



Predecir repuntes en el mercado para la industria de boletos en los Estados Unidos utilizando técnicas modernas de ingeniería y análisis de datos

Máster Universitario de Ingeniería de Sistemas de Decisión

Diego Iglesias Bayo

Dirigido por:

- Ana Elizabeth García Sipols
- Clara Simón De Blas
- Felipe Ortega



Universidad
Rey Juan Carlos

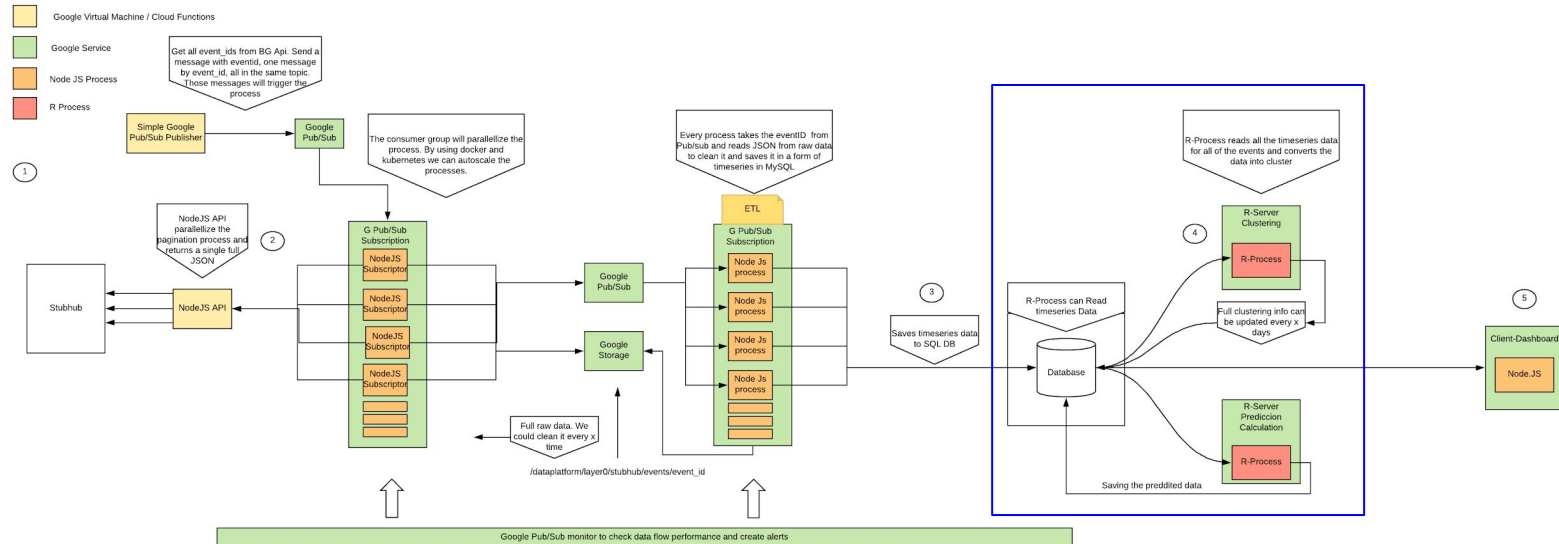
Introducción

- En EEUU existe un mercado secundario de entradas para eventos socioculturales.
- Revendedores tienen necesidad de contar con información en tiempo real.
- Broker Genius Inc ofrece una plataforma de información centralizada de plataformas de venta de entradas (ejemplo: Stubhub).
- Si los usuarios de la plataforma de Broker Genius pueden contar con una estimación del precio futuro de las entradas de sus eventos, podrán tomar decisiones más óptimas.
- División de los eventos por:
 - Zonas
 - Sección

