Universidad de Lima

Facultad de Ingeniería y Arquitectura

Carrera de Ingeniería de Sistemas

PREDICCIÓN DE LA DEMANDA DE PASAJEROS A CLÚSTERES DE ESTACIONES DEL METROPOLITANO USANDO MÉTODOS DE DATA MINING, LA METODOLOGÍA BOX-JENKINS Y SARIMA

Tesis para optar Título Profesional de Ingeniería de Sistemas

Edwin Roque Rojas

Código 20152307

Asesor

José Antonio Cárdenas Garro

Lima – Perú

Junio de 2022

Predicción de la demanda de pasajeros en clústeres de estaciones del Metropolitano usando métodos de Data Mining, la Metodología Box-Jenkins y SARIMA

Edwin Roque Rojas
20152307@aloe.ulima.edu.pe
Universidad de Lima

Resumen: El nivel de demanda de pasajeros del servicio del Metropolitano ha aumentado y la planificación llamada JICA, actualmente utilizada, no es suficiente, causando saturación de pasajeros en sus 38 estaciones. Según expertos e informes realizados por la Municipalidad Metropolitana de Lima y ProTransporte en el 2018, afirman que se sobrepasó la capacidad máxima de estaciones de 700 mil pasajeros diariamente que se planificó, siendo el doble que el año 2010 y sugieren actualizar la planificación de demanda. Por lo que se propuso predecir la demanda de pasajeros en clústeres de estaciones usando SARIMA a partir de un análisis espacio-temporal usando dos métodos de data mining y la metodología Box-Jenkins para obtener el mejor modelo por clúster. Los resultados del análisis espacio-temporal mostraron un comportamiento similar entre estaciones al agruparlos en clústeres con estacionalidad semanal. Los modelos no realizaron una predicción correcta para los días festivos anuales, ya que fueron interpretados como valores outliers, por lo que se reemplazó la demanda que registraron estas fechas para que los modelos fueran más precisos; obteniendo finalmente buenos resultados con un RMSPE, MAPE y R^2 entre un 6.37% - 8.13%, 4.19% - 5.93% y 0.91 - 0.98 respectivamente entre los modelos, estando por debajo del límite máximo en cada métrica de pronóstico que se propusieron como objetivos. A pesar del problema, las predicciones de los modelos pueden ser usados para optimizar los recursos del Metropolitano en la distribución de sus buses atendiendo adecuadamente la demanda que satura sus estaciones, sin contar los días festivos anuales.

Palabras Clave: demanda de pasajeros, estación, Metropolitano, minería de datos, patrones espacio-temporales, Box-Jenkins, modelo SARIMA.

Abstract: The level of passenger demand for the Metropolitan service has increased and the planning called JICA, currently used, is not enough, causing the saturation of passengers in their 38 stations. According to experts and reports made by the Metropolitan Municipality of Lima and ProTransporte in 2018, claim that the maximum station capacity of 700,000 passengers was exceeded daily, which was planned, being twice as much as 2010 and suggesting updating demand planning. So, it was proposed to predict the passenger demand of station clusters using SARIMA from a spatio-temporal analysis using two data mining methods and the Box-Jenkins methodology to get the best possible model for cluster. The results of the spatio-temporal analysis showed similar behavior between stations when grouped into clusters with weekly seasonality. The models didn't make a correct prediction for the annual holidays, as they were interpreted as outlier´s values, so the demand that recorded these dates was replaced to make the models more accurate; finally getting good results with a RMSPE, MAPE and R^2 between 6.37% - 8.13%, 4.19% - 5.93% y 0.91 - 0.98 respectively between the four models, below the ceiling for each forecast metric that was proposed as targets. Despite the problem, model predictions can be used to optimize the Metropolitan's resources in the distribution of its buses, adequately taking care of the demand that saturates its stations, not counting the annual holidays.

Keywords: passenger demand, station, Metropolitano, data mining, space-time patterns, Box-Jenkins, SARIMA model

1. INTRODUCCIÓN

Según ProTransporte (2018), el servicio actual del Metropolitano atiende actualmente una demanda creciente de forma diaria de 700 mil viajes en las rutas troncales y alimentadoras, siendo el doble que el 2010 cuando empezó a operar el servicio. Este servicio de transporte tan solo cuenta con la información de horarios de atención de buses y estaciones. Según la Municipalidad Metropolitana de Lima (MML) (2017), en un día de semana, sábados y domingos fue de 437148, 313644 y 161477 pasajeros respectivamente, en estos tres casos, el número de pasajeros fue mayor que el año anterior en un 28,2%, 29,1% y 33,3% respectivamente; además, menciona que la flota de buses fue aumentando en 74 unidades, pero esto no ha mejorado la atención a los pasajeros y como consecuencia, en una encuesta realizada por la MML, indicó que un 70% la población piensa que el sistema de transporte está igual o peor cada año. ProTransporte (2018), menciona que actualmente de forma diaria se realizan más de 700 mil viajes en las rutas troncales y alimentadoras, el doble desde 2010, cuando empezó a operar el servicio. Esta problemática del transporte público es una línea de investigación que es abordada principalmente por modelos estadísticos, pero la planificación actual que se proyectó hasta el 2025 iniciado en el 2005 de la Agencia de Cooperación Internacional del Japón (JICA) realizado por el Corredor Segregado de Alta Capacidad Fase I (COSAC1) para el 2010 (JICA, 2005), ya no es eficiente al cambio de comportamiento que presentan los usuarios en las estaciones. En la noticia del diario

El Comercio “Caos en el Metropolitano por mala programación de buses” (Malpartida, 2018) Alexandre Almeida, director de Ingeniería Civil de la Universidad de Lima, recomienda actualizar los estudios de demanda del Metropolitano, pues no se toma en cuenta a las personas que pueden realizar más de un viaje y a los buses que respaldan a las personas que se quedan esperando por mucho tiempo a partir de las 7 p.m. Las estaciones ya llegaron a su capacidad máxima tras ocho años que se planteó el servicio de 750 mil personas al día, según Linio de la Barrer, ex asesor del Ministerio de Transportes y Comunicaciones que explica para El Comercio (Paz, 2018. Esto genera cada vez más disconformidad por parte de los usuarios, ya que observan que la cantidad de buses en las horas pico, entre los horarios de 6:00 am - 8:30 am y 6:00pm - 7:30pm, no cumplen con la capacidad de movilizar a todos los usuarios de forma rápida y fluida.

Por ello este estudio de investigación propone dos técnicas de Data Mining para encontrar patrones espaciales, temporales y facilitar el análisis del comportamiento de la demanda de pasajeros para finalmente desarrollar modelos adecuados de predicción diariamente usando SARIMA con un RMSPE y MAPE del 10% como máximo para ambas mediciones y R^2 entre 0.80 y 0.90. Esto permitirá al servicio del Metropolitano una mejor y actualizada planificación de la demanda y así una toma de decisiones anticipada hacia la optimización de sus recursos en la distribución de buses a sus estaciones durante el año conforme al comportamiento de la demanda actual y las predicciones de los modelos SARIMA por clúster. En este documento se explicarán marcos teóricos de métodos que se usaron en el análisis espacio temporal, después se procederá a explicar la metodología de la investigación desde el tratamiento de los datos, recogidos de ProTransporte diariamente por estación de tres años consecutivos hasta el modelado y predicción. Luego se realizará la validación de predicciones por clúster de estaciones usando métricas de pronóstico y, por último, se incluirán conclusiones y propuestas de trabajos futuros.

2. ESTADO DEL ARTE

2.1 Obtención y análisis de patrones en la demanda de pasajeros usando Clustering y Series de Tiempo

Briand et al al., (2016) propusieron un modelo mixto de dos niveles que dividió a los pasajeros según sus perfiles por temporada, usando marcas de tiempo de transacciones de los pasajeros en la red de transporte público. En el primer nivel particionaron los datos en un conjunto reducido de grupos y el segundo nivel capturó la distribución en el tiempo de los viajes realizados por cada grupo de pasajeros. Los datos fueron tickets recopilados de la red de transporte urbano de Rennes Métropole en Francia, durante el mes de abril del 2014 del transporte público Service de Transport en Commun de l'Agglomération Rennaise (STAR), cada registro contiene una identificación de tarjeta anónima, tiempo de embarque, ubicación, bus o línea de metro abordado y tipo de tarifa de la tarjeta inteligente. Su metodología aplicó clustering a los pasajeros en función a sus actividades temporales, con el objetivo de descubrir patrones de viaje más frecuentes en una determinada red de transporte público con lo cual contribuyó a una mejor caracterización de la demanda. Obtuvieron curvas de perfil emitidas temporalmente que ofrecieron una mejor visión de la actividad de los pasajeros; en lugar de solo probabilidades de actividad discretas por hora; también formaron una probabilidad de actividad de tiempo continuo que no sufre potencial sesgo de agregación de observaciones en intervalos de tiempo y obtuvieron una actividad espacial del número de viajes por estación según la hora del día.

Lee et al., (2018), propusieron mejorar mecanismos de embarque de pasajeros y procesos de descarga, así como la influencia del tiempo de permanencia y operación de una estación de metro más concurrida de Seúl, Corea. Recopilaron datos de 54 trenes de dos direcciones norte y sur, estos datos registraron un tiempo entre que un pasajero entra y sale de una estación a un bus, derivaron una curva de tasa de servicio temporal única de cada ubicación de puerta y demostraron que las curvas de tarifas del servicio de pasajeros son únicas en cada ubicación de la puerta. Usaron Dynamic Time Warping (DTW) para evaluar similitudes entre las curvas estirando el eje del tiempo y efecto del volumen de pasajeros mientras sobresale el efecto de interferencia. Asignaron las puertas del tren en pocos grupos basados en la evaluación similitud, separaron en plataforma norte y sur para aplicar clustering a las puertas de los buses aplicando el criterio de clustering cúbico (CCC) para validar y obtener el número óptimo de clústeres. Mostraron dos dendogramas jerárquicos basados en la distancia del clustering de puertas, luego realizaron un análisis de regresión lineal para verificar el patrón de tasa del servicio versus su impacto en el tiempo al pasajero usando una variable dependiente y cuatro explicativas, que incluyen el número de pasajeros que bajan, embarque de pasajeros, peso del vagón de tren y grupo de puertas del tren. Finalmente, realizaron modelos para cada plataforma, demostraron que la inclusión de la variable clúster mejoran los modelos y el R^2 ajustado para predecir en cada plataforma.

Zhang et al, (2019) observaron la relación entre los patrones de movimiento de los pasajeros y la demografía social a través de tarjetas inteligentes (TI) a partir de características de viaje a pasajeros en espacio-tiempo, forma y frecuencia de viaje, para identificar patrones a largo plazo, su estacionalidad y análisis de cómo viajan en las ciudades. Usaron un algoritmo clustering llamado Affinity Propagation (AP) para detectar grupos de patrones de viajes calculando el índice Dunn y número de grupos predefinidos que van de 2 a 20, obteniendo 15 clústeres pues fue el óptimo. A través de encuestas a hogares, clasificaron a pasajeros en varios grupos en función de sus características sociodemográficas, como la edad y estado de trabajo, para identificar la homogeneidad de pasajeros y comprender "quién viaja" en el transporte público y, por último, exploraron relaciones importantes entre patrones de viaje y grupos demográficos que explican cuándo, dónde y con qué frecuencia viajan las personas y qué modo de viaje usan.

2.2 Modelación y pronóstico de demanda de pasajeros usando Series de Tiempo

Anvari et al., (2015), propusieron predecir series de tiempo basado en el método de Box- Jenkins para el transporte público basado en la verificación de la estacionariedad aplicado para predecir el tráfico de pasajeros de un metro de Estambul. Desarrollaron modelos por semanal y diariamente. seleccionaron el mejor modelo tomando el menor AIC, RMSPE (1.02%) y MAPE (10%), estos dos últimos en promedio entre las cuatro semanas.

Cyril et al., (2018), usaron un modelo de series de tiempo univariadas llamado ARIMA para predecir la demanda de viajes de transporte público entre distritos de la ciudad Trivandrum a otros cinco distritos de Kerala. Los datos que usaron fueron extraídos por la Corporación Estatal de Transporte por Carretera de Kerala (CETCK) que están en un repositorio central de la misma ciudad entre el 2010 y 2013. Los valores de demanda pronosticados fueron comparados con valores reales del año 2013 y demostraron que es precisa para zonas que dependen unas de otras y para la predicción a corto plazo.

