INSTITUTO POLITÉCNICO NACIONAL
ESCUELA SUPERIOR DE CÓMPUTO
PROGRAMA ACADÉMICO DE LICENCIATURA EN CIENCIA DE DATOS

Practica 5.0 Construcción modelo ARIMA

PRESENTA: GERARDO SUÁREZ SANTOS

TITULAR DEL CURSO: DANIEL JIMÉNEZ ALCANTAR

NOMBRE DEL CURSO: ANALISIS DE SERIES DE TIEMPO

Contenido

Introducción

El análisis de series de tiempo es una herramienta fundamental en la ciencia de datos que permite modelar datos que evolucionan con el tiempo. En la práctica, muchas áreas como la economía, la meteorología, la salud, las finanzas y, en este caso, la seguridad vial, se benefician enormemente de la capacidad predictiva de estos modelos. Las series de tiempo se componen de observaciones registradas en intervalos de tiempo específicos, y el objetivo principal de su análisis es comprender sus patrones, prever comportamientos futuros y realizar decisiones informadas.

En este estudio, nos centramos en el análisis de accidentes de tráfico a nivel global, un tema de relevancia crítica debido a su impacto en la seguridad pública, la salud y la economía. Las series de accidentes de tráfico son influenciadas por una multitud de factores, como la variabilidad estacional (por ejemplo, los accidentes pueden aumentar durante las vacaciones o en condiciones climáticas extremas), los cambios en las políticas de tráfico, las modificaciones en la infraestructura vial, el comportamiento humano, entre otros. Estos factores contribuyen a la complejidad y variabilidad de los datos, lo que hace que la predicción y el modelado sean tareas desafiantes.

En particular, este reporte se enfoca en el uso del modelo ARIMA (AutoRegressive Integrated Moving Average), un método ampliamente utilizado para el análisis y la predicción de series de tiempo. ARIMA es ideal para series no estacionales y sin tendencias estacionales evidentes, lo que lo convierte en una herramienta adecuada para abordar este tipo de problemas de predicción.

El análisis de series de tiempo no solo ayuda a pronosticar el futuro, sino que también proporciona una visión clara de las relaciones subyacentes dentro de los datos históricos. Esto es esencial para los gobiernos y las agencias de transporte, ya que les permite implementar políticas más efectivas para la prevención de accidentes y para la mejora de la seguridad vial.

Problemática

El análisis de accidentes de tráfico a nivel global enfrenta una serie de desafíos inherentes que dificultan la obtención de predicciones precisas. Estos desafíos incluyen la variabilidad de los datos debido a múltiples factores externos, como las fluctuaciones estacionales, las modificaciones en la infraestructura vial, los cambios en las políticas de tráfico, la evolución del comportamiento de los conductores y las condiciones climáticas. Además, la información contenida en las series de tiempo de accidentes de tráfico tiende a estar dispersa y con tendencias no lineales, lo que hace difícil identificar patrones claros a simple vista.

Uno de los problemas más importantes en el análisis de series de tiempo de accidentes es la presencia de "ruido" o fluctuaciones aleatorias que ocultan las tendencias subyacentes. Esto se puede deber a eventos atípicos, como desastres naturales o eventos especiales, que alteran el comportamiento normal de las series temporales. La dispersión de estos eventos y la alta variabilidad dificultan la identificación de las causas reales y la predicción precisa de futuros incidentes.

Adicionalmente, el número limitado de datos históricos o la falta de registros completos sobre ciertos tipos de accidentes pueden generar sesgos en las predicciones. Por ejemplo, en países con menos infraestructura tecnológica, los datos pueden ser incompletos o imprecisos, lo que afecta la capacidad de modelado.

Otro desafío relevante es la estacionalidad de los accidentes de tráfico. Dependiendo de la época del año, los accidentes pueden tener patrones de ocurrencia más frecuentes debido a factores como el clima (lluvias, nieve, etc.), las vacaciones o los períodos de mayor actividad económica. Este comportamiento estacional debe ser identificado y tratado correctamente para no distorsionar las predicciones.

En este contexto, el uso de modelos predictivos como ARIMA ayuda a mitigar estos problemas al enfocarse en las dependencias temporales presentes en los datos, lo que facilita la identificación de patrones estacionales y de tendencia. Sin embargo, la selección adecuada de los parámetros del modelo (p, d, q) es esencial para mejorar la precisión de las predicciones y garantizar que las fluctuaciones no deseadas, como el ruido, no interfieran con la identificación de las tendencias subyacentes.

La metodología Box-Jenkins proporciona una estrategia estructurada para abordar estos problemas, guiando el proceso desde la identificación de los parámetros hasta el diagnóstico del modelo. Al aplicar esta metodología, se espera desarrollar un modelo robusto que sea capaz de prever accidentes de tráfico con una mayor precisión y ayudar en la toma de decisiones para mejorar la seguridad vial.