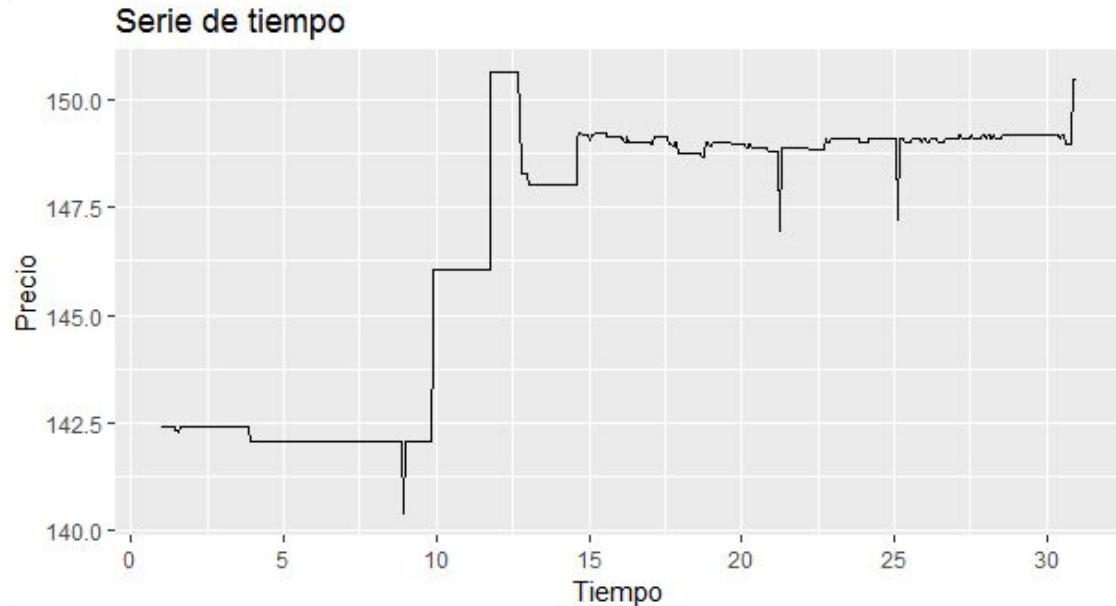


Planteamiento: Arquitectura de implementación

Uptick Price Prediction, Data Pipeline and Architecture



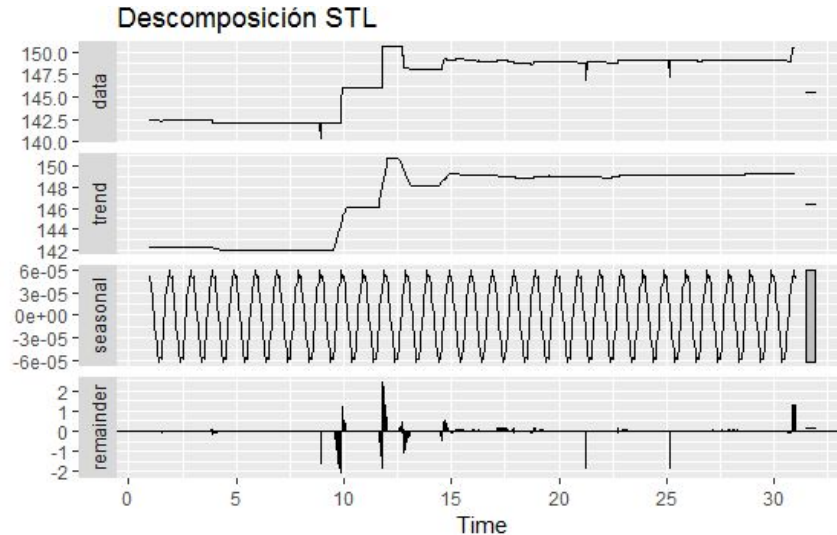
Metodología utilizada: Series Temporales



Descomposición de las series temporales:

Para el análisis de las series temporales se utilizaron diferentes metodologías de descomposición de series temporales:

- Descomposición clásica
- Descomposición STL



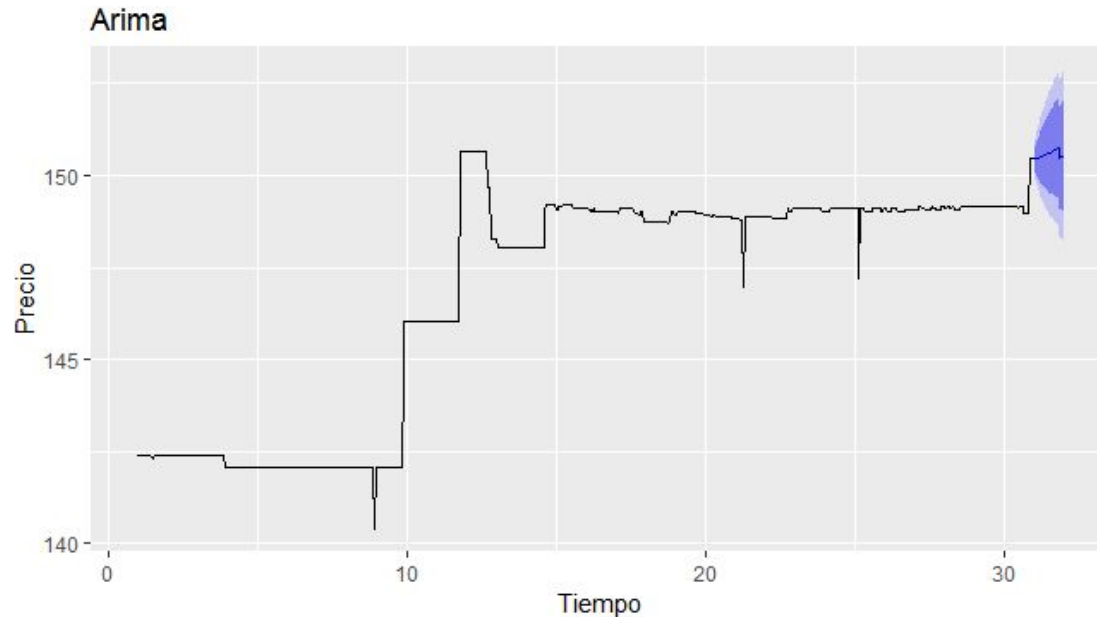


Métodos predictivos utilizados

Para el pronóstico de las series de tiempo se utilizaron diferentes metodologías con el fin de seleccionar aquella que mejor se ajuste a los datos:

- Método naïve
- Método de alisado exponencial simple
- Método de alisado exponencial con tendencia suavizada
- Método ARIMA
- Método KNN
- Método redes neuronales autorregresivas

Ejemplo de pronóstico:





Medidas de precisión

El error en la estimación viene definido por la diferencia entre el valor observado y el valor pronosticado:

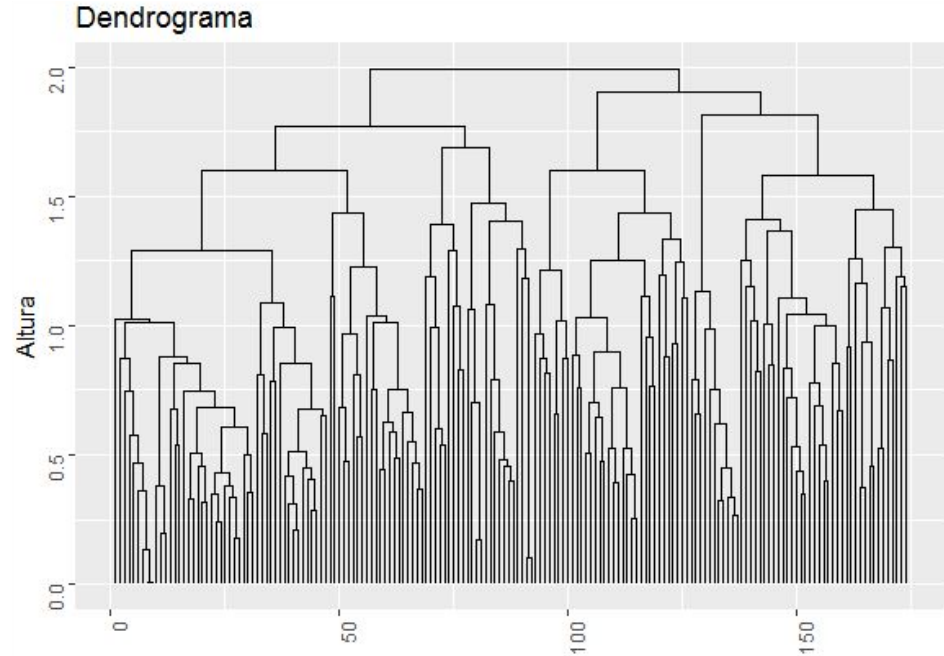
$$e_t = y_{t+h} - \hat{y}_{t+h}$$

Para la medición de la precisión de los pronósticos, se utilizaron dos medidas:

- **MSE** (Mean Square Error): $\text{promedio}(e_t^2)$
- **MAE** (Mean Absolute Error): $\text{promedio}(|e_t|)$

Método de clustering utilizado

Para optimizar el proceso de pronóstico, y que este se realice de una manera más precisa para cada serie de tiempo aplicando el método de pronóstico más adecuado, se realiza un proceso de clustering jerárquico con la metodología de *agglomerative nesting*.





Medidas de distancia

Se consideraron las siguientes medidas de distancia entre series de tiempo:

- Distancia euclídea: $d_{euc}(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \dots + (p_n - q_n)^2}$
- Distancia basada en correlación: $d_{cor} = \sqrt{2 \times (1 - cor(p, q))}$
- Distancia basada en Dynamic Time Warping



Selección del número óptimo de clusters

Para la selección del número óptimo de clusters se consideraron dos métodos:

- Método Elbow
- Método Average Silhouette

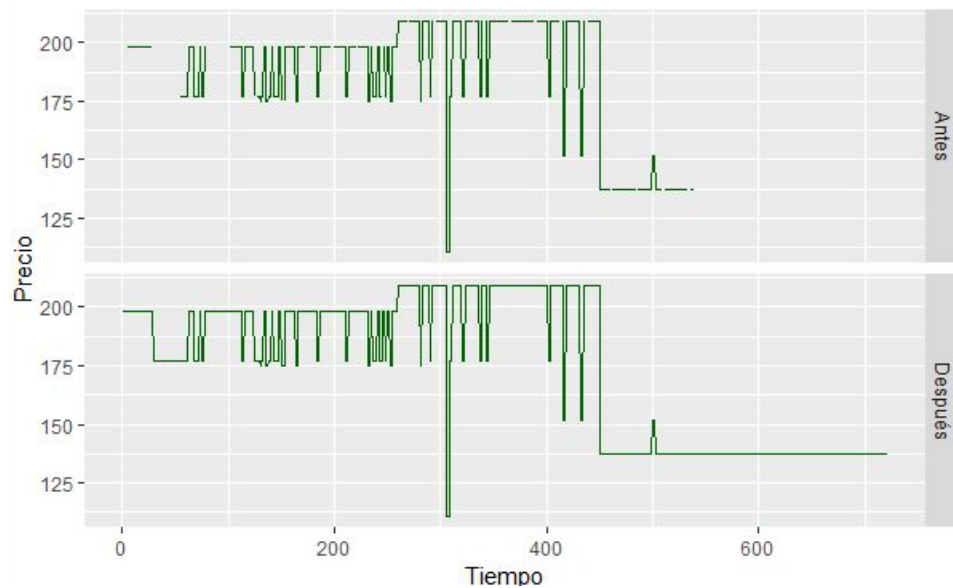


Presentación de los datos

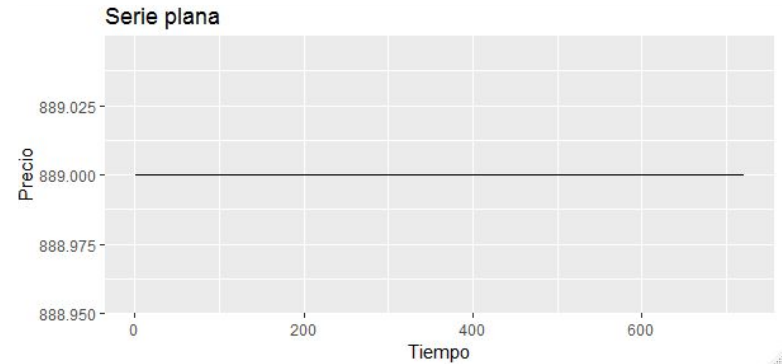
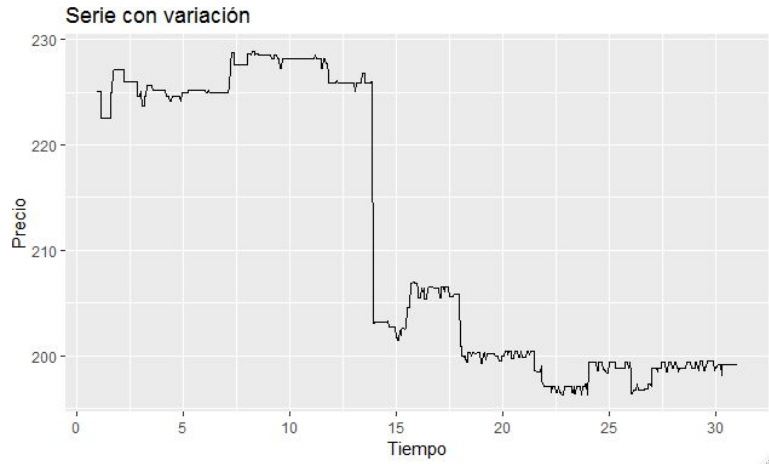
Para efectos de esta TFM se contó con una muestra de:


- 184 series temporales correspondientes a 12 zonas y 172 secciones de un solo evento.
- Cada serie cuenta con 721 observaciones que se corresponden a una hora del día. En total son 31 días.
- Cada observación se corresponde con el precio promedio de las entradas restantes (el precio de las entradas ya vendidas no se considera) para cada sección y zona.

En la extracción de la información, se presentan valores omisos, por lo que se estiman para el proceso de pronóstico

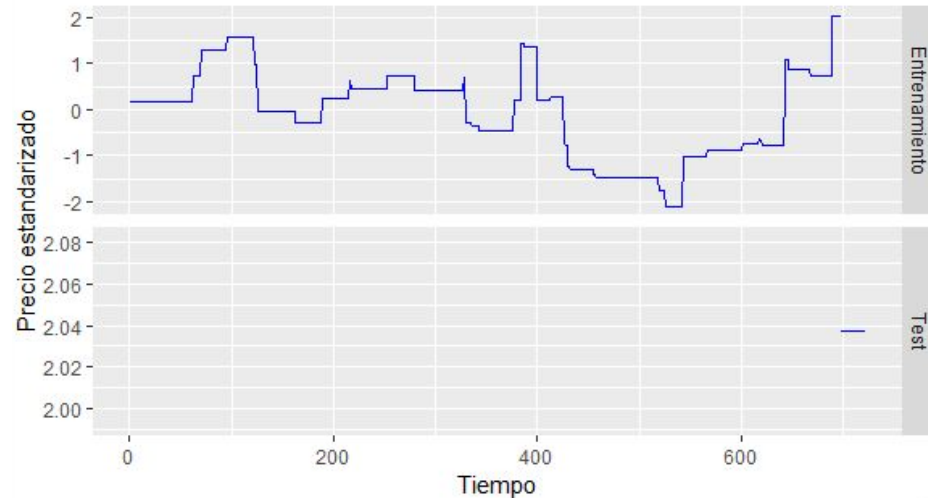



Se dividen las series temporales en aquellas que tienen variación (desviación estándar > 0) y las que no tienen variación (desviación estándar $= 0$). Las series sin variación con separadas del posterior análisis, y para efectos de predicción se aplica el método naïve.





Finalmente se dividen las series temporales en muestra de entrenamiento y muestra de validación. Para la muestra de validación se toman las últimas 24 observaciones (24 horas) dado que el horizonte de pronóstico para la plataforma de Broker Genius se fija en un día.