Cyril et al., (2019), realizaron un novedoso enfoque para generar un modelo de pronóstico de demanda de pasajeros de buses públicos, los datos de entrada fueron tomados de Máquinas de Boletos Electrónicos (MBE) que emiten boletos, estos datos fueron de vital importancia para el modelo que ellos generaron, estos datos fueron indexados con fechas y por ellos los autores pudieron usar series tiempo para su exploración. Como objetivos de este artículo fueron el estudio y la aplicación de método de Series de Tiempo para el pronóstico de demanda de pasajeros de buses públicos, el enfoque temporal constó del uso de cuatro métodos de Holt-Winters usando ACF y PACF con y sin amortiguación para que hicieran comparaciones. Como resultados de comparación del pronóstico al aplicar los cuatro métodos fue el error porcentual absoluto medio (EPAM) o (MAPE) y la bondad de ajuste del modelo. Concluyeron que los modelos con y sin amortiguamiento explican mejor las variaciones estacionales.

Cyprich et al., (2013), aplicaron un modelo de predicción a la demanda de pasajeros de series de tiempo de transporte de buses suburbanos que toman de base el modelo de parámetros de y aleatoriedad de residuos de Box-Jenkins y el suavizado exponencial de regresión lineal múltiple, pues quisieron diseñar un modelo más preciso y confiable. La confiabilidad del modelo fue evaluada por algunos indicadores como media absoluta calculada y errores porcentuales. En específico, desarrollaron y aplicaron el modelo ARIMA y como resultados obtuvieron un error porcentual promedio del 12% mensual.

Dou et al, (2014), presentan un modelo de predicción de demanda de pasajeros con un modelo difuso óptimo de predicción, a partir del análisis de las características complejas, el flujo de pasajeros durante las vacaciones y la teoría de conjuntos difusos de series temporales. Tuvieron como objetivo minimizar el costo de operación de trenes y el volumen de pasajeros no atendidos. Finalmente, validaron el modelo ARIMA minimizando el RMSE (4.56), MAE (8.64) y MAPE (9.6%) con los datos de un ferrocarril de Beijing-Shanghái, China; demostraron que los resultados ayudaron a optimizar el despacho de un número pequeño de trenes para que atiendan a pasajeros sin servicio eficazmente.

Gong et al., (2014), propusieron un marco de pronóstico para el flujo de pasajeros en paraderos de buses, este marco constó de tres etapas, en la primera usaron un método basado en ARIMA estacional (SARIMA) para predecir la llegada de pasajeros y el espacio vacío en un bus cuando llega a sus paraderos, también actualiza el flujo de pasajeros cuando llega el bus a su paradero; en la segunda etapa desarrollaron un método basado en eventos que predijo el conteo de pasajeros que salen del paradero. En la última etapa aplicaron el filtro Kalman para predecir la cantidad de personas de la primera y segunda etapa. Los resultados confirmaron que el marco que propusieron y el algoritmo de solución son eficientes en la precisión de la predicción del flujo de pasajeros con un 2.95% y 3,02% de error relativo porcentual medio.

Ma et al., (2014), propusieron un enfoque híbrido de patrones basado en modelos múltiples interactivos (MMI) para predecir la demanda de pasajeros a corto plazo, este enfoque maximiza lo que contiene la información de los modelos de patrones de datos de datos históricos y la interacción entre estos observándolos en tiempo real, Además, puede este modelo puede estimar la combinación de modelos por adelantado para el próximo intervalo. Los datos que usaron fueron recopilados de tarjetas inteligentes de una ruta de un bus durante un año. Primero generaron series de tiempo semanales, diarios y por hora para capturar patrones temporales, para desarrollar modelos usando MA, ARIMA, SARIMA respectivamente usando ACF y PACF y aplicaron el algoritmo MMI para combinar los modelos de patrones y así generaron el pronóstico de demanda final. El error porcentual absoluto medio (MAPE) de los modelos fueron 9.57%, 9.89% y 23.1%, indicando un mejor rendimiento aplicado a la demanda de pasajeros.

Milenković et al, (2016), usaron SARIMA para predecir los flujos de pasajeros en metros, los datos que usaron fueron datos históricos de enero del 2004 al junio del 2014 de demanda mensual de la red de metros de Serbia. Se apoyaron de la metodología Box-Jenkins para encontrar los parámetros adecuados del modelo final usaron ACF y PACF. El modelo final fue $SARIMA(0,1,0) \times (0,1,1)^{12}$ con un BIC de 7.056 y MAPE del 7.13%, este modelo fue el más indicado y considerado para pronosticar el flujo mensual de pasajeros.

Ni et al, (2016), hicieron un análisis de pasajeros de un metro a través y las tasas de las publicaciones de las redes sociales (tweets), demostrando que hay una correlación positiva entre estos dos elementos. Posteriormente hicieron un modelo SARIMA y una función de pérdida híbrida (OPL), este último les ayudó para mejorar la relación que encontraron y a predecir mejor. Los datos de entrada fueron de la línea de 7 de Nueva York de abril hasta octubre

en horas y fueron separados en entrenamiento (70%) y prueba (30%). El rendimiento del modelo tuvo un R^2 de 0.616 y un MAPE de 33.08% y cuando aplicaron OPL el MAPE fue de 11.4%.

Tsai et al., (2013), usaron ARIMA y un modelo de ajuste parcial dinámico (PAM) que estimó la elasticidad de la demanda incluyendo la tarifa del transporte público, la socio-demografía de los usuarios y el nivel del servicio. ARIMA se puede modeló, en este caso, con datos mensuales de abordaje de trenes y buses desde el 2007 hasta 2011. El modelo PAM usó datos de un pseudo panel en Sydney Household Travel de 1997 hasta 2009 y predijo la demanda del transporte público de Sydney. Los resultados mostraron que ARIMA puede lograr mejores resultados de precisión a corto plazo, mientras que prefirieron el modelo PAM porque se quiere pronosticar la demanda en varios escenarios.

Yan et al, (2018), aplicaron el modelo ARIMA para predecir el flujo de pasajeros a corto plazo en el metro de Guangzhou, China. Primero se centraron en seleccionar los parámetros más apropiados del modelo (p,q) basado en la prueba de estacionariedad Dickey-Fuller, luego reconocieron el modelo, la estimación de parámetros y el mejor modelo usando AIC o parcelas SACF y SPACF. Identificación dos posibles modelos $ARIMA(1,0,0)$ y $ARIMA(1,1,1)$ con un MAPE de 23.57% y 20.23% respectivamente. Por último, compararon el MAPE con modelos SVM y demostraron que los modelos ARIMA son mejores pues arrojan un menor MAPE.

Wang et al, (2015), utilizaron el modelo estacional SARIMA para predecir el flujo de pasajeros del metro de Beijing a partir de datos de datos de tráfico desde mayo hasta julio del 2013, describieron la tendencia de cambio del flujo de tráfico en esta estación. Este modelo fue el adecuado pues arrojó un error promedio del 0.3%, esto les permitió analizar las características del volumen del flujo de pasajeros entrantes, ayudando a una mejor optimización del diseño, operación y seguridad de las estaciones.

Wei y Chen (2012), propusieron un modelo de predicción híbrido (DME-NRP) con el modelo SARIMA, en descomposición modo empírico (DME) y redes neuronales de retro-propagación (NRP), estas redes neuronales fueron desarrolladas para predecir el flujo de pasajeros a corto plazo de un metro. Los autores explicaron las tres etapas del modelo híbrido; la primera, descompone los datos de la serie de tiempo de los pasajeros a corto plazo en otra serie de componentes de Función de Modo Intrínseco (FMI); la segunda, identificó los FMI importantes como entradas para el modelo DME y el tercer aplicó DME para predecir el flujo de pasajeros, Los datos de entrada fueron indexados por día de la semana y periodos del día. Usaron el modelo $SARIMA(2,0,3) \times (2,1,2)^{75}$ siguiendo la metodología Box-Jenkins. Los resultados mostraron, en este caso, que el modelo SARIMA no es el adecuado para predecir el flujo de pasajeros debido al supuesto de linealidad.

Xue et al., (2015), propusieron un modelo basado en el algoritmo de filtro de modelo múltiple interactivo (MMI) con el objetivo de predecir la demanda de pasajeros a corto plazo, los datos de entrada fueron recopilados de una ruta de buses durante cuatro meses para luego desarrollar tres series de tiempo semanales, diarias y de 15 minutos. Luego hicieron un análisis de correlación, periodicidad y estacionalidad y después se construyeron los modelos de predicción de series de tiempo ARMA, ARIMA y SARIMA. Los modelos que usaron fueron: $ARMA(2,2)$, $ARIMA(2,1,0)$ para predecir semanalmente, $SARIMA(2,0,3) \times (1,0,0)^{24}$ para predecir diariamente. Se enfocaron en la heterocedasticidad de estas series para un mejor rendimiento y aplicaron el modelo GARCH para mejorar la precisión del modelo ARIMA. Por último, aplicaron el filtro MMI para combinar modelos con la demanda de pasajeros pronosticada para el próximo intervalo de tiempo y compararon con índices de error el análisis de los modelos individuales e híbridos. Los resultados que arrojaron las comparaciones fueron que los modelos híbridos son superiores a los individuales en precisión.

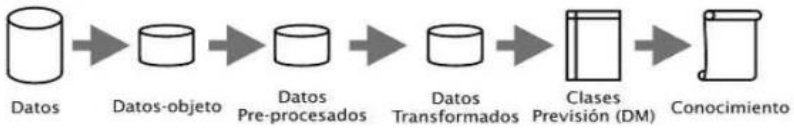
3. ANTECEDENTES

3.1 Data Mining para descubrir patrones:

En Data Mining los patrones sirven para desarrollar modelos predictivos que estimen valores futuros o desconocidos que son importantes pues generan conocimiento para la toma de decisiones proactivamente. En la Figura 3.1 se puede mostrar un esquema de cómo funciona el proceso de Data Mining (Viera Brag et al., 2009).

Figura 3.1

Esquema de Data Mining



Fuente: Viera Braga, Ortiz Valencia, y Ramírez Carvajal (2009).

Para este estudio de investigación, se describirán dos métodos de Data Mining para encontrar patrones de espaciales y temporales que serán de gran ayuda para poder predecir finalmente.

a. Método de Data Mining para encontrar patrones espaciales

Clustering

Según Alboukadel Kassambara (2017), clustering es uno de los métodos importantes de Data Mining para descubrir conocimiento en datos multidimensionales, el objetivo de este método es identificar patrones o grupos de objetos

similares dentro de un conjunto de datos de interés. En cuanto a los métodos de clúster se pueden clasificar por Separación o Agrupación Jerárquica (Ryu & Eick, 2005).

- Clustering Jerárquico:

Según Esling y Agaon (2012), este tipo el clúster combina datos en grupos y estos en grupos más grandes, y así sucesivamente, generando una jerarquía; fluyen de un árbol binario que tiene muchos nodos segundarios y un solo nodo principal (Amutha & Renuka, 2015). Esta jerarquía de grupos se representa en dendogramas mostrando objetos de datos individuales llamados hojas del árbol y nodos interiores que son grupos no vacíos. Los nodos hermanos dividen los puntos que abarca el padre único, permitiendo explorar distintos niveles de granularidad. Entre los beneficios de su uso son la movilidad y flexibilidad al nivel de granularidad, manejo sencillo de cualquier forma de similitud o distancia y se aplica a cualquier tipo de atributo (Esling & Agon, 2012).
Existen dos tipos de clustering jerárquico divisible y aglomerativo este último sigue la dirección de abajo hacia arriba cada nodo se considera como un clúster individual y tienen sus propias características fusionándose con cada iteración con sus grupos similares mediante un enlace único.

