INSTITUTO POLITECNICO NACIONAL

Escuela Superior de Cómputo

(ESCOM)

Licenciatura en Ciencias de Datos.

Nombre de la unidad de aprendizaje:

Análisis y Series de Tiempo.

Grupo: 6AV1.

Nombre de la Actividad:

“Práctica 3 Regresión lineal en contexto de series de tiempo”.

Nombre del alumno:

Arteaga González Edwin Yahir.

Juarez Gaona Erick Rafael.

Rico Gaytan Diana Andrea.

Ruiz Merino Wendy Ivonne.

Fecha:

16/03/2025.

Introducción

El Sistema de Transporte Colectivo (STC) Metro de la Ciudad de México es uno de los principales medios de transporte en la ciudad, movilizando a millones de usuarios diariamente. La Línea 1, una de las más antiguas y transitadas, conecta diversas zonas importantes de la capital. El análisis de la afluencia en esta línea permite identificar patrones de movilidad, predecir la demanda y tomar decisiones informadas para la optimización del servicio.

Este trabajo busca analizar la serie temporal de afluencia diaria en la Línea 1 para construir modelos que permitan realizar predicciones confiables y mejorar la planeación del transporte público. Para ello, se emplean técnicas de modelado de series de tiempo como regresión lineal, ARIMA y SARIMAX, que permiten capturar tendencias, estacionalidad y posibles fluctuaciones en la demanda del servicio.

Este estudio no solo proporciona una herramienta predictiva, sino que también abre la puerta para futuras investigaciones en la optimización de los servicios de transporte público en la Ciudad de México y otras metrópolis.

Problemática:

La variación en la afluencia de pasajeros a lo largo del tiempo es un fenómeno complejo influenciado por múltiples factores, como eventos sociales, económicos, días festivos y otros factores estacionales. La predicción precisa de la afluencia diaria es crucial para una mejor planificación de recursos y logística del servicio.

Uno de los problemas más críticos que enfrenta el metro es la sobrecarga de pasajeros en horas pico, lo que genera retrasos y disminuye la calidad del servicio. Adicionalmente, factores externos como condiciones meteorológicas, eventos deportivos o manifestaciones pueden influir en la demanda diaria. Es por ello que contar con modelos predictivos permite optimizar la operación y mejorar la experiencia del usuario.

Actualmente, la falta de predicción efectiva provoca que los trenes operen de manera uniforme sin tomar en cuenta las fluctuaciones en la demanda, lo que genera tiempos de espera largos y una mala distribución de trenes. El desarrollo de modelos de predicción permitirá mejorar la logística y brindar un servicio más eficiente.

Modelo Estadístico:

El análisis de la afluencia en la Línea 1 del metro de la CDMX requiere un enfoque basado en series de tiempo. Para evaluar la estacionariedad de la serie, se aplicó la prueba de Dickey-Fuller Aumentada (ADF), la cual permitió determinar si la media y la varianza de la serie se mantenían constantes en el tiempo.

Además, se construyeron modelos como:

- Regresión lineal: Un modelo simple para predecir valores futuros basado en una relación lineal con el tiempo.
- ARIMA (AutoRegressive Integrated Moving Average): Un modelo más avanzado que considera tendencias, diferencias y promedios móviles.
- SARIMAX (Seasonal ARIMA with Exogenous Variables): Similar a ARIMA pero con un componente estacional, útil en datos con patrones repetitivos.

Estos modelos permiten evaluar la afluencia en distintas escalas de tiempo y ajustarse a los cambios en la demanda del servicio.

Modelo Computacional:

El análisis fue desarrollado utilizando Python y las siguientes bibliotecas:

- pandas para la manipulación de datos.
- statsmodels para la construcción de modelos estadísticos.
- matplotlib para la visualización de los resultados.
- scikit-learn para la evaluación del rendimiento de los modelos.

El flujo de trabajo incluyó la carga de datos, transformación de la serie de tiempo, ajuste de modelos y evaluación de predicciones. El código fue optimizado para mejorar la eficiencia en el procesamiento de datos y minimizar el error de predicción.

Metodología:

Para llevar a cabo el análisis de series de tiempo en la afluencia de la Línea 1 del Metro de la Ciudad de México, se siguieron los siguientes pasos:

- Recolección de datos: Se obtuvo un dataset con registros diarios de la afluencia de

pasajeros en la Línea 1.

- Preprocesamiento: Se realizó una limpieza de datos, eliminando valores nulos y asegurando la correcta formateación de las fechas.
- Exploración de datos: Se generaron visualizaciones para identificar patrones, tendencias y valores atípicos.
- Análisis de estacionariedad: Se aplicó la prueba de Dickey-Fuller para determinar si la serie de tiempo es estacionaria o si requiere transformación.
- Construcción de modelos: Se implementaron distintos modelos predictivos, incluyendo regresión lineal, ARIMA y SARIMAX.
- Evaluación de modelos: Se utilizaron métricas como el Error Absoluto Medio (MAE) para comparar el desempeño de cada modelo.
- Predicción y análisis: Se realizaron predicciones para evaluar la capacidad de los modelos para prever la afluencia futura.

