Metodología de pronóstico escalable con aprendizaje autónomo, integración en la nube y reportes automatizados

Scalable forecasting methodology with machine learning and integrated cloud-based reports

Luis D. Chavarría-Múnera[1], Juan M. Cogollo-Flórez[2], Alexander A. Correa-Espinal[3]

[1]Ingeniero Industrial. Departamento de Ingeniería de la Organización, Facultad de Minas, Universidad Nacional de Colombia. Medellín, Colombia.
[2]Magíster en Ingeniería Administrativa. Profesor Asociado. Departamento de Calidad y Producción, Instituto Tecnológico Metropolitano – ITM. Medellín, Colombia.
[3]Doctor en Estadística e Investigación Operativa. Profesor Titular. Departamento de Ingeniería de la Organización, Facultad de Minas, Universidad Nacional de Colombia. Medellín, Colombia
Correo autor para correspondencia: juancogollo@itm.edu.co

Recibido: 07/01/2020
Aceptado: 08/04/2020

***Cite this article as*: L. D. Chavarría-Múnera, J. M. Cogollo-Flórez, A. A. Correa-Espinal "Metodología de pronóstico escalable con aprendizaje autónomo, integración en la nube y reportes automatizados", *Prospectiva*, Vol 18, N° 2, 2020.**

http:://doi.org/10.15665/rp.v18i2.2243

RESUMEN

El análisis de series de tiempo es una de las herramientas más utilizadas para hacer predicciones basándose en los datos del pasado. En este trabajo se desarrolló una metodología de pronóstico escalable que supera las dificultades del análisis tradicional de series de tiempo, utilizando nuevas herramientas y estructuras de datos computacionales que facilitan la integración con las aplicaciones empresariales y disminuye la curva de aprendizaje necesaria para obtener buenos pronósticos. La metodología consta de cinco etapas: (1) Importar datos desde la nube o el dispositivo del usuario, (2) Ordenar y transformar, (3) Visualizar (4) Modelar automáticamente y validar resultados, y (5) Comunicar pronósticos obtenidos mediante un reporte automatizado. La metodología se utilizó en un caso aplicado considerando diez series de tiempo de índices de ventas reales de comercio minorista en Colombia, mostrando mejoras apreciables con un promedio de disminución del error de pronóstico medio absoluto (MAPE) del 50.56%.

Palabras clave*:* Aprendizaje autónomo, Integración en la nube, Pronóstico escalable, Series de tiempo, Reportes automatizados.

ABSTRACT

Time series analysis is one of the most used tools to forecast based on past data. This work develops a scalable forecasting methodology that attempts to overcome the difficulties of traditional time series analysis; utilizing new computational tools and data structures that facilitate integration with business applications and reduce the learning curve needed to obtain right forecasts. The methodology consists of five phases: (1) Importing data directly from the cloud or the user's device, (2) Tidying and transforming, (3) Visualization, (4) Automatically model and validate the results, and (5) Communicate the obtained forecasts with an automated report. The methodology was used in an applied case considering ten time series from real retail sales indexes in Colombia, showing appreciable improvements with an average decrease on the Mean Absolute Percentage Error (MAPE) of 50.56%.

Keywords*:* Machine Learning, Cloud-based integration, Scalable forecasting, Time series, Automated reports.

1. INTRODUCCIÓN

Los pronósticos tienen como objetivo principal determinar con anterioridad el resultado más probable de una variable, con el fin de tomar decisiones de planificación y control orientadas a responder adecuadamente a los requerimientos del mercado de los sistemas productivos de bienes y servicios. Un adecuado pronóstico, especialmente cuando se gestionan múltiples productos, promueve y facilita la colaboración en los flujos de productos, servicios, información y dinero en la cadena de suministro [1].

Una de las herramientas más utilizadas para realizar pronósticos basándose en los datos del pasado es el análisis de series de tiempo [2]. Este tipo de análisis requiere alto rigor matemático-estadístico para garantizar su confiabilidad, por lo que, en ocasiones, las empresas prefieren utilizar métodos cualitativos o empíricos con mayor facilidad de uso. El proceso de pronóstico basado en series de tiempo se divide en tres etapas principales: i) Preprocesamiento de los datos, para identificar y eliminar valores atípicos que puedan afectar la predicción; ii) Elección del método de pronóstico más adecuado y de sus parámetros, cuando así se requiera; y, iii) Evaluación de la exactitud del pronóstico, calculando el error una vez se conozca el valor real de la variable pronosticada [2].

Debido a que las empresas tienen múltiples referencias, productos o familias de productos, es de interés práctico el pronóstico escalable basado en el análisis simultáneo de series de tiempo. Para ello, en la literatura se han propuesto diferentes enfoques, como el uso de modelos de regresión, análisis de componentes principales, análisis factorial y modelos GARCH (*Generalized Autoregressive Conditional Heteroscedasticity*) [3] [4]. En cuanto a las aplicaciones en sectores empresariales específicos, por ejemplo, se ha desarrollado una metodología de pronóstico aplicando un modelo autorregresivo que caracteriza la correlación en series de tiempo geofísicas a través de un agrupamiento espacio-temporal [5].

También, se han realizado pronósticos de precios del carbón usando modelos *AutoRegressive Integrated Moving Average* (ARIMA), modelos tradicionales de series de tiempo y redes neuronales tales como *Generalized Regression Neural Networks* (GRNNs) y Multi-layer Feed-forward Networks (MLFNs) [6]. González-Vidal *et al* [7] propusieron una metodología para el pronóstico de consumo de energía en edificios inteligentes usando series de tiempo multivariadas, transformando la base de datos dependiente del tiempo en una estructura que pueden procesar algoritmos estándar de aprendizaje autónomo. Barthel *et al* [8] modelaron y pronosticaron series temporales de volatilidad multivariada aplicando un enfoque basado en correlaciones parciales.

Asimismo, Karmy y Maldonado [9] desarrollaron una metodología para el pronóstico de ventas de viajes utilizando regresión de vectores de soporte y series de tiempo jerárquicas. Este enfoque permite obtener pronósticos con más alta precisión a nivel de producto, comparados con los modelos ARIMA y Holt-Winters. Niu *et al* [10] desarrollaron una metodología de pronóstico en series financieras, utilizando un modelo de selección de características de dos etapas, un modelo de aprendizaje profundo y un modelo de corrección de errores que captura efectivamente la no linealidad. Du *et al* [11] proponen un modelo codificador-decodificador para el pronóstico de series de tiempo multivariadas, basado en una estructura de aprendizaje profundo que integra el vector de codificación tradicional y el vector de atención temporal, usando capas de memoria bidireccionales.

