donde se almacena la información y características de los automóviles, las cuales se obtienen de los diferentes portales y/o anuncios clasificados, como por ejemplo, www.mercadolibre.com.mx, www.soloautos.mx, www.autocosmos.com.mx, entre otros.

La recopilación de información de mercado se obtuvo entre los meses de enero y julio del 2017 mediante el uso del *web scraping*, la cual es una técnica utilizada para el rastreo y extracción de datos de la *web*, actualmente ejecutada mediante procesos ETL (Extracción, Transformación y Carga según sus siglas en inglés) los cuales estan desarrollados en *pentaho - software* de inteligencia de negocios (BI) que ofrece productos de código abierto proporcionando integración de datos, informes, tableros de información y extracción de datos-. La implementación del proyecto (tanto la clasificación de grupos como la predicción de precios) se ejecutó con el *software* libre R.

3.1 EXCLUSIONES DENTRO DE LA BASE DE DATOS

Luego de recopilar la información necesaria, se excluyó de la base de datos aquellos vehículos que contengan los siguientes estatus en la variable SKU (*Stock-keeping unit*) en español número de referencia es un identificador usado en el comercio con el objeto de permitir el seguimiento sistémico de los productos y servicios ofrecidos a los clientes:

- No identificada: son aquellos vehículos que cumplen con ciertas condiciones (autos con más de noventa mil kilometros, autos chocados, taxis, ambulancias, grúas, autos blindados, entre otros).
- No activa: son todos aquellos anuncios que ya expiraron o el artículo ha sido vendido y no se encuentran activos en el portal.
- No listada: son aquellas versiones de autos que no están dentro de la lista que permite clasificarlos al momento de capturar las muestras.

- No contable: son aquellos vehículos que no se conoce a ciencia cierta la versión correcta del vehículo expuesto.

Una vez teniendo el conjunto de datos que se analizará, se procedió a separarlo por meses para realizar la clasificación.

3.2 ÁRBOL DE CLASIFICACIÓN

El objetivo del análisis mediante árboles de clasificación, es agrupar diferentes modelos de vehículos para conocer la distribución de precios dentro de cada grupo. En primera instacia se realizó un particionamiento por las marcas de los autos observando características como año, tipo de carrocería, precio, entre otras; para luego realizarse una agrupación de modelos dentro de cada grupo de marcas que se obtuvieron previamente.

Para realizar la clasificación de los vehículos se utilizó la librería *rpart*, la cual se enfoca en particionamiento recursivo y se llevó a cabo con un porcentaje internodo de al menos 15% del total de vehículos reportados en la base de datos.

3.3 *K - MEANS*

Una vez obtenida las agrupaciones con los árboles de clasificación se llevó a cabo un nuevo análisis de conglomerados. Se realizó igualmente considerando las marcas de los vehículos y sus características como año, tipo de carrocería y precio, sin embargo, en esta oportunidad se partió de una segmentación inicial realizada a nivel experto, es decir, se agruparon las marcas de los autos en base al análisis lógico del negocio.

Una vez teniendo dicha segmentación se procedió a realizar la clasificación final por el algoritmo de *k-means*, para esto se utilizó el comando *bigkmeans,* el cual realiza comparaciones entre el valor a asignar y el centroide de cada grupo, para luego asignar dicho valor al *cluster* donde se encontró la mayor homogeneidad.

3.4 SERIES TEMPORALES

Para predecir los precios de los vehículos en el mercado mexicano, se realizó un modelo empírico utilizando una serie de cuatro meses con los precios de los vehiculos en inventario, en donde se graficó para observar la tendencia y estacionalidad. Así mismo se realizó el análisis para conocer si se está frente a un modelo multiplicativo o un modelo aditivo, y así obtener las predicciones deseadas.

En relación al análisis de los métodos autoregresivos, se le realizó un análisis de series temporales a cada grupo obtenido de los árboles de clasificación y k-means tomando el promedio de precios por período. Los pasos y comandos utilizados para cada grupo fueron:

- Se estudió si la serie era estacionaria con la ayuda del comando *kpss.test* el cual realiza la prueba de hipótesis de Kwiatowski, Phillips, Schmidt y Shin.
- Se analizó las gráficas de autocorrelación y autocorrelación parcial, con los comandos *acf* y *pacf* para obtener los posibles parámetros p y q a utilizar en los modelos autoregresivos. La gráfica de autocorrelación nos facilitará con el parámetro de media móvil q mientras que la autocorrelación parcial nos dará un indicio del parámetro autorregresivo p.
- Basándose en el resultado anterior, se evaluaron tres modelos para cada serie, AR(1), MA(1) y ARMA(1,1). Todos con la ayuda del comando *arima* el cual se ejecutó bajo el método de máxima verosimilitud.
- Una vez obtenido los modelos se les calculó el criterio de información de Akaike con el comando *AIC,* el Akaike corregido con *AICc* y el criterio de información Bayesiana con el comando *BIC*. El modelo con menor valor en al menos dos de estos tres criterios se consideró como el mejor que se adapta a los datos en cuestión.
- Para el estudio de los residuos del modelo "ganador" se consideró:
  - El análisis de normalidad utilizando los comandos *qqnorm* y *qqline* para observar a nivel gráfico si los mismos seguían una distribución normal. Así mismo se calculó la prueba de hipótesis de Kolmogorov Smirnov con la ayuda del comando *ks.test* y para corroborar resultados, se observó su media con el comando *mean* y la desviación estándar con *sd*.
  - El análisis de independencia basado en la prueba de hipótesis de Ljung-Box utilizando el comando *Box.test.*

- Una vez superado todos los pasos anteriores, se procedió a la predicción utilizando el modelo "ganador". Dicha predicción se generó con el comando *sarima* en una ventana de 6 meses.

CAPÍTULO 4

ANÁLISIS DE RESULTADOS

En este capítulo se presentan los resultados obtenidos en la segmentación realizada con los árboles de clasificación. Para clasificar los vehículos, se utilizó la base de datos con información recopilada de los diferentes portales y anuncios clasificados; y como variables de entrada se utilizaron: marca del vehículo, modelo del mismo, precio, año, clase (camioneta, compacto, lujo o subcompacto) y tipo de carrocería (convertible, coupé, *hatchback*, mini *vans*, *pick ups*, sedan o *suvs*). Por último, se muestra los resultados del modelo empírico y las predicciones obtenidas para los precios de los vehículos en el mercado mexicano.

4.1 RESULTADOS OBTENIDOS DE LOS ÁRBOLES DE CLASIFICACIÓN

4.1.1 SEGMENTACIÓN POR MARCA DE VEHÍCULOS

Para llevar a cabo la segmentación, se realizó en primera instancia una agrupación donde la marca representó la variable dependiente y las variables precio, año, clase y tipo de carrocería las variables independientes. Así mismo, se consideró el 15% del total de vehículos reportados en la base de datos, como la mínima cantidad dentro de cada nodo. Los grupos obtenidos se detallan a continuación.

Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5
Audi	Hyundai	Acura	Chevrolet	BMW
Infinity	Mazda	Chrysler	Suzuki	Fiat
KIA	Mini Cooper	Dodge		Smart
Land Rover	Mitsubishi	Ford		Volkswagen
Mercedes	Seat	Honda		
Benz				
		Jeep		
		Nissan		
		Renault		
		Toyota		

Tabla 4.1.1.1 Segmentación por Marcas de Vehículos

Al obtener esta segmentación se pudo notar que no cumplía con el resultado esperado, ya que, presenta ciertas incosistencias en las marcas resultantes en cada grupo. En otras palabras, se esperaba que marcas como Audi, BMW y Mercedes Benz quedarán asignadas en el mismo grupo debido a que son competencia en el mercado y sus modelos son similares tanto en precio como en características. De igual forma se puede notar que marcas como Mazda y Toyota o Chevrolet y Ford están en grupos diferentes.

Visto esto, se decidió atacar el inconvenite de la segmentación con el algoritmo de K-medias. Sin embargo, se presenta la segmentación realizada a nivel modelo para cada uno de estos grupos de marca.

4.1.2 SEGMENTACIÓN POR MODELO DE VEHÍCULOS

Una vez segmentado por marca, se tomó cada grupo anterior y se clasificó por modelo, es decir, en cada grupo anterior la nueva variable dependiente será el modelo de los vehículos mientras que las variables precio, año, clase y tipo de carrocería las variables independientes. Igualmente, se consideró el 15% del total de vehículos reportados en la base de datos, como la mínima cantidad dentro de cada nodo.

Grupo 1 - Marca				
Grupo 1 - Modelo	**Grupo 2 - Modelo**	**Grupo 3 - Modelo**	**Grupo 4 - Modelo**	**Grupo 5 - Modelo**
A6	A1 Sportback	A1	A5	Clase B
A8 Quattro	A3 Sportback	A3	Clase G	Clase C
Clase CLA Coupe	A5 Sportback	A4	Clase GL	Forte
Clase CLS	A7	Clase A	Clase GLA	Optima
Clase E	A7 Sportback Quattro		Clase GLE	Q3
Clase S	Clase SL		Clase GLK	Q3 Quattro Rio
G37	Clase SLK		Clase M	Sorento
M37	FX50		Clase R	Soul
M56	Q60		Defender	
Q50	R8 Quattro		Discovery	
Q70	RS5		Discovery Sport	
S4	S3		LR2	
S6	S5		LR4	
	S5 Sportback		Q5 Quattro	
	S7		Q7 Quattro	
	S7 Sportback Quattro		QX56	
	TT		QX60	
			QX70	
			QX80	
			Range Rover	
			Range Rover Evoque	
			Range Rover Sport	
			Sportage	
			Sprinter Wagon L4	
			Viano Pasajeros	
			Vito Pasajeros	

Tabla 4.1.2.1 Segmentación por Modelos de Vehículos para el Grupo 1 de Marcas

Grupo 2 - Marca		
Grupo 6 - Modelo	**Grupo 7 - Modelo**	**Grupo 8 - Modelo**
ASX	CX-5	Altea XL
Ibiza	CX-7	CX-3
Mazda 2	CX-9	Eclipse
	ix35	Elantra
	L200	Endeavor
	Mazda 6	Exeo
	Montero	Freetrack
	MX-5	Grand i10
	Outlander	Lancer
	Sonata	Leon
	Tucson	Mazda 3
		Mazda 5
		Mini Cooper
		Mirage
		Toledo

Tabla 4.1.2.2 Segmentación por Modelos de Vehículos para el Grupo 2 de Marcas

Grupo 3 - Marca				
Grupo 9 - Modelo	**Grupo 10 - Modelo**	**Grupo 11 - Modelo**	**Grupo 12 - Modelo**	**Grupo 13 - Modelo**
Armada	Avanza	200	Avenger	Aprio
Cherokee	Compass	300 C	Caliber	Atos
Durango	CR-V	370Z	Cirrus	Attitude
Edge	Duster	Accord	Civic	City
Expedition	Escape	Altima	Dart	Clio
Explorer	Journey	Camry	Eco Sport	Corolla
FJ Cruiser	Koleos	Challenger	Fluence	Fiesta
Grand Cherokee	Liberty	Charger	Focus	Figo
Highlander	Murano	CRZ	Juke	Fit
HRV	Nitro	Ecoline Wagon	MDX	i10
Land Cruiser	Patriot	Frontier	Neon	Leaf
Pathfinder	Rogue	Fusion	PT Cruiser	Logan
Pilot	X Trail	H 100	RDX	March
RAV4		Hiace	Sentra	Matrix
Sequoia		Hilux	Tiida	Note

Wrangler	ILX	ZDX	Platina
	Kangoo		Prius
	Lobo		RL
	Maxima		Tsuru
	Mustang		Versa
	NP 300		Vision
	Odyssey		Yaris
	Ram 700		
	Ranger		
	Ridgeline		
	Safrane		
	Sandero		
	Scala		
	Sienna		
	Stepway		
	Tacoma		
	TL		
	TLX		
	Town & Country		
	Transit		
	TSX		
	Urvan NV350 Pasejeros		
	Urvan Pasajeros		

Tabla 4.1.2.3 Segmentación por Modelos de Vehículos para el Grupo 3 de Marcas

Grupo 4 - Marca				
Grupo 14 - Modelo	**Grupo 15 - Modelo**	**Grupo 16 - Modelo**	**Grupo 17 - Modelo**	**Grupo 18 - Modelo**
Aveo	Camaro	Captiva Sport	Chevy	Spark
Chevy	Corvette	Equinox	75 años Matiz	Swift
Optra	Cruze	Express Van		Volt
	Malibu	S-Cross		
	Sonic	Silverado 1500 V6		
		Suburban		
		Tahoe		
		Traverse		
		Trax		

Tabla 4.1.2.4 Segmentación por Modelos de Vehículos para el Grupo 4 de Marcas

Grupo 5 - Marca				
Grupo 19 - Modelo	**Grupo 20 - Modelo**	**Grupo 21 - Modelo**	**Grupo 22 - Modelo**	**Grupo 23 - Modelo**
Beetle	500	Bora	Jetta A6	Passat
Crossfox	Golf A6	Gol	Serie 2	Routan
Golf A7	Palio	Jetta Clasico A4	Serie 3	Serie 5
GTI Golf A7	Panda	Linea	Serie 4	Serie 6
i3	Polo	Smart		Serie 7
Serie 1	Punto	Vento		SportVan
	Uno			Tiguan
	Up			Touareg
				Transporter
				X1
				X3
				X4
				X5
				X6

Tabla 4.1.2.5 Segmentación por Modelos de Vehículos para el Grupo 5 de Marcas

4.2 RESULTADOS OBTENIDOS DE K - MEDIAS

Una vez realizada la segmentación por árboles de clasificación, se realizó una clusterización por el método de K-medias con el fin de comparar con la segmentación previa los resultados obtenidos. Para llevar a cabo la segmentación de K-medias se tomó la siguiente agrupación de marcas de vehículo como agrupación inicial:

Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5
Audi	Buick	Renault	Honda	Dodge
BMW	Chevrolet	Seat	Mazda	Jeep
Mercedes Benz	Ford	Volkswagen	Mitsubishi	Mini
	Nissan		Toyota	Chrysler
				Fiat
				Suzuki
				Acura
				Land Rover
				Infiniti
				Smart
				Hyundai
				KIA

Tabla 4.2.1 Segmentación a nivel experto por Marca de Vehículos

Luego de formar la segmentación a nivel experto se realizó la clusterización por el método de K-medias con información del modelo, marca, año, precio y número de puertas de los vehículos, obteniéndose lo siguiente:

Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5
Audi	Ford	Ford	Acura	Audi
BMW	KIA	Honda	Chrysler	BMW
Chevrolet	Mazda	Jeep	Dodge	Chevrolet
Dodge	Mercedes Benz	Land Rover	Fiat	Ford
Fiat	Mini	Mazda	Honda	Mercedes Benz
Ford	Mitsubishi	Mercedes Benz	Hyundai	Nissan
Honda	Nissan	Mini	Infiniti	
Infiniti	Renault	Mitsubishi	Jeep	
Jeep	Seat	Nissan	KIA	
Land Rover	Smart	Renault	Land Rover	
Mercedes Benz	Suzuki	Seat	Mazda	
Nissan	Toyota	Smart	Mini	
Toyota	Volkswagen	Suzuki	Mitsubishi	
Volkswagen		Toyota		
		Volkswagen		

Tabla 4.2.2 Segmentación por el método de K-medias por Marca de Vehículos

Especificamente, la segmentación por modelo de vehículos resultó de la siguiente manera:

• **Grupo 1**: Clase C, A3, Serie 1, Grand Cherokee, A1, X3, X4, X5, Mustang, Suburban, Camaro, Serie 3, Serie 6, Clase E, Clase S, Clase CLS, Clase GL, Spark, Lobo, A5, Vento, Palio, 370Z, Clase M, Challenger, A6, A7, X6, A5 Sportback, Range Rover Sport, Serie 4, Serie 2, QX80, S7 Sportback Quattro, Clase SLK, Clase G, Clase SL, Serie 7, Range Rover, Serie 5, Express Van, Corvette, TT, A8 Quattro, Q7 Quattro, R8 Quattro, S5, Q5 Quattro, RS5, A7 Sportback Quattro, Econoline Wagon, Volt, A4, Discovery, S5 Sportback, Expedition, S3, Tahoe, Clase GLE, Charger, Clase GLA, Viano Pasajeros, Defender, Range Rover Evoque, Clase GLK, Q70, Malibu, S6, Clase CLA, Coupe Altima, Silverado 1500 V6, Touareg, Explorer, Civic, Wrangler, Clase A, Edge, i3, S7, Q60, Versa, Q3, Focus, Sequoia.

• Grupo 2: March, Vento, Ibiza, Corolla, Sentra, RAV4, Mazda 3, Highlander, Tiguan, Pathfinder, NP 300, Tacoma, Passat, Duster, Hilux, Yaris, Avanza, Golf A7, Camry, Beetle, MX-5, Logan, Jetta, A6, Toledo, Mini Cooper, Crossfox, L200, Outlander, Frontier, Jetta Clasico, A4, Kangoo, Maxima, Urvan NV350 Pasajeros, Mazda 2, Smart, Touareg, Altima, Juke, Transporter, Note, S-Cross, Up, Swift, Hiace, Tsuru, X Trail, Mirage, Stepway, Fluence, Koleos, Viano Pasajeros, Clio, Sandero, Safrane, Murano, Rogue, Gol, Lancer, Mazda 5, Leaf, Leon, Land, Cruiser, Armada, Tiida, Versa, GTI Golf A7, Polo, Sienna, Mazda 6, Prius, Freetrack, Montero, Ranger, Rio, Soul.

• Grupo 3: Versa, Ibiza, Highlander, Camry, Sienna, Altima, Avanza, Tiguan, NP 300, Odyssey, Yaris, Hilux, Clase C, Tiida, Jetta A6, Mazda 6, Swift, Crossfox, Jetta Clasico A4, Mazda 3, Mazda 5, Matrix, Juke, Sequoia, FJ Cruiser, Mazda 2, Rogue, Leon, Touareg, CX-7, L200, Toledo, MX-5, Hiace, Murano, Sprinter Wagon L4, Smart, CX-9, Safrane, Outlander, Exeo, Lancer, Golf A6, Tsuru, Liberty, Eclipse, Maxima, X Trail, RAV4, Urvan Pasajeros, Ridgeline, Stepway, Sandero, Fluence, Scala, Endeavor, Koleos, Routan, Kangoo, Clase E, Mustang, Frontier, Armada, Aprio, LR2, Freetrack, Transporter, Transit, Bora, Polo, Platina, Clio, SportVan, City, March, Civic, Sentra, Vito pasajeros, Tacoma, Patriot, Montero, Pilot, Wrangler, Altea XL, Lobo, Ranger, CR-V.

• Grupo 4: Journey, Cherokee, Grand Cherokee, Fit, Mazda 3, Accord, Patriot, Compass, Town & Country, 500, Pilot, Uno, Dart, Nitro, Atos, Charger, CX-3, Avenger, H 100, Palio, Punto, ILX, Caliber, Panda, Vision, Mazda 5, Discovery, CX-9, i10, Linea, Range Rover Evoque, Range Rover Sport, Mazda 2, Liberty, CRZ, RDX, 200, Grand i10, ASX, Q70, Soul, 300 C, PT Cruiser, Sonata, TSX, Attitude, TL, G37, Range Rover, RL, M37, QX56, M56, QX80, QX70, MDX, LR2, Discovery Sport, QX60, FX50, LR4, Q60, ZDX, Sportage, TLX, Q50, Mazda 6, Durango, HRV, CX-5, Elantra, Sorento, Ram 700, Cirrus, Ridgeline, Forte, Optima, ix35, CX-7, Neon.

• Grupo 5: Explorer, Aveo, X3, Fiesta, Focus, Clase C, Serie 1, Serie 3, A4, X5, A1, Q3, A3, Ranger, Clase A, Cruze, Suburban, Equinox, Traverse, Spark, Tahoe, Clase GLK, Fusion, Matiz, Clase M, Clase R, Optra, Q3 Quattro, Silverado 1500 V6, Eco Sport, X6, Escape, Edge, Clase E, Expedition, Chevy, Serie 5, A6, Clase B, Viano Pasajeros, X4, Clase GLA, Clase GL, S4, A1 Sportback, X1, Serie 4, Serie 2, Clase S, Chevy 75 Años, Express Van, A5 Sportback, A5, Clase CLS, Vito pasajeros, A7 Sportback Quattro, Q7 Quattro, Serie 7, Figo, Q5 Quattro, Armada, Sonic, S5 Sportback, S3, A3 Sportback, Trax, Lobo, Captiva Sport.

De esta forma se puede notar que la segmentación resultante del algoritmo de K-medias cumple de mejor forma con las expectativas, pues, modelos que se consideran comptencia (Ej. Mercedes-Benz Calse C, Audi A3 y BMW Serie 3) quedan dentro del mismo grupo.

4.3 RESULTADOS OBTENIDOS DE LAS SERIES TEMPORALES

En esta sección se mostrarán las series temporales para los diferentes análisis que se realizaron: el empirico, los grupos por modelos de vehículos obtenidos por los árboles de clasificación y los grupos obtenidos por el análisis de K-medias.

4.3 RESULTADOS OBTENIDOS DE LAS SERIES TEMPORALES

En esta sección se mostrarán las series temporales para los diferentes análisis que se realizaron: el empirico, los grupos por modelos de vehículos obtenidos por los árboles de clasificación y los grupos obtenidos por el análisis de K-medias.

