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

**Luis D. Chavarría1 y Juan M. Cogollo 1**

(1) Facultad de Minas, Dpto. Ingeniería de la Organización, Universidad Nacional de Colombia.

E-mail: ldchavarriam@unal.edu.co; jmcogollof@unal.edu.co

Resumen

Parámetros trabajo: <http://congreso.pucp.edu.pe/caip2019/envio-trabajos/trabajos-completos/>

Palabras clave: elementos finitos; difusión; Algor; modelado de procesos; extracción; aceite de clavo de olor

## Multivariate forecast methodology with integrated cloud-based reports

## Abstract

Resumen en inglés.

Keywords: finite elements; diffusion; Algor; process modeling; extraction; clove oil

### Introducción

La estimación correcta de la demanda desagregada para múltiples productos es fundamental para el buen funcionamiento de la mayoría de los modelos de gestión industrial y financiera (MPS, MRP, EOQ, flujos de caja, entre otros). Ampliar y Referenciar

El análisis de series de tiempo es una de las herramientas más utilizadas en este campo y tiene como objetivo hacer predicciones basándose en los datos del pasado. Este tipo de análisis requiere alto rigor matemático-estadístico para ser confiable; por lo que no es escalable de manera factible a n productos por su alto costo, por lo que, en ocasiones, las empresas prefieren utilizar métodos cualitativos o empíricos. Ampliar y Referenciar con base en los métodos tradicionales

Los problemas de series de tiempo temporales se agrupan típicamente en dos tipos de análisis, de series de tiempo y longitudinales. A pesar de que sus entradas sean los mismos datos, sus acercamientos al modelamiento difieren. (Wang, Cook, & Hyndman, 2019)

Los programas ERP *(Enterprise Resource Planning)* generalmente permiten realizar exportaciones de sus datos, los cuales posteriormente pueden ser analizados para llegar a conclusiones útiles. Estimar correctamente los valores futuros de una variable es con frecuencia uno de los análisis más requeridos.

Argumentar las dificultades de los métodos tradicionales y sus limitaciones

Posteriormente, finalizar explicando las bondades (se presenta), motivación (en la revisión de literatura no se encontró una metodología que lo resolviera) y aporte (por qué es novedoso con respecto a los demás) de la nueva metodología y por qué se desarrolla. Utilizar el enfoque del congreso, que es aplicado de la computación a la industria de procesos.

### Métodos de pronóstico cuantitativos tradicionales

Existen múltiples maneras de generar pronósticos, entre estos métodos se encuentran los cualitativos y los cuantitativos. Estos últimos requieren que exista información del pasado, y además, que sea razonable asumir que algunos patrones anteriores continuarán en el futuro. (R. J. Hyndman & Athanasopoulos, 2014)

Entre los métodos cuantitativos, hay tres muy utilizados en la industria que corresponden a los métodos más sencillos: el promedio, el método de *Naïve* y la deriva.

**Método del promedio**

Consiste en tomar el promedio de las observaciones en un periodo particular y asumir el futuro como este resultado. Si denotamos los datos históricos como entonces podemos escribir los pronósticos como:

|  |  |
| --- | --- |
|  | (1) |

Dónde la notación es una abreviación para la estimación de dados los datos históricos , es el horizonte de pronóstico utilizado y 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.

**Método de Naïve**

Para producir pronósticos se utiliza únicamente el valor de la última observación, es decir:

|  |  |
| --- | --- |
|  | (2) |

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

|  |  |
| --- | --- |
|  | (3) |

Dónde 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:

|  |  |
| --- | --- |
|  | (4) |

Dónde 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, los valores futuros para marzo serán iguales al valor de marzo del año anterior.

**Método de la deriva**

Es una variación del método de Naïve que considera que los pronósticos crezcan 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:

|  |  |
| --- | --- |
|  | (5) |

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

### Estructura de datos

Existen dos estructuras principales utilizadas para el almacenamiento de datos temporales: longitudinales o de series temporales. La configuración longitudinal tiene en sus columnas puntos temporales 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, por ejemplo (como se presenta en la 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 como se expone en la Tabla 2.

