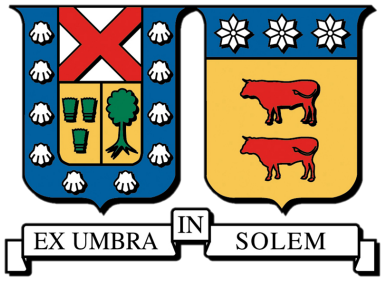

Universidad Técnica Federico Santa María

Departamento de Matemática

MAT-282

Laboratorio de Modelación I

Predicciones de combinación óptima para series temporales jerárquicas

Autores:
David Rivas
Rodrigo Serrano
Diego Vivallo

Profesor: Francisco Cuevas *Experto:* Sebastián Torres

Noviembre 2022

Índice

1. Motivación

En negocios y las empresas de Retail es muy común querer predecir la demanda de sus productos. Para esto, existen distintos métodos aplicables, algunos con un carácter más cualitativo, y otros con un carácter cuantitativo. En este documento se verá el caso aplicado del método de series temporales aplicado a la empresa mundialmente conocida Walmart para la predicción de las demandas de sus productos, los cuales han sido jerarquizados por distintos niveles.

2. Introducción al problema

Walmart es una corporación multinacional de tiendas de origen estadounidense, que opera cadenas de grandes almacenes a lo largo de varios continentes. Sin embargo, toda empresa quiere poder predecir sus demandas para ciertos productos. En este documento se propondrán dos métodos de series temporales para poder realizar dicha predicción.

Es común que las series de tiempo a pronosticar para el caso multidisciplinario tengan una estructura jerárquica, por ejemplo, una salsa de tomates de una marca particular pertenece a una categoría de salsas de tomate, y la categoría de salsas de tomate pertenece a un departamento de despensa, y el departamento de despensa al segmento de abarrotes y consumibles. Es por esto, que desde Walmart plantearon la necesidad de poder realizar una predicción de la demanda para los productos en particular, y también de forma general a la cadena completa.

Objetivos

- Con los modelos planteados, realizar una predicción usando los datos de entrenamiento para compararlos con los datos de validación.
- Realizar una comparación de ambos métodos y ver cual es más cercano a los datos de validación.
- Concretar un ajuste del método seleccionado en comparación al dato de validación.
- Realizar una predicción para los eventos futuros.

3. Herramientas matemáticas

Modelo autorregresivo integrado de media móvil (ARIMA)

El modelo ARIMA es una metodología basada en modelos dinámicos que utiliza datos de series temporales. Se trata de un modelo autorregresivo integrado de promedio móvil. En particular, un modelo estadístico que utiliza variaciones y regresiones de datos estadísticos con el fin de encontrar patrones para una predicción hacia el futuro. Las estimaciones futuras vienen explicadas por los datos del pasado y no por variables independientes.

Un modelo ARIMA estacional se forma incluyendo términos estacionales adicionales en los modelos ARIMA que hemos visto hasta ahora. Se escribe como sigue:

$$\text{ARIMA} \quad (p,d,q) \quad (P,D,Q)_m$$

donde (p,d,q) es la parte no estacional del modelo y $(P,D,Q)_m$ es la parte estacional, con m el periodo estacional (por ejemplo, el número de observaciones por año). Utilizamos las mayúsculas para las partes estacionales del modelo y las minúsculas para las partes no estacionales.

La parte estacional del modelo consiste en términos similares a los componentes no estacionales del modelo, pero que implican desplazamientos hacia atrás del periodo estacional. Por ejemplo, un modelo ARIMA$(1,1,1)(1,1,1)_4$ (sin constante) es para datos trimestrales.

Series temporales jerárquicas

En muchas aplicaciones, hay múltiples series de tiempo que están organizadas jerárquicamente y se pueden agregar en varios niveles diferentes de grupos basados en productos, geografía u otras características. En este documento proponemos un nuevo método estadístico para pronosticar series temporales jerárquicas que:

i. Brinda pronósticos puntuales que se concilian en todos los niveles de jerarquía.

ii. Permite las correlaciones e interacciones entre las series en cada nivel de la jerarquía.

iii. Proporciona estimaciones de la incertidumbre del pronóstico que se reconcilian a través de los niveles de la jerarquía.

iv. Es lo suficientemente flexible como para que se puedan incorporar ajustes ad hoc, además se puede permitir información sobre series individuales, y se pueden incluir covariables importantes.