- Metodología Clustering:

Para cualquier tipo de clustering es necesario una serie de pasos que ayudan a generar una agrupación óptima y de fácil interpretación. Cada paso cuenta con técnicas que fueron usadas en el trabajo de investigación. Chávez y López (2005), proponen técnicas a seguir para tener un buen análisis clustering en cualquier tipo de clúster:

1. Selección de Variables
2. Medida de Distancia: llamada medida de disimilitud o semejanza, miden la distancia entre dos datos con el objetivo de diferenciar los datos según sus distancias y dar una menor probabilidad de que los métodos de clustering los pongan en el mismo clúster. La distancia euclidiana cuadrada se calcula empleando el teorema de Pitágoras y puede aplicarse en vectores de *n* valores, para los vectores *i* y *j* de la siguiente forma:

$$d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + \cdots + (x_{in} - x_{jn})^2} \quad (1)$$

3. Elegir tipo y algoritmo de clustering correspondiente
4. Elegir un tipo de enlace de clustering: se definirá un tipo de enlace para el algoritmo de clustering, estos enlaces son procesos iterativos que agrupan a objetos y que definen a los vecinos en las ramas de los clústeres.
   El Método de Ward constituye los clústeres para que, al momento de comparar dos elementos, la pérdida de información sea la mínima posible, hace esto posible al sumar distancias al cuadrado (SCE o SSW) de cada elemento respecto al centroide del clúster que pertenece (Ward, 1963).

$$ssw_r = \sum_{m=1}^{n_r} \sum_{j=1}^{p} (X_{rjm} - \underline{X_{rj}})^2 \quad (2)$$

   Donde: X_{rjm} es el valor de la variable X_j en el m-ésimo elemento del grupo r.

5. Validar el número de clústeres: ANOVA está basado en un análisis de varianza que toma como datos de entrada grupos o clústeres que se tomarán como un factor y cada variable incluida será tomada como una dependiente (Califiski & Corsten, 1985). Ayuda a verificar que el número de clústeres ya obtenidos sean los correctos u óptimos, esto compara cada clúster con otro con un 95% de confianza, si el *p_adj* o el *p_value* es menor que 0.05 significa que los clústeres tienen una correcta agrupación.

b. Método de Data Mining para encontrar patrones temporales

Series de Tiempo

Esling y Agon (2012), indican que una serie temporal equivale a una colección de valores obtenidos de mediciones consecutivas en el tiempo ordenadas cronológicamente representándose como una serie temporal "T" de *n* variables:

$$T = (t_{1,}, \ldots, t_{n,}), t_i \in R \quad (3)$$

El objetivo principal del modelado de series temporales es recolectar cuidadosamente y estudiar rigurosamente las observaciones pasadas para desarrollar un modelo apropiado que describa la estructura de la serie y así generar valores futuros, es decir, hacer pronósticos (Raicharoen, Lursinsap & Sanguanbhokai, 2003).

Análisis de series de tiempo:

Según Adhikari y Agrawal (2013), una serie de tiempo en general se ve afectada por dos componentes principales: tendencia y estacionalidad:

- Tendencia: la tendencia generalmente puede aumentar, disminuir o estancarse durante un largo período del tiempo denominada tendencia secular o simplemente tendencia. Por lo tanto, se puede decir que la tendencia es, a largo plazo, el movimiento de una serie temporal.
- Estacionalidad son fluctuaciones dentro de un año durante una temporada (días, meses, trimestres, etc.). Los factores importantes que causan variaciones estacionales son: las condiciones climáticas, climáticas, las costumbres, los hábitos tradicionales, etc.

El patrón temporal está relacionado con un evento del cual es necesario ~~para~~ predecir, para ello es necesario considerar una serie de tiempo como un todo. La naturaleza de las series de tiempo hace que los métodos comunes de Data Mining sean diferentes; entre sus características se tiene: alta numerosidad, gran número de dimensiones y una constante actualización el transcurrir el tiempo (González Castellanos & Soto Valero, 2013).

3.2 Modelos de pronóstico de Series de Tiempo

Por lo general los modelos de predicción pueden ser de muchas formas y tener procesos estocásticos, existen dos modelos muy usados, el modelo autorregresivo (AR) y promedio móvil (PM) o (MA), al combinan estos modelos se obtiene el modelo medio móvil autorregresiva (MMAR) o (ARMA) y el modelo móvil integrado autorregresivo (MMIAR) o (ARIMA) (Adhikari & Agrawal, 2013).

a. Modelo medio móvil e integrado autorregresivo (MMIAR) o (ARIMA)

Es un modelo de series temporales usado para predecir, son usados para predecir la demanda, a diferencia de los demás modelos, ARIMA pronóstica series de temporales con un comportamiento no estacional (Hamzaçebi, 2008), su formulación matemática es $ARIMA(p,d,q)$ (Hipel & McLeod, 1994), por lo general d es 1 pero cuando es 0 el modelo se reduce a ARMA (p, q).

b. Modelo de media móvil integrada autorregresiva estacional (MMIAE) o (SARIMA)

Box y Jenkins (1970) generalizaron el modelo ARIMA para tratar la estacionalidad y propusieron el modelo estacional ARIMA (SARIMA). Este modelo aplica una diferenciación del orden correcto, esto consiste en la observación del año anterior y se calcula como: $z_t = y_t - y_{t-s}$, la s es el número de periodos de un año, por ejemplo, si la serie de tiempo es mensual s=12 o si es trimestral s=4; este modelo se expresa generalmente como el modelo:

$$SARIMA(p,d,q) \times (P,D,Q)^s \quad (4)$$

Donde: p es el parámetro asociado a la parte autorregresiva del modelo, d es el parámetro del orden de diferenciación mínima para que la serie sea estacionaria y q está relacionado al promedio móvil del modelo; los parámetros P, D, Q son similares a los tres primeros, pero son componentes estacionales del modelo.

3.3 Metodología Box y Jenkins

Para encontrar un óptimo modelo y predecir, Box y Jenkins (1970) propusieron una metodología a partir de la estacionariedad de una serie temporal. La estacionariedad se refiere a que una serie temporal tiene una media y varianza constante en el tiempo (Anvari, Tuna, Canci & Turkay, 2015). A continuación, se presentan las fases de la metodología de Box y Jenkins (1970):

1. Identificación: se basa en un análisis de patrones temporales de una serie temporal, luego se debe verificar si los datos son estacionarios, se puede usar la prueba de Aumento Dickey Fuller (ADF), si el ADF muestra un valor p menor a 0.05 (valor mínimo de significancia) significa que la serie es estacionaria y no es necesario un orden de diferenciación, es decir, el parámetro $d=0$. También se puede apoyar de dos criterios: Función de Autocorrelación (ACF) y Función Parcial de Autocorrelación (PACF) para verificar que la media y la varianza sean constantes siendo efectivos en los datos de entrenamiento y deben coincidir con los valores teóricos o reales (Hipel & McLeod, 1994), esto es considerado luego a los pedidos estacionales apropiados para los modelos ARIMA y SARIMA para restringir la importancia a los retrasos estacionarios (Ramesh Reddy, Ganesh, Venkateswaran, & Reddy, 2017).
2. Estimaciones y pruebas: las estimaciones se basan en la selección el modelo ARIMA o SARIMA a partir del análisis de la estacionalidad de la serie de tiempo que se está analizando, para luego estimar parámetros: usando ACF y PACF para definir los parámetros p y q respectivamente a partir de los valores altos que se repiten cada retardo o *lag* representados gráficamente. Luego se realiza un diagnóstico a partir de la partición de la serie original (entramiento y prueba) para seleccionar el mejor modelo, se pueden usar las autocorrelaciones de los residuos del ACF y PACF de datos de prueba, pero también existen indicadores como el Criterio de Información de Akaike (AIC) y el Criterio de Información Bayesiano (BIC) que se usan para elegir los mejores subconjuntos de predictores de un modelo comparando modelos no anidados cosa que las pruebas estadísticas ordinarias no pueden hacer (Ramesh Reddy, Ganesh, Venkateswaran, & Reddy, 2017). La interpretación de valores de los criterios es que el AIC más bajo de un modelo significa que un modelo se considera más cercano a la verdad y para el BIC es similar solo que trabaja con probabilidades, ambos pueden ser usados para una mejor decisión al escoger un modelo de predicción, pero el AIC es mejor en situación en las que un resultado falso negativo se consideraría más engañoso que un falso positivo. Estos dos criterios, en esencia, usan el Error Cuadrado Medio (ECM) o (MSE) y el número de parámetros del modelo (Yaffee & McGee, 2000):

$$\text{AIC} = \text{T} \ln(\text{MSE}) + 2k \quad (5)$$

$$\text{BIC} = \text{T} \ln(\text{MSE}) + k(\ln(T)) \quad (6)$$

Donde: T es el número de observaciones y k es el número de parámetros (p, d, q, P, D, Q) más uno.

Por último, se procede elegir el mejor modelo al que tenga el menor AIC y error cuadrático medio (RMSE) pues este es más sensible al error que otras métricas de rendimiento debido al error cuadrático (Anvari, Tuna, Canci & Turkay, 2015).

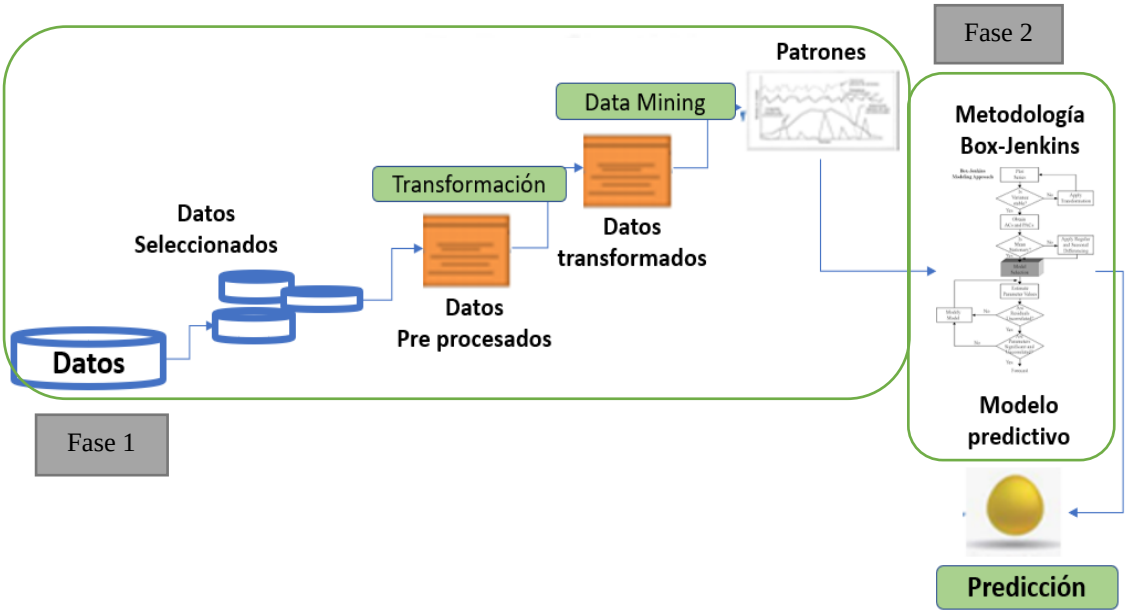
3. Aplicación del modelo de predicción:
   Cuando el mejor modelo se identifica, se define la fórmula del modelo y se realiza el pronóstico, en el caso de elegir SARIMA se debe definir también el número de periodos (*s*).

4. METODOLOGÍA

Esta investigación propone predecir la cantidad de pasajeros usando el modelo de predicción SARIMA a partir de la aplicación de dos métodos de Data Mining, pues ofrecen técnicas para obtener patrones espaciales y temporales de la demanda de pasajeros y luego seguir la metodología Box-Jenkins para definir adecuadamente modelos de predicción con una buena precisión. A continuación, se explicará secuencialmente la metodología de la investigación que consta de dos fases como se puede apreciar en la Figura. 4.1:

Figura 4.1

Flujo de aplicación de Data Mining (Fase 1) y modelación para la predicción (Fase 2)



4.1 Fase 1

1. Obtención y selección de los datos:

Los datos se extrajeron de la página web de ProTransporte (http://www.protransporte.gob.pe/datos-abiertos/), en una de sus secciones llamada "Datos Abiertos- Estadísticas", esta cuenta con datos estadísticos e históricos diaria, mensual y anualmente de la demanda de pasajeros de sus servicios Troncales y Alimentadoras desde el 2010 hasta la actualidad. Para este trabajo de investigación solo se tomó desde el 1 de enero del 2016 hasta el 31 de diciembre del 2018 de la sección "Validaciones por Estaciones", estás validaciones son registradas en el momento que la persona pasa a la estación usando una tarjeta que el Metropolitano da a cada uno de sus usuarios.

a. Variables:

Tabla 4.1

Resumen de variables del dataset

Variable	Tipos de entrada	Rango	Detalle
Estaciones	Texto	38 estaciones	Nombre de estación
Fecha	Fecha	01/01/16 - 31/12/18	Días de registro de demanda de pasajeros
Demanda	Numérico	n > 0	Demanda de pasajeros

b. Tipos de estaciones:
   Central: Estación Central (conecta el norte y sur de Lima Metropolitana)
   Intermedias: 35 estaciones
   Terminales: Matellini (Norte) y Naranjal (Sur)

c. Demanda de pasajeros entre el 2016 y 2018:
   Se puede observar que algunas estaciones presentan un comportamiento similar en la demanda cada año, indicando que se podría tomar grupos de estaciones con comportamientos similares, luego hacer un análisis temporal para contar con una mejor capacidad de análisis entre estos grupos de estaciones. Además, se pudo encontrar comportamientos en días festivos anormales donde la demanda baja drásticamente.