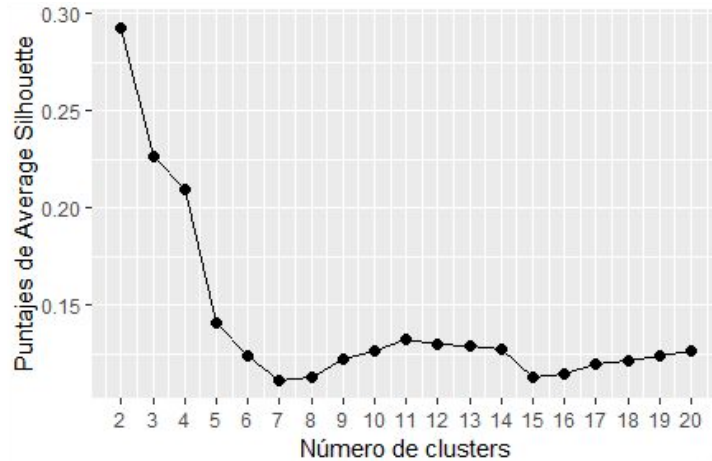




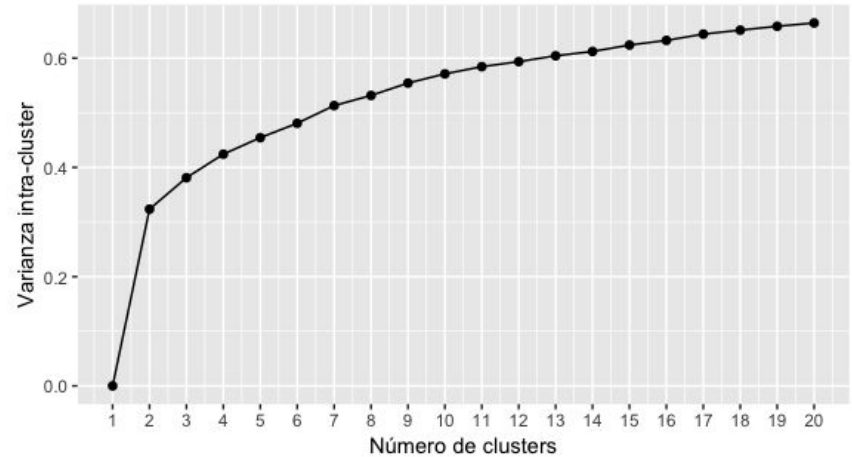
Para el proceso de clustering, se consideran 3 medidas de distancia. Finalmente se selecciona la medida basada en correlación ya que tiene un costo computacional similar a la distancia euclídea y muy inferior a la distancia basada en Dynamic Time Warping, y porque se considera más apropiada para las series temporales que la distancia euclídea.

Distancia Euclídea	1.000000
Distancia Correlación	1.236901
Distancia Dynamic Time Warping	950.811490

Para la selección del número óptimo de clusters el método Average Silhouette propone 2 particiones de los datos (máxima media de silhouettes) y a través del método Elbow se consideran 11 clusters (decremento inferior al 1% de la varianza intra-cluster total).



Silhouette




Elbow



A continuación, se detalla la aplicación de cada modelo y la selección de parámetros escogidos:

- El **método Naïve** se aplica de **manera individual** a todas las series temporales en cada clúster.
- El **método Alisado Exponencial Simple** también se aplica de **manera individual** a todas las series temporales. Además, para cada serie, el algoritmo que implementa el alisado selecciona I y α a través de **optimización no lineal**.
- El **método de Alisado Exponencial con tendencia suavizada** se aplica de **manera individual** a todas las series temporales. Además, para cada serie, el algoritmo que implementa el alisado selecciona I , α , β y ϕ a través de optimización no lineal.
- El método **Arima** se aplica, en **primer lugar**, de **manera individual** a todas las series temporales. En **segundo lugar**, se aplica **escogiendo al azar una serie de tiempo dentro del clúster, entrenando un modelo con esa serie de tiempo, y aplicando el modelo** entrenado a **todas** las series temporales del **clúster**. El algoritmo de Arima considera primero aplicar diferenciación a la serie de tiempo en función de un test de raíz unitaria, y los órdenes de AR y MA en función del que tenga el menor AICc. Cabe mencionar, que no se realiza una validación de que los residuos no estén auto-correlacionados y se comporten como ruido blanco.