Aunque las técnicas descritas en los trabajos anteriores permiten combinar el análisis de múltiples series de tiempo y clústeres y obtener pronósticos con un adecuado desempeño, tienen ciertas limitaciones prácticas dado que su aplicación se realiza usando el mismo peso entre la distancia espacial e igual diferencia temporal, lo cual requiere del cumplimiento de determinadas condiciones en el formato y estructura de los datos, tales como tener intervalos de tiempo únicos y regulares, datos homogéneos, no agregación de eventos puntuales, entre otras [12]. Aunque los sistemas ERP (*Enterprise Resource Planning*) generalmente

permiten realizar exportaciones de sus datos para ser analizados y llegar a conclusiones útiles, se requiere que los datos almacenados cumplan las condiciones anteriores, lo cual siempre no se da en la práctica empresarial.

Con el fin de contribuir a la generación de conocimiento en el área de pronóstico usando análisis de series de tiempo, en este artículo se desarrolla y aplica una metodología de pronóstico escalable la cual permite superar las dificultades tradicionales del modelamiento de series de tiempo, especialmente en lo concerniente al grado de escalabilidad y de fácil uso en la gestión de procesos administrativos y productivos con disminución de las curvas de aprendizaje a través de la integración en la nube y reportes automatizados.

La organización del artículo es la siguiente: en la sección 2 se describen los métodos de pronóstico cuantitativos para series de tiempo como fundamento conceptual del trabajo realizado, en la sección 3 se describe la metodología desarrollada, en la sección 4 se muestran los resultados obtenidos y, se finaliza, con las conclusiones y recomendaciones en la sección 5.

2. MÉTODOS DE PRONÓSTICO PARA SERIES DE TIEMPO

Los métodos de pronóstico pueden ser cualitativos o cuantitativos. Los métodos cualitativos aplican cuando no hay datos históricos disponibles y se basan en información secundaria o en el juicio de expertos para predecir el valor futuro de la variable de interés. Algunos métodos cualitativos son el método Delphi, pronóstico por analogía, pronóstico basado en escenarios, agregación de la fuerza de ventas, opinión de los ejecutivos y encuestas de intenciones de los clientes [13].

Los métodos cuantitativos requieren que exista información del pasado con datos almacenados en una serie y, además, que sea razonable asumir que algunos patrones anteriores continuarán en el futuro [12]. Existen dos estructuras principales utilizadas para el almacenamiento de los datos de la serie: longitudinales o temporales. La configuración longitudinal tiene en sus columnas puntos individuales y en sus filas variables a observar correspondientes, por ejemplo, a productos o índices. Esta estructura se presenta generalmente en el sector productivo, ya que suelen ser tableros que se actualizan constantemente (Tabla 1). De manera opuesta, los datos de series temporales tienen en sus columnas variables a observar y en sus filas puntos temporales iguales para cada una de las variables (Tabla 2).

Tabla 1. Ejemplo de estructura de datos longitudinal.
Table 1. An example of longitudinal data structure.

Código	01/01/2016	01/02/2016	01/03/2016	01/04/2016	01/05/2016	01/06/2016
47622	170	626	249	179	332	242
18425	10	10	10	13	15	13
18404	3.063	2.798	2.868	2.866	2.644	2.951
48402	265	266	260	127	351	426
18211	2.611	2.535	2.457	2.763	2.324	2.478

Tabla 2. Ejemplo de estructura de datos para tres series de tiempo.
Table 2. An example of data structure for three time series.

Fecha	Comercio minorista	Alimentos y bebidas no alcohólicas	Bebidas alcohólicas y productos del tabaco
1/1/2003	53.8	74.7	51.7
1/2/2003	51.1	75.2	44.9
1/3/2003	54.8	85.9	55.1

A continuación, se describen algunos métodos cuantitativos utilizados en la industria para realizar pronósticos usando análisis de series de tiempo [13].

2.1 Método del promedio

Consiste en tomar el promedio de las observaciones en un periodo particular y asumir el futuro como este resultado. Si se denotan los datos históricos como $y_1, \dots, y_T$ entonces se puede escribir los pronósticos como:

$$\hat{y}_{(T+h)|T} = \bar{y} = \frac{(y_1 + \cdots + y_T)}{T} \tag{1}$$

donde la notación $\hat{y}_{(T+h)|T}$ es una abreviación para la estimación de $y_{(T+h)}$ dados los datos históricos $y_1, \dots, y_T$, h es el horizonte de pronóstico utilizado y T es el número de datos. Este método presenta múltiples limitaciones, entre ellas que no se tiene en cuenta la tendencia de la serie o su componente estacional.

2.2 Método de Naïve

En este método, el pronóstico se realiza utilizando únicamente el valor de la última observación, es decir:

$$\hat{y}_{(T+h)|T} = y_T \tag{2}$$

Este es un método que funciona especialmente bien en estructuras económicas y financieras, dado que en estas los datos se ajustan a un modelo de caminata aleatoria, es decir:

$$y_t - y_{t-1} = \varepsilon_t \tag{3}$$

Dónde ε_t es un ruido blanco. Esto ocurre porque los movimientos futuros son impredecibles y tienen una probabilidad igual de ser superiores o inferiores al valor actual, por lo que el mejor pronóstico se define con el método de Naïve. Similarmente, para series estacionales, también es posible utilizar el método Naïve estacional, que toma el mismo valor del periodo estacional anterior como pronóstico, es decir:

$$\hat{y}_{(T+h)|T} = y_{T+h-m(k+1)} \tag{4}$$

donde *m* es el periodo estacional y *k* es la parte entera de $(h-1)/m$; esto significa que, si por ejemplo se está tratando con datos mensuales, el valor futuro para marzo en el próximo año será igual al valor de marzo del año actual.

2.3 Método de la deriva

Es una variación del método de Naïve considerando que los pronósticos crecen o no a través del tiempo, donde la tasa de cambio se denomina deriva, que es equivalente al cambio promedio visto entre dos puntos que contienen los datos, por lo que el pronóstico es equivalente a:

$$\hat{y}_{(T+h)|T} = y_T + \frac{h}{T-1}\sum_{t=2}^{T}(y_t - y_{t-1}) = y_t + h\left(\frac{y_T - y_1}{T-1}\right) \quad (5)$$

Esto corresponde a trazar una línea recta entre la primera y la última observación y realizar una extrapolación a futuro con la misma pendiente.