2. Preprocesamiento de datos:
Se acotaron los filtros de fechas que se encuentran en la página web mencionada de ProTransporte para obtener la demanda por día pues si se acotaba desde el 1 de enero del 2016 al 31 de diciembre del 2018 solo mostraba el acumulado. Esta demanda fue copiada en un archivo para ordenar los datos donde las columnas fueron las 38 estaciones y las filas los días. Se usaron dos archivos uno para aplicar clustering jerárquico y otro para las series de tiempo de clústeres; pero antes se observaron datos faltantes y atípicos, estos fueron reemplazos por el promedio de la demanda del día de la semana anterior y posterior en la que se encontraron estos valores.

3. Aplicación y análisis de métodos de data Mining para encontrar patrones espacio-temporales del 2016 al 2018:
Se usó RStudio para el análisis espacial aplicando clustering jerárquico de tipo Aglomerativo a las estaciones, pues se quiso ver la independencia de la jerarquía de los clústeres de estaciones según el comportamiento de la demanda, se usó la distancia euclidiana, el método de enlace fue Ward y, por último, se validó el número de clústeres usando ANOVA en un archivo aparte (ver Tabla 4.2).

a. Distancia euclidiana entre las estaciones se expresaron como pesos según el nivel de demanda, estás se restaron entre cada una hasta llegar a la última estación. La distancia euclidiana de estaciones es representada de la siguiente manera:

$$d_{(p,q)} = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + (p_3 - q_3)^2 + \ldots + (p_{38} - q_{38})^2} = \sqrt{\sum_{i=1}^{n=38} (p_i - q_i)^2} \quad (7)$$

Donde *p* y *q* representan a estaciones diferentes y *n* a la cantidad de estaciones que hay.

b. Clustering Jerárquico Aglomerativo de estaciones: para obtener los clústeres fue necesario que primero se forme una matriz de proximidad en la cual se encuentran las distancias euclidianas de estaciones, luego se obtuvo la proximidad entre cada par de estaciones usando el método Ward y, por último, se usó el algoritmo de clustering jerárquico Aglomerativo que formará los clústeres de estaciones juntando estaciones con distancias más próximas y ordenando según pesos que el algoritmo da a cada clúster evidenciando la jerarquía .

Algoritmo 1: Aplicación del Clustering Jerárquico Aglomerativo de estaciones

1. *Entrada:* Matriz de proximidad o similitud de distancias euclidianas
2. *Salida: Clústeres Jerárquicos*
3. ***Desde*** *i ← 1* ***hasta*** *n ← 38*
4. *Enlace Ward*
5. *Algoritmo Clustering Jerárquico Aglomerativo*
6. *Unir estaciones con distancias de similitud más próxima*
7. *Ordenar por pesos*
8. *n ← n – 1*
9. *Actualizar la matriz de proximidad*
10. ***Return*** *Clúster*
11. *Clúster Final*

c. Se validó el número de clústeres que se obtuvieron con el fin de que la agrupación sea la adecuada. Se usó ANOVA para esta validación que usa el modelo Tukey, este modelo tiene como datos de entrada el clúster aginado a la estación y la demanda promedio de cada estación.

Tabla 4.2
Variables para validar el número clúster de estaciones

Variable	Tipos de entrada	Rango	Detalle
Clúster	Numérico	1-5	Número de clústeres obtenido de estaciones
Demanda	Numérico	n > 0	Promedio de demanda de estaciones

Finalmente, en un nuevo archivo se calculó la media de la cantidad pasajeros de los elementos (estaciones) que cada clúster agrupó. Como se puede ver en la Tabla 4.3, se puso como primera columna "Fecha" y las demás tendrán el promedio de demanda por cada clúster de estaciones diariamente Este nuevo archivo servirá para el análisis temporal de estos a partir de series de tiempo en Python y su descomposición en tendencia y estacionalidad. Se analizaron los patrones temporales encontrados a partir de la tendencia y estacionalidad de cada serie de tiempo, para saber si es estacional o no estacional, pues es necesaria esta información para los modelos de predicción.

Tabla 4.3
Variables del dataset final

Variables	Tipos de entrada	Rango	Detalle
Fecha	Fecha	01/01/16 - 31/12/18	Días donde se registra demanda de pasajeros
c2, c3, c4, c5	Numérico	n > 0	Promedio de demanda de clúster

4.2 Fase 2

1. Desarrollar de modelos predictivos SARIMA:

En primer lugar, se tomó en cuenta todo lo del apartado anterior para la elección de los modelos para luego usar el enfoque de Box-Jenkins (ver Figura 4.2):

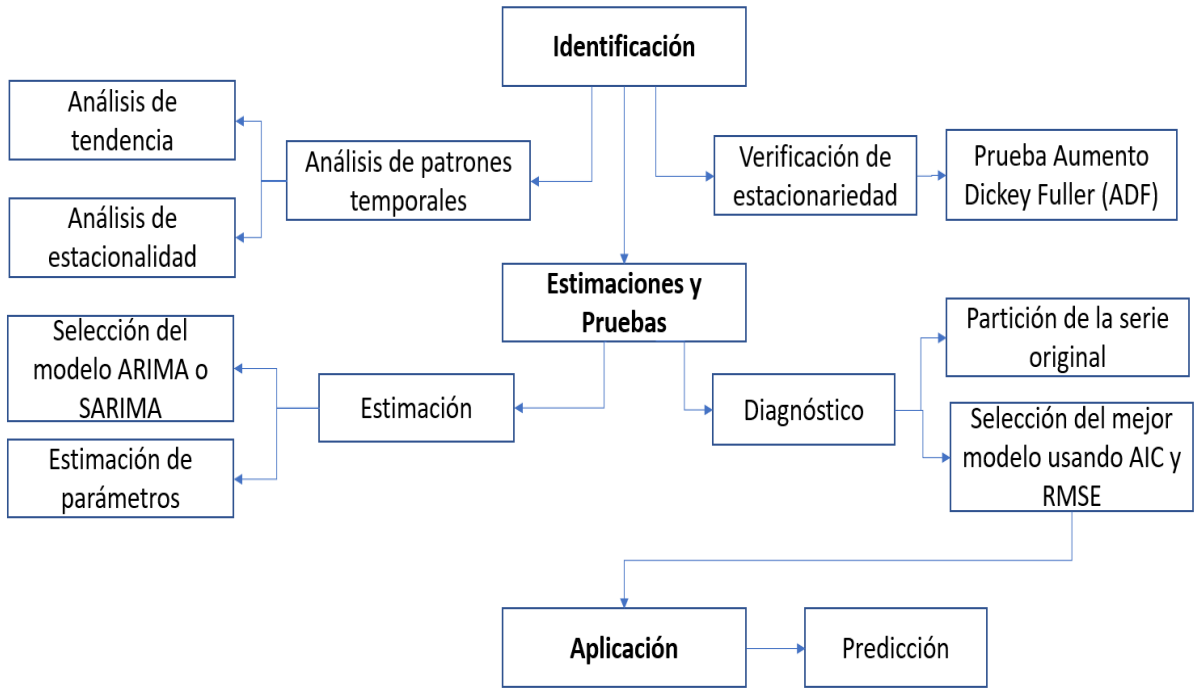
a. Se tomó en cuenta el análisis temporal, luego se verificó la estacionariedad usando la prueba ADF por clúster.
b. Luego se estimaron los parámetros usando los criterios ACF y PACF. Adicionalmente, se desarrolló un script que automáticamente evalúa los modelos con los posibles valores de los parámetros *p, d, q* que el modelo SARIMA puede tener con el indicador AIC, para encontrar el mejor modelo por clúster al seleccionar el menor AIC de las posibles combinaciones y seleccionar el que tenga el menor RMSE también.

   Algoritmo 2: SARIMA automático

   1. Separar en 70 % y 30% de entrenamiento y prueba respectivamente de la serie de tiempo
   2. Se crean las variables que serán los parámetros del modelo final: *p, q* con un rango de 0 hasta 8 y *d* de 0 hasta 2.
   3. Los posibles valores serán almacenados en una lista producto *p, d, q.*
   4. Luego serán almacenados en una ecuación que se representará el modelo SARIMA $seasonal\_pdq = [(x[0], x[1], x[2], 12)]$ hasta x en un lista producto *p, d. q*.
   5. Se creará un bucle introduciendo la lista de posibles parámetros al modelo generando posibles modelos con sus respectivos AIC.

c. Se formuló el modelo final para luego predecir y se mostró gráficamente una comparación entre la demanda de prueba real y predicha de cada clúster con su respectivo indicador RMSE.

Figura 4.2.
Representación de la metodología de Box-Jenkins usado en la investigación



2. Validación de modelos predictivos SARIMA:

Por último, se realizó un análisis estadístico sobre los resultados de las predicciones del año 2019 por cada clúster, luego se validaron los modelos por cada clúster con algunos indicadores usados en trabajos mencionados en la sección "Estado del Arte" como el RMSE, RMSPE, MAE, MAPE, R^2 y finalmente con el fin de ampliar aún más el análisis se realizó un análisis más granulado analizando los modelos semanalmente con los mismos indicadores mencionados.

5. RESULTADOS

5.1 Clustering Jerárquico de Estaciones del 2016 al 2018:

A continuación, se mostrará el dendograma de estaciones al usar clustering jerárquico (ver Figura. 5.1), una tabla resumen de las estaciones que se encuentran en cada clúster (ver Tabla 5.1) y se validarán el número de clústeres encontrados usando ANOVA (ver Tabla 5.2):

Figura. 5.1
Dendograma de estaciones

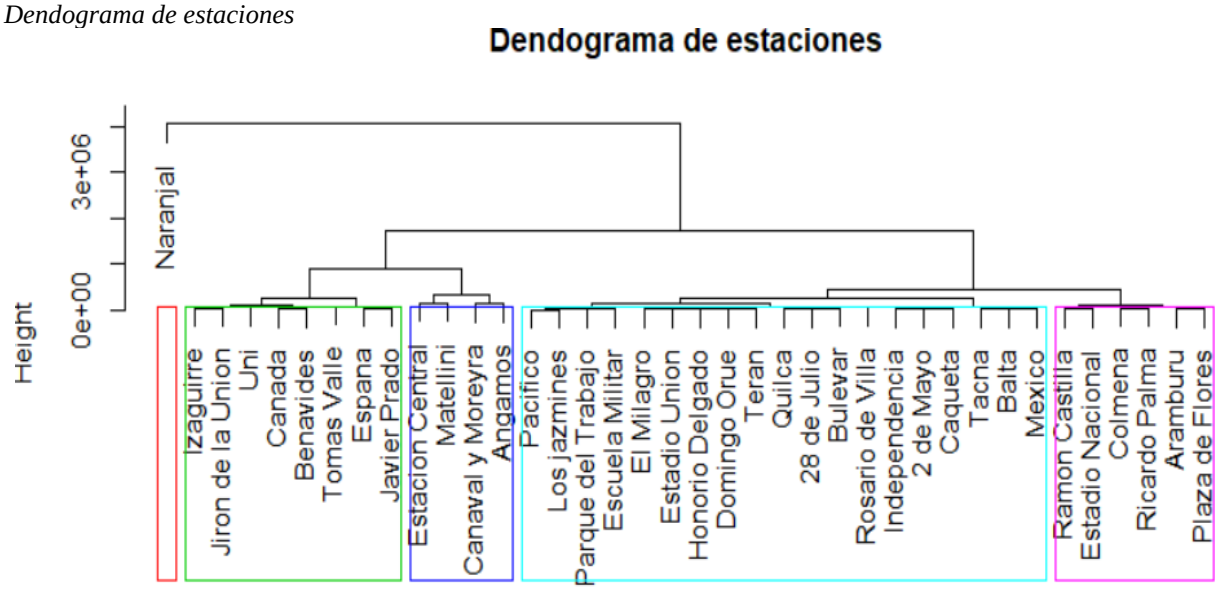


Tabla 5.1.
Estaciones según el clúster que pertenecen

Clúster	2016 - 2018
1	**Naranjal**
2	Izaguirre, Tomás Valle, Uni, Jirón de la Unión, España, Canadá, Javier Prado, Benavides
3	Pacífico, Independencia, Los jazmines, El Milagro, Honorio Delgado, Parque del Trabajo, Caquetá, Tacna, 2 de Mayo, Quilca, México, Domingo Orué, 28 de Julio, Balta, Estadio Unión, Escuela Militar, Terán, Rosario de Villa
4	Ramón Castilla, Colmena, Estadio Nacional, Aramburú, Ricardo Palma, Plaza de Flores
5	Estación Central, Canaval y Moreyra, Angamos, Matellini

Tabla 5.2.
Validación del número de clúster de estaciones

p adj	Comparación entre clústeres
0.0000000	(c2 – c1) (c3 – c1) (c4 – c1) (c5 – c1) (c3 – c2) (c5 – c3) (c5 – c4)
0.0002119	(c4 – c2)
0.0004454	(c4 - c3)

5.2 Componentes de las Series de tiempo de pasajeros por clúster entre el 2016 y 2018:

A continuación, se mostrarán componentes de las series de tiempo semanalmente relevantes en esta investigación por clúster, no se seleccionó el clúster 1, pues solo tiene un elemento y no es considerado como un clúster.