Notación para el pronóstico jerárquico

Considere una jerarquía de varios niveles, donde el nivel 0 denota la serie completamente agregada, el nivel 1 el primer nivel de desagregación, hasta el nivel K que contiene la serie temporal más desagregada. Utilizamos una secuencia de letras para identificar las series individuales y el nivel de desagregación. Por ejemplo: A indica la serie A en el nivel 1; AF

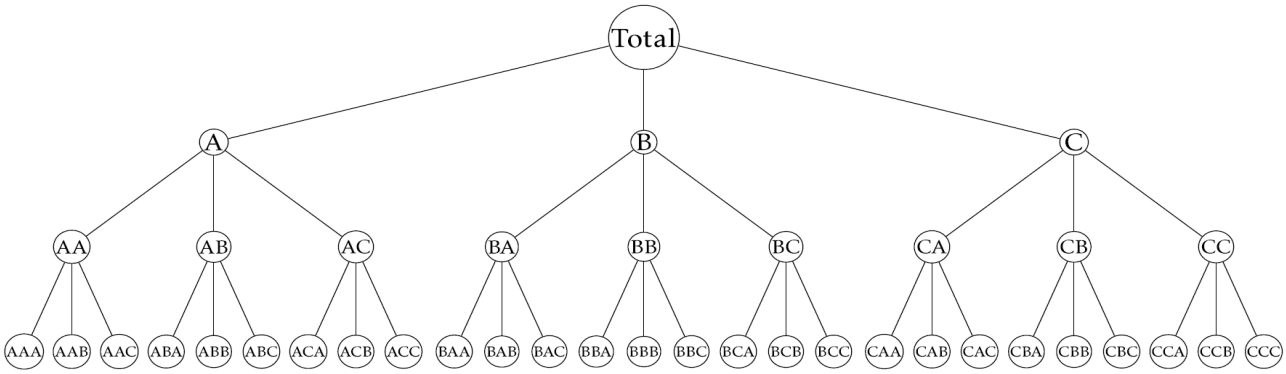

Figura 1: Un diagrama de árbol jerárquico de 3 niveles

indica la serie F en el nivel 2 dentro de la serie A en el nivel 1; AFC denota la serie C en el nivel 3 dentro de la serie AF en el nivel 2; y así sucesivamente.

En concreto, supongamos que tenemos tres niveles en la jerarquía y que cada grupo de cada nivel está formado por tres series. En este caso, $K = 3$ y la jerarquía tiene la estructura de árbol que se muestra en la Figura1.

Se supone que las observaciones se registran en tiempos $t = 1, 2, ..., n$, y que nos interesa pronosticar cada serie en cada nivel en tiempos $t = n + 1, n + 2, ..., n + h$. A veces será conveniente utilizar la notación X para referirse a una serie genérica dentro de la jerarquía. Las observaciones de la serie X se escriben como $Y_{X,t}$. Así, $Y_{AF,t}$ es el valor de la serie AF en el momento t. Utilizamos Y_t para el agregado de todas las series en el momento t. Por lo tanto

$$Y_t = \sum_i Y_{i,t}, \quad Y_{i,t} = \sum_j Y_{ij,t}, \quad Y_{ij,t} = \sum_k Y_{ijk,t}, \quad Y_{ijk,t} = \sum_\ell Y_{ijk\ell,t}$$

y así sucesivamente. En el caso de la Figura1, tenemos que $(i, j, k) \in \{A, B, C\}^3$ Así, las observaciones de los niveles superiores pueden obtenerse sumando las series inferiores.

Sea m_i el número total de series en el nivel i, $i = 0, 1, 2, ..., K$. Por tanto, $m_i > m_{i-1}$ y el número total de series en la jerarquía es $m = m_0 + m_1 + m_2 + ... + m_K$. En el ejemplo anterior, $m_i = 3^i$ y $m = 40$.

Será conveniente trabajar con expresiones matriciales y vectoriales. Dejamos que $Y_{i,t}$ denote el vector de todas las observaciones en el nivel i y el tiempo t y $Y_t = [Y_t, Y_{1,t}, ..., Y_{K,t}]^T$. Note que

$$Y_t = \boldsymbol{S} Y_{K,t} \tag{1}$$

donde $\boldsymbol{S}$ es una matriz "sumadora" de orden $m \times m_K$ utilizada para agregar las series de nivel más bajo. En el ejemplo anterior, $[Y_t, Y_{\mathrm{A},t}, Y_{\mathrm{B},t}, Y_{\mathrm{C},t}, Y_{\mathrm{AA},t}, Y_{\mathrm{AB},t}, \ldots, Y_{\mathrm{CC},t}, Y_{\mathrm{AAA},t}, Y_{\mathrm{AAB},t}, \ldots, Y_{\mathrm{CCC},t}]^T$

y la matriz suma es de orden 40×27 y viene dada por

$$\boldsymbol{S} = \begin{bmatrix} 111111111111111111111111111 \\ 111111111000000000000000000 \\ 000000000111111111000000000 \\ 000000000000000000111111111 \\ 111000000000000000000000000 \\ 000011100000000000000000000 \\ \vdots \\ 000000000000000000000000111 \\ 100000000000000000000000000 \\ 010000000000000000000000000 \\ \vdots \\ 000000000000000000000000001 \end{bmatrix}$$

Está claro que la matriz $\boldsymbol{S}$ se puede dividir por los niveles de la jerarquía y que $rg(\boldsymbol{S}) = m_K$. La fila superior es un vector unitario de longitud m_K y la parte inferior es una matriz de identidad $m_K \times m_K$. Las partes centrales de $\boldsymbol{S}$ son matrices rectangulares diagonales vectoriales.