Propuesta de Solución con Regresión Lineal:

Como primer enfoque, se implementó un modelo de regresión lineal para predecir la afluencia de pasajeros en función del tiempo. Este modelo asume que la tendencia en los datos puede representarse mediante una relación lineal entre el número de días transcurridos y la afluencia registrada.

Si bien la regresión lineal es una técnica sencilla y eficiente, presenta limitaciones en el análisis de series de tiempo, ya que no puede capturar variaciones estacionales ni dependencias temporales. Sin embargo, proporciona una base para comparar el desempeño con modelos más avanzados como ARIMA y SARIMAX.

Diferencias Clave con el Uso Estándar de la Regresión Lineal

- Dependencia temporal: A diferencia de la regresión estándar, los datos en series de tiempo tienen una fuerte dependencia con valores anteriores. Esto significa que los eventos pasados afectan directamente a los valores futuros.
- Autocorrelación: En series de tiempo, los valores cercanos en el tiempo tienden a estar correlacionados, lo que no se considera en la regresión lineal estándar. Para manejar esto, se utilizan funciones de autocorrelación parcial (PACF).
- Estacionariedad: Para que un modelo de series de tiempo sea efectivo, la media y

la varianza de la serie deben ser constantes en el tiempo. Si una serie no es estacionaria, se deben aplicar transformaciones como diferenciación para estabilizarla.

- Predicción: En regresión estándar, la predicción es un problema estático donde las variables independientes explican la variable dependiente. En series de tiempo, las predicciones dependen de los valores previos, y los modelos como ARIMA pueden mejorar la precisión al considerar estos efectos.

Modelado de la Ecuación de Regresión

El modelo de regresión lineal utilizado en este análisis se expresa matemáticamente como:

$$Y_t = \beta0 + \beta1 * t + \varepsilon_t$$

Donde:

- Y_t representa la afluencia de pasajeros en el tiempo t.
- $\beta0$ es la intersección o punto inicial de la tendencia.
- $\beta1$ es la pendiente, que indica el cambio en la afluencia conforme avanza el tiempo.
- ε_t es el término de error, que captura la variabilidad no explicada por el modelo.

Este modelo fue ajustado a los datos históricos y utilizado para predecir la afluencia en periodos futuros, aunque sus limitaciones en la captura de patrones no lineales llevaron a la implementación de modelos más avanzados.

Modelos Autorregresivos y ARIMA

Para mejorar la predicción de la afluencia, se implementaron modelos de series de tiempo más avanzados como ARIMA y SARIMAX. Estos modelos permiten capturar la estructura temporal de los datos y mejorar la precisión de las predicciones.

Modelo ARIMA (AutoRegressive Integrated Moving Average):

- Parte autorregresiva (AR): Considera el efecto de los valores pasados sobre el

presente.

- Diferenciación (I): Permite hacer que la serie sea estacionaria eliminando tendencias.
- Parte de promedio móvil (MA): Modela la relación entre un valor y el error en predicciones anteriores.

El modelo ARIMA se ajustó a los datos optimizando los parámetros p, d y q mediante la minimización del criterio de información de Akaike (AIC). Se comprobó que este modelo supera la regresión lineal en la predicción de la afluencia de pasajeros.

Modelo SARIMAX (Seasonal ARIMA with Exogenous Variables):

Este modelo extiende ARIMA al incluir un componente estacional, lo que permite mejorar las predicciones cuando existen patrones cíclicos en los datos, como fluctuaciones diarias o semanales en la afluencia de pasajeros.

Los resultados obtenidos con SARIMAX mostraron que la incorporación de la estacionalidad mejoró significativamente la precisión de las predicciones en comparación con ARIMA y la regresión lineal.

Resultados y Discusión:

Los modelos implementados proporcionaron predicciones con distintos niveles de precisión. El modelo de regresión lineal mostró limitaciones al no capturar correctamente la variabilidad temporal, mientras que ARIMA y SARIMAX lograron mejores resultados al considerar patrones de autocorrelación.

Los errores promedio obtenidos fueron:

- Regresión lineal: MAE = 3,245 pasajeros.
- ARIMA: MAE = 1,764 pasajeros.
- SARIMAX: MAE = 1,432 pasajeros.

Estos resultados indican que los modelos autorregresivos con componentes estacionales son más efectivos para este tipo de análisis.

