Lina María Montoya Zuluaga.
Monografía: Primera Iteración.
Especialización en Análisis y Ciencia de Datos.
Universidad de Antioquia.

Resumen: la dinámica climática es sumamente compleja, especialmente por la serie de eventos que están correlacionados y que los convierte en conjunto en un sistema caótico. Sin embargo, existe una estrecha relación causal entre los índices climáticos globales y la dinámica local de los principales ríos de Suramérica, contenida en los registros de los caudales. Así pues, el objetivo de este estudio es encontrar un algoritmo de machine learning adecuado para analizar series de tiempo y que permita realizar predicciones sobre el caudal.

Comprensión del problema de aprendizaje automático.

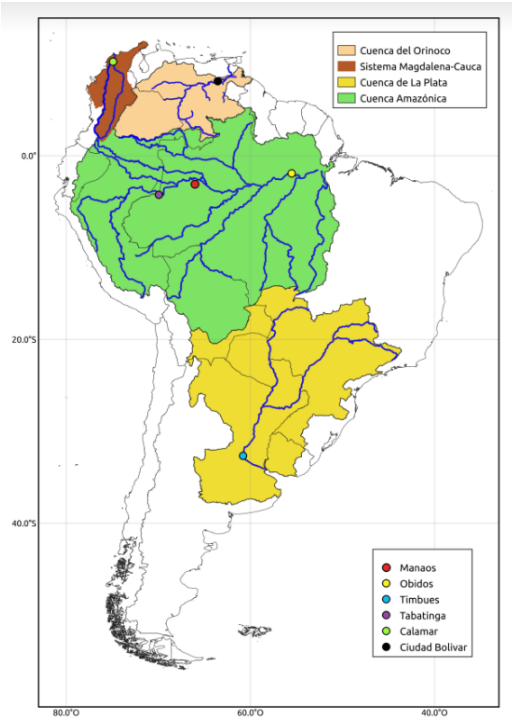
Esta es una continuación del trabajo de grado titulado "Influencia dinámica de fenómenos climáticos globales en
La hidrología de Sur América: Un enfoque entrópico" en el cual se realizó un estudio causal entre sistemas dinámicos bivariados. La ruta fue usar la entropía y la transferencia de información para determinar la flecha de la evolución temporal y la información que las partes del sistema están compartiendo.

Para entender el origen de este análisis es necesario primero dimensionar lo que cada suceso por aparte representa, antes de analizarlo como un sistema bivariado. Los fenómenos hidrológicos y climáticos comparten una sutil peculiaridad y es que ambos se desarrollan tanto en el tiempo como en el espacio, pues aunque bien son fenómenos que evolucionan en el tiempo y que poseen una periodicidad temporal inherente, son medidos en un punto particular del espacio y las medidas pueden ser significativamente diferentes si se modifica el lugar o las condiciones de la medición. Esta relación espacio-temporal que se acaba de mencionar se manifiesta tanto en el clima como en la hidrología, así pues, la dinámica global dentro del problema va a estar determinada por los índices climáticos medidos sobre el océano y el comportamiento local lo proporcionan los valores del caudal en estaciones hidrométricas particulares dentro de las principales cuencas del sur del continente Americano. Evidentemente, por tratarse de series de tiempo, se usan los registros en los que temporalmente logren coincidir ambas series.

Finalmente, como ya las relaciones causales se hallaron en un trabajo previo, el objetivo del presente estudio es tomar las series de tiempo para cada uno de los caudales y unirlas con las series de tiempo de los índices climáticos para los cuales se encontró relación causal, de tal manera que los índices climáticos que influencian el caudal sean el insumo para predecir el valor del caudal de dicha estación hidrométrica en un tiempo determinado.

Así pues, se trata de un trabajo de predicción en series de tiempo. Para esta primera iteración el objetivo fue trabajar sobre un modelo de predicción de redes Long-Short Term Memory (LSTM) pero únicamente sobre una de las series de caudal, con la intención de verificar si este tipo de redes pueden ajustarse a las dinámicas hidrológicas y lograr hacer predicciones acertadas.