Aunque la agregación es nuestro principal interés aquí, los resultados que siguen son lo suficientemente generales como para que $\boldsymbol{S}$ pueda ser un operador lineal general, y no sea necesario restringirlo a la agregación.

Predicción jerárquica general

Supongamos que primero calculamos las predicciones de cada serie en cada nivel, dando m predicciones base para cada uno de los periodos $n+1, ..., n+h$, a partir de la información disponible hasta el momento n inclusive. Denotamos estas predicciones base por $\hat{Y}_{X,n}(h)$, donde X denota la serie que se pronostica. Así pues, $\hat{Y}_n(h)$ denota la predicción base del total con h pasos de antelación, $\hat{Y}_{A,n}(h)$ denota la predicción de la serie A, $\hat{Y}_{AC,n}(h)$ denota la predicción de la serie AC, y así sucesivamente. Dejamos que $\hat{\boldsymbol{Y}}_n(h)$ sea el vector formado por estas predicciones base, apiladas en el mismo orden de las series que para Y_t.

Todos los métodos de predicciones jerárquicas existentes pueden entonces escribirse como

$$\tilde{\boldsymbol{Y}}_n(h) = \boldsymbol{S}\boldsymbol{P}\hat{\boldsymbol{Y}}_n(h) \tag{2}$$

para alguna matriz $\boldsymbol{P}$ adecuadamente elegida de orden $m_K \times m$. Es decir, los métodos existentes implican combinaciones lineales de las predicciones de base. Estas combinaciones lineales se "reconcilian" en el sentido de que las predicciones de nivel inferior se suman para dar las predicciones de nivel superior. El efecto de la matriz $\boldsymbol{P}$ es extraer y combinar los elementos relevantes de las predicciones base $\hat{\boldsymbol{Y}}_n(h)$, que se que se suman por $\boldsymbol{S}$ para obtener las predicciones jerárquicas finales, $\tilde{\boldsymbol{Y}}_n(h)$.

Por ejemplo, las predicciones ascendentes se obtienen utilizando

$$\boldsymbol{P} = \left[\boldsymbol{0}_{m_K \times (m-m_K)} \mid \boldsymbol{I}_{m_K}\right] \tag{3}$$

donde $\mathbf{0}_{\ell\times k}$ es una matriz nula de orden $\ell \times k$ e $\boldsymbol{I}_k$ es una matriz de identidad de orden $k \times k$. En este caso, la matriz $\boldsymbol{P}$ extrae sólo las predicciones de nivel inferior de $\hat{\boldsymbol{Y}}_n(h)$, que luego se suman por $\boldsymbol{S}$ para obtener las predicciones ascendentes.

Las predicciones de nivel superior se obtienen mediante

$$\boldsymbol{P} = \left[\boldsymbol{p} \mid \mathbf{0}_{m_K\times(m-1)}\right] \tag{4}$$

donde $\boldsymbol{p} = [p_1, p_2, \ldots, p_{m_K}]^T$ es un vector de proporciones que suman uno. El efecto de la matriz $\boldsymbol{P}$ aquí es distribuir la predicción del agregado a las series de menor nivel. Diferentes métodos de previsión descendente conducen a diferentes vectores de proporcionalidad $\boldsymbol{p}$.

Las variaciones, como las predicciones intermedias, son posibles si se define la matriz $\boldsymbol{P}$ de forma adecuada. Esto sugiere que otros (nuevos) métodos de predicciones jerárquicas pueden definirse eligiendo una matriz $\boldsymbol{P}$ diferente, siempre que pongamos algunas restricciones a $\boldsymbol{P}$ para obtener predicciones razonables.

Si suponemos que las predicciones base (independientes) son insesgadas (es decir, $E[\hat{\boldsymbol{Y}}_n(h)] = E[\boldsymbol{Y}_n(h)]$), y que queremos que las predicciones jerárquicas finales también sean insesgadas, entonces debemos exigir que $E[\tilde{\boldsymbol{Y}}_n(h)] = E\left[\boldsymbol{Y}_n(h)\right] = \boldsymbol{S}E\left[\boldsymbol{Y}_{K,n}(h)\right]$. Supongamos que $\boldsymbol{\beta}_n(h) = \mathrm{E}\left[\boldsymbol{Y}_{K,n+h} \mid \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n\right]$ es la media de los valores futuros del nivel inferior K. Entonces $E[\tilde{\boldsymbol{Y}}_n(h)] = \boldsymbol{SP}E[\hat{\boldsymbol{Y}}_n(h)] = \boldsymbol{SPS}\boldsymbol{\beta}_n(h)$. Por tanto, la insesgadez de la predicción final se mantendrá siempre que

$$\boldsymbol{SPS} = \boldsymbol{S} \tag{5}$$

Esta condición es cierta para el método ascendente con $\boldsymbol{P}$ dado por (3). Sin embargo, utilizando el método descendente con $\boldsymbol{P}$ dado por (4), encontramos que $\boldsymbol{SPS} \neq \boldsymbol{S}$ para cualquier elección de $\boldsymbol{p}$. Por tanto, el método descendente nunca puede dar predicciones insesgadas, aunque las predicciones base sean insesgadas.