2.4 Modelos ARIMA

Un modelo ARMA (*AutoRegressive Moving Average*) cuenta con un componente autorregresivo (AR) y un componente de medias móviles (MA) [14]. Considérese inicialmente la ecuación que describe un modelo ARMA(P,Q):

$$y_t - \phi_1 y_{t-1} - \cdots - \phi_P y_{t-P} = c + a_t - \theta_1 a_{t-1} - \cdots - \theta_Q a_{t-Q} \quad (6)$$

donde a_t generalmente se asume que cumple los supuestos de media cero y varianza constante, $\phi_p$representa el coeficiente autorregresivo de orden P y θ_Q el componente de medias móviles de orden Q.

Para un modelo ARIMA, se incluye adicionalmente una componente de diferenciación que modela los retrasos para los efectos de las variables regresoras sobre la respuesta, como se presenta en la ecuación (7).

$$\nabla^d y_t - \varphi_1 \nabla^d y_{t-1} - \cdots - \varphi_p \nabla^d y_{t-1} = c + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q} \quad (7)$$

donde $d = 0, 1, 2, \ldots, \nabla^d y_t$ y los parámetros $(\phi_1, \ldots, \phi_p)$ y $(\varphi_1, \ldots, \varphi_p)$ están funcionalmente relacionados y se asume un modelo ARMA para $\nabla^d y_t$ que además debe ser estacionario. En este trabajo, el paquete *fable* utiliza elementos del paquete *forecast* como la función *auto.arima()* para ajustar el mejor modelo con base en los criterios AIC *(Akaike's Information Criterion),* AICc *(Corrected AIC)* o BIC *(Bayesian Information Criterion).*

2.5 Modelos globales con variables indicadoras

En el caso más simple, la regresión permite hallar una relación lineal utilizando estimación de mínimos cuadrados, esto da como resultado la tendencia de la serie, utilizando un parámetro β_0 para definir el intercepto con el eje vertical y β_p para incrementar el número de parámetros en el modelo en la tendencia. Para modelar la estacionalidad se estima un parámetro para cada periodo δ_i, el cual se activa cuando es necesario utilizando variables indicadoras. Se especifica además un término de error e_t que debe distribuirse idénticamente independiente con media cero y varianza constante, como se muestra en la ecuación (8).

$$y_t = \beta_0 + \beta_1 t^1 + \cdots + \beta_p t^p + \sum_{i=1}^{11} \delta_i I_{i,t} + e_t \qquad e_t \sim iid\ N(0, \sigma^2) \quad (8)$$

donde la variable $I_{i,t}$ es una variable indicadora que tiene valor 1 si y_t en el tiempo t se encuentra en la estación i y 0 en otro caso, y adicionalmente δ_i es su parámetro estimado correspondiente.

2.6 Modelos de suavizamiento exponencial

Los suavizamientos exponenciales fueron propuestos a finales de 1950 y han inspirado o motivado algunos de los métodos de pronóstico más exitosos [13]. Los pronósticos producidos con este método son medias ponderadas de observaciones pasadas, con sus pesos decayendo exponencialmente a medida que las observaciones se vuelven más antiguas. En otras palabras, las observaciones más recientes tienen el mayor peso asociado.

Dado que estos modelos son extremadamente amplios, se presenta el modelo Holt-Winters [15], que utiliza tres ecuaciones de suavizamiento: una para el nivel l_t, otra para la tendencia b_t y otra para la componente estacional γ_t. La ecuación para el pronóstico en el periodo $\hat{y}_{(t+h)}$ es la que se presenta en la ecuación (9).

$$\hat{y}_{(t+h)|t} = (l_t + hb_t)s_{t+h-m(k+1)} \tag{9}$$

donde:

$$l_t = \alpha \frac{y_t}{s_{t-m}} + (1-\alpha)(l_{t-1} + b_{t-1}) \tag{10}$$

$$b_t = \beta^*(l_t - l_{t-1}) + (1-\beta^*)b_{t-1} \tag{11}$$

$$s_t = \gamma \frac{y_t}{(l_{t-1} + b_{t-1})} + (1-\gamma)s_{t-m} \tag{12}$$

Los parámetros α, β^* y γ en las ecuaciones (10), (11) y (12) son constantes a estimar utilizando algoritmos determinísticos que converjan a una solución óptima que minimice el error de ajuste; en este caso, se utiliza estimación por mínimos cuadrados que minimice el AIC, AICc y BIC.

2.7 Modelos NNETAR

Los modelos de redes neuronales están basados en modelos matemáticos simples del funcionamiento cerebral y permiten construir relaciones complejas no lineales entre la variable respuesta y sus predictoras [16]. Para los propósitos de generar pronósticos, se utiliza un tipo especial de redes neuronales, denominadas redes neuronales autorregresivas (*NNETAR o NNAR*, por sus siglas en inglés).

A grandes rasgos, el modelo NNAR tiene 2 entradas: el número de periodos retrasados (*lags*) en el tiempo p a tener en cuenta y el número de nodos en la capa oculta de neuronas k. Un modelo $NNAR(p,0)$ es equivalente a un $ARIMA(p,0,0)$, pero sin las restricciones en los parámetros para garantizar estacionariedad.

Un modelo de ajuste NNAR (3,5) se puede escribir como se muestra en la ecuación (13).

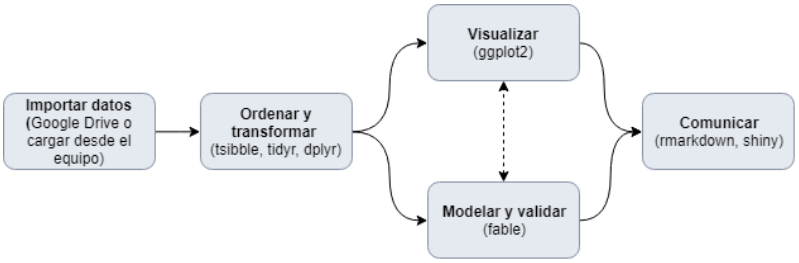
$$y_t = f(y_{t-1}, y_{t-2}, y_{t-3}) + e_t \qquad e_t \sim iid\ N(0, \sigma^2) \tag{13}$$

Dónde f es una red neuronal con cinco nodos ocultos en una sola capa. En este tipo de modelos no es posible derivar analíticamente intervalos de predicción, por lo que se utiliza simulación para generar múltiples trayectorias y construirlos [16].