Tabla 1. Ejemplo de estructura de datos longitudinal

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 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 de series de tiempo

|  |  |  |  |
| --- | --- | --- | --- |
| 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 |

### Metodología

En el flujo de trabajo de esta metodología se propone una serie de pasos claves ilustrados 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.

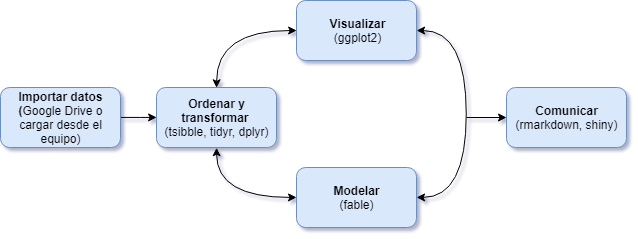


Figura 1. Flujo de trabajo de la metodología propuesta. Adaptado de: (Wang et al., 2019)

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 (FedCSIS, 2016) y que más demanda el mercado (Coleman & Blankenship, 2017). En el software desarrollado y acorde a la metodología propuesta, se permite al usuario importar sus datos directamente desde Google Drive, que es un servicio web gratuito de almacenamiento popular (Sadik, 2017) o subir directamente el archivo en formato *.xlsx.*

(Wang et al., 2019) proponen una nueva estructura temporal para los datos denominada *“tsibble”* fundamentada en los principios de “*tidy data*” propuestos por (Wickham, 2015), con el propósito de facilitar el modelamiento utilizando programas computacionales.

El segundo paso de la metodología desarrollada es ordenar y transformar las estructuras longitudinales y de series de tiempo al formato de *tidy data.* El programa computacional desarrollado utiliza las herramientas de las librerías “*dplyr*” (Wickham, Francois, Henry, & Müller, 2019) y *“tidyr”* (Wickham, Henry, & RStudio, 2017) de R (Chambers & Al, 2012) para realizar este cambio de formato sin perder información en el proceso. La nueva estructura de datos correspondiente a la tabla 2 sería la siguiente:

Tabla 3. Estructura de datos en *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 |

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.

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

Para ilustrar la metodología y validar los resultados, los siguientes pasos se explicarán con 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 17 series diferentes con datos temporales capturados desde el año 2003 hasta la actualidad (DANE, 2015).

Gracias al trabajo de transformación que se realizó en los pasos anteriores, utilizando la capacidad computacional de la librería *ggplot2* (Wickham, 2011) el programa es capaz de generar automáticamente gráficas individuales para todas las series a analizar; para propósitos de simplificación se presentan únicamente 10 en la Figura 2.