Sea la varianza de las predicciones base, $\hat{\boldsymbol{Y}}_n(h)$, venga dada por $\boldsymbol{\Sigma}_h$. Entonces la varianza de las predicciones finales vienen dadas por

$$\mathrm{Var}[\tilde{\boldsymbol{Y}}_n(h)] = \boldsymbol{SP}\boldsymbol{\Sigma}_h\boldsymbol{P}^T\boldsymbol{S}^T \tag{6}$$

Así, se pueden obtener intervalos de predicción sobre las predicciones revisadas siempre que se pueda estimar $\boldsymbol{\Sigma}_h$ de forma fiable. Obsérvese que el resultado (6) se aplica a todos los métodos existentes que pueden expresarse como (2), incluidos los métodos bottom-up, top-down y middle-out.

Predicciones óptimas mediante regresión

Podemos escribir las predicciones de base como

$$\hat{\boldsymbol{Y}}_n(h) = \boldsymbol{S}\boldsymbol{\beta}_n(h) + \boldsymbol{\varepsilon}_h \tag{7}$$

donde $\boldsymbol{\beta}_n(h) = \mathrm{E}\left[\boldsymbol{Y}_{K,n+h} \mid \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n\right]$ es la media desconocida del nivel inferior K , y $\boldsymbol{\varepsilon}_h$ tiene media cero y matriz de covarianza $\mathrm{Var}(\boldsymbol{\varepsilon}_h) = \Sigma_h$. Esto sugiere que podemos estimar $\boldsymbol{\beta}_n(h)$ tratando (7) como una ecuación de regresión, y así obtener todos los niveles de

la jerarquía. Si se conociera $\mathbf{\Sigma}_h$, podríamos utilizar la estimación por mínimos cuadrados generalizado para obtener la estimación insesgada de varianza mínima de $\boldsymbol{\beta}_n(h)$ como

$$\hat{\boldsymbol{\beta}}_n(h) = \left(\boldsymbol{S}^T\mathbf{\Sigma}_h^{\dagger}\boldsymbol{S}\right)^{-1}\boldsymbol{S}^T\mathbf{\Sigma}_h^{\dagger}\hat{\boldsymbol{Y}}_n(h) \tag{8}$$

donde $\mathbf{\Sigma}_h^{\dagger}$ es la inversa generalizada de Moore-Penrose de $\mathbf{\Sigma}_h$. Utilizamos una inversa generalizada porque $\mathbf{\Sigma}_h$ es a menudo (casi) singular debido a la agregación que implica $\boldsymbol{Y}_n$. Esto nos lleva a las siguientes predicciones finales

$$\tilde{\boldsymbol{Y}}_n(h) = \boldsymbol{S}\hat{\boldsymbol{\beta}}_n(h) = \boldsymbol{SP}\hat{\boldsymbol{Y}}_n(h)$$

donde $\boldsymbol{P} = (\boldsymbol{S}^T\mathbf{\Sigma}_h^{\dagger}\boldsymbol{S})^{-1}\boldsymbol{S}^T\mathbf{\Sigma}_h^{\dagger}$. Claramente, esto satisface la propiedad de insesgadez (5). La varianza de estas predicciones viene dada por

$$\text{Var}[\tilde{\boldsymbol{Y}}_n(h)] = \boldsymbol{S}(\boldsymbol{S}^T\mathbf{\Sigma}_h^{\dagger}\boldsymbol{S})^{-1}\boldsymbol{S}^T$$

La dificultad de este método es que requiere conocer $\mathbf{\Sigma}_h$, o al menos una buena estimación de la misma. En una jerarquía grande con miles de series, esto puede no ser posible.

Sin embargo, podemos simplificar mucho los cálculos suponiendo que el error en (7) puede expresarse como $\boldsymbol{\varepsilon}_h \approx \boldsymbol{S}\boldsymbol{\varepsilon}_{K,h}$, donde $\boldsymbol{\varepsilon}_{K,h}$ es el error de predicción en el nivel inferior. Es decir, suponemos que los errores satisfacen la misma restricción de agregación que los datos originales (1). Esta suposición será cierta siempre que las predicciones también satisfagan aproximadamente esta restricción de agregación lo que debería ocurrir para cualquier conjunto razonable de predicciones. En algunos casos, es posible que los errores satisfagan la expresión exactamente. Por ejemplo, si se utiliza un método de previsión lineal (por ejemplo, un modelo **ARIMA**) con parámetros fijos para todas las series, entonces los errores serían exactamente aditivos de esta manera. Por lo tanto, la suposición es una aproximación razonable de lo que sucede en la práctica.