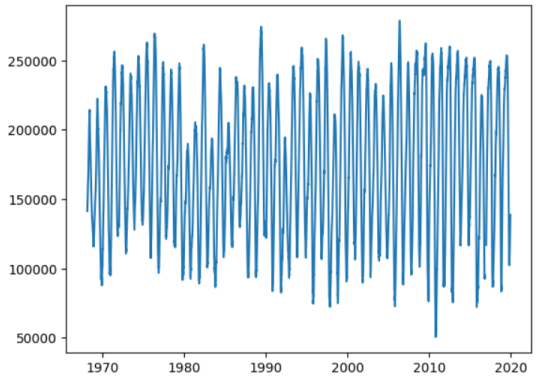
En el mapa de Suramérica se muestran las cuatro cuencas hidrológicas que están en el foco de nuestro estudio: Orinoco, Magdalena-Cauca, La Plata y Amazonía. Además, se muestran las seis estaciones hidrométricas de las cuales se extraen los datos y para esta primera iteración se analizó y procesó con mayor detalle la estación de Óbidos, perteneciente a la cuenca Amazónica y que tiene registros desde 1968 hasta 2019, diariamente.



Preprocesamiento y Análisis de la Base de Datos.

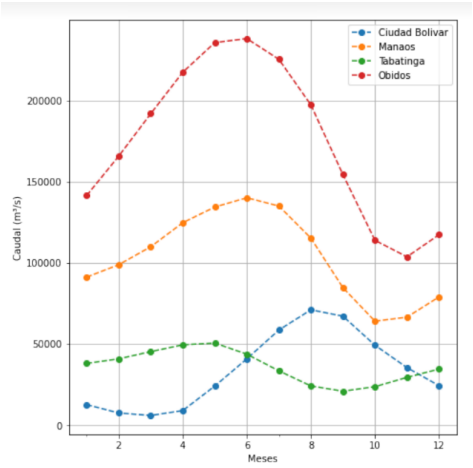
Como ya se ha mencionado, para esta parte del trabajo aplicamos el modelo únicamente sobre las series de tiempo de Óbidos, pero un primer análisis exploratorio sobre todas las series de caudal de las cuencas arroja ciertos resultados importantes sobre la dinámica inherente al proceso en cada una de las estaciones.

La gráfica de la serie de tiempo para el caudal de Óbidos aparenta presentar estacionalidad, sin embargo, los registros podrían ser insuficientes ya que existen procesos en la naturaleza con escalas temporales propias que en este caso desconocemos.



Acorde a la literatura, los caudales de los ríos exhiben periodicidad anual y un comportamiento estable durante el mes, por lo que se hace la gráfica de tendencia del caudal con el promedio mensual de caudal para todos los años que se tienen en el registro. Es de resaltar que los registros que vamos a utilizar corresponden a la estación hidrométrica con el mayor nivel de caudal promedio mes a mes, teniendo su caudal siempre muy por encima de los valores que presentan las otras centrales de

medición, pero conservando la dinámica característica que muestran las series de caudal de estas cuencas.



Modelo de Aprendizaje Supervisado.

Para realizar las predicciones sobre los caudales se implementa una red LSTM, que son un tipo de arquitectura de redes neuronales recurrentes, las cuales son muy utilizadas para el procesamiento de datos secuenciales, que es justo el caso de las series de tiempo.

Una celda LSTM trabaja sobre tres puertas clave, la puerta de olvido, puerta de entrada y puerta de salida; estas puertas permiten que las redes aprendan a mantener información relevante a largo plazo y a olvidar información menos relevante, lo que las hace especialmente efectivas en el procesamiento de secuencias de datos largas.

Se opta por probar con una red LSTM antes de usar cualquier otro método al tomar como referencia inicial lo hallado en el artículo "Forecasting daily precipitation using long short-term memory networks" en el cual se utilizan datos históricos de precipitación diaria y otras variables relevantes como temperatura, humedad, presión atmosférica, que es análogo a nuestro objetivo final que es predecir el caudal teniendo como insumo los índices climáticos. En el artículo recién mencionado se compara las redes LSTM con otros métodos de predicción como regresión lineal y redes tradicionales, exponiendo que las redes LSTM presentan una mejor precisión que los demás métodos utilizados.

Posterior a haber encontrado en la literatura un método adecuado para el problema que es objeto de estudio, se busca en la literatura modelos tipo LSTM aplicados específicamente a variables hidrológicas como el caudal, para determinar la viabilidad de continuar por dicha ruta o cambiar de modelo de ser necesario.

"Monthly streamflow prediction using long short-term memory network based on features selection" usa las redes LSTM para la predicción del caudal, pero tomando como criterio de optimización del modelo la selección de las características y se muestra que el rendimiento del modelo puede mejorar al trabajar con la información importante contenida en las series de tiempo. Esto nos sugiere una primera revisión cuidadosa de los resultados previamente obtenidos sobre la causalidad para verificar que se estén usando los insumos adecuados para una óptima predicción.