3. METODOLOGÍA

La metodología propuesta en este trabajo consta de una serie de pasos claves, como se muestra en la Figura 1. Inicialmente, se realiza la importación de los datos, se procede con su limpieza y, posteriormente, se realiza un proceso iterativo de modelamiento, visualización y transformación, concluyendo con la comunicación a las partes interesadas.

Figura 1. Etapas de la metodología propuesta. Adaptado de [12].
Figure 1. Stages of the proposed methodology. Adapted from [12].



La metodología pretende disminuir considerablemente las dificultades usuales de importación de los datos, facilitando la integración con las aplicaciones basadas en hojas de cálculo, que actualmente utilizan en gran proporción las empresas y que más demanda el mercado [17] [18]. En la metodología propuesta se le permite al usuario importar sus datos directamente desde la API (*Application Programming Interface*) de Google Drive, que es un servicio web gratuito de almacenamiento popular [19] o subir directamente el archivo en formato *.xlsx.*

El segundo paso de la metodología es ordenar y transformar las estructuras longitudinales y de series de tiempo al formato de *tidy data.* Wang *et al*. [12] propusieron una nueva estructura temporal para los datos denominada *tsibble*, fundamentada en los principios de *tidy data* propuestos por Wickham [20], con el propósito de facilitar el modelamiento utilizando programas computacionales. El programa computacional desarrollado utiliza las herramientas de las librerías *dplyr* [21] y *tidyr"* del software R [22] [23] para realizar este cambio de formato sin perder información en el proceso, obteniendo una estructura de datos como la mostrada en la Tabla 3. En términos generales, se identifica que esta estructura no facilita el ingreso manual de los datos, ya que implica agregar filas y digitar fechas, por lo que generalmente se evita; pero es altamente conveniente para los programas ya que cada fila representa una observación única y cada columna representa una variable diferente.

Tabla 3. Ejemplo de estructura de datos en *tidy data.*
Table 3. An example of data structure in *tidy data.*

Fecha	Serie	Valor
1/1/2003	Comercio minorista	53.8
1/2/2003	Comercio minorista	51.1
1/3/2003	Comercio minorista	54.8
1/1/2003	Alimentos y bebidas no alcohólicas	74.7
1/2/2003	Alimentos y bebidas no alcohólicas	75.2
1/3/2003	Alimentos y bebidas no alcohólicas	85.9
1/1/2003	Bebidas alcohólicas y productos del tabaco	51.7
1/2/2003	Bebidas alcohólicas y productos del tabaco	44.9
1/3/2003	Bebidas alcohólicas y productos del tabaco	55.1

Una vez los datos están en el formato deseado *(tidy data)*, el tercer paso de la metodología corresponde a visualizar las series de tiempo con el propósito de identificar si es razonable asumir que el comportamiento pasado defina las características del futuro esperado.

La validación de la metodología se llevó a cabo mediante un caso aplicado a los índices empalmados de las ventas en valores reales de la Encuesta Mensual de Comercio Minorista (EMCM) del Departamento Administrativo Nacional de Estadística (DANE) de Colombia, que cuenta con diferentes series de datos temporales capturados desde el año 2003 hasta la actualidad, para diferentes productos como alimentos, bebidas, textiles, entre otros [24]. Cada serie cuenta con 198 datos tomados con una frecuencia mensual.

Resultado de la transformación que se realizó en los pasos anteriores, utilizando la capacidad computacional de la librería *ggplot2* [16], el programa es capaz de generar automáticamente gráficas individuales para todas las series a analizar. Para propósitos ilustrativos y dada la restricción de la extensión del artículo, se presentan solo diez reportes generados en la Figura 2, donde se evidencian diferentes tipos de series, algunas con componente estacional marcado, como la serie de papelería e útiles escolares, especialmente al final de cada año y otras con componentes cíclicas (cambios en la tendencia) evidentes, como la serie de repuestos y lubricantes para vehículos.

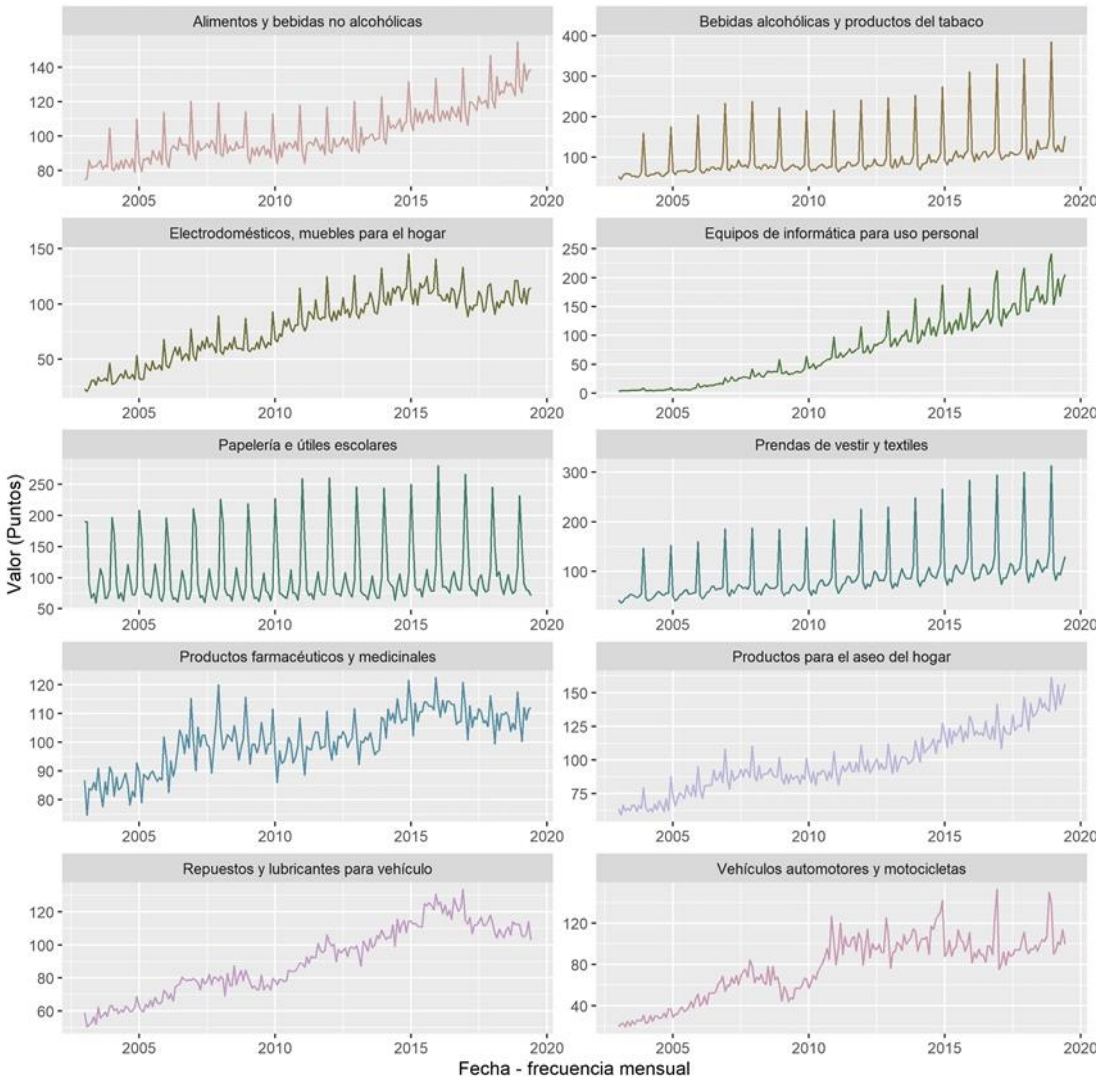
Existen múltiples análisis más detallados para la identificación de series y selección de los mejores modelos posibles, como la gráfica de la función de autocorrelación (ACF, por sus siglas en inglés), que mide las relaciones lineales de una misma serie con respecto a los k periodos anteriores, o utilizando pruebas formales como Box-Pierce [25] y Ljung-Box [26].