4.3.1 SERIES TEMPORALES - MODELO EMPÍRICO

Se predijo los precios de los vehículos en el mercado mexicano, a través de un modelo empírico utilizando una serie de cuatro meses con los precios de los vehiculos en inventario. Este modelo se llevó a cabo mediante un código implementado en R donde se realizó el análisis para conocer si se está frente a un modelo multiplicativo o un modelo aditivo. Entre las predicciones obtenidas se tienen las siguientes:

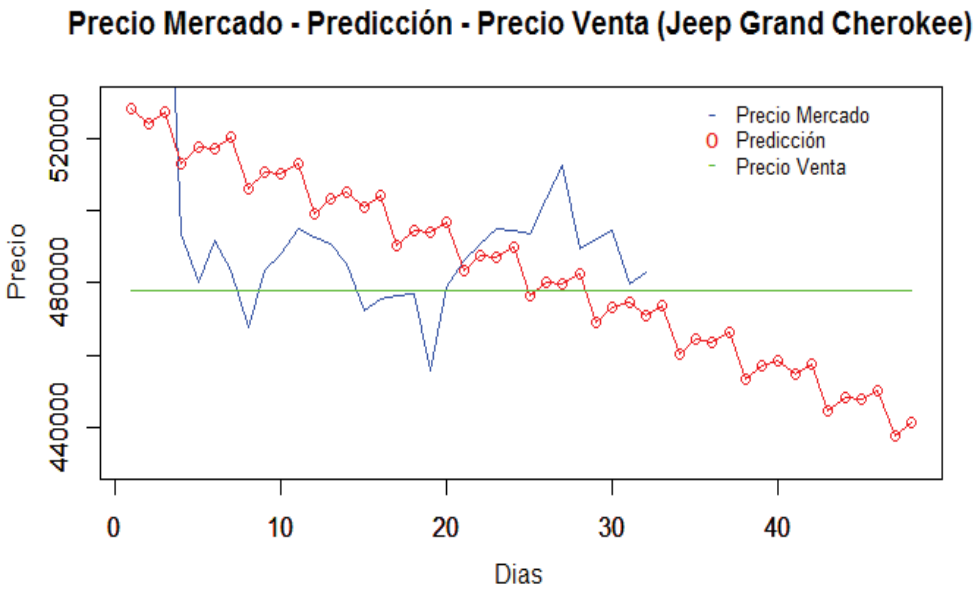

Figura 4.2.1 Predicción para los Precios de Jeep - Grand Cherokee

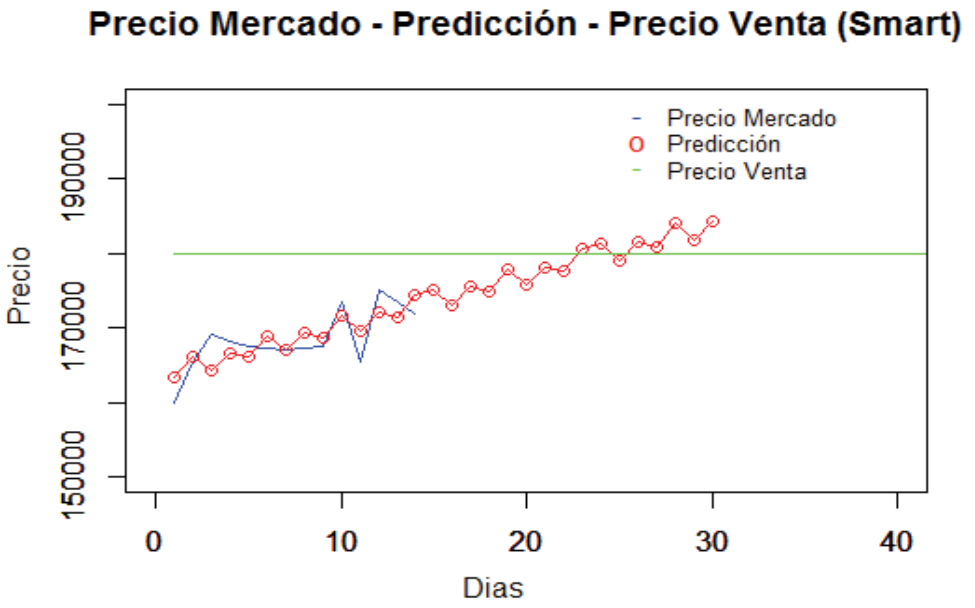

Figura 4.2.2 Predicción para los Precios de Smart - Smart

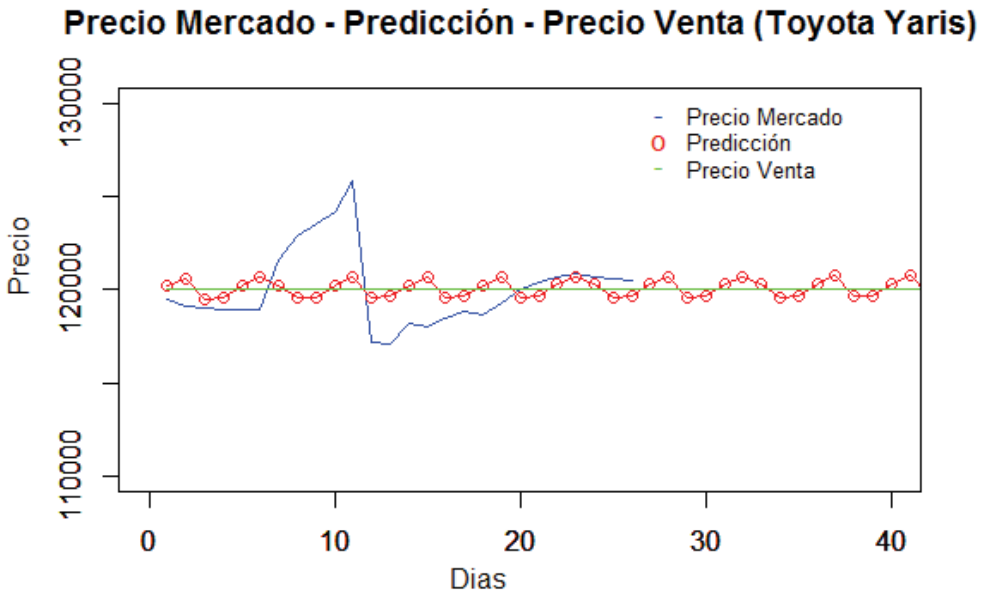

Figura 4.2.3 Predicción para los Precios de Toyota - Yaris

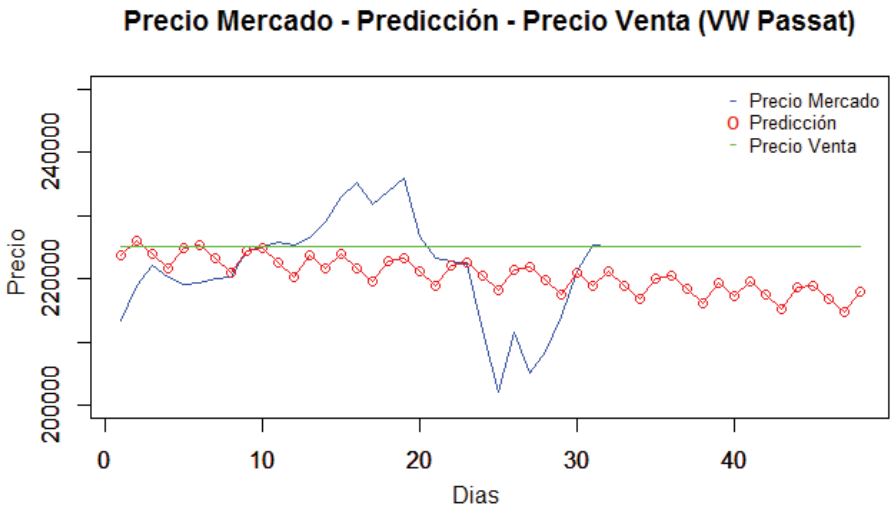

Figura 4.2.4 Predicción para los Precios de Volkswagen - Passat

Para los ejemplos mostrados anteriormente los modelos elegidos fueron todos multiplicativos, sin embargo, las predicciones solo eran acertadas durante los primero 5 días. Por tratarse de predicciones unicamente para un auto en específico se decidío implementar una metodología que englobará a más de un auto. Por ende se trabajarían con modelos de series temporales (AR, MA o ARMA) sobre los grupos arrojados por la segmentación de Arboles de Clasificación y K-medias.

4.3.2 SERIES TEMPORALES - ÁRBOLES DE CLASIFICACIÓN

A continuación se muestran las predicciones de los precios por grupo de los vehículos obtenido de las diferentes segmentaciones del árbol de clasificación. Para cada grupo se estimó un AR(1), un MA(1) y un ARMA(1,1), escogiéndose aquel que arrojara menor AIC, BIC y AICc; todo el análisis se llevó a cabo mediante un código implementado en R. Entre las predicciones obtenidas se tienen las siguientes:

4.3.2.1 GRUPO DE MODELO #4