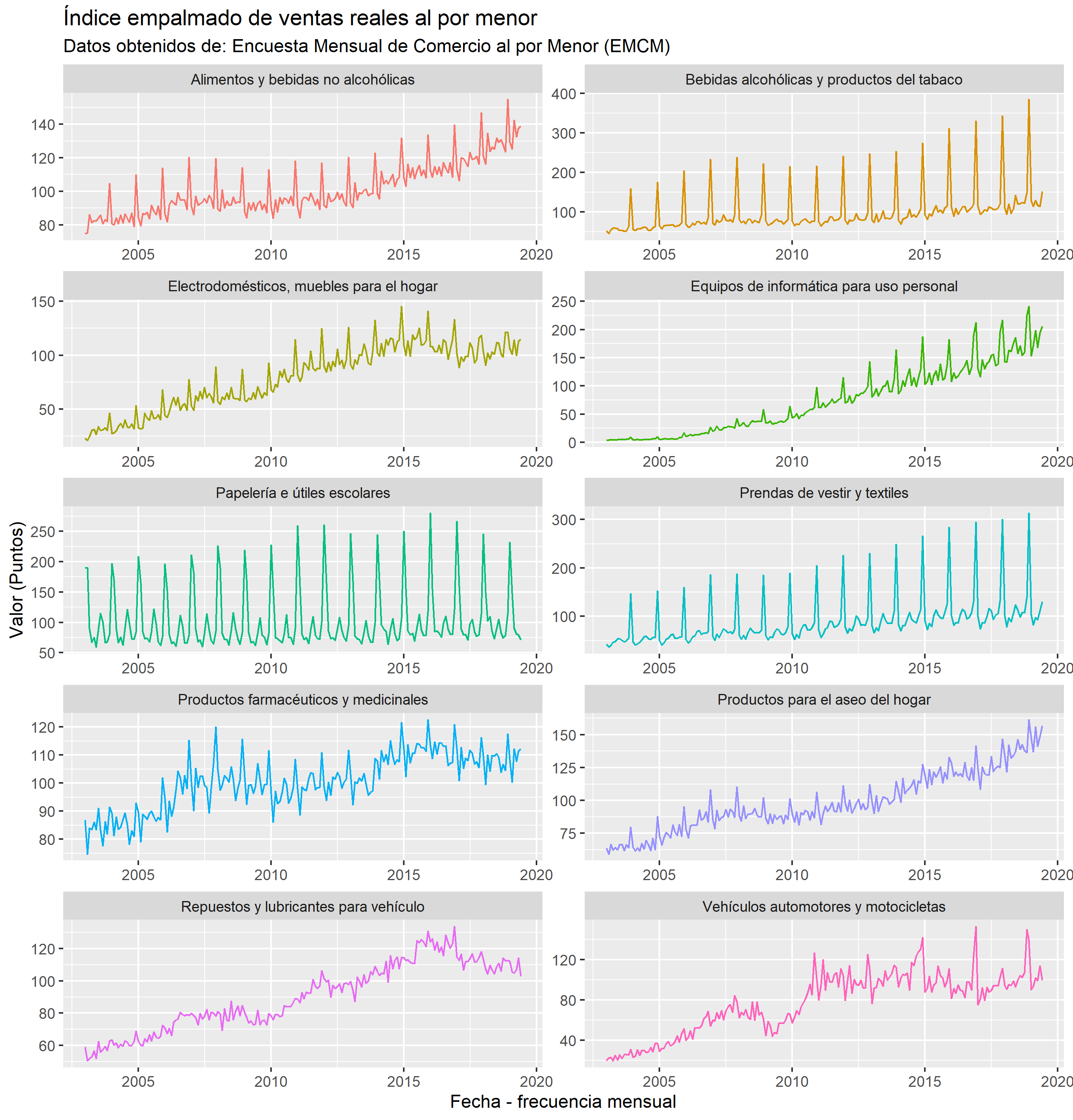


Figura 2. Series del índice empalmado de ventas reales del comercio minorista colombiano.

En la figura 2 se evidencian diferentes tipos de series, algunas con componente estacional marcada, 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 las de Box-Pierce (Box & Pierce, 1970) o las de Ljung-Box (Ljung & Box, 1978), entre otras herramientas esenciales.

En esta metodología, asegurarse de hacer eficiente el trabajo y ajustar únicamente los mejores modelos no es una preocupación gracias a la estructura de datos; ya que se pasa directamente a la validación cruzada de los pronósticos y los modelos que generan malas predicciones son fácilmente descartados utilizando las medidas de error; lo que generalmente se haría antes de pronosticar descartándolos por incumplir los supuestos de los modelos mediante análisis.

El siguiente paso es pasar directamente a 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 ejemplo de la Figura 2 y utilizando el paquete de R *“fable”* (R. Hyndman, O’Hara-Wild, & Wang, 2019), se utilizarán 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 simples, únicamente la deriva para propósitos de comparación. Adicionalmente, se realizarán pronósticos para 24 meses en el futuro y la validación cruzada se hará con los primeros 12 meses de éstos.