En esta metodología se facilita la eficiencia del proceso de pronóstico y el ajuste de los mejores modelos debido a la estructura de datos, la cual permite pasar directamente a la validación cruzada de los pronósticos. Así, los modelos que generan predicciones con errores altos o incumplen los supuestos, son fácilmente identificados y descartados.

El siguiente paso es generar los pronósticos con los modelos deseados y descartar los que generan medidas de error de pronóstico inadecuadas; esto dependerá del propósito del pronóstico y de las medidas de error aceptadas por las partes interesadas. Las entradas que requiere el programa para generar los pronósticos son las siguientes: los modelos a utilizar, la cantidad de periodos a pronosticar y el número de datos recortados de la serie original para la validación cruzada.

Dándole continuidad al caso de aplicación de la Figura 2 y utilizando el paquete *fable* del software R [27], se utilizaron los siguientes modelos: ARIMA *(AutoRegressive Integrated Moving Average)*, Globales con variables indicadoras, suavizamientos exponenciales (Holt, Winters, Holt-Winters), NNETAR *(Neural Network AutoRegressive models)* y de los métodos tradicionales, la deriva para propósitos de comparación. Adicionalmente, se realizó pronósticos para 24 meses en el futuro y la validación cruzada se hizo con los primeros 12 meses de estos.

Figura 2. Series del índice empalmado de ventas reales del comercio minorista colombiano. Fuente [24].
Figure 2. Series of the spliced index of real sales of Colombian retail trade. Source [24].



4. RESULTADOS Y DISCUSIÓN

La validación de la metodología se llevó a cabo a través de la evaluación de las medidas de error de pronóstico, para así seleccionar el mejor de los modelos y utilizar los pronósticos para su propósito específico, dado que cumplan los parámetros definidos por el usuario final (frecuencia de los datos temporales, estructura de entrada, modelos a utilizar y medidas de error). En la Figura 4 se presentan los resultados utilizando el MAPE (*Mean Absolute Percentage Error*) como criterio de selección para el mejor modelo de pronóstico, dado que es la medida más representativa porque considera desviaciones porcentuales [6]. Se recomienda tener en cuenta las consideraciones sobre esta medida de error expuestas en [28], la cual se calcula como se muestra en la ecuación (14).

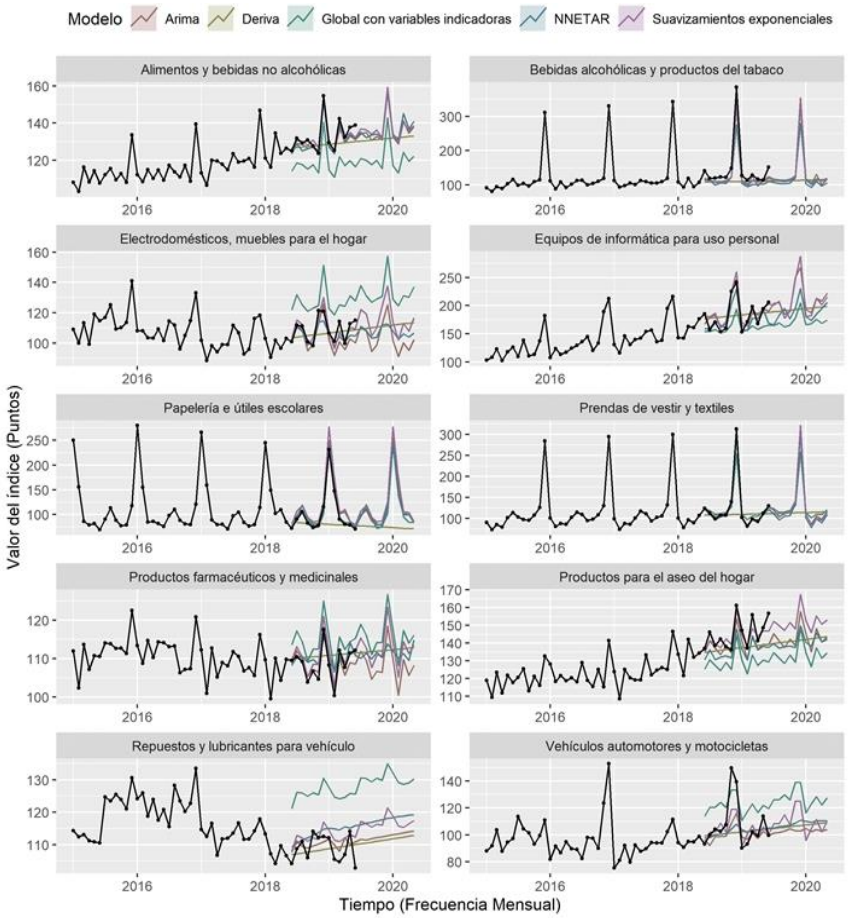
$$MAPE = mean\left(\left|\frac{100e_t}{y_t}\right|\right) \quad (14)$$

donde e_t corresponde a la diferencia entre el valor observado $y_{(T+h)}$ y su pronóstico $\hat{y}_{(T+h)|T}$, de acuerdo a la ecuación (15) y, además, $mean$ se refiere a la media aritmética.