Tendencia:

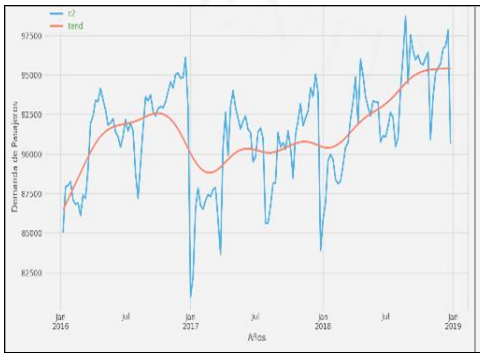


Figura 5.2 Tendencia clúster 2.

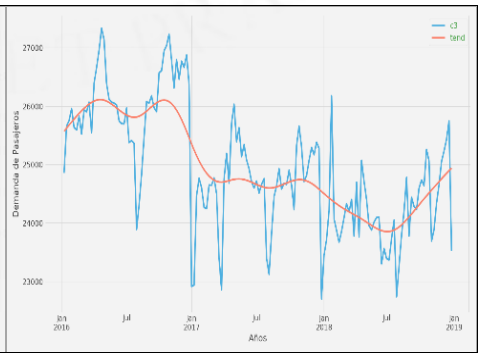


Figura 5.3 Tendencia clúster 3.

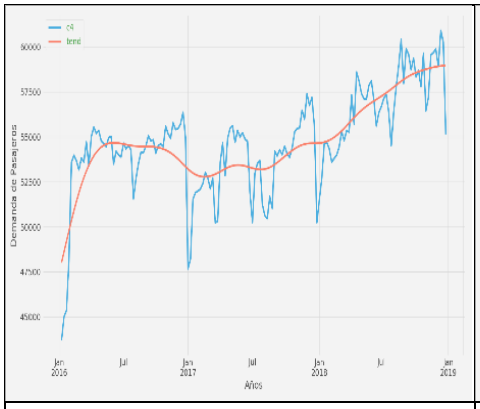


Figura 5.4 Tendencia clúster 4

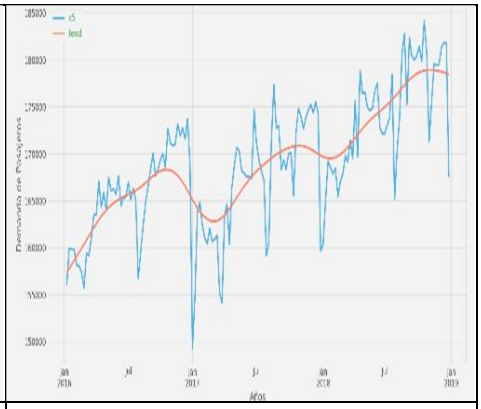


Figura 5.5 Tendencia clúster 5.

Estacionalidad:

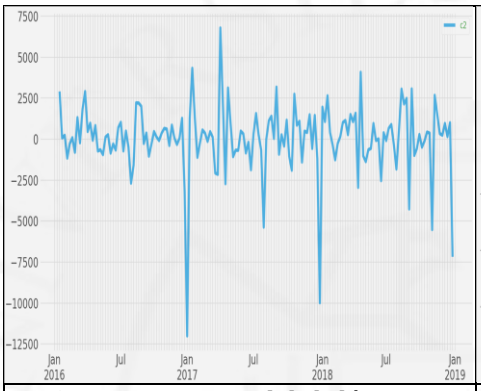


Figura 5.6 Estacionalidad clúster 2.

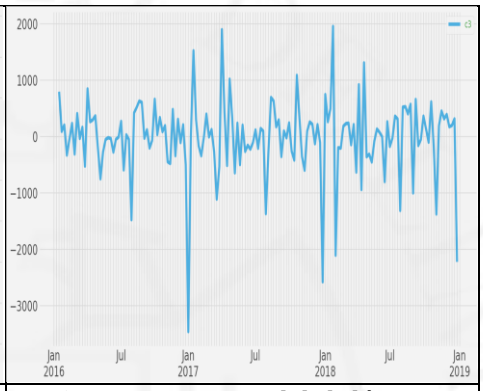


Figura 5.7. Estacionalidad clúster 3.

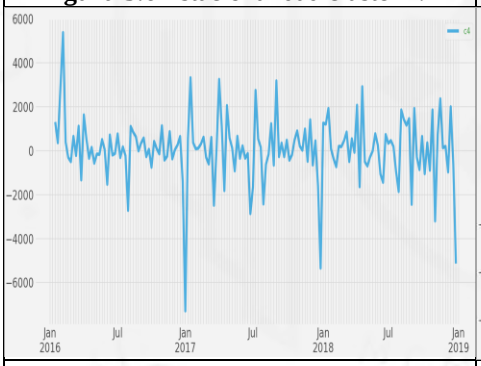


Figura 5.8 Estacionalidad clúster 4

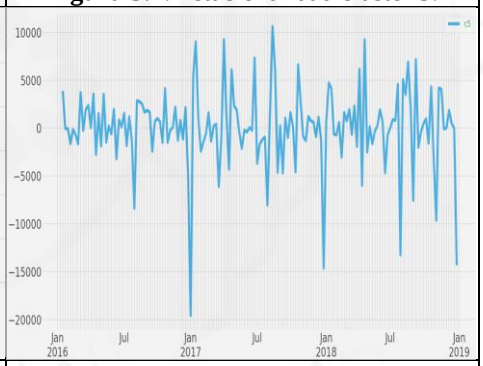


Figura 5.9 Estacionalidad clúster 5.

5.3 Predicción de pasajeros usando SARIMA

A continuación, se mostrarán los resultados al seguir la metodología de Box-Jenkins para conseguir el modelo adecuado SARIMA por cada clúster; también se mostrarán las predicciones de pasajeros de la demanda de prueba por clúster:

a. Prueba ADF para verificar la estacionariedad

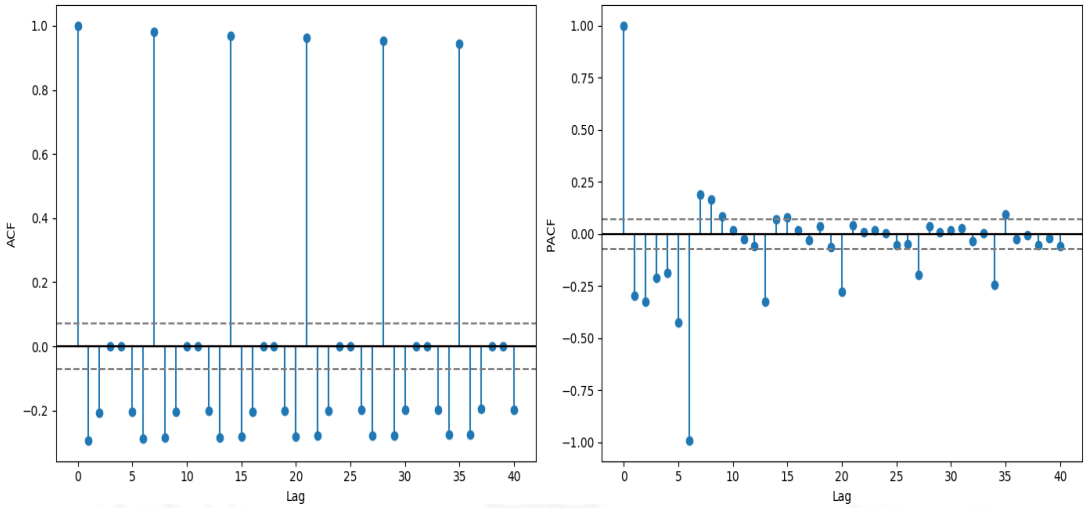
Tabla 5.3

Verificación de estacionariedad por clúster usando ADF

Clúster	p valor
2	0.006000
3	0.016269
4	0.000391
5	0.040948

b. ACF y PACF para encontrar parámetros

Figura 5.10
ACF y PACF de clústeres



c. Modelos y predicciones de cada clúster de pasajeros usando SARIMA

Tabla 5.4
Modelos de predicción SARIMA por clúster

Clúster	Modelos	AIC	RMSE - prueba
2	$SARIMA(7,0,7) \times (7,0,7)^{12}$	10138.94	1270.91
3	$SARIMA(7,0,7) \times (7,0,7)^{12}$	8816.46	197.13
4	$SARIMA(7,0,7) \times (7,0,7)^{12}$	9763.72	675.50
5	$SARIMA(7,0,7) \times (7,0,7)^{12}$	11230.50	1533.15