A continuación, se presenta una breve descripción de los modelos utilizados:

**Modelos ARIMA**

De acuerdo a tiao, arima methods un modelo ARMA cuenta con un componente autorregresivo (Parte AR) y un componente de medias móviles (MA). Tengamos inicialmente en cuenta la ecuación que describe un modelo ARMA(P,Q)

|  |  |
| --- | --- |
|  | (6) |

Dónde generalmente se asume que cumple los supuestos de media cero y varianza constante, representa el coeficiente autorregresivo de orden P y 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.

|  |  |
| --- | --- |
|  | (7) |

Dónde y los parámetros y están funcionalmente relacionados y se asume un modelo ARMA para 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).*

**Modelos globales con variables indicadoras**

En el caso más simple, la regresión permite hallar una relación lineal utilizando estimación de mínimos cuadrados, esto daría como resultado la tendencia de la serie, utilizando un parámetro para definir el intercepto con el eje y para incrementar el número de parámetros en el modelo en la tendencia.

Para modelar la estacionalidad se estima un parámetro para cada periodo el cuál se activa cuando es necesario utilizando variables indicadoras. Se especifica además un término de error que debe distribuirse idénticamente independiente con media cero y varianza constate, como se muestra en la ecuación 8.

|  |  |
| --- | --- |
|  | (8) |

Dónde la variable es una variable indicadora que tiene valor 1 si en el tiempo se encuentra en la estación y 0 en otro caso, y adicionalmente es su parámetro estimado correspondiente.

**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. 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, presentamos aquí el modelo *Holt-Winters* multiplicativo Holt-Winters, que utiliza 3 ecuaciones de suavizamiento, una para el nivel , otra para la tendencia y otra para la componente estacional .

La ecuación para el pronóstico en el periodo es la que se presenta en la ecuación 9.

|  |  |
| --- | --- |
|  | (9) |

Dónde:

|  |  |
| --- | --- |
|  | (10) |
|  | (11) |
|  | (12) |

Los parámetros 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.

**Modelos NNETAR**

Los modelos de redes neuronales están basados en modelos matemáticos simples del funcionamiento cerebral, permiten construir relaciones complejas no-lineares entre la variable respuesta y sus predictoras. 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, su número de periodos retrasados *(lags)* en el tiempo a tener en cuenta y el número de nodos en la capa oculta de neuronas . Un modelo es equivalente a un , 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.

|  |  |
| --- | --- |
|  | (13) |

Dónde es una red neuronal con 5 nodos ocultos en una sola capa. Una particularidad de este tipo de modelos es que no es posible derivar analíticamente intervalos de predicción, por lo que se utiliza simulación iterativa para generarlos hasta que con base en los resultados de estas, es posible estimar sus intervalos de predicción. fpp2

### Resultados y discusión

Una vez todas las entradas han sido correctamente definidas, se construyen las gráficas de la serie original en color negro y los pronósticos generados por los modelos seleccionados utilizando colores en la Figura 3, con el propósito de evaluarlos en el periodo de validación cruzada y visualizar el comportamiento futuro esperado.

El siguiente paso es evaluar 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. 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, se recomienda tener en cuenta las consideraciones sobre esta medida de error expuestas por (de Myttenaere, Golden, Le Grand, & Rossi, 2016) el cual se calcula como se muestra en la ecuación 14.

|  |  |
| --- | --- |
|  | (14) |

Dónde corresponde a la diferencia entre el valor observado y su pronóstico , de acuerdo a la ecuación 15 y además, se refiere a la media aritmética.

|  |  |
| --- | --- |
|  | (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)*, MASE *(Mean Absolute Scaled Error)* y sus demás medidas derivadas a voluntad del usuario.