La aproximación conduce al resultado $\mathbf{\Sigma}_h \approx \boldsymbol{S}\mathbf{\Omega}_h\boldsymbol{S}^T$, donde $\mathbf{\Omega}_h = \text{Var}\left(\boldsymbol{\varepsilon}_{K,h}\right)$. Ahora estamos preparados para enunciar nuestro resultado principal.

Teorema

Sea $\boldsymbol{Y} = \boldsymbol{S}\boldsymbol{\beta}_h + \varepsilon$ con $\text{Var}(\varepsilon) = \mathbf{\Sigma}_h = \boldsymbol{S}\mathbf{\Omega}_h\boldsymbol{S}^T$ y $\boldsymbol{S}$ una matriz "sumadora". Entonces la estimación de $\boldsymbol{\beta}$ por mínimos cuadrados generalizados, obtenida mediante la inversa generalizada de Moore-Penrose, es independiente de $\mathbf{\Omega}_h$:

$$\hat{\boldsymbol{\beta}}_h = (\boldsymbol{S}^T\mathbf{\Sigma}_h^{\dagger}\boldsymbol{S})^{-1}\boldsymbol{S}^T\mathbf{\Sigma}_h^{\dagger}\boldsymbol{Y} = \left(\boldsymbol{S}^T\boldsymbol{S}\right)^{-1}\boldsymbol{S}^T\boldsymbol{Y}$$

con la matriz de varianza $\text{Var}(\hat{\boldsymbol{\beta}}) = \mathbf{\Omega}_h$. Además, esta es la estimación lineal insesgada de mínima varianza.

Demostración: Sea $\mathbf{\Sigma}_h = \boldsymbol{BC}$, donde $\boldsymbol{B} = \boldsymbol{S}\mathbf{\Omega}_h$ y $\boldsymbol{C} = \boldsymbol{S^T}$, entonces por el hecho 6.4.8 de **Bernstein** (2005, p.216-217), la inversa generalizada de Moore-Penrose de $\mathbf{\Sigma}_h$ es

$$\mathbf{\Sigma}_h^{\dagger} = \boldsymbol{C}^T(\boldsymbol{CC^T})^{-1}\left(\boldsymbol{B}^T\boldsymbol{B}\right)^{-1}\boldsymbol{B}^T = \boldsymbol{S}\left(\boldsymbol{S}^T\boldsymbol{S}\right)^{-1}\left(\mathbf{\Omega}_h^T\boldsymbol{S}^T\boldsymbol{S}\mathbf{\Omega}_h\right)^{-1}\mathbf{\Omega}_h^T\boldsymbol{S}^T \tag{9}$$

luego $(\boldsymbol{S}^T\boldsymbol{\Sigma}_h^\dagger\boldsymbol{S})^{-1}\boldsymbol{S}^T\boldsymbol{\Sigma}_{\mathrm{h}}^\dagger = (\boldsymbol{S}^T\boldsymbol{S})^{-1}\boldsymbol{S}^T$. La varianza se obtiene sustituyendo (9) en $(\boldsymbol{S}^T\boldsymbol{\Sigma}_h^\dagger\boldsymbol{S})^{-1}$. Por otro lado, **Tian y Wiens** (2006, Teorema 3) mostraron que el estimador GLS será el estimador insesgado de varianza mínima si y sólo si

$$\boldsymbol{S}\boldsymbol{S}^\dagger\boldsymbol{\Sigma}_h^\dagger\boldsymbol{\Sigma}_h\left(I - \boldsymbol{S}\boldsymbol{S}^\dagger\right) = \mathbf{0}$$

donde $\boldsymbol{S}^\dagger = \left(\boldsymbol{S}^T\boldsymbol{S}\right)^{-1}\boldsymbol{S}^T$. Utilizando (9), se puede demostrar que esta condición se cumple.■

4. Análisis exploratorio de los datos

Explorando 2 archivos .csv (un archivo con las fechas, ventas, categorías y tiendas y otro con tiendas, segmentos y departamentos, con fechas correspondientes a los 722 días hábiles entre 2015 y 2016), se comenzó uniendo los datos de tal manera de respetar una jerarquía, luego al manejar los datos, existían series de categorías que carecían de fechas, por lo que se decidió eliminar las que les faltaban más del 20 % de días, resultando ser 1.977 series y por otro lado se rellenaron con 0 a las series que les faltaban días (esto tiene como interpretación que la categoría no vendió nada ese día).

Luego por temas de tiempo de ejecución se tomaron el 20 % de las series que más vendían, y además solo se trabajó con la serie total y con las tiendas A y B.

5. Factibilidad de resultados

Tiempos de ejecución:

Al ejecutar el código de **SARIMA** y **combinación óptima**, se obtuvo un tiempo de respuesta de $6638,39$ segundos los cuales son aproximadamente 110 minutos. Donde un gran volumen del tiempo fue destinado a la ejecución del método **SARIMA**.