Figura 5.11
Comparativa entre los datos y predicciones de prueba del clúster 2

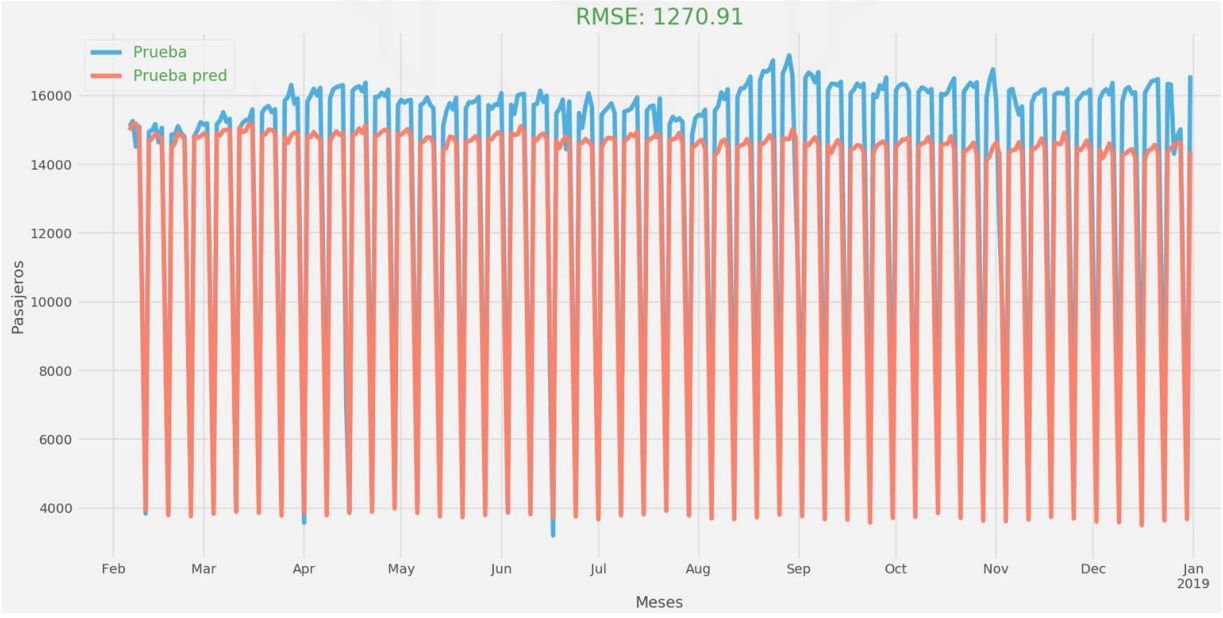


Figura 5.12
Comparativa entre los datos y predicciones de prueba del clúster 3

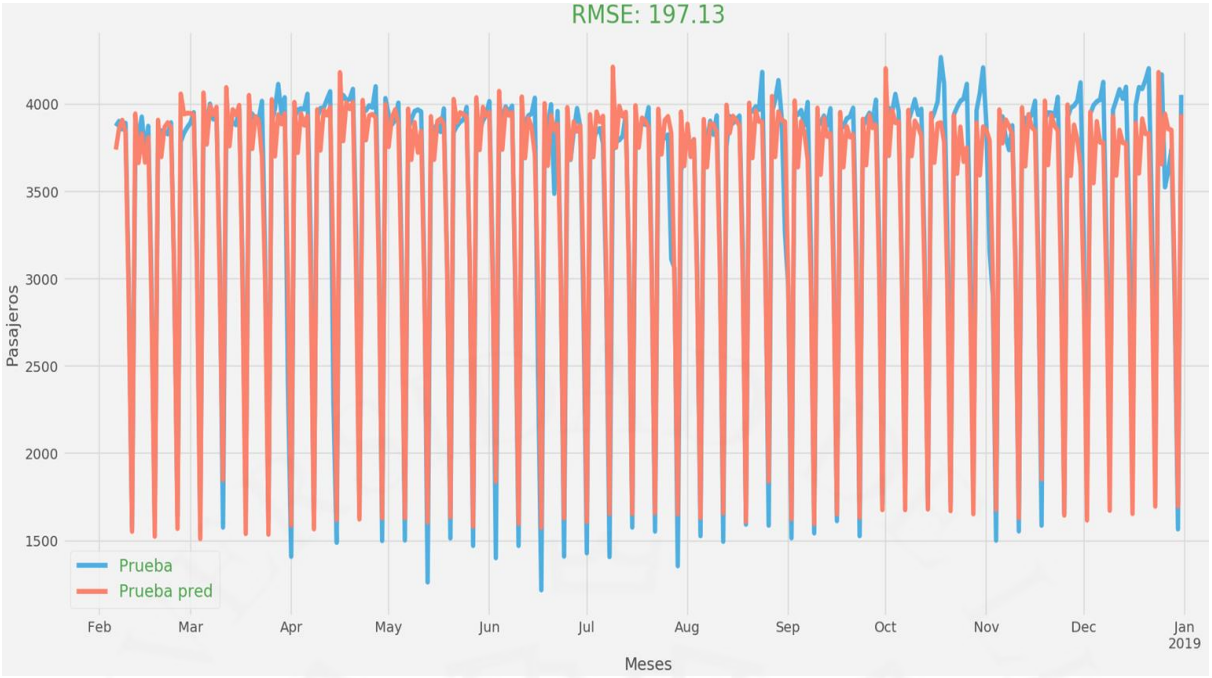


Figura 5.13
Comparativa entre los datos y predicciones de prueba del clúster 4

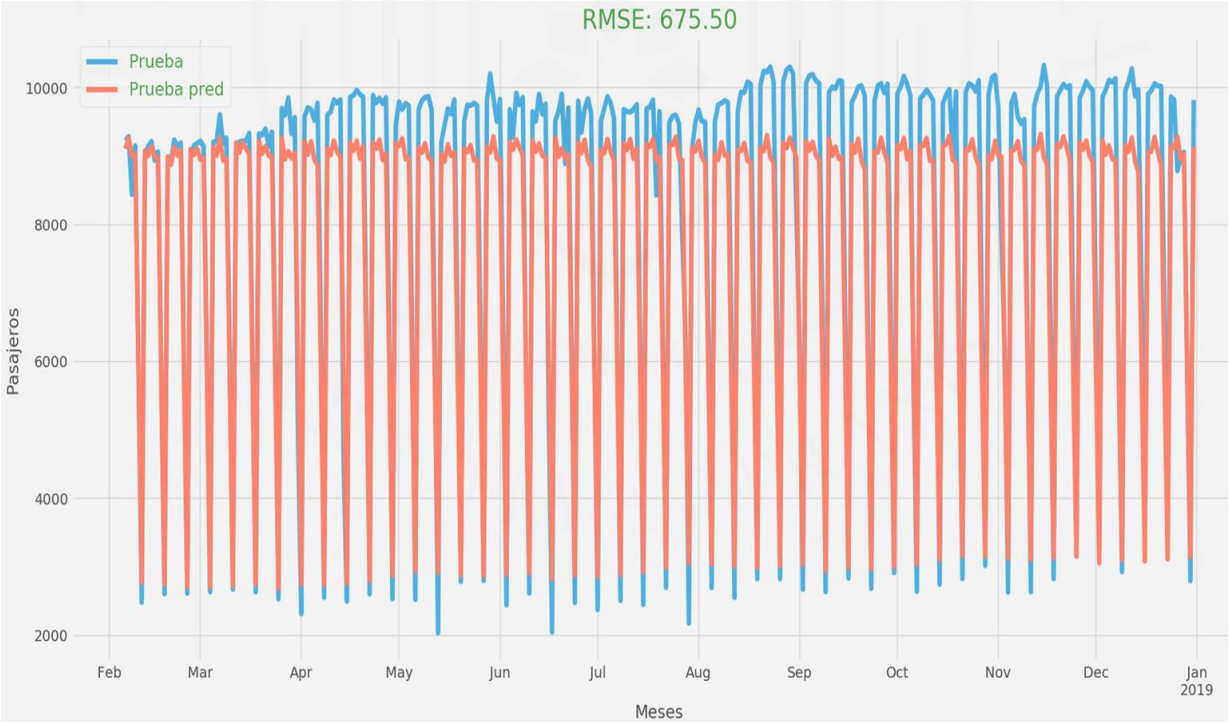
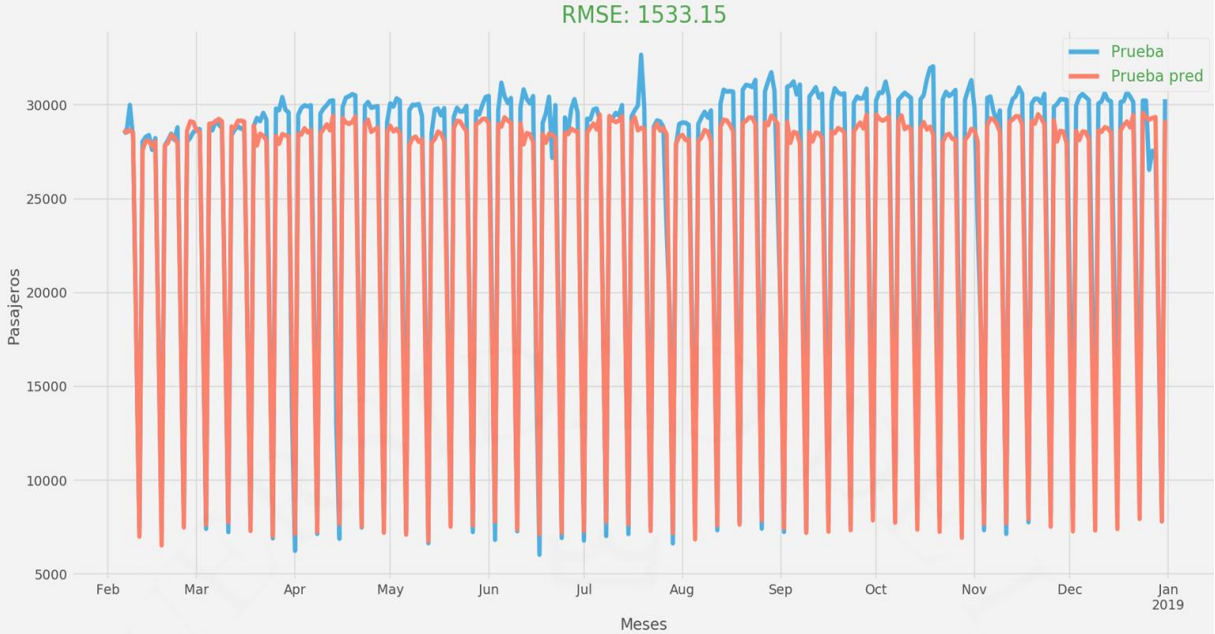


Figura 5.14
Comparativa entre los datos y predicciones de prueba del clúster 5



6. DISCUSIÓN

6.1 Análisis de patrones de clúster entre desde el 2016 y 2018:

A continuación, se presentará un análisis y la variación porcentual anual de pasajeros por clúster de estaciones:

Figura 6.1
Representación gráfica del número de pasajeros en cada clúster por año

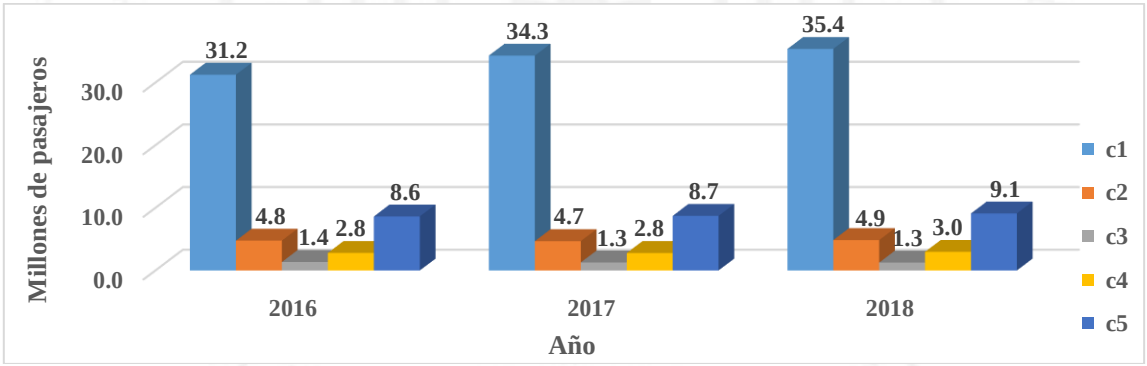


Tabla 6.1
Variación porcentual anual de pasajeros por clúster

	Naranjal	Cluster 2	Cluster 3	Cluster 4	Cluster 5
2016 al 2017	9.76%	-1.83%	-5.40%	-0.56%	1.00%
2017 al 2018	3.10%	3.70%	-1.66%	6.76%	4.52%

- **Estación Naranjal:** esta estación tiene una demanda muy alta y su comportamiento a través de los años indica que ninguna estación de las 37 presenta algo similar. Una de las razones es la capacidad de esta estación, ya que esta es más grande pues es una estación "Terminal", es decir, para los pasajeros que van al norte es la última estación, mientras los que van al sur es la primera. Otra razón es que se ubica entre Independencia y Comas, siendo este último uno de los cuatro distritos con mayor población de Lima Metropolitana, según el Instituto Nacional de Estadística e Informática (INEI) en el 2017, por esto llegan muchísimas personas de Coma a esta estación al no contar con estaciones en Comas o lugares cercanos.
- **Estaciones del clúster 2:** Las estaciones que agrupa este clúster están ubicadas en su mayoría en el centro de Lima y otras no como Canadá y Javier Prado, mostrando una disminución de pasajeros del 2016 al 2017 y un aumento del 2017 al 2018 como se muestra en la Tabla 6.1.
- **Estaciones del clúster 3:** Las estaciones que están en el clúster 3 están en una constante diminución de pasajeros

entre cada año de pasajeros llegando hasta menos de un millón de pasajeros por año como se muestra la Tabla 6.1.

- **Estaciones del clúster 4:** Las estaciones que agrupa este clúster se encuentran igual que el clúster 2 pero muestran una disminución menor del 2016 al 2017 y un aumento mayor entre del 2017 al 2018 como se muestra la Tabla 6.1.
- **Estaciones del clúster 5:** Las estaciones que agrupa este clúster tienen una gran demanda (ver Figura 6.1) en comparación a los demás clústeres pues agrupa estaciones con un nivel de demanda alto como Estación Central y Matellini y que aumentan al pasar los años, según la Tabla 6.1.