Del análisis de estacionaridad a un nivel de significancia de 0.01

```
KPSS Test for Trend Stationarity

data:  Precio_dife
KPSS Trend = 0.14866, Truncation lag parameter = 2, p-value = 0.04778
```

Figura 4.3.2.1.1 KPSS Test para la estacionaridad

con lo que se comprueba que la serie es estacionaria. Luego se comprobaron las gráficas de autocorrelación y autocorrelación parcial, y se obtuvo lo siguiente:

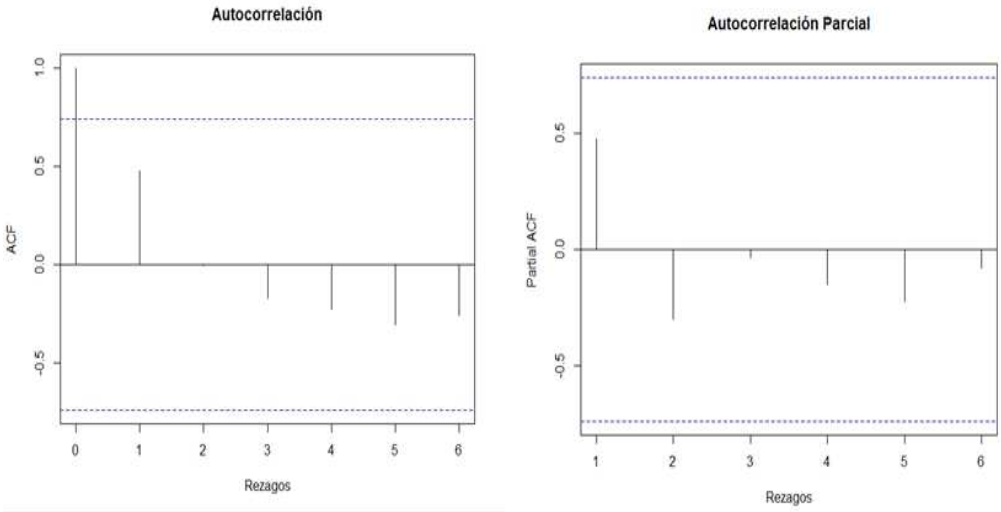

Figura 4.3.2.1.2 Autocorrelación y Autocorrelación parcial

Basados en las gráficas anteriores se probó ajustando modelos AR(1), MA(1) y ARMA(1,1) y se obtuvo lo siguiente:

	Estadísticos		
Modelos	**AIC**	**BIC**	**AICc**
AR(1)	180,06	179,9	188,06
MA(1)	179,07	178,91	187,07
ARMA(1,1)	179,42	179,2	199,42

Tabla 4.3.2.1.1 Estadísticos para el Grupo 4 de Modelos

De aquí se decide ajustar un MA(1) ya que para los tres criterios es el modelo que presenta un mejor ajuste. Luego se analizaron los residuales y se obtuvieron los siguientes resultados:

```
        One-sample Kolmogorov-Smirnov test

data:  modelo2$residuals
D = 0.71429, p-value = 0.0003959
alternative hypothesis: two-sided


        studentized Breusch-Pagan test

data:  Precio_dife ~ modelo2$residuals
BP = 0.49634, df = 1, p-value = 0.4811


        Box-Ljung test

data:  modelo2$residuals
X-squared = 1.2062, df = 4, p-value = 0.8771
```

Figura 4.3.2.1.3 Análisis de Residuales.

de la imagen anterior se puede concluir que estos residuos a pesar de no ser ruido blanco, son independiente y hocedásticos. Por otra parte, se puede ver que los residuales se sajustan a una normal como se muestra a continuación:

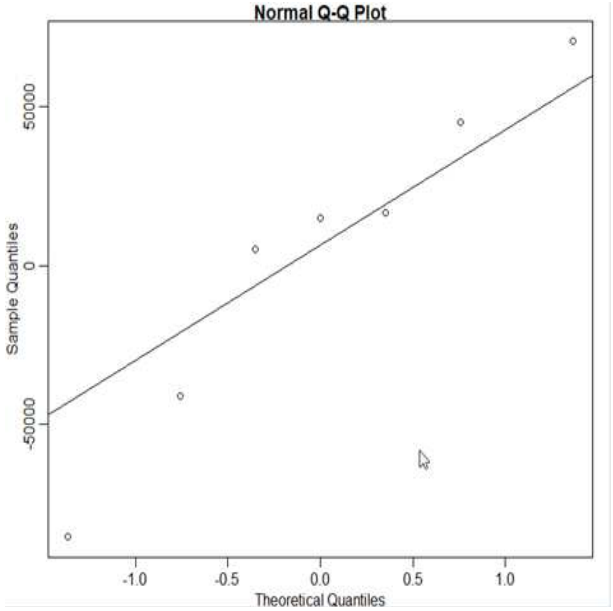

Figura 4.3.2.1.2 Gráfico de Normalidad.

Por último se presenta el gráfico del pronóstico de la serie, donde el areá gris representa $\mp$ 1 o 2 desviaciones de error con respecto a la predicción.