$$e_t = e_{T+h} = y_{T+h} - \hat{y}_{(T+h)|T} \quad (15)$$

La metodología permite además utilizar en lugar del MAPE otras medidas de error, como el ME (*Mean Error*), RMSE (*Root Mean Square Error*), MAE (*Mean Absolute Error*), MPE (*Mean Percentage Error*) o MASE (*Mean Absolute Scaled Error*). Luego del modelado realizado aplicando las técnicas descritas en la sección anterior, se lleva a cabo la fase de comunicación mediante la construcción de las gráficas de la serie original y los pronósticos generados por los modelos seleccionados en la Figura 3, con el propósito de evaluarlos en el periodo de validación cruzada y visualizar el comportamiento futuro esperado.

Figura 3. Pronósticos generados y validación cruzada para el índice de ventas reales de comercio minorista.
Figure 3. Forecasts generated and cross-validation for the real retail sales index.



En general, los suavizamientos exponenciales y los modelos ARIMA capturaron apropiadamente las componentes de tendencia y estacionalidad de la mayoría de las series. Sin embargo, los modelos utilizados no capturaron gran parte de los cambios estructurales, ya que esto requiere intervención directa de un analista especializado. Los modelos globales adicionalmente tuvieron resultados particularmente buenos cuando existían tendencias lineales, como en la serie de productos farmacéuticos y medicinales. Nótese que en la Figura 3 cada una de las series tiene una escala vertical diferente, por lo que no son comparables y el análisis debe realizarse de manera independiente.

En la Tabla 4 se muestran los resultados comparativos del MAPE obtenido para cada serie utilizando el método de la deriva y el método seleccionado automáticamente, al igual que el porcentaje de mejora obtenido. Se evidencia que el método de pronóstico seleccionado por esta metodología presenta un porcentaje de mejora significativa del 50.56% en promedio, con respecto al uso de un método tradicional de pronóstico como la deriva.

En la Figura 4 se identifican los mejores modelos para cada una de las series teniendo como criterio el MAPE. En este caso, la serie de productos farmacéuticos y medicinales fue la que menores medidas de error de pronóstico presentó con los modelos utilizados. Dependiendo del propósito del pronóstico, aceptar o no un error en los rangos obtenidos puede implicar decisiones diferentes; por ejemplo, si el propósito es realizar un presupuesto, los resultados pueden ser considerablemente mejores que si se realiza un promedio simple o un cálculo con deriva, como se muestra en la Figura 5 donde se presenta el MAPE para este método de pronóstico.

Para la generación de los reportes automatizados (Figura 6) de aspectos de interés de los modelos seleccionados, se utilizó el paquete *broom* [29], que entrega estructuras tabulares limpias de los resultados que se pueden conectar posteriormente a otros aplicativos, como tablas dinámicas en hojas de cálculo. Los reportes automatizados pueden ser programados con las salidas de interés utilizando la herramienta *rmarkdown* [30].

Tabla 4. Resultados del MAPE y porcentaje de mejora para cada serie respecto al método seleccionado.
Table 4. MAPE and percentage improvement results for each time series according to the selected method.

Serie	MAPE del método de la deriva (%)	MAPE del método seleccionado por la metodología (%)	Porcentaje de mejora (%)
Alimentos y bebidas no alcohólicas	4.40	1.65	62.56
Bebidas alcohólicas y productos del tabaco	17.33	11.29	34.84
Electrodomésticos, muebles para el hogar	6.28	3.51	44.12
Equipos de informática para uso personal	11.79	5.36	54.56
Papelería e útiles escolares	17.36	6.68	61.52
Prendas de vestir y textiles	15.70	1.56	90.06
Productos farmacéuticos y medicinales	3.13	1.39	55.70
Productos para el aseo del hogar	5.41	2.48	54.19
Repuestos y lubricantes para vehículo	3.48	3.13	10.14
Vehículos automotores y motocicletas	9.24	5.74	37.92
		Mejora Promedio Total:	**50.56**

Figura 4. Resultados del modelo con menor MAPE para cada serie pronosticada.
Figure 4. Results of the model with the lowest MAPE for each forecasted series.

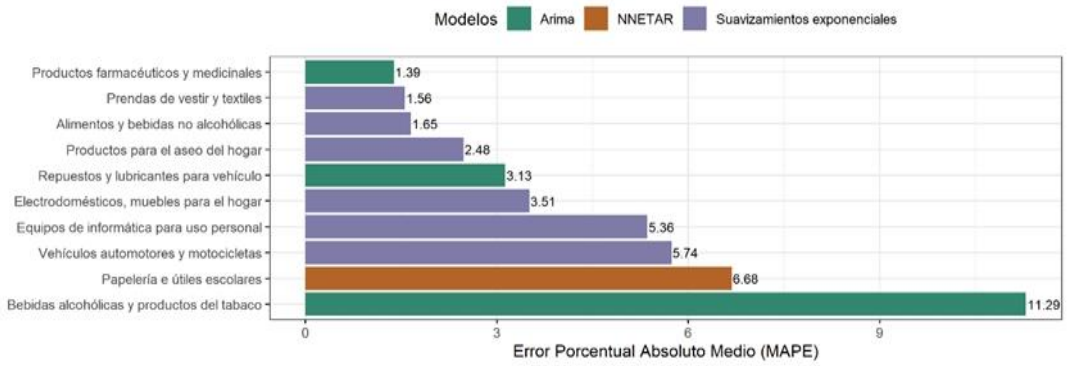


Figura 5. Resultados de MAPE utilizando un método de pronóstico simple - Deriva.
Figure 5. MAPE results using a simple forecasting method - Drift.