6.2 Análisis de patrones de Series de Tiempo de clústeres:

En las figuras 5.2 a la 5.5 se pueden apreciar las tendencias de cada clúster desde el 2016 hasta el 2018, la tendencia semanalmente entre los clústeres sigue un comportamiento de subida entre enero-julio y luego de julio-diciembre la tendencia es decreciente, este comportamiento se da en todos los clústeres cada año; la diferencia radica que, si se toman las series como un todo en los clústeres 2, 4 y 5 la tendencia es creciente mientras que en el clúster 3 es opuesta. En la Tabla 6.2 se aprecia la variación porcentual semanal por clúster mostrando que existe una estacionalidad semana a lo largo de los años, este comportamiento semanal tiende a disminuir entre un 0.43% y 1.85% de miércoles a jueves, pero disminuye aún más entre el viernes y domingo (fin de semana) entre un 23.89% hasta un poco más del 60%.

Tabla 6.2

Variación porcentual de pasajeros entre días de semana por clúster

	c2	c3	c4	c5
Lunes - Martes	0.25%	0.23%	0.56%	0.48%
Martes-Miércoles	0.46%	0.72%	0.81%	0.21%
Miércoles-Jueves	-0.80%	-0.84%	-1.85%	-0.43%
Jueves -Viernes	1.65%	2.57%	1.27%	0.47%
Viernes -Sábado	-34.90%	-23.89%	-34.30%	-36.62%
Sábado - Domingo	-59.57%	-49.83%	-58.43%	-60.35%

6.3 Análisis de resultados de modelamiento y predicción por clúster

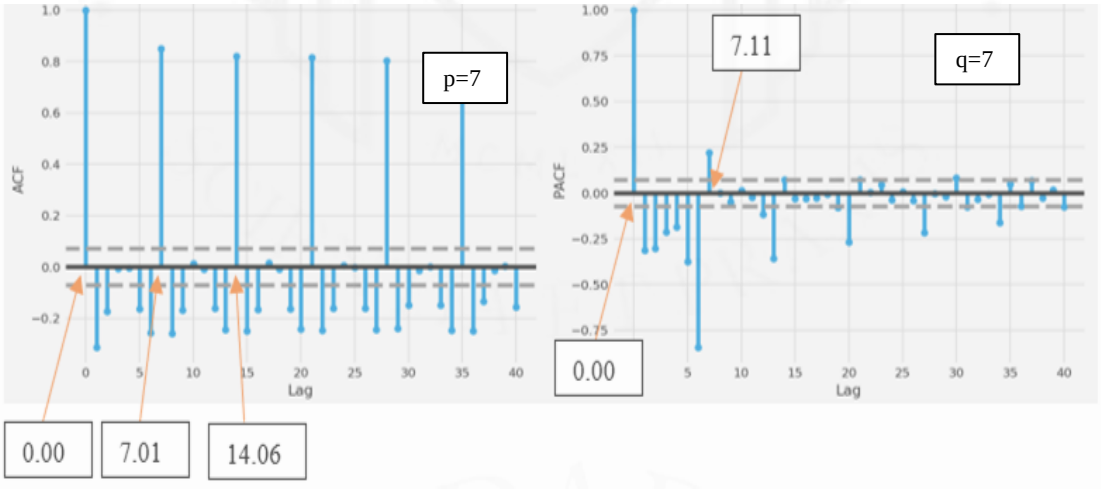
Verificación de estacionariedad: se puede apreciar en la Tabla 5.2 que todas las series de tiempo de clústeres de estaciones presentan estacionariedad al usar la prueba ADF, pues los valores de p son menores que 0.05 por lo tanto no hay la necesidad de hacer diferenciaciones a las series, es decir el parámetro *d* es 0.

Estimación de parámetros y mejores modelos por clúster:

Según lo mostrado en la Figura 5.10, en todos los clústeres en el ACF el parámetro *p* vale 7 porque se obtienen valores altos de líneas azules cada 7 *lag* o retrasos y en el PACF *q* vale también 7, se usó el SARIMA automático para elegir los mejores modelos (ver Tabla 5.4), pues obtuvieron el menor AIC y RMSE por clúster.

Figura 6.2

ACF y PACF de clústeres



Análisis de predicciones de pasajeros por clúster:

Se puede apreciar en las figuras 5.11 y 5.12 que la demanda de prueba y real de los clústeres 2 y 4 no muestran una buena semejanza gráficamente, es decir, no son precisos, ya que estos agrupan estaciones ubicadas en el centro de Lima y estaciones como Canadá o Javier Prado que son afectados por eventos deportivos o días festivos que alteran el comportamiento normal de la semana como se pueden apreciar en las figuras 5.2 y 5.4 donde en los meses festivos hay un cambio drástico en la demanda; mientras que en los clústeres 3 y 5 agrupan estaciones que están lejos del centro de Lima y estos eventos no las afectan mostrando mucha similitud entre los valores reales y predichos de

prueba. En la Tabla 6.3 se muestran algunos datos estadísticos entre la demanda de prueba real y predicha, corroborando lo que se ha mencionado anteriormente, mostrando que los clústeres 3 y 5 una menor variación de los datos estadísticos entre la demanda de prueba real y predicha que los otros clústeres.

Tabla 6.3
Datos estadísticos de la demanda de prueba real y predicha por clúster

Clúster	Demanda de prueba	Predicción de prueba
2	Promedio: 13364.83 Desviación estándar: 4276.98	Promedio: 12313.61 Desviación estándar: 4015.76
3	Promedio: 3459.11 Desviación estándar: 840.25	Promedio: 3427.46 Desviación estándar: 798.54
4	Promedio: 8198.55 Desviación estándar: 2582.05	Promedio: 7743.74 Desviación estándar: 2256.10
5	Promedio: 25060.83 Desviación estándar: 8173.67	Promedio: 24107.70 Desviación estándar: 7816.36

6.4 Validación de Resultados

Análisis de precisión de los modelos:
Se validaron los modelos como se muestra en la Tabla 6.4, con el fin de asegurarse de que los valores predichos, en este caso demanda de pasajeros del 2019, sean confiables. Para esto se unió la demanda de pasajeros del 2019 con la que ya se tuvo inicialmente, esta demanda ya estará por clúster de estaciones, así como se trabajó en un principio. Se usaron métricas de pronóstico como el RMSE, RMSE porcentual (RMSPE), el error absoluto medio (MAE), MAPE y R^2 para verificar la precisión de los modelos, estos fueron considerados ya que en varios artículos presentados en el estado del arte los utilizaron para validar los modelos que implementaron.

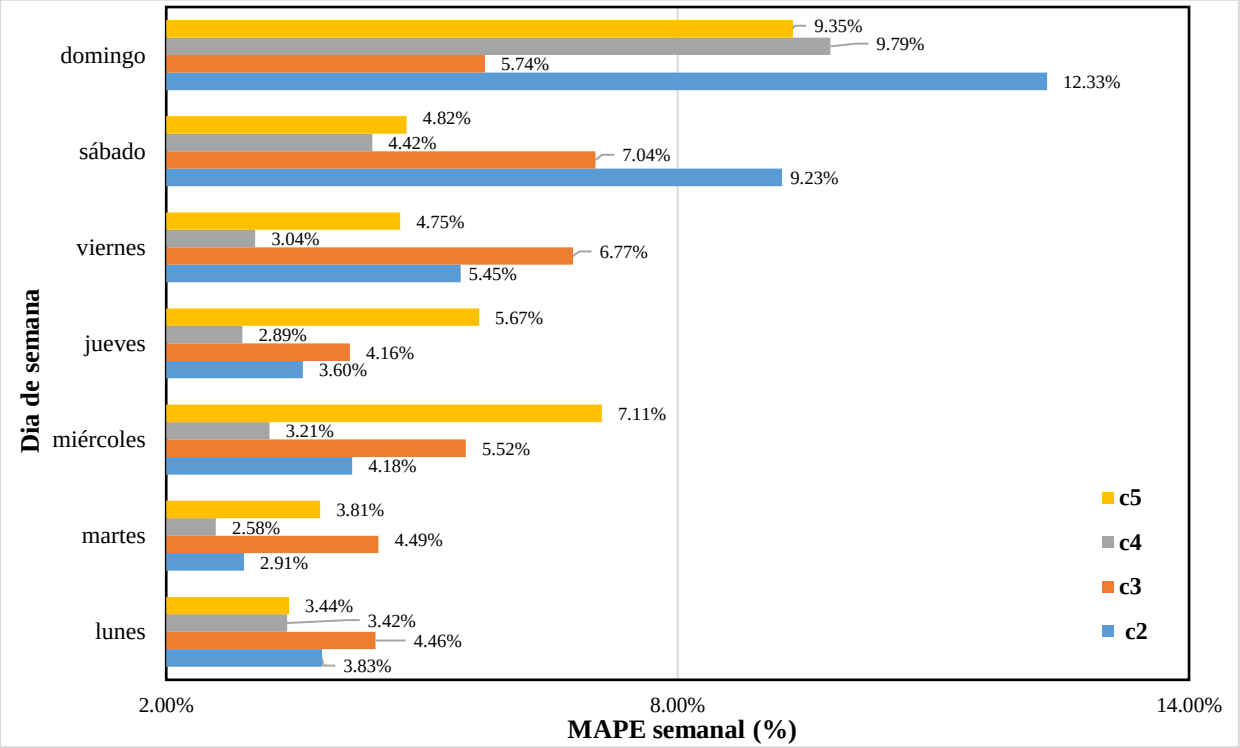
Tabla 6.4
Métricas de pronóstico al predecir el 2019 por clúster

Clúster	RMSE	RMSPE	MAE	MAPE	R^2
2	849.08	8.13 %	681.93	5.93 %	0.96
3	248.84	6.92 %	192.49	5.45 %	0.91
4	377.00	6.37 %	288.86	4.19 %	0.97
5	1588.49	6.99 %	1285.65	5.56 %	0.96

En la Tabla 6.4, se puede apreciar que los modelos presentan una buena precisión siendo el de mayor precisión el modelo del clúster 4 pues tiene un MAPE y R^2 de 4.19 % y 0.97 respetivamente. En la evaluación el RMSE y MAE no se pueden definir valores límites porque varían según el clúster pues cada uno tiene diferente nivel de demanda, el valor alto del RMSE que tiene el clúster 5 no significa que el modelo no sea el adecuado, sino que esto se debe a que cuenta con una demanda superior a los demás clústeres como se puede apreciar en la Figura 6.2 y es lógico que cuente con un RMSE y MAE mayor que los demás clústeres. Para evitar esta confusión el RMSPE, MAPE y R^2 ayudaron a disipar toda duda sobre el rendimiento de los modelos, ya que las dos primeras métricas muestran lo que representa el error porcentualmente y el último la correlación entre lo real y predicho. Estas tres métricas de pronóstico fueron bajos en comparación con los estudios realizados por Milenković et al, (2016), Ma et al., (2014), Gong et al., (2014) y Ni et al., (2016), ya que los valores del RMSPE estuvieron entre un 6.37% y 8.13%, el MAPE entre un 4.19% y 5.93% y el R^2 entre el 0.91 y 0.97 mostrando una buena correlación entre la demanda real y predicha del 2019.

Análisis del error porcentual absoluto medio (MAPE) semanal por clúster:
Como se puede apreciar en la Figura 6.3, el MAPE semanal llega un máximo del 12.33% en el domingo del 2019 siendo el modelo del clúster 2 que arroja este valor y el menor lo registra el modelo del clúster 4 en el viernes con un 2.58%, también se puede observar que el MAPE tiene un gran nivel en el fin de semana, ya que se detectó que la gran mayoría de días festivos durante los años 2016 al 2018 se dieron en los fines de semana y días cercanos a estos alterando su comportamiento normal, a pesar de que se hayan reemplazado estos valores indicando que aún hay fechas que muestran un comportamiento que los modelos aún lo consideran como atípicos y como consecuencia no pueden predecir con exactitud con un MAPE entre el 9.23% y 12.33%. Este análisis también es aplicado en específico en los clústeres 2 y 4 mostrando un mayor MAPE que los demás, pues como se explicó en la sección "Análisis de resultados de modelamiento y predicción por clúster", estos agrupan estaciones que son afectadas por días festivos, partidos de fútbol o conciertos alterando el comportamiento semanal incluso así se hayan reemplazado muchos de estos valores atípicos que los modelos consideraron, no solo estos clústeres agrupan estas estaciones sino el clúster 3 y 5 pero en menor medida por ello también registra algunos días un MAPE alto.