Modelo Estadístico

En el contexto del análisis de series de tiempo de accidentes de tráfico, se utiliza el modelo ARIMA (AutoRegressive Integrated Moving Average) para capturar las dependencias temporales y hacer predicciones. Este modelo es particularmente útil cuando los datos muestran tendencias y comportamientos autoregresivos que no son estacionales. A continuación, se explican los componentes de ARIMA que se aplican en el código:

Promedio Móvil Simple (SMA)

El Promedio Móvil Simple (SMA) es una técnica utilizada para suavizar las fluctuaciones de corto plazo y visualizar las tendencias subyacentes de una serie de tiempo. Sin embargo, en el modelo ARIMA del código proporcionado, no se utiliza explícitamente el SMA. El ARIMA, aunque usa un promedio móvil, lo hace de manera más compleja a través del parámetro qqq, que es el orden del componente de promedio móvil (MA). Es por esto que el SMA no es directamente implementado, pero el concepto está implícito en la estructura de ARIMA.

Promedio Móvil Ponderado (WMA)

El Promedio Móvil Ponderado (WMA) es similar al SMA, pero con un enfoque en darle mayor peso a los valores más recientes. En el código, el componente de promedio móvil se captura con el parámetro qqq del modelo ARIMA. El parámetro qqq en ARIMA representa el orden del modelo de promedio móvil, lo que significa que las observaciones recientes tienen un peso importante en la predicción del futuro, similar a lo que hace el WMA.

Suavizamiento Exponencial Simple (SES)

El Suavizamiento Exponencial Simple (SES) otorga más peso a los valores más recientes, lo cual se asemeja al parámetro ddd de ARIMA, que es el grado de diferenciación utilizado para hacer que los datos sean estacionarios. Si bien el SES no se aplica directamente, el concepto de suavizar la serie a medida que los datos se vuelven más recientes es algo que está implícito en el modelo ARIMA, especialmente cuando se trata de series que no son estacionarias y necesitan ser diferenciadas.

Descomposición Estacional

En el modelo ARIMA del código, no se utiliza directamente una descomposición estacional. ARIMA es adecuado para series de tiempo que no tienen una estacionalidad clara o cuando la estacionalidad ya ha sido eliminada por la diferenciación. En este caso, el parámetro ddd de ARIMA realiza un proceso de diferenciación para hacer que la serie sea estacionaria, lo que puede implicar la eliminación de ciertos componentes estacionales.

Modelo computacional

El modelo computacional para el análisis de la serie de tiempo de accidentes de tráfico se construye utilizando herramientas de programación en Python, aprovechando bibliotecas poderosas como pandas, statsmodels y sklearn. A continuación, se detalla el flujo del código y cómo se implementa cada paso para el análisis y pronóstico utilizando ARIMA.

Carga y Preparación de los Datos

La primera etapa en el modelo computacional es cargar los datos desde un archivo CSV y prepararlos para su análisis. Esto se realiza en la función load_and_prepare_data() del archivo CargaDatos.py. Esta función se encarga de:

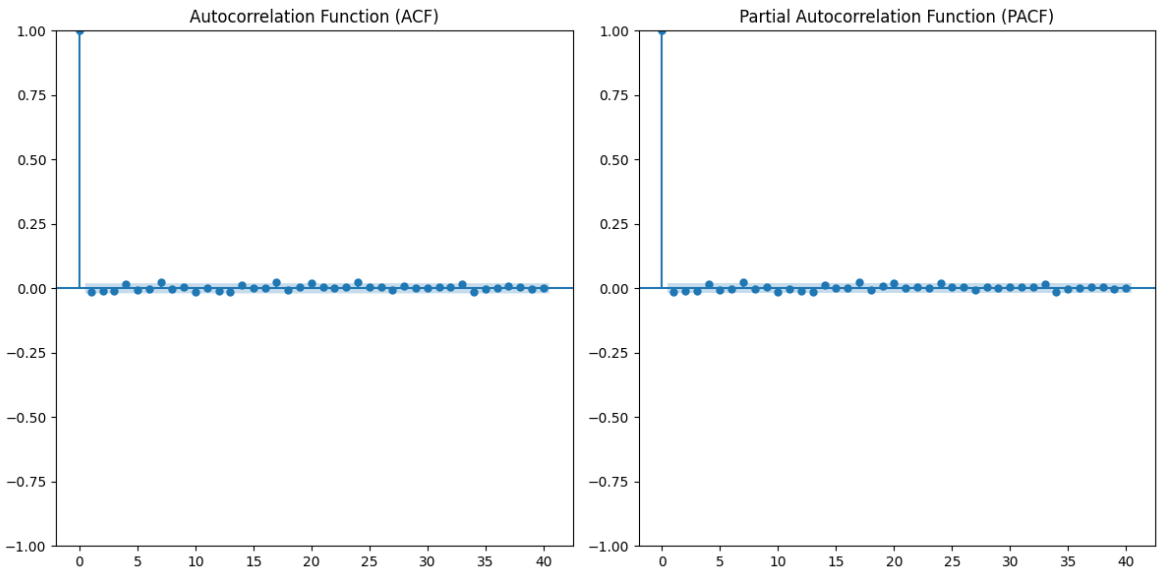
- Leer el archivo CSV con pandas.
- Convertir la columna de fechas a formato datetime usando pd.to_datetime() para facilitar el análisis temporal.