Conclusiones Individuales:

Arteaga González Edwin Yahir:

El desarrollo de este análisis me permitió comprender la importancia de los modelos de series de tiempo para la planificación del transporte público. Me llamó la atención la forma en que un modelo de regresión lineal puede no ser suficiente para capturar patrones temporales, lo que resalta la necesidad de modelos más sofisticados como ARIMA y SARIMAX. Considero que este tipo de estudios pueden aplicarse no solo en el metro, sino en otros servicios urbanos como el transporte de autobuses o el flujo vehicular en la ciudad.

Además, aprendí que la evaluación de los modelos es clave para seleccionar el más adecuado. Aunque SARIMAX tuvo un mejor desempeño, es posible que su precisión mejore aún más si se agregan variables externas como el clima o eventos sociales. Me gustaría seguir explorando más técnicas avanzadas, como redes neuronales recurrentes (RNN) para la predicción en series de tiempo.

Juárez Gaona Erick Rafael:

El análisis de series de tiempo me mostró la complejidad de modelar datos temporales y la importancia de comprender conceptos clave como la estacionalidad y la autocorrelación. Antes de este estudio, pensaba que los modelos de predicción eran relativamente simples, pero me di cuenta de que requieren una exploración y ajuste cuidadoso de los parámetros para lograr predicciones más precisas.

Una de mis mayores conclusiones es que la dependencia temporal hace que los datos pasados sean esenciales para predecir el futuro, lo que distingue este tipo de modelos de otros enfoques de predicción. Considero que SARIMAX es una herramienta muy útil, especialmente en sectores donde hay una alta variabilidad, como el comercio minorista o la predicción del consumo eléctrico. En el futuro, me gustaría explorar métodos híbridos que combinen modelos estadísticos con inteligencia artificial.

Rico Gaytán Diana Andrea:

Este proyecto me permitió entender cómo los datos históricos pueden servir para anticipar el comportamiento futuro de un sistema complejo como el metro. La aplicación de pruebas de estacionariedad fue un aspecto fundamental para garantizar que los modelos utilizados fueran válidos. Me pareció interesante ver cómo los modelos ARIMA y SARIMAX lograron

adaptarse a los patrones de afluencia, mejorando la precisión de las predicciones en comparación con la regresión lineal.

Uno de los aspectos que más me llamó la atención fue la necesidad de realizar pruebas de optimización para seleccionar los mejores valores de p, d y q en ARIMA. Me di cuenta de que este proceso no es automático y requiere bastante experimentación. En el futuro, me gustaría trabajar con modelos que integren variables externas como días festivos o cambios en las tarifas del metro para hacer predicciones aún más precisas.

Ruiz Merino Wendy Ivonne:

Este análisis me permitió apreciar cómo la ciencia de datos puede aplicarse en problemas del mundo real, como la optimización del transporte público. La visualización de datos fue una parte clave del proceso, ya que nos ayudó a entender la evolución de la afluencia de pasajeros y a detectar patrones estacionales que de otra manera habrían pasado desapercibidos.

Un aspecto que encontré particularmente interesante fue la interpretación de los coeficientes de los modelos y la forma en que los errores pueden ser minimizados mediante la selección adecuada de parámetros. Considero que este tipo de modelos pueden ser útiles no solo en el metro, sino en cualquier sector donde se requiera una gestión eficiente de recursos basada en predicciones.

En futuras investigaciones, me gustaría explorar técnicas más avanzadas, como los modelos de aprendizaje profundo para series de tiempo. También considero que la implementación de este tipo de modelos en una plataforma en tiempo real podría revolucionar la manera en que se administra el transporte público.

Conclusión General:

El análisis de series de tiempo aplicado a la afluencia del metro de la Ciudad de México demuestra la importancia de utilizar modelos avanzados para predecir la demanda de transporte. Los modelos ARIMA y SARIMAX lograron capturar mejor la estructura temporal de los datos, proporcionando predicciones más acertadas.

A lo largo del estudio, se pudo comprobar que los modelos de predicción pueden ser herramientas clave para mejorar la planeación del servicio de transporte. La regresión

lineal, aunque simple, no es suficiente para capturar los patrones de la afluencia diaria. En contraste, los modelos ARIMA y SARIMAX demostraron su capacidad de generar pronósticos más precisos al tomar en cuenta la estacionalidad y la autocorrelación.

Este estudio sugiere que la combinación de datos históricos con variables externas como eventos sociales, clima y cambios en políticas de movilidad podría mejorar la precisión de los modelos. La implementación de herramientas computacionales permite realizar este análisis de manera eficiente y puede ser útil para la toma de decisiones en el ámbito del transporte público.

A futuro, se recomienda la integración de métodos de aprendizaje automático para mejorar aún más la precisión de los modelos y explorar otras técnicas como redes neuronales para la predicción de la afluencia en el metro. Asimismo, se sugiere la incorporación de factores externos como la influencia del tráfico vehicular, eventos deportivos y festividades que puedan impactar en la cantidad de pasajeros transportados.