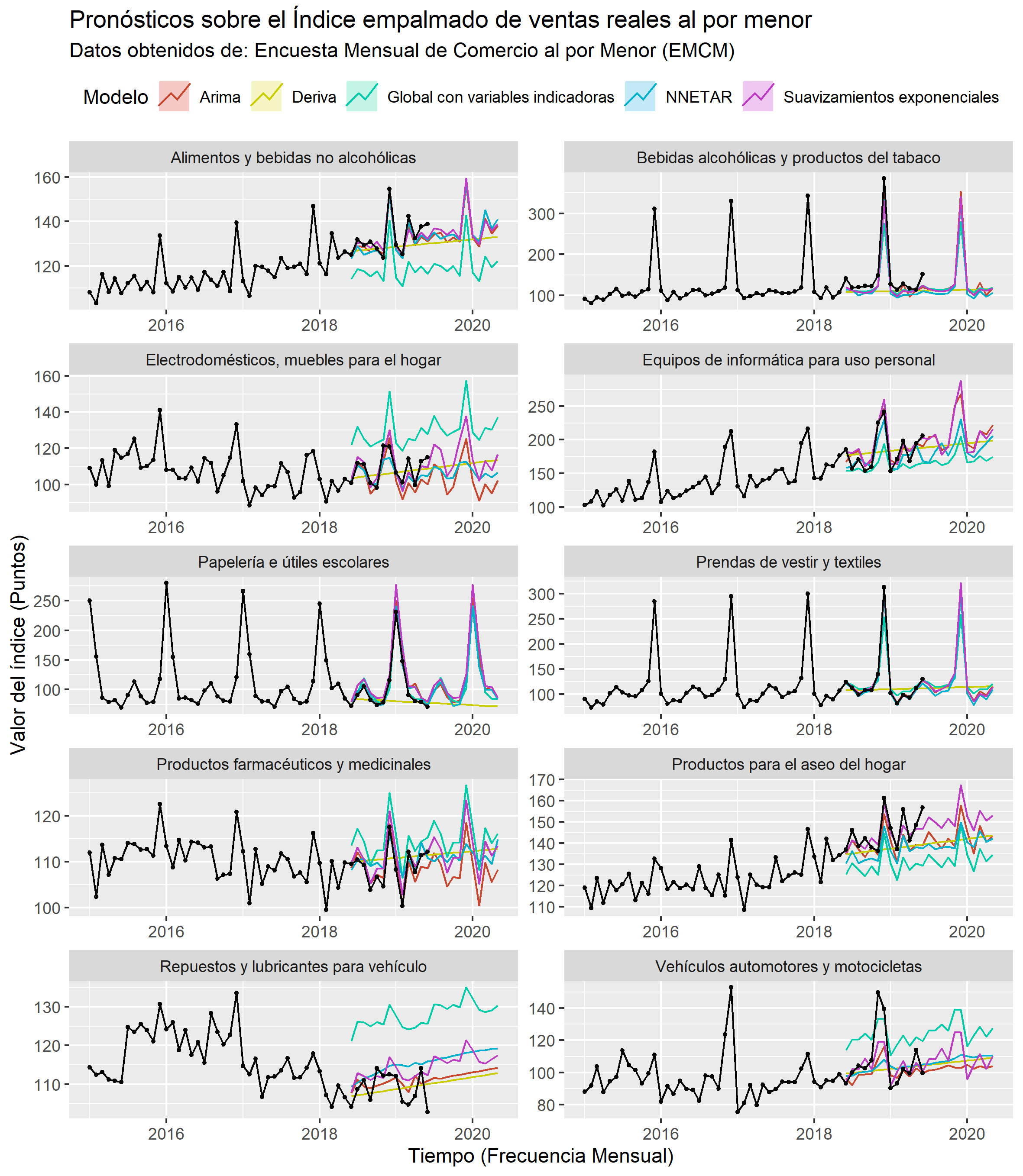


Figura 3. Pronósticos generados y validación cruzada para el índice de ventas reales de comercio minorista

En general, los Suavizamientos exponenciales y 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.