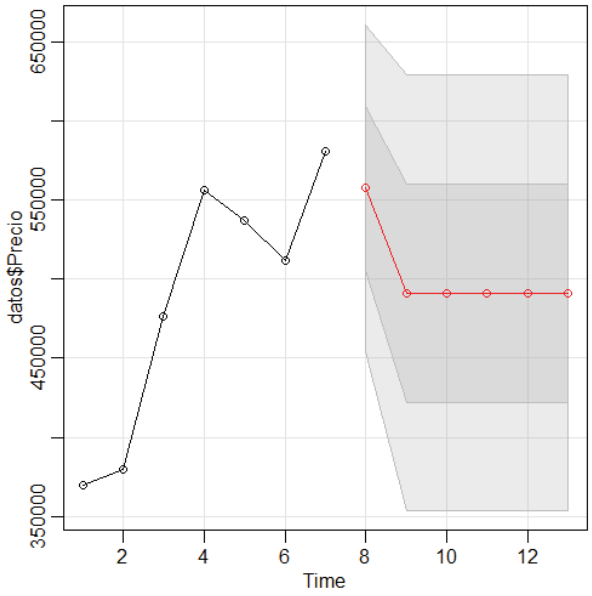

Figura 4.3.2.1.3 Pronóstico de la Serie

4.3.2.2 GRUPO DE MODELO #23

Del análisis de estacionaridad a un nivel de significancia de 0.01

```
KPSS Test for Trend Stationarity

data:  Precio_dife
KPSS Trend = 0.13026, Truncation lag parameter = 2, p-value = 0.07915
```

Figura 4.3.2.2.1 KPSS Test para la estacionaridad

con lo que se comprueba que la serie es estacionaria. Luego se comprobaron las gráficas de autocorrelación y autocorrelación parcial, y se obtuvo lo siguiente:

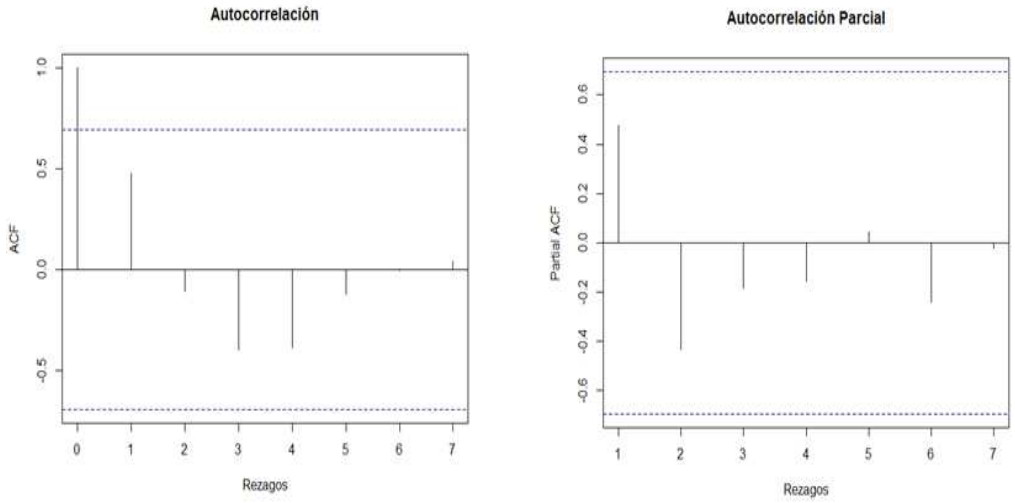

Figura 4.3.2.1.2 Autocorrelación y Autocorrelación parcial

Basados en las gráficas anteriores se probó ajustando modelos AR(1), MA(1) y ARMA(1,1) y se obtuvo lo siguiente:

	Estadísticos		
Modelos	**AIC**	**BIC**	**AICc**
AR(1)	204,66	204,09	210,66
MA(1)	202,89	203,13	208,89
ARMA(1,1)	204,87	205,19	218,21

Tabla 4.3.2.2.1 Estadísticos para el Grupo 23 de Modelos

De aquí se decide ajustar un MA(1) ya que para los tres criterios es el modelo que presenta un mejor ajuste. Luego se analizaron los residuales y se obtuvieron los siguientes resultados:

```
        One-sample Kolmogorov-Smirnov test

data:  modelo2$residuals
D = 0.625, p-value = 0.001509
alternative hypothesis: two-sided

        studentized Breusch-Pagan test

data:  Precio_dife ~ modelo2$residuals
BP = 0.26609, df = 1, p-value = 0.606


        Box-Ljung test

data:  modelo2$residuals
X-squared = 4.7745, df = 4, p-value = 0.3112
```

Figura 4.3.2.1.3 Análisis de Residuales.

de la imagen anterior se puede concluir que estos residuos a pesar de no ser ruido blanco, son independiente y hocedásticos. Por otra parte, se puede ver que los residuales se sajustan a una normal como se muestra a continuación:

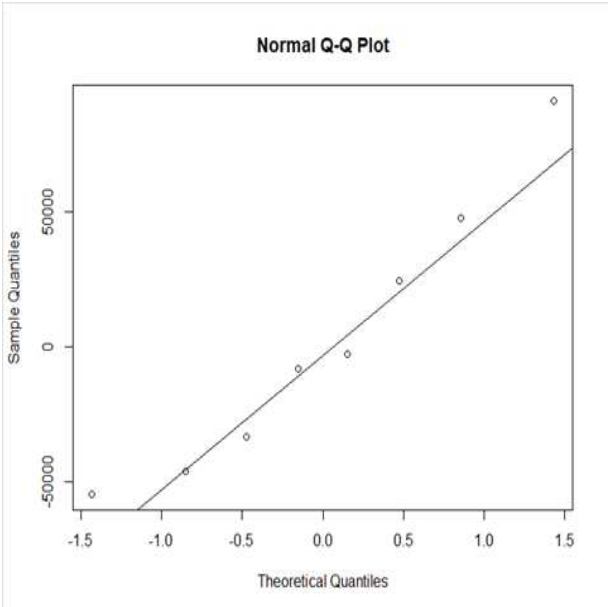

Figura 4.3.2.1.4 Gráfico de Normalidad.

Por último se presenta el gráfico del pronóstico de la serie, donde el areá gris representa $\mp$ 1 o 2 desviaciones de error con respecto a la predicción.

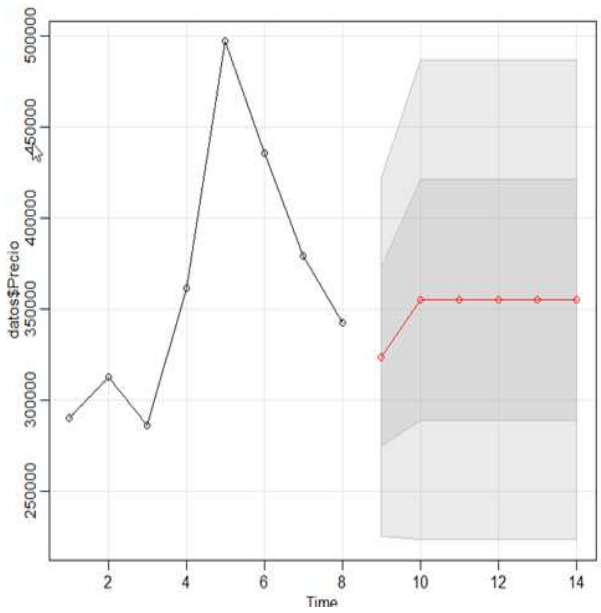

Figura 4.3.2.1.5 Pronóstico de la Serie

4.3.3 SERIES TEMPORALES - K - MEDIAS

4.3.3.1 AGRUPACIÓN POR MARCAS #1

Del análisis de estacionaridad a un nivel de significancia de 0.01

```
        KPSS Test for Trend Stationarity

data:  Precio_dife
KPSS Trend = 0.12937, Truncation lag parameter = 2, p-value = 0.0808
```