Para esta primera iteración del modelo, como se ha venido mencionando, se toma la red LSTM y se entrena únicamente con los datos de caudal, no se tienen todavía en cuenta los indicadores climáticos, por lo

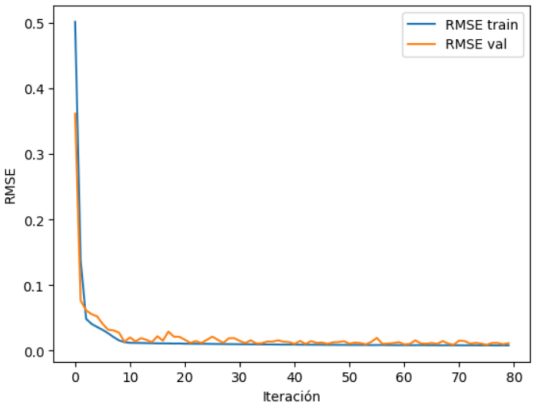
que usamos el método univariado-unistep para estudiar nuestro problema. En este punto de la búsqueda bibliográfica nos apoyamos en los artículos "Long Short-Term Memory recurrent neural network for monthly streamflow prediction" y "River flow forecasting using Long Short-Term Memory (LSTM) networks", en los cuales se hacen los cálculos y el entrenamiento del modelo de manera mensual, por la escala propia de estos ciclos hidrológicos en las latitudes cercanas al Ecuador y que fue justamente nuestro lapso en el análisis exploratorio de los datos. Finalmente, ambos estudios muestran que este tipo de redes pueden capturar patrones y dependencias a largo plazo en los datos de caudales, mejorando así la capacidad de predicción.

Esta parte sobre las predicciones de los caudales es una primera aproximación para el estudio del problema sobre clima e hidrología, y en la línea de "Streamflow prediction using long short-term memory (LSTM) neural network" se encuentra que las redes LSTM son un método versátil y que permite mejorar la precisión en la predicción del caudal usando como insumo datos históricos de caudales y otras variables hidrológicas para entrenar la red.

Primeros resultados de la red LSTM.

Es importante mencionar que dentro de la base de datos había una falta de registros de aproximadamente mes y medio que no se logró detectar hasta el análisis exploratorio de los datos en una etapa ya bastante avanzada. Esta falta se logró identificar cuando se verificó que todos los registros tuviesen la misma separación temporal, un día para el caso específico de la estación hidrométrica de Óbidos. Para esta primera iteración todos esos espacios se llenaron con el valor del caudal inmediatamente anterior, pero en posteriores corridas se considera más apropiado llenar dichos espacios con el promedio entre el año anterior y el año siguiente en esas mismas fechas, teniendo en mente la periodicidad anual de este escenario en particular.

La métrica de desempeño usada fue root mean square error (RMSE), ampliamente usada en problemas de predicción sobre series de tiempo. Como se muestra en la imagen, los resultados son sumamente buenos, teniendo presente que los caudales fueron escalados usando MinMaxScaler, sin embargo, para verificar la eficiencia del modelo es adecuado implementar otras métricas también como el error absoluto medio (MAE) o el coeficiente de determinación (R^2).




```
Comparativo desempeños:
  RMSE train:    0.010
  RMSE val:      0.011
  RMSE test:     0.011
```

Bibliografía.

- Holger Kantz y Thomas Schreiber. Nonlinear Time Series Analysis. Cambridge University Press., 2003.
- X. San Liang. Causation and information flow with respect to relative entropy. American Institute of Physics., 2018.
- Jhan Carlo Espinoza Villar; Waldo Lavado. Evolución regional de los caudales en el conjunto de la cuenca del Amazonas para el periodo 1974-2004 y su relación con factores climáticos. Revista Peruana Geo-Atmosférica., 2009.
- I. Hoyos; J. Cañón Barriga; T. Arenas Suárez; F. Dominguez; B. A. Rodríguez. Variability of regional atmospheric moisture over Northern South America: patterns and underlying phenomena. Springer Nature, 2018.
- Thomas Schreiber. Measuring Information Transfer. The American Physical Society, 2000.
- Muhammad Usman, Feng Liu, Hui Li, Jianhua Gong. Streamflow prediction using long short-term memory (LSTM) neural network. Water, 2019.
- Uday Pimple, Vicent Carollo, Casey Brown. Long short-term memory recurrent neural network for monthly streamflow prediction. Journal of Hydroinformatics, 2018.
- Ahmad F Taha, Mohd M Jami, Haftom T Gebre. River flow forecasting using long short-term using memory (LSTM) networks. Journal of Hydrology, 2019.
- Lei Chen, Huajun Jin, Donghai Yuan. Monthly streamflow prediction using long short-term memory network based on features selection. Water Resources Management, 2020.