Figura 6.3
MAPE por día de la semana por clúster



7. CONCLUSIONES

Los modelos SARIMA mostraron buenos resultados pues el RMSPE, MAPE y R^2 estuvieron entre un 6.37% - 8.13%, 4.19% - 5.93% y 0.91-0.98 respectivamente entre los cuatro modelos estando dentro de los límites que se propusieron como objetivos en el trabajo de investigación. Estos modelos se basaron a partir de un análisis de patrones en espacio y tiempo usando dos métodos de data mining, estos métodos fueron de gran ayuda en el trabajo de investigación pues la aplicación del clustering facilitó a la observación de comportamientos similares de demanda de estaciones en el tiempo, descartando el supuesto que solo estaciones ubicadas en lugares cercanos tienen un comportamiento similar, ya que al algoritmo de clustering agrupó estaciones que no necesariamente estaban muy cerca. Además, este método redujo y fue tomado como la mejor forma que abordaría el trabajo en el análisis temporal pues si se tomaban las 38 estaciones individualmente se hubieran desarrollado un modelo SARIMA por cada una lo que significa un trabajo mucho más arduo a predecir y analizar el comportamiento del error y, por último, estos patrones de comportamiento por clúster, ayudó luego a encontrar razones del comportamiento del MAPE semanal de los modelos de algunos clústeres en específico. El segundo método ayudó a encontrar el comportamiento semanal en tendencia y estacionalidad de cada clúster, también se pudieron encontrar comportamientos atípicos que en un primer vistazo en la limpieza de datos no se pudieron detectar y fueron reemplazados.

Ahondando en los modelos de predicción, se tuvieron algunos problemas con algunos de los valores atípicos que no fueron considerados en un principio como tales para los modelos, ya que estos valores se dan en fechas festivas fijas anualmente, por lo tanto no se pueden considerar como como outliers; sin embargo, los modelos no lo interpretaron de esta forma y no pudieron predecir estas fechas correctamente, el motivo fue que las series de tiempo tienen una estacionalidad semanal y no anual, y los modelos no tomaron en cuenta estos días festivos fijos anuales, por lo tanto se procedió a reemplazar la demanda de estos días para que los modelos obtuvieron buenos resultados. A pesar de que los modelos muestran buenos resultados es importante resaltar este problema porque el modelo debería asimilar este comportamiento en cada año y no solo semanalmente, pues estos “outliers” son fijos en cada año. Finalmente, las predicciones de estos modelos pueden ser usadas por cualquier sistema de transporte público como el Metropolitano para la actualización de la planificación de demanda de pasajeros semanalmente para redistribuir sus buses de acuerdo con el comportamiento semanal actual que la demanda de pasajeros por clúster cuenta y así atender adecuadamente la demanda que satura sus estaciones en todo el año, pero en lo que respecta a los días festivos los modelos no serán precisos.

8. TRABAJOS FUTUROS

Los modelos SARIMA desarrollados mostraron ser muy precisos, pero tuvieron algunos problemas con estacionalidades múltiples de las series temporales, anual y semanalmente. Ante este problema existen modelos

alternativos que han sido aplicados en la última década para la predicción de la cantidad pasajeros en el transporte público y que podrían abordar el problema de los días festivos anuales que se dieron en simultáneo con la estacionalidad semanal en las series de tiempo de cada clúster. Estos modelos tienen un enfoque basado en redes neuronales (ANN) y han mostrado un buen rendimiento comparado a los modelos ARIMA y sus variantes, pues no solo muestran un bajo nivel de error, sino que pueden detectar regularidades y patrones en los datos aprendiendo a través de iteraciones consiguiendo experiencia para luego pronosticar basados en el conocimiento previo (Adhikari, R., & Agrawal, R.,2013). Existe un tipo de redes neuronales artificiales estacionales (RNAS) o (SANN) que no requiere un pre-procesamiento de datos extenso y sería de gran ayuda pues detectaría rápidamente patrones estacionales diferentes, en este caso semanal y anual que se encontraron en las series de tiempo de cada clúster sin la necesidad de eliminar valores atípicos.

Por otro lado, con el objetivo contar con un análisis más granulado se debería contar con información de demanda diaria por horas, como algunas investigaciones citadas tuvieron, esto ayudaría a un análisis en el tiempo más detallado pues dependiendo de la hora y día no se registra el mismo nivel de pasajeros; además se podrían encontrar patrones por hora en espacio y tiempo al usar los métodos de data mining propuestos.

AGRADECIMIENTOS

Dedicado a mi madre, tíos y pareja, por todo el esfuerzo y apoyo que me dieron, y a todos aquellos amigos que, de una forma u otra, me ayudaron en el camino tan largo que ha significado la realización de este trabajo con mucha pasión y entrega.

También agradezco de forma especial a la profesora Rosa Fátima Medina Merino, que fue mi primera asesora de tesis y la que me impulsó en aplicar los métodos y ahondar más en el análisis para que este trabajo de investigación. Que en paz descanse.

REFERENCIAS

Adhikari, R., y Agrawal, R. (2013). An Introductory Study on Time Series Modeling and Forecasting. *ArXiv, abs/1302.6613*, 67.

Agencia de Cooperación Internacional del Japón (JICA). (2005). *Plan Maestro de Transporte Urbano para el Área Metropolitano de Lima y Callao en la República del Perú (Fase 1)* (pp. 1- 86). Lima-Perú. Recuperado de https://openjicareport.jica.go.jp/pdf/11798261_01.pdf

Amutha, R., y Renuka, K. (2015). Different Data Mining Techniques and Clustering Algorithms. *International Journal of Technology Enhancements and Emerging Engineering Research*, *3*(11), 15-17.

Anvari, S., Tuna, S., Canci, M., y Turkay, M. (2015). Automated Box–Jenkins forecasting tool with an application for passenger demand in urban rail systems. *Journal of Advanced Transportation, 50(1),* 25-49. doi:10.1002/atr.1332

Briand, A., Côme, E., El Mahrsil, M. K., y Oukhellou, L. (2016). A mixture model clustering approach for temporal passenger pattern characterization in public transport. *International Journal of Data Science and Analytics, 1*(1), 37-50. doi: 10.1007/s41060-015-0002-x

Box, G. y Jenkins, G. (1970). *Time Series Analysis, Forecasting and Control.* San Francisco.

Califiski, T., y Corsten, L. (1985). Clustering Means in ANOVA by Simultaneous Testing. *BIOMETRIC, 41*(1), 39.48. doi:10.2307/2530641

Cyprich, O., Konečný , V. y Kiliánová, K. (2013). Short-Term Passenger Demand Forecasting Using Univariate Time Series Theory. *Promet-Traffic & Transportation, 25*(6), 533-541. doi:10.7307/ptt.v25i6.338

Cyril, A., Raviraj, H., y Varghese, G. (2018). Modelling and Forecasting Bus Passenger Demand. *7th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, (págs. 460-466). Noida-India. doi:10.1109/ICRITO.2018.8748443

Cyril, A., Raviraj , M., y Varghese, G. (2019). Bus Passenger Demand Modelling Using Time-Series Techniques Big Data Analytics. *The Open Transportation Journal, 13*(1), 41-47. doi:10.2174/1874447801913010041

Dou, F. D., Xu, J., Wang, L., & Jia, L. (2013). A train dispatching model based on fuzzy passenger demand forecasting during holidays. *Journal of Industrial Engineering and Management*, *6*(1), 320-335. doi:10.3926/jiem.699

Gong, M., Fei, X., Wang, Z., y Qiu, Y. (2014). Sequential Framework for Short-Term Passenger Flow Prediction at Bus Stop. *Transportation Research Record Journal of the Transportation Research Board, 2417*(1), 58-66. doi:10.3141/2417-07

González Castellanos, M., y Soto Valero, C. (2013). Minería de Datos para series Temporales. Santa Clara: Feijó.

Hamzaçebi, C. (2008). Improving artificial neural networks performance in seasonal time series forecasting. *Information Sciences, 178*(23), 4550-4559. doi:10.1016/j.ins.2008.07.024

Hipel, K. y McLeod, A. (1994). *Time series modelling of water resources and environmental systems*. Amsterdam New York: Elsevier.

Lee, J., Yoo, S., Kim, H., y Chung, Y. (2018). The spatial and temporal variation in passenger service rate and its impact on train dwell time: A time-series clustering approach using dynamic time warping. *International Journal of Sustainable Transportation, 12*(10), 1-12. doi:10.1080/15568318.2018.1432731

Ma, Z., Xing, J., Mesbah, M., y Ferreira, L. (2014). Predicting short-term bus passenger demand using a pattern hybrid approach. *Transportation Research Part C: Emerging Technologies, 39*, 148-163. doi:10.1016/j.trc.2013.12.008

Milenković, M., Švadlenka, L., Melichar, V., Bojović, N., y Avramović, Z. (2016). SARIMA modelling approach for subway passenger flow forecasting. *Transport, 33*(5), 1113-1120. doi:https://doi.org/10.3846/16484142.2016.1139623

Municipalidad Metropolitana de Lima (MML). (2017). *Lima Como Vamos - Observatorio Ciudadano.* Obtenido de http://www.limacomovamos.org/cm/wp-content/uploads/2018/03/EncuestaLimaCómoVamos_2017.pdf

Ni, M., He, Q., y Gao, J. (2017). Forecasting the Subway Passenger Flow Under Event Occurrences With Social Media. *IEEE Transactions on Intelligent Transportation Systems, 18*(6), 1623-1632. doi:10.1109/TITS.2016.2611644

ProTransporte. (2018), *ProTransporte*. Obtenido de http://www.protransporte.gob.pe/metropolitano/

Raicharoen, T., Lursinsap, C., y Sanguanbhokai, P. (2003). Application of critical support vector machine to time series prediction. *Conference: Circuits and Systems, 2003. ISCAS '03. Proceedings of the 2003 International Symposium, 5,* págs. 741-744. Bangkok, Thailand. doi:10.1109/ISCAS.2003.1206419

Ramesh Reddy, J., Ganesh, T., Venkateswaran, M., y Reddy, P. (2017). Forecasting of Monthly Mean Rainfall in. *International Journal of Current Research and Review, 7*(4), 197-204. doi:10.5923/j.statistics.20170704.01

Ryu, T., y Eick, C. (2005). A database clustering methodology and tool. *Information Sciences: An International Journal*, *171(1),* 29-59, doi: 10.1016/j.ins.2004.03.016

Tsai, C., Mulley, C., y Clifton, G. (2013). Forecasting public transport demand for the Sydney Greater Metropolitan Area: a comparison of univariate and multivariate methods. *Australasian Transport Research Forum (ATRF), 36th, 2013, Brisbane, Queensland, Australia.* Obtenido de https://www.australasiantransportresearchforum.org.au/

Viera, L., Ortiz, L., y Ramírez, S. (2009). Introducción a la Minería de Datos. Río de Janeiro - Brazil: Ltda.

Ward, J. H. (1963) Hierarchical grouping to optimize an objective function. . J. Amer. Statist. Assoc., 58, 236-244.

Wang, Y., Han, B., Zhang, Q., y Li, D. (2015). Forecasting of Entering Passenger Flow Volume in Beijing Subway Based on SARIMA Model. *Journal of Transportation Systems Engineering and Information Technology, 15*(6)*,* 205-211

Wei, Y., y Chen, M. (2012). Forecasting the short-term metro passenger flow with empirical mode decomposition and neural networks. *Transportation Research Part C: Emerging Technologies, 21*(1), 148-162. doi:10.1016/j.trc.2011.06.009

Xue, R., Sun, D., y Chen, S. (2015). Short-Term Bus Passenger Demand Prediction Based on Time Series Model and Interactive Multiple Model Approach. *Discrete Dynamics in Nature and Society, 2015*, 1-11. doi:10.1155/2015/682390

Yaffee, R., y McGee, M. (2000). An Introduction to Time Series Analysis and Forecasting: with Applications of SAS and. Academic.

Yan, D., Zhou, J., Zhao, Y., y Wu, B. (2018). *Short-Term Subway Passenger Flow Prediction Based on ARIMA. Geo-Spatial Knowledge and Intelligence, 464–479.* doi:10.1007/978-981-13-0893-2_49

Zhang, Y., Cheng, T., y Sari Aslam, N. (2019). Exploring the relationship between travel pattern and social-demographics using smart card data and household survey. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *42*, págs. 1375-1382. Enschede, The Netherlands doi:10.5194/isprs-archives-XLII-2-W13-1375-2019