Figura 4.3.3.1.1 KPSS Test para la estacionaridad

con lo que se comprueba que la serie es estacionaria. Luego se comprobaron las gráficas de autocorrelación y autocorrelación parcial, y se obtuvo lo siguiente:

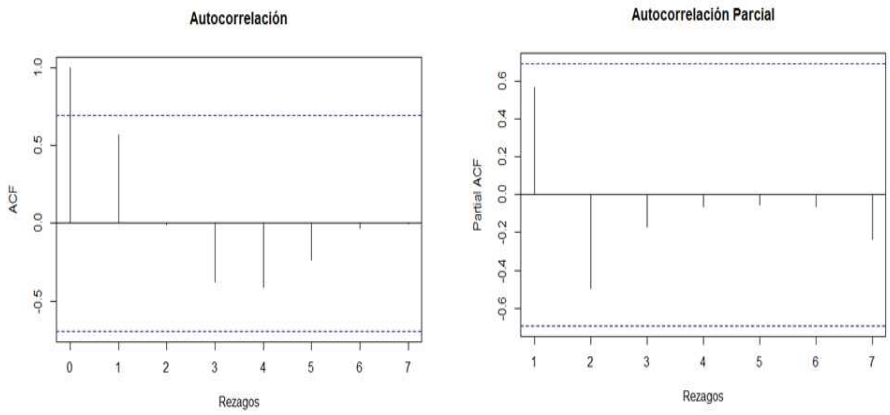

Figura 4.3.3.1.2 Autocorrelación y Autocorrelación parcial

Basados en las gráficas anteriores se probó ajustando modelos AR(1), MA(1) y ARMA(1,1) y se obtuvo lo siguiente:

	Estadísticos		
Modelos	**AIC**	**BIC**	**AICc**
AR(1)	180,31	180,14	188,31
MA(1)	179,8	179,63	187,8
ARMA(1,1)	181,78	181,57	201,78

Tabla 4.3.3.1.1 Estadísticos para el Grupo 1 de Marcas

De aquí se decide ajustar un MA(1) ya que para los tres criterios es el modelo que presenta un mejor ajuste. Luego se analizaron los residuales y se obtuvieron los siguientes resultados:

```
        One-sample Kolmogorov-Smirnov test

data:  modelo2$residuals
D = 0.57143, p-value = 0.01089
alternative hypothesis: two-sided
```

```
        studentized Breusch-Pagan test

data:  Precio_dife ~ modelo2$residuals
BP = 0.29744, df = 1, p-value = 0.5855
```

```
        Box-Ljung test

data:  modelo2$residuals
X-squared = 4.8859, df = 4, p-value = 0.2992
```

Figura 4.3.3.1.3 Análisis de Residuales.

de la imagen anterior se puede concluir que estos residuos a pesar de no ser ruido blanco, son independiente y hocedásticos. Por otra parte, se puede ver que los residuales se sajustan a una normal como se muestra a continuación:

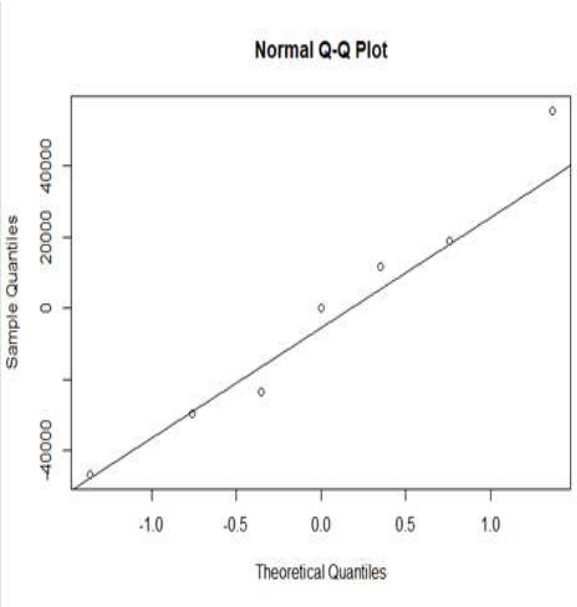

Figura 4.3.3.1.4 Gráfico de Normalidad.

Por último se presenta el gráfico del pronóstico de la serie, donde el areá gris representa $\mp$ 1 o 2 desviaciones de error con respecto a la predicción.

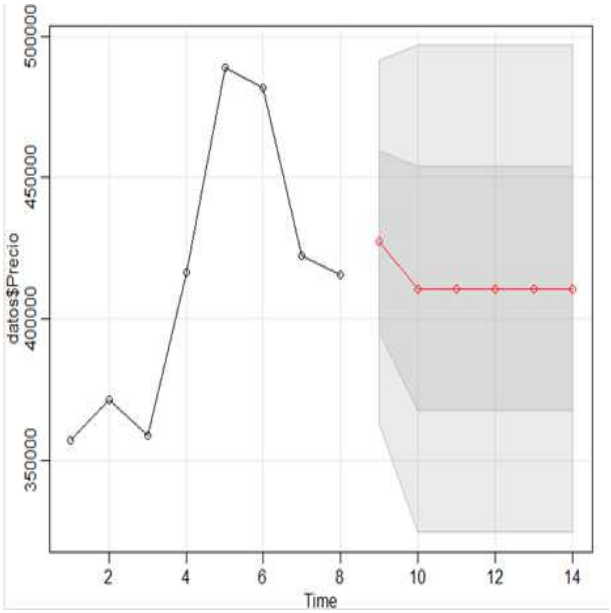

Figura 4.3.3.1.5 Pronóstico de la Serie

4.3.3.2 AGRUPACIÓN POR MARCAS #3

Del análisis de estacionaridad a un nivel de significancia de 0.01

```
        KPSS Test for Trend Stationarity

data:  Precio_dife
KPSS Trend = 0.1538, Truncation lag parameter = 2, p-value = 0.0435
```

Figura 4.3.3.2.1 KPSS Test para la estacionaridad

con lo que se comprueba que la serie es estacionaria. Luego se comprobaron las gráficas de autocorrelación y autocorrelación parcial, y se obtuvo lo siguiente:

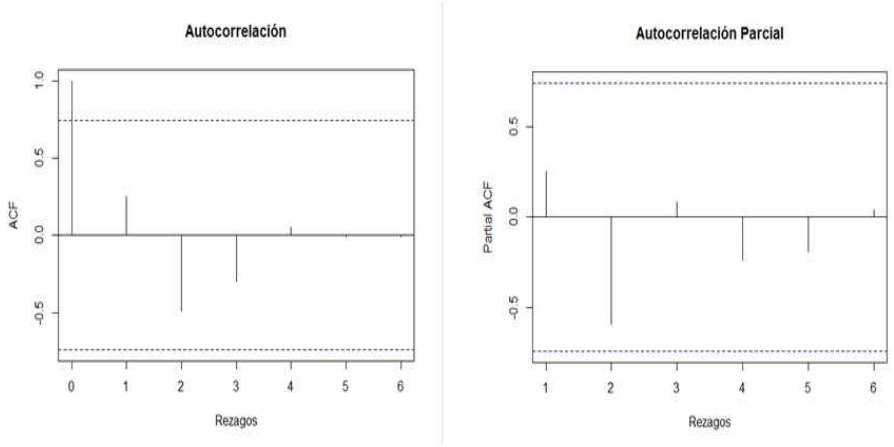

Figura 4.3.3.2.2 Autocorrelación y Autocorrelación parcial

Basados en las gráficas anteriores se probó ajustando modelos AR(1), MA(1) y ARMA(1,1) y se obtuvo lo siguiente:

	Estadísticos		
Modelos	**AIC**	**BIC**	**AICc**
AR(1)	158,73	158,57	166,73
MA(1)	158,79	158,63	166,79
ARMA(1,1)	160,39	160,17	180,39

Tabla 4.3.3.2.1 Estadísticos para el Grupo 3 de Marcas

De aquí se decide ajustar un MA(1) ya que para los tres criterios es el modelo que presenta un mejor ajuste. Luego se analizaron los residuales y se obtuvieron los siguientes resultados:

```
        One-sample Kolmogorov-Smirnov test

data:  modelo2$residuals
D = 0.71429, p-value = 0.0003959
alternative hypothesis: two-sided

        studentized Breusch-Pagan test

data:  Precio_dife ~ modelo2$residuals
BP = 0.26609, df = 1, p-value = 0.606


        Box-Ljung test

data:  modelo2$residuals
X-squared = 4.8859, df = 4, p-value = 0.2992
```

Figura 4.3.3.2.3 Análisis de Residuales.

de la imagen anterior se puede concluir que estos residuos a pesar de no ser ruido blanco, son independiente y hocedásticos. Por otra parte, se puede ver que los residuales se sajustan a una normal como se muestra a continuación:

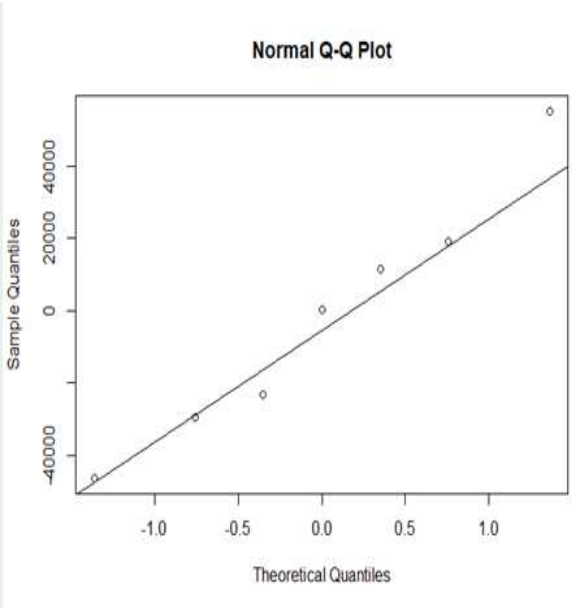

Figura 4.3.3.2.4 Gráfico de Normalidad.

Por último se presenta el gráfico del pronóstico de la serie, donde el areá gris representa $\mp$ 1 o 2 desviaciones de error con respecto a la predicción.

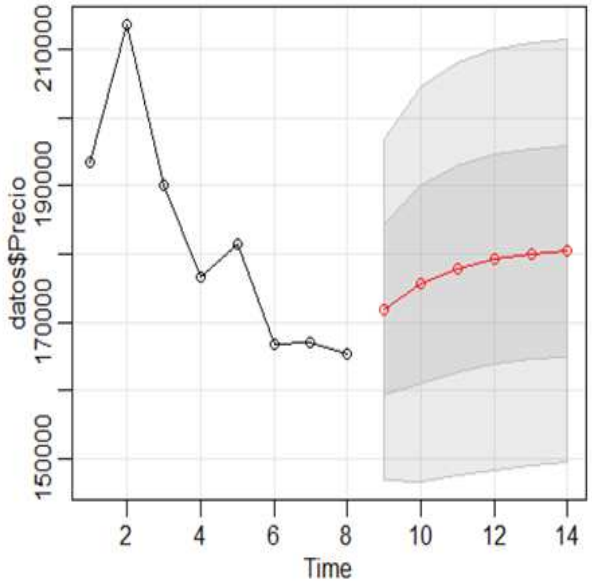

Figura 4.3.3.2.5 Pronóstico de la Serie

CONCLUSIONES Y RECOMENDACIONES

La segmentación de los vehículos para la posterior predicción de sus precios, se realizó basándose en los precios recopilados de los portales y/o anuncios clasificados y su análisis fue implementado con el software R.

En relación a la segmentación obtenida de los árboles de clasificación se pudo observar que se presentó incongruencias en las agrupaciones a nivel de marca, ya que marcas consideradas competencia directa como por ejemplo, Mercedes Benz, BMW ó Audi, resultaron en grupos diferentes. Así mismo, se podría comentar que se hubiese esperado que Ford y Chevrolet estuviesen en el mismo grupo por presentar modelo de vehículos similares y no resultó de esa manera.

Este inconveniente se intentó resolver con el método de K-medias, sin embargo, luego de realizar las predicciones se encontraron varios problemas con el supuesto de estacionaridad, por lo que se tuvo que realizar ciertos tratamientos a los datos y diferenciación para lograr el cumplimiento de esta condición. Por otra parte, en la segmentación obtenida de los árboles de clasificación se pudo observar que aquellos grupos con mayor cantidad de registros, es decir, mayor cantidad de modelos de vehículos obtuvieron una mejor predicción de los precios en comparación al resto de los clusters.

Sin embargo, al realizar el modelo empírico se concluye que su precio de mercado coincide en tendencia con las predicciones. En general se recomienda:

- Poder establecer una metodología para la segmentación de grupos, ya que para el mercado mexicano se manejan mas de 15000 versiones de coches y no es escalable hacer el analisis y predicción para todas las versiones.
- Generar modelos de series temporales (AR, MA ó ARMA) a partir de la segmentación obtenida con el nuevo análisis seleccionado.
- Automatizar todo el proceso de predicciones de precios y genear pruebas de validación para asegurar la confiabilidad de las predicciones.

REFERENCIAS

[1] Velázquez Garcia, Leticia, **Principales características de la Reestructuración de la Industria Automotriz**. Distrito Federal, México, 2004. Disponible en Internet:
http://www.redalyc.org/articulo.oa?id=32512815

[2] Página Oficial de Kavak México. Disponible en Internet: https://www.kavak.com

[3] Marin JM, **Análisis de Cluster y Arboles de Clasificación**. Disponible en Internet:
http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/DM/tema6dm.pdf

[4] González Casimiro MP, **Análisis de series temporales: Modelos ARIMA**. Pais Vasco, España. Disponible en Internet:
https://addi.ehu.es/bitstream/handle/10810/12492/04-09gon.pdf;jsessionid= ...
... D3F95794B99BC20110D74C0702C98A32?sequence=1

[5] Departamento de Métodos Cuantitativos e Informáticos, **Series Temporales**. Universidad Politécnica de Cartagena, Colombia.
Disponible en Internet: http://metodos.upct.es/Asignaturas/Diplomatura/Introduccion_estadistica/...
...2008_2009/material_didactico/apuntes/TEMA5SERIESTEMPORALES.pdf

6. Shumway