Note que en la Figura 3, las escalas para cada gráfica son diferentes para que se puedan diferenciar individualmente, por lo que no son comparables, además de ser datos diferentes. Para propósitos de comparación entre modelos se recomienda ir directamente las medidas de error expuestas en la Figura 4.

En la Figura 4 se identifican los mejores modelos para cada una de las series de acuerdo al 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 dependiendo de la escala; 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.

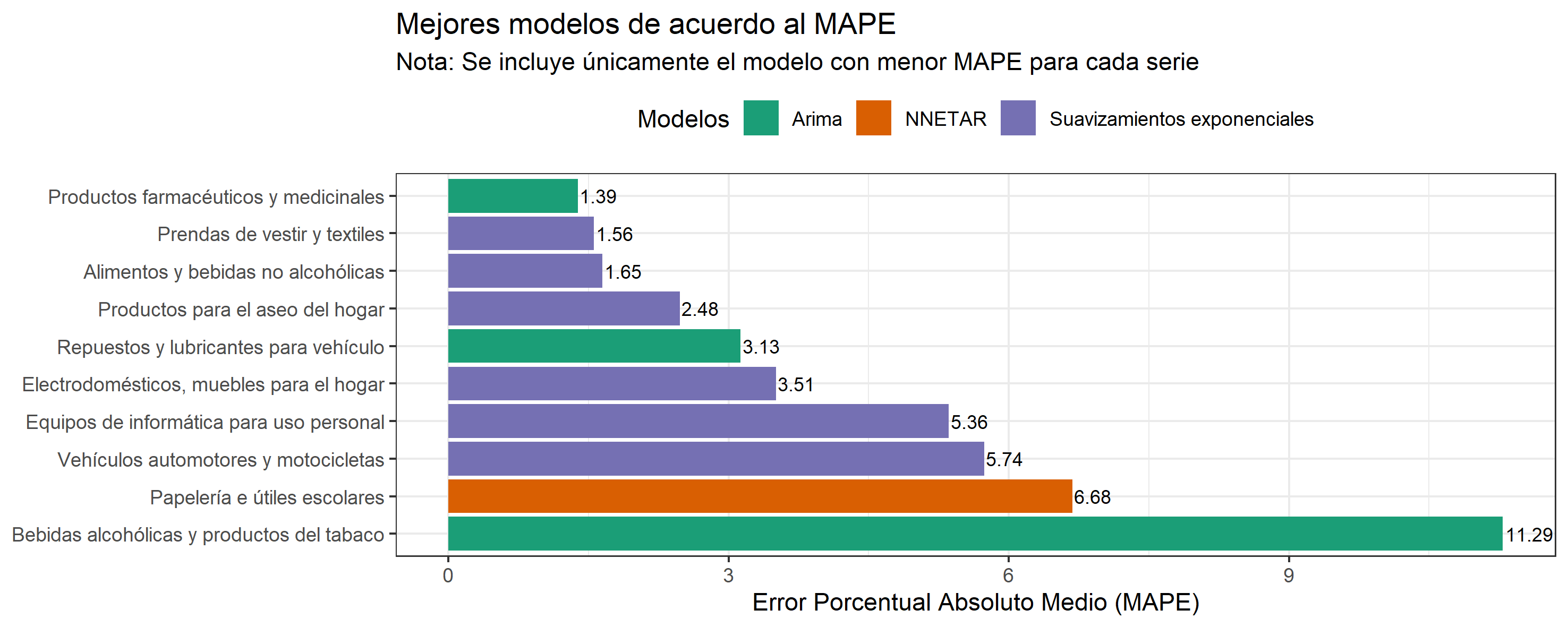


Figura 4. Mejores modelos de acuerdo al MAPE

En la Figura 5 se evidencia que los modelos ajustados llevaron a mejoras sustanciales con respecto a un método de pronóstico simple popular, como lo es la extrapolación de una línea de tendencia entre la primera y última observación (Deriva).