Resultados SARIMA:

Para la serie total el mejor modelo fue ARIMA$(1,0,2)(0,1,0)$, para la serie de la tienda A fue ARIMA$(1,0,3)(0,1,0)$ y para la tienda B fue de ARIMA$(0,1,1)(0,1,0)$, con $m = 362$, que experimentalmente fue el que dió los mejores resultados.

Matriz Sumadora

$$\boldsymbol{S} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Resultados Obtenidos

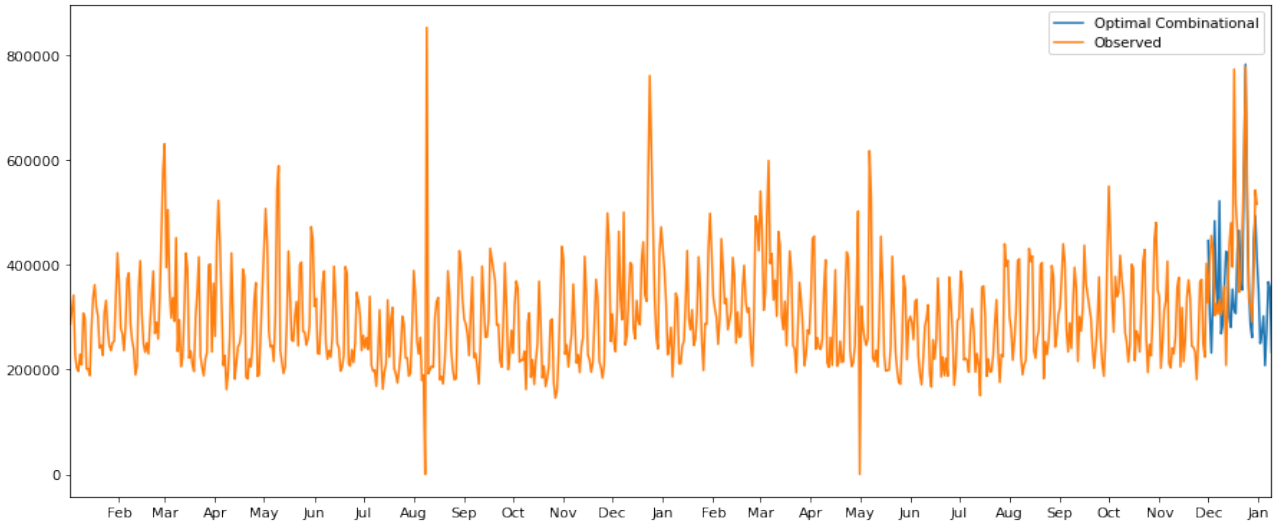

Figura 2: Ventas totales 2015-2016

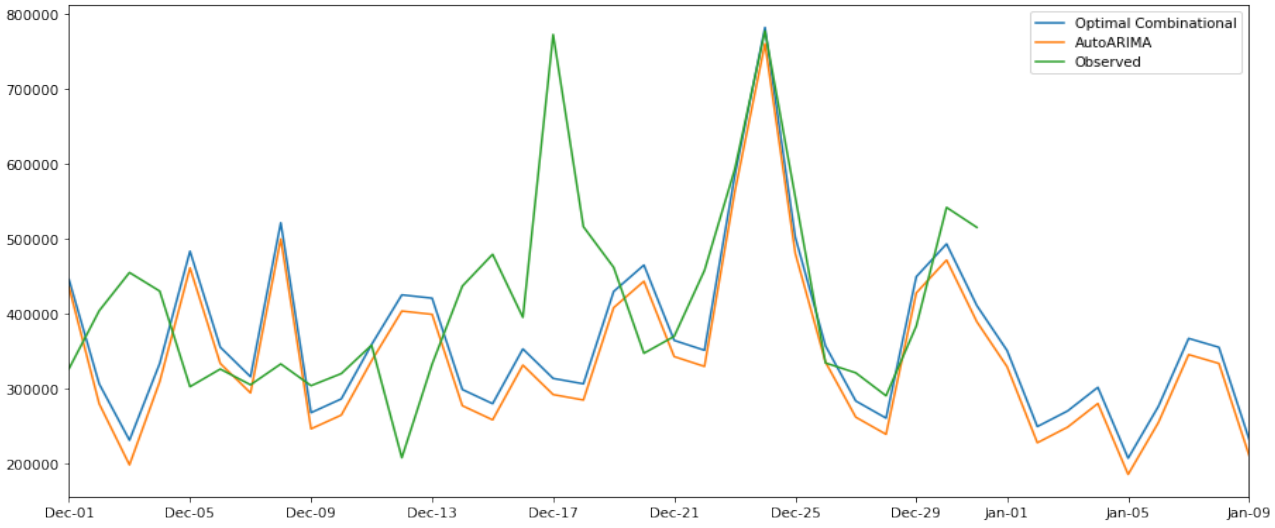

Figura 3: Ventas totales December 2016 and prediction

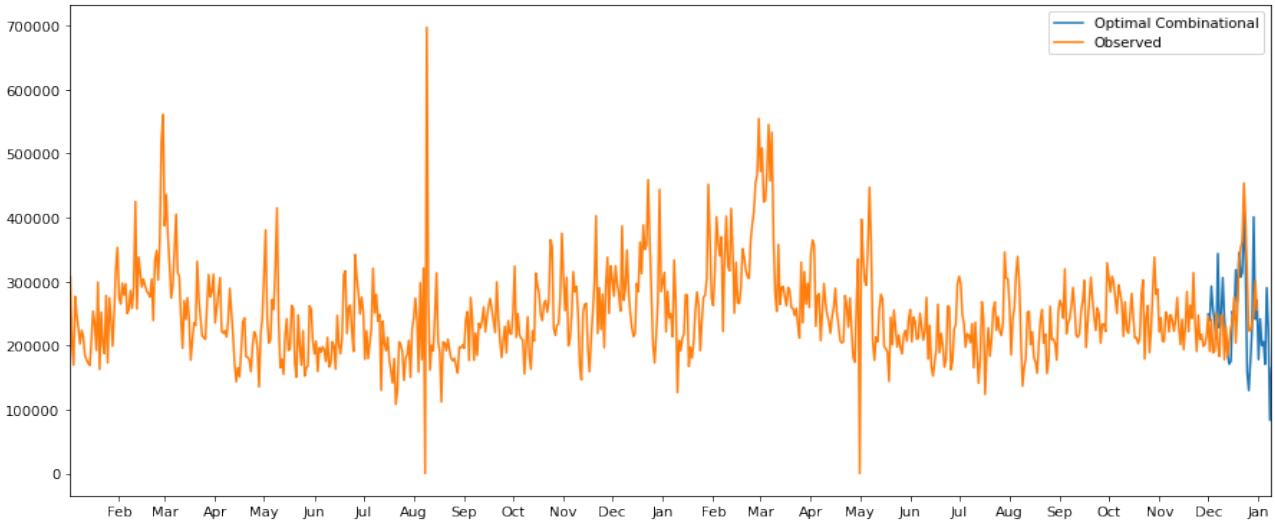

Figura 4: Ventas tienda A 2015-2016