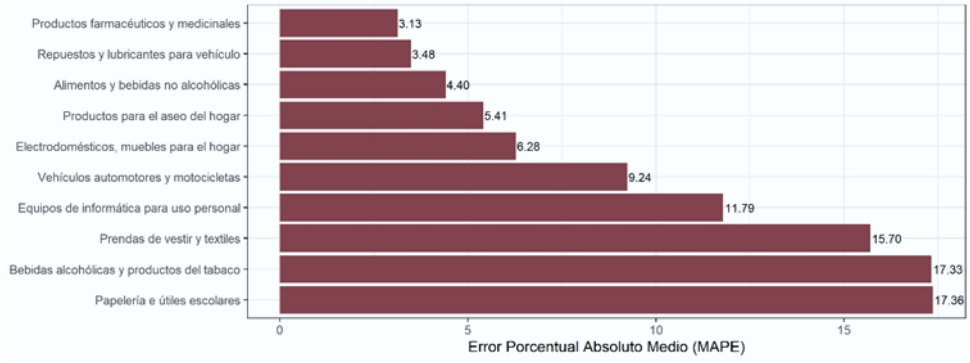
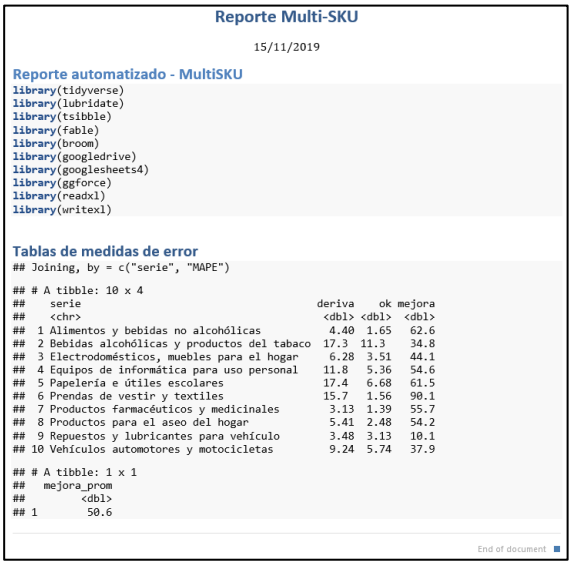


Figura 6. Ejemplo de reporte automatizado programable obtenido con la metodología.
Figure 6. Example of an automated customizable report obtained with the methodology.



Reporte Multi-SKU

15/11/2019

Reporte automatizado - MultiSKU

```
library(tidyverse)
library(lubridate)
library(tsibble)
library(fable)
library(broom)
library(googledrive)
library(googlesheets4)
library(ggforce)
library(readxl)
library(writexl)
```

Tablas de medidas de error

```
## Joining, by = c("serie", "MAPE")

## # A tibble: 10 x 4
##    serie                                deriva    ok mejora
##    <chr>                                <dbl> <dbl>  <dbl>
##  1 Alimentos y bebidas no alcohólicas    4.40  1.65   62.6
##  2 Bebidas alcohólicas y productos del tabaco  17.3  11.3   34.8
##  3 Electrodomésticos, muebles para el hogar     6.28  3.51   44.1
##  4 Equipos de informática para uso personal    11.8   5.36   54.6
##  5 Papelería e útiles escolares                17.4   6.68   61.5
##  6 Prendas de vestir y textiles                15.7   1.56   90.1
##  7 Productos farmacéuticos y medicinales        3.13  1.39   55.7
##  8 Productos para el aseo del hogar             5.41  2.48   54.2
##  9 Repuestos y lubricantes para vehículo        3.48  3.13   10.1
## 10 Vehículos automotores y motocicletas         9.24  5.74   37.9

## # A tibble: 1 x 1
##   mejora_prom
##         <dbl>
## 1        50.6
```

End of document

5. CONCLUSIONES

La metodología de pronóstico desarrollada en este trabajo se aplicó en un caso real considerando diez series de tiempo simultáneamente; sin embargo, es posible utilizarla en un número mayor de series temporales sin restricciones, debido a sus características estructurales y de escalabilidad, siempre y cuando sea razonable asumir que los comportamientos anteriores de las series son útiles para predecir los valores futuros. En esta metodología se superan dificultades tradicionales del modelamiento de series de tiempo, principalmente relativas al grado de escalabilidad y se disminuyen considerablemente las curvas de aprendizaje, facilitando así su uso en ambientes empresariales en el sector real o de servicios.

Esto se logra en gran parte debido a un cambio en el proceso de modelamiento utilizando nuevas herramientas computacionales y ajustando las estructuras de datos tradicionales para series de tiempo en estructuras *tidy*. Las empresas que deseen implementar la metodología sólo requieren disponer de personal capacitado en el manejo de los paquetes *tsibble* y *fable* de R o sus análogos en otros lenguajes y no necesitan de compras adicionales de software.

La metodología desarrollada permite presentar las gráficas de las series analizadas sin tener que realizarlas de manera individual y pasar directamente a los pronósticos y su validación. La aplicación de la metodología mostró mejoras apreciables con un promedio de disminución del MAPE del 50.56% en el caso aplicado. El rigor en la toma de datos, su consolidación y manejo adecuado son factores determinantes para el buen funcionamiento de la metodología; teniendo presente además que la cantidad de observaciones juega un papel fundamental para el funcionamiento de algunos modelos, como es el caso de los suavizamientos exponenciales, los modelos ARIMA y los modelos de tendencia global con variables indicadoras.

Este trabajo es un resultado de una línea de investigación centrada en el desarrollo de modelos de pronóstico más cercanos a la realidad de los procesos y las operaciones. Los trabajos futuros están orientados a aplicaciones en otros sectores industriales y a la incorporación de otras técnicas de inteligencia artificial que permitan mejorar el desempeño global de los pronósticos y faciliten la interacción del usuario. También, es de interés la incorporación de múltiples variables predictoras que mejoren los pronósticos mediante metodologías multivariadas.

REFERENCIAS

[1]. S. Chopra, *Supply Chain Management: Strategy, Planning, and Operation*. 7 ed. New York: Pearson, 2018.

[2]. G. Ghiani, G. Laporte, R. Musmanno, *Introduction to Logistics Systems Management*. 2 ed. Chichester, UK: John Wiley & Sons, 2013.

[3]. W. Wei, *Multivariate Time Series Analysis and Applications*. Hoboken, NJ: John Wiley & Sons, 2019.

[4]. R. Tsai, *Multivariate Time Series Analysis with R and Financial Applications*. Hoboken, NJ: John Wiley & Sons, 2014.