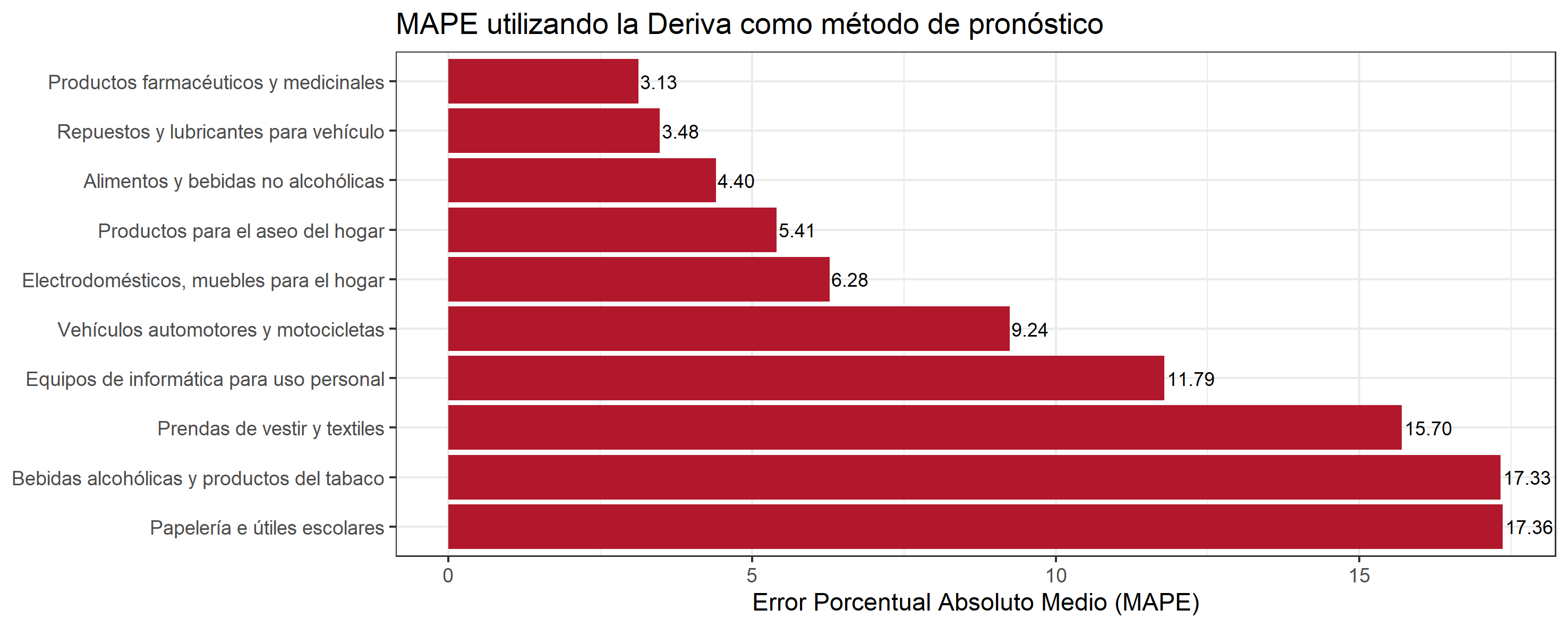


Figura 5. MAPE utilizando un método de pronóstico simple - Deriva.

### Conclusiones

En la metodología propuesta se superan dificultades tradicionales del modelamiento de series de tiempo, principalmente relativas al grado de escalabilidad, se mejora el énfasis gráfico y se disminuyen considerablemente las curvas de aprendizaje, facilitando así su uso en ambientes empresariales y productivos. Esto se logra en gran parte gracias 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”*; este flujo de trabajo se presentó en la Figura 1*.*

En la Figura 2 se presentaron las gráficas de las series analizadas sin tener que realizarlas de manera individual, lo que confirma el funcionamiento adecuado del trabajo previo realizado y en la Figura 3 se presentan los pronósticos, los cuales fueron generados para todas las series utilizando como entrada únicamente los modelos a utilizar (De un repertorio de modelos aplicables definido por el paquete fable), la duración del período de validación cruzada (12 meses en este caso) y el horizonte de pronóstico (24 meses en este trabajo).