- 
- El **método KNN** se aplica dividiendo las series temporales en paquetes de **25** observaciones consecutivas. La observación más reciente funge como variable dependiente (y) y el resto de las observaciones fungen como variables independientes (x).
 - El **método ARNN** sigue un **esquema similar** al segundo **modelo de Arima**, seleccionando una de las series temporales para cada clúster, entrenando el modelo con esa serie, y aplicando ese modelo al resto de series de mismo clúster con el fin de obtener una predicción. La razón de implementar el modelo de redes neuronales artificiales autorregresivas de esta manera se debe también al hecho de que, dado que el objetivo es evaluar un producto que se comercialice, la necesidad de tiempo computacional para el cálculo individual de las ARNNs a las series temporales no haría posible la entrega del producto en un tiempo razonable.



Resultados

Para cada cluster, se selecciona el modelo que menor MSE resulta:

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10	Cluster 11
Método Naive	4.241217	10.36390	0.7370843	23.73745	0.5274337	0.6896258	27.79959	0.6149385	2.7395181	4.948318	0.5875755
Método Alisado Exponencial Simple	4.245279	10.24117	0.6570326	24.67596	0.5440777	0.6896361	27.69714	0.6104647	2.7201267	4.888415	0.4629017
Método Holt Damped	4.243603	10.22957	0.6542543	24.69941	0.5442201	0.6896171	27.69889	0.6105250	2.7293867	4.888821	0.4624285
Método Arima	4.307043	10.13553	0.5306384	24.48166	0.5571464	0.3638251	26.43569	0.5681920	0.8321867	4.584828	0.3863961
Método Arima con Selección	4.241217	10.36390	1.0086612	24.51391	0.5274337	2.4314685	28.51567	0.6159708	2.7395181	4.872913	0.3744578
Método KNN	6.519901	25.83200	0.6520420	21.80329	3.1124393	0.7082422	33.32174	1.6156095	2.8625751	5.436640	0.5659239
Método ARNN	13.723443	39.61503	3.8249203	33.98640	388.0754537	2.7381724	104.00887	3.7939437	5.3650900	13.458609	14.4367901



Para MAE los resultados son:

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10	Cluster 11
Método Naive	4.448601	5.423058	0.9898015	6.031241	0.5408376	0.5660366	6.149756	1.041217	1.701399	3.362119	0.8846209
Método Alisado Exponencial Simple	4.497717	5.420082	0.8820891	6.564478	0.5541675	0.5660843	6.112117	1.034925	1.695896	3.338979	0.7392350
Método Holt Damped	4.534194	5.419400	0.8788398	6.699516	0.5543243	0.5659942	6.116068	1.035479	1.699208	3.339243	0.7389744
Método Arima	4.928838	5.547104	1.0975294	6.578246	0.6763309	0.6584683	6.686067	1.057842	1.109144	3.374444	0.8022289
Método Arima con Selección	4.448601	5.423058	2.2253131	6.550936	0.5408376	2.7357558	6.918726	1.042677	1.701399	3.332130	0.8852800
Método KNN	8.368127	12.105063	1.1759485	7.443537	2.2460646	0.6200676	11.462530	1.815421	2.478851	4.004127	1.0228924
Método ARNN	17.656306	24.705883	4.9837017	15.028681	37.4769860	2.7513963	35.967859	4.246473	4.695558	9.627942	10.8853072



MSE promedio

Método Naive	10.049631
Método Alisado Exponencial Simple	10.116105
Método Holt Damped	10.117410
Método Arima	9.793494
Método Arima con Selección	10.290225
Método KNN	13.475379
Método ARNN	46.502798

MAE promedio

Método Naive	4.008802
Método Alisado Exponencial Simple	4.063237
Método Holt Damped	4.089842
Método Arima	4.287827
Método Arima con Selección	4.312971
Método KNN	7.105522
Método ARNN	18.682552

Costo computacional

Método Naive	1.000000
Método Alisado Exponencial Simple	1.695508
Método Holt Damped	7.014285
Método Arima	763.141101
Método Arima con Selección	38.572306
Método KNN	4.002029
Método ARNN	2205.278900