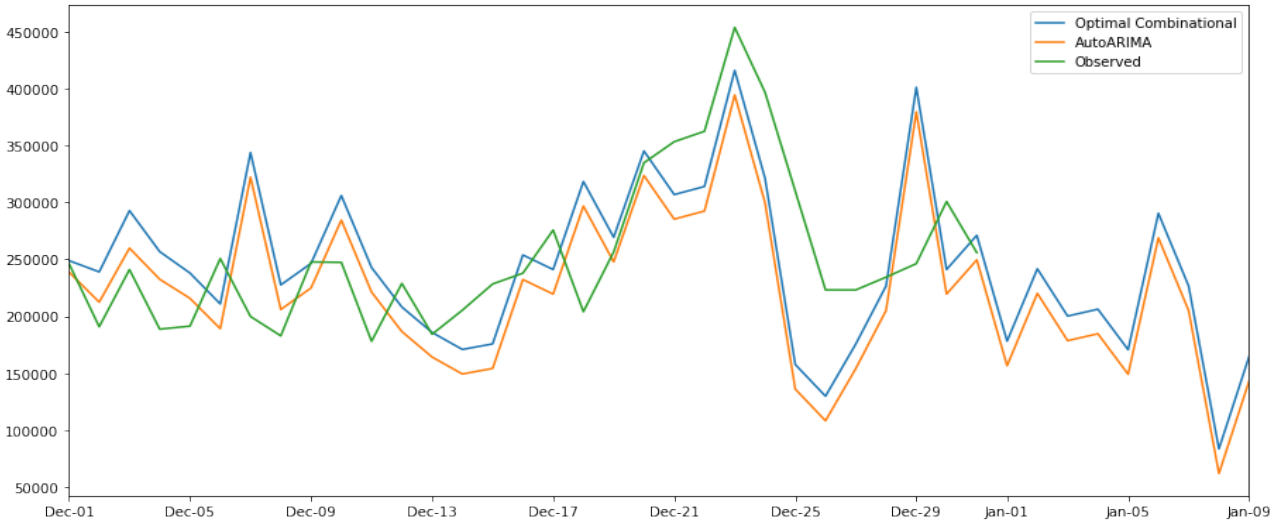

Figura 5: Ventas tienda A Diciembre 2016 y predicción

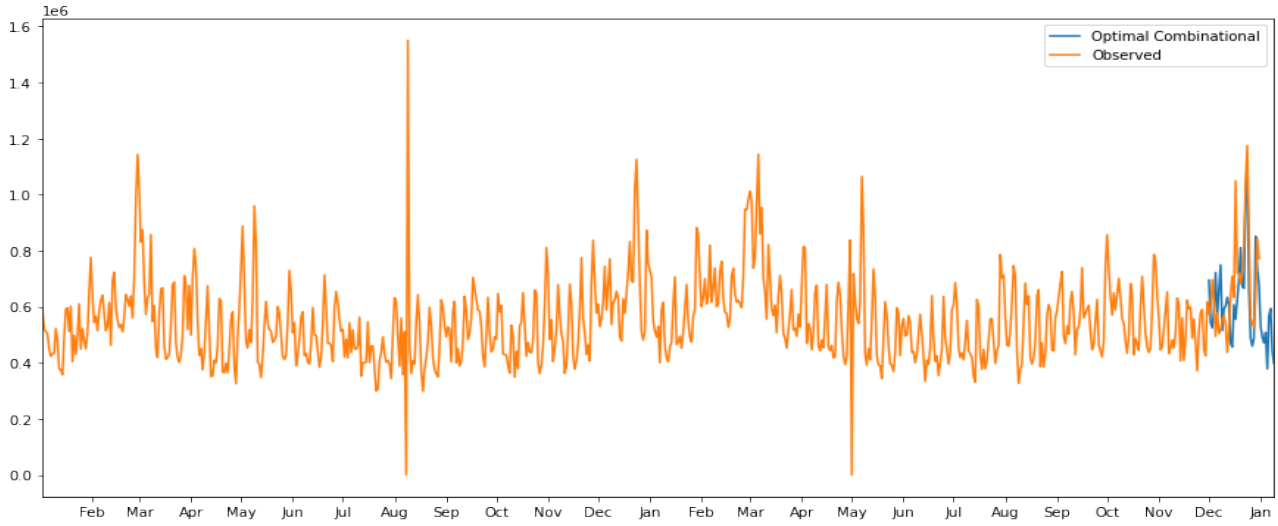

Figura 6: Ventas tienda B 2015-2016

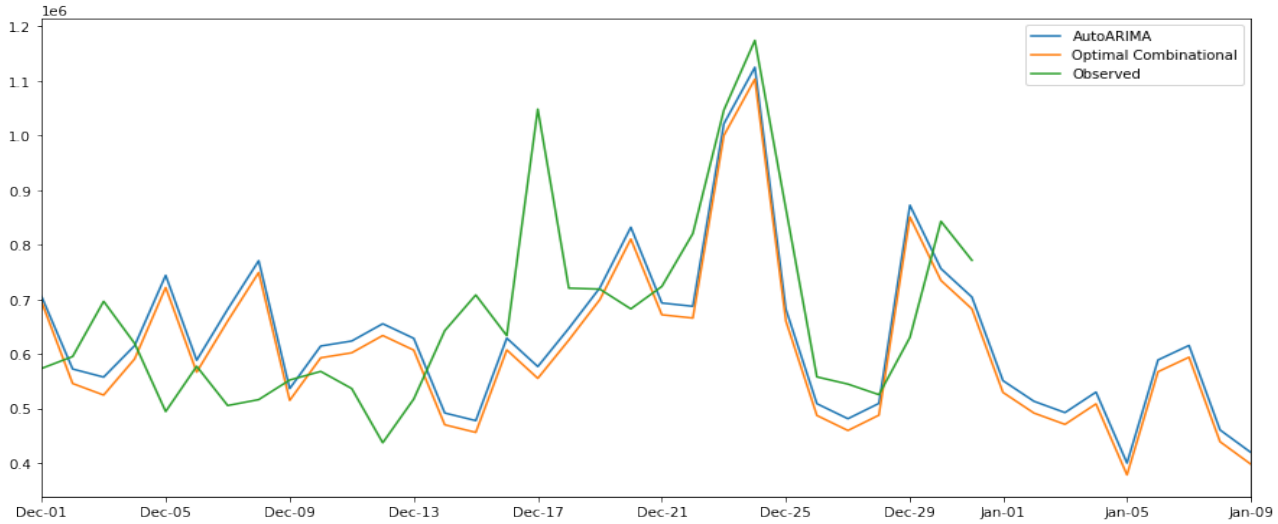

Figura 7: Ventas tienda B Diciembre 2016 y predicción

Análisis del error

Para el análisis del error se utilizó *Root Mean Squared Error* (RMSE), el cual se calcula mediante la fórmula:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{10}$$

Que al aplicar al modelo, arrojó los siguientes resultados:

RMSE	total	A	B
SARIMA	147690.35	62762	77333,25
Combinación optima	150526,9	84347,41	55747,8

Referencias

[1] Rob J.Hyndman, Roman A.Ahmed, George Athanasopoulos and Han LinShang, *«Optimal combination forecasts for hierarchical time series»*, *Computational Statistics Data Analysis*, Volumen 55, Tema 9, 2001, Páginas 2579-2589.

[2] Bernstein, D.S., *«Matrix Mathematics: Theory, Facts, and Formulas with Applications to Linear Systems Theory»*, *Princeton University Press, Princeton, Nueva Jersey.*, 2005.

[3] Tian, Y., Wiens, D.P., *«On equality and proportionality of ordinary least squares, weighted least squares and best linear unbiased estimators in the general linear model. Statistics and Probability Letters 76 (12)»*, 2006, Páginas 1265-1722.