Posteriormente, en la Figura 4 se presentaron los resultados del error de pronóstico MAPE para los mejores modelos de la metodología y en la Figura 5 los mismo resultados de un método de pronóstico simple, como el cálculo de extrapolaciones a futuro con base en la última y primera observación de los datos (Deriva), mostrando mejoras apreciables en todas las series con un promedio de disminución del MAPE del 50.6%.

De la misma manera que en este trabajo se realizó el proceso para 10 series de tiempo de manera simultánea, es posible realizarlo para cuantas series temporales sea necesario siguiendo los mismos pasos; siempre y cuando sea razonable asumir que los comportamientos anteriores de las series son útiles para predecir los valores futuros.

El rigor en la toma de datos, su consolidación y manejo adecuado son 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.

### Referencias

Box, G. E. P., & Pierce, D. A. (1970). Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *Journal of the American Statistical Association*. https://doi.org/10.1080/01621459.1970.10481180

Chambers, J., & Al, E. (2012). R-project. https://doi.org/10.1007/978-3-540-74686-7

Coleman, P., & Blankenship, R. (2017). What Spreadsheet and Database Skills Do Business Students Need?. *Journal of Instructional Pedagogies*, *19*, 1–8.

DANE. (2015). *Ficha Metodológica Encuesta Mensual de Comercio al por Menor y Comercio de Vehículos - EMCM*. Retrieved from https://www.dane.gov.co/index.php/estadisticas-por-tema/comercio-interno/encuesta-emcm#informacion-emcm-junio-2019

de Myttenaere, A., Golden, B., Le Grand, B., & Rossi, F. (2016). Mean Absolute Percentage Error for regression models. *Neurocomputing*, *192*, 38–48. https://doi.org/10.1016/j.neucom.2015.12.114

FedCSIS. (2016). Spreadsheet-based Business Process modeling. *2016 Federated Conference on Computer Science and Information Systems (FedCSIS), Computer Science and Information Systems (FedCSIS), 2016 Federated Conference On*, p. 1355. Polish Information Processing Society.

Hyndman, R. J., & Athanasopoulos, G. (2014). Forecasting : principles and practice. *Monash University*.

Hyndman, R., O’Hara-Wild, M., & Wang, E. (2019). *Fable R Package*. Retrieved from https://fable.tidyverts.org/index.html

Ljung, G. M., & Box, G. E. P. (1978). On a measure of lack of fit in time series models. *Biometrika*. https://doi.org/10.1093/biomet/65.2.297

Sadik, A. (2017). Students’ acceptance of file sharing systems as a tool for sharing course materials: The case of Google Drive. *Education and Information Technologies*. https://doi.org/10.1007/s10639-016-9556-z

Wang, E., Cook, D., & Hyndman, R. J. (2019). *A new tidy data structure to support exploration and modeling of temporal data*. (February). Retrieved from http://arxiv.org/abs/1901.10257

Wickham, H. (2011). ggplot2. *Wiley Interdisciplinary Reviews: Computational Statistics*. https://doi.org/10.1002/wics.147

Wickham, H. (2015). Tidy Data. *Journal of Statistical Software*. https://doi.org/10.18637/jss.v059.i10

Wickham, H., Francois, R., Henry, L., & Müller, K. (2019). Package ‘dplyr’. A Grammar of Data Manipulation. *R Package Version 0.8.0.1*.

Wickham, H., Henry, L., & RStudio. (2017). R: Package ‘tidyr.’ *Cran*.