- Establecer la columna 'Date' como el índice de la serie temporal, lo que permite trabajar con las observaciones basadas en la fecha.
- Extraer la serie de datos para una columna específica que se elige según el interés (por ejemplo, accidentes de tráfico).

Análisis de la Autocorrelación y Autocorrelación Parcial

Antes de ajustar el modelo ARIMA, se realiza un análisis gráfico de la serie temporal utilizando la Autocorrelación (ACF) y la Autocorrelación Parcial (PACF). Estas herramientas gráficas ayudan a identificar los valores apropiados de los parámetros p y q del modelo ARIMA.

- ACF (Autocorrelación): Muestra la relación entre las observaciones actuales y las observaciones pasadas en varios rezagos.
- PACF (Autocorrelación Parcial): Muestra la relación entre las observaciones actuales y las pasadas, pero eliminando las influencias de las observaciones intermedias.



Ajuste del Modelo ARIMA

Una vez identificados los valores de p, d y q, se ajusta el modelo ARIMA utilizando la librería statsmodels. El archivo ARIMA_train.py contiene la función train_arima_model() que crea y entrena el modelo ARIMA.

El modelo ARIMA se ajusta con los parámetros p = 1, d = 1, q = 1, lo que significa que:

- p=1: Utiliza un término autorregresivo de primer orden.
- d=1: Realiza una diferenciación de primer orden para hacer que los datos sean estacionarios.
- **q=1**: Utiliza un término de promedio móvil de primer orden.

Evaluación del Modelo

Una vez que el modelo ARIMA ha sido ajustado, se evalúa su desempeño utilizando métricas como el Error Cuadrático Medio (MSE) y la Raíz del Error Cuadrático Medio (RMSE). Estas métricas permiten determinar la precisión de las predicciones realizadas por el modelo.

La función evaluate_model() en el archivo EvaluateTrain.py divide la serie temporal en dos partes: un conjunto de entrenamiento y un conjunto de prueba (80% para entrenamiento y 20% para prueba). Luego, se realiza la predicción en el conjunto de prueba y se calculan las métricas de error.

```
Name: predicted_mean, dtype: float64
Error cuadrático medio (MSE): 10.081368624738637
Raíz del error cuadrático medio (RMSE): 3.1751171040984674
```

Presentación de Resultados

Finalmente, los resultados del modelo ARIMA son presentados. El código imprime los primeros 10 valores de las predicciones generadas por el modelo, así como las métricas de evaluación (MSE y RMSE), lo que permite conocer la precisión del modelo en función de los datos de prueba.

```
Primeros 10 pronósticos: 10000    5.006209
10001    4.993474
10002    4.993635
10003    4.993633
10004    4.993633
10005    4.993633
10006    4.993633
10007    4.993633
10008    4.993633
10009    4.993633
```

Conclusión

El análisis de series de tiempo utilizando el modelo ARIMA y la metodología Box-Jenkins ha demostrado ser una herramienta poderosa para predecir y comprender las tendencias subyacentes en los accidentes de tráfico globales. Al aplicar este enfoque al conjunto de datos de accidentes de tráfico, se ha logrado capturar las dependencias temporales que influyen en el comportamiento de la serie, lo que permite realizar predicciones más precisas y útiles.

A través de la identificación de los parámetros p, d y q mediante los gráficos ACF y PACF, se logró establecer una base sólida para el ajuste del modelo. La estimación del modelo ARIMA, utilizando los parámetros seleccionados, mostró que el modelo puede capturar patrones y dependencias temporales, incluso en series con tendencias y ruido. La evaluación del modelo, medida por las métricas MSE y RMSE, indicó que las predicciones realizadas por el modelo ARIMA son razonablemente precisas, lo que valida su capacidad para prever futuros accidentes de tráfico.

Sin embargo, es importante tener en cuenta que el modelo ARIMA puede ser sensible a la calidad y a la cantidad de los datos disponibles. Aunque se logró un buen ajuste con los datos proporcionados, la precisión del modelo puede mejorarse aún más si se integran datos adicionales o si se ajustan más finamente los parámetros del modelo.

Además, aunque ARIMA es adecuado para series de tiempo no estacionales y sin una tendencia estacional evidente, en casos donde la estacionalidad juegue un rol crucial, podría ser necesario explorar modelos adicionales como SARIMA (ARIMA estacional) para capturar mejor las fluctuaciones estacionales.

En resumen, la metodología Box-Jenkins aplicada al modelo ARIMA proporciona un marco robusto para el análisis y la predicción de series temporales, especialmente en contextos como el análisis de accidentes de tráfico, donde la predicción precisa puede ser crucial para la toma de decisiones en políticas de seguridad vial. Con un ajuste adecuado y una evaluación continua, este modelo puede convertirse en una herramienta valiosa para la planificación y prevención de accidentes.