[5]. S. Pravilovic, M. Bilancia, A. Appice, D. Malerba, "Using multiple time series analysis for geosensor data forecasting", *Information Sciences*, 380, 31-52, 2017. http://dx.doi.org/10.1016/j.ins.2016.11.001.

[6]. M. Matyjaszek, P. Riesgo, A. Krzemień, K. Wodarski, G. Fidalgo, "Forecasting coking coal prices by means of ARIMA models and neural networks, considering the transgenic time series theory", *Resources Policy*, 61, 283-292, 2019. https://doi.org/10.1016/j.resourpol.2019.02.017.

[7]. A. González-Vidal, F. Jiménez, A. Gómez-Skarmeta, "A methodology for energy multivariate time series forecasting in smart buildings based on feature selection", *Energy & Buildings*, 196, 71-82, 2019. https://doi.org/10.1016/j.enbuild.2019.05.021.

[8]. N. Barthel, C. Czado, Y. Okhrin, "A partial correlation vine based approach for modeling and forecasting multivariate volatility time-series", *Computational Statistics and Data Analysis*, 142, 1-29, 2020. https://doi.org/10.1016/j.csda.2019.106810.

[9]. J. Karmy, S. Maldonado, "Hierarchical time series forecasting via Support Vector Regression in the European Travel Retail Industry", *Expert Systems With Applications*, 137, 59-73, 2019. https://doi.org/10.1016/j.eswa.2019.06.060.
[10]. T. Niu, J. Wang, H. Lu, W. Yang, P. Du, "Developing a deep learning framework with two-stage feature selection for multivariate financial time series forecasting", *Expert Systems With Applications*, 148, 1-17, 2020. https://doi.org/10.1016/j.eswa.2020.113237.
[11]. S. Du, T. Li, Y. Yang, S. Horng, "Multivariate time series forecasting via attention-based encoder–decoder framework", *Neurocomputing, In Press*, 2020. https://doi.org/10.1016/j.neucom.2019.12.118.
[12]. E. Wang, D. Cook, R.J. Hyndman, (2019) A new tidy data structure to support exploration and modeling of temporal data [on line], (February), p. 1-28. Disponible desde: https://arxiv.org/pdf/1901.10257.pdf. [Acceso 10 de noviembre 2019].
[13]. R.J. Hyndman, G. Athanasopoulos, *Forecasting: principles and practice*. 2 ed. Melbourne, Australia: OTexts, 2018.
[14]. G.C. Tiao, "Time Series: ARIMA Methods", *International Encyclopedia of the Social & Behavioral Sciences*, 23, 316-321, 2015. https://doi.org/10.1016/B978-0-08-097086-8.42182-3.
[15]. P.R. Winters, "Forecasting Sales by Exponentially Weighted Moving Averages", *Management Science*, 6 (3), 324-342, 1960. https://doi.org/10.1287/mnsc.6.3.324.
[16]. H. Wickham, "ggplot2", *Wiley Interdisciplinary Reviews: Computational Statistics*, 3 (2), 180-185, 2011. https://doi.org/10.1002/wics.147.
[17]. K. Kluza, P. Wiśniewski, "Spreadsheet-based Business Process modeling", *2016 Federated Conference on Computer Science and Information Systems (FedCSIS)*, Gdansk, Poland, 2016, pp. 1355-1358.
[18]. P. Coleman, R. Blankenship, "What Spreadsheet and Database Skills Do Business Students Need?" *Journal of Instructional Pedagogies*, 19, 1–8, 2017.
[19]. A. Sadik, "Students' acceptance of file sharing systems as a tool for sharing course materials: The case of Google Drive", *Education and Information Technologies*, 22 (5), 2455-2470, 2017. https://doi.org/10.1007/s10639-016-9556-z.
[20]. H. Wickham, "Tidy Data", *Journal of Statistical Software*, 59 (10), 1-23, 2014. https://doi.org/10.18637/jss.v059.i10.
[21]. H. Wickham, R. Francois, L. Henry, K. Müller, (2019) *dplyr: A Grammar of Data Manipulation* [on line]. Disponible desde: https://cloud.r-project.org/web/packages/dplyr/index.html. [Acceso 12 de noviembre 2019].
[22]. H. Wickham, L. Henry, (2019) *tidyr: Tidy Messy Data* [on line]. Disponible desde: https://cran.r-project.org/web/packages/tidyr/index.html. [Acceso 12 de noviembre 2019].
[23]. J.M. Chambers, *Software for Data Analysis: Programming with R*. New York: Springer, 2008.
[24]. DEPARTAMENTO ADMINISTRATIVO NACIONAL DE ESTADÍSTICAS. (2019). *Ficha Metodológica Encuesta Mensual de Comercio al por Menor y Comercio de Vehículos – EMCM* [Internet], Bogotá, DANE. Disponible desde: https://www.dane.gov.co/index.php/estadisticas-por-tema/comercio-interno/encuesta-emcm#informacion-emcm-junio-2019. [Acceso 14 de noviembre 2019].
[25]. G.E.P. Box, D. Pierce, "Distribution of Residual Autocorrelations in Autoregressive-Integrated Moving Average Time Series Models", *Journal of the American Statistical Association*, 65 (332), 1509-1526, 1970. https://doi.org/10.1080/01621459.1970.10481180.
[26]. G.M. Ljung, G.E.P. Box, "On a measure of lack of fit in time series models", *Biometrika*, 65 (2), 297-303, 1978. https://doi.org/10.1093/biomet/65.2.297.
[27]. M. O'Hara-Wild, R. Hyndman, E. Wang, (2019) *Fable R Package* [on line]. Disponible desde: https://fable.tidyverts.org/index.html. [Acceso 12 de noviembre 2019].
[28]. A. de Myttenaere, B. Golden, B. Le Grand, F. Rossi, "Mean Absolute Percentage Error for regression models", *Neurocomputing*, 192 (5), 38–48, 2016. https://doi.org/10.1016/j.neucom.2015.12.114.
[29]. D. Robinson, A. Hayes, (2019) broom: Convert Statistical Analysis Objects into Tidy Tibbles [on line], CRAN. Disponible desde: https://cran.r-project.org/web/packages/broom/index.html. [Acceso 12 de noviembre 2019].
[30]. B. Baumer, D. Udwin. "R Markdown". *Wiley Interdisciplinary Reviews: Computational Statistics*, 7 (3), 167-177, 2015. https://doi.org/10.1002/wics.1348.