Plan de Proyecto de Titulación

Cristian Pachacama

May 29, 2018

1 INFORMACIÓN BÁSICA

1. **Propuesto por:** Cristian David Pachacama Simbaña
2. **Línea de Investigación:** Estadística Aplicada
3. **Fecha:** 11 de mayo de 2018

2 RELACIÓN

1. **Nombre del Proyecto de Investigación:** Forecast and Impact of extreme low levels of streamflow in hydropower plants.
2. **Director del Proyecto de Investigación:** PhD. Adriana Uquillas

3 INFORMACIÓN DEL TRABAJO DE TITULACIÓN

3.1 Título del Trabajo de Titulación

ANÁLISIS CLUSTER DE SERIES DE TIEMPO.

3.2 Planteamiento del Problema

Brasil tiene una de los sistemas hidrológicos más complejos, diversos y extensos del mundo. A diferencia de la gran mayoría de los países desarrollados, Brasil tiene en los ríos su principal fuente de generación de electricidad , ocupando el tercer lugar dentro de los más grandes productores hidroeléctricos del mundo. Debido a la importancia del sector hidroeléctrico, buscar formas de facilitar y mejorar el modelamiento de datos asociados a este sector es un problema prioritario. Provocado por la dificultad que supone lidiar con la enorme cantidad de datos asociados a mediociones de Caudales de los ríos que componen este sistema, que cuenta con alrededor de 150 estaciones de medición repartidas en todo Brasil. Dichos datos se presentan en forma de Series de Tiempo que posee tres carácteristicas que dificultan su análisis, la primera es que estas series de tiempo poseen observaciones diarias de los caudales en un periodo de tiempo de alrededor de 30 años, es decir, son series muy extensas. La segunda característica es que estas series de tiempo son estacionales, y por último existe evidencia de que el ruido o error asociado a estas series no se distribuye normalmente sino que su distribución posee colas más pesadas como las analizadas en teoría de valores extremos.

En ese contexto, notamos que es posible disminuir la dimensión del problema a travéz la identificación de clústers o zonas representativas (no necesariamente geográficas) que resuman el comportamiento temporal que poseen los caudales de los ríos. Esto en términos de modelamiento esto se traduce en pasar del problema de modelar el nivel de caudal en todas las 150 estaciones, al problema de modelar unicamente 1 estación por cada clúster.

3.3 Justificación

Ya que el problema se basa en identificar grupos de ríos cuyos Caudales se comportan de manera similar en el tiempo, se propone la utilización de el "Análisis Clúster de Series de Tiempo", que es una técnica de agrupamiento que considera una función de "disimilitud" entre las series de tiempo (que mide que tan distintas son un par de series) y a partir de ella crea grupos de series, cada grupo contiene series de tiempo "parecidas". Al elegir adecuadamente la función de disimilitud (diseñada para series de tiempo) es posible agrupar a los ríos en grupos basados en el comportamiento temporal de sus caudales. Esto con la finalidad de lidiar con la complejidad que supone analizar y modelar esta enorme cantidad de series de tiempo de caudales, pasando de analizar alrededor de 150 series a unas pocas (una serie por Clúster), sin dejar de lado la estructura y comportamiento estacional de cada una de ellas, partiendo de una adecuada elección de la función de disimilitud.

Hay que destacar que el modelamiento de caudales juega un rol trascendental en la creación de políticas que adopta sector energético de Brasil, que como mencionamos anteriormente está alimentado en su mayoría por el sector hidroeléctrico en donde el análisis que planteamos permitirá profundizar en la planificación de las operaciones de plantas hidroeléctricas que depende directamente del comportamiento temporal de los ríos que las alimentan, esta planificación podría evitar por ejemplo eventos de déficit energético provocados por una deficiencia estructural de la disponibilidad de energía, que a la larga tiene impactos económicos y sociales mayores que los cortes de energía.

3.4 Objetivo General

El principal objetivo del proyecto es utilizar el Análissi Clúster para agrupar estaciones (asociadas a ríos en Brasil), basandonos en el comportamiento temporal del caudal de los rios que se mide en dichas estaciones, y posteriormente modelar los caudales (1 por clúster) usando variables micro y macro climáticas.

3.5 Objetivos Específicos

1. Comparar una gama de técnicas tanto de clusterización, así como la elección de distintas métricas o funciones de disimilitud a fin de identificar aquella combinación (disimilitud/algotimo de clusterización) que permita un adecuado agrupamiento de las series de tiempo asociadas a los caudales de los principales ríos de Brasil, que tienen la peculiaridad de que dichas series son usualmente estacionales, y este tipo de series no han sido analizadas con esta técnica previamente.

2. Comparar el modelamiento de estas series de tiempo usando el análisis cluster versus el modelamiento sin un previo análisis, a fin de validar la metodología de clústerización planteada, y además mostrar posibles ventajas en cuanto a tiempo y eficiencia de la metodología planteada.

3. Encontrar los factores determinantes del nivel de caudal para cada clúster y comparar como varían sus efectos entre uno clúster y otro.

4. Automatizar la descomposición y Análisis Clúster de series de tiempo de Caudales a través de la utilización de software estadístico y programación.

5. Crear una Plataforma de Análisis (Visualización y descomposición) de las Series de Tiempo de Clima, Caudales, y Contaminación, que permita a más investigadores análizar toda la información recolectada durante este proyecto.

3.6 Metodología

El Análisis Clúster es un técnica de aprendizaje no supervisada que tiene como objetivo dividir un conjunto de objetos en grupos homogéneos (clústers). La partición se realiza de tal manera que los objetos en el mismo clúster son más similares entre sí que los

objetos en diferentes grupos según un criterio definido. En muchas aplicaciones reales, el análisis de clúster debe realizarse con datos asociados a series de tiempo. De hecho, los problemas de agrupamiento de series de tiempo surgen de manera natural en una amplia variedad de campos, incluyendo economía, finanzas, medicina, ecología, estudios ambientales, ingeniería y muchos otros. Con frecuencia, la agrupación de series de tiempo desempeña un papel central en el problema estudiado. Estos argumentos motivan el creciente interés en la literatura sobre la agrupación de series de tiempo, especialmente en las últimas dos décadas, donde se ha proporcionado una gran cantidad de contribuciones sobre este tema. En [Liao, 2005] se puede encontrar un excelente estudio sobre la agrupación de series de tiempo, aunque posteriormente se han realizado nuevas contribuciones significativas. Particularmente importante en la última década ha sido la explosión de documentos sobre el tema provenientes tanto de comunidades de minería de datos como de reconocimiento de patrones. [Fu, 2011] proporciona una visión general completa y exhaustiva de las últimas orientaciones de minería de datos de series de tiempo, incluida una gama de problemas clave como representación, indexación y segmentación de series de tiempo, medidas de disimilitud, procedimientos de agrupamiento y herramientas de visualización.

Una pregunta crucial en el Análisis Clúster es establecer lo que queremos decir con objetos de datos "similares", es decir, determinar una medida de similitud (o disimilitud) adecuada entre dos objetos. En el contexto específico de los datos asociados a series de tiempo, el concepto de disimilitud es particularmente complejo debido al carácter dinámico de la serie. Las diferencias generalmente consideradas en la agrupación convencional no podrían funcionar adecuadamente con los datos dependientes del tiempo porque ignoran la relación de interdependencia entre los valores.

De esta manera, diferentes enfoques para definir una función de disimilitud entre series de tiempo han sido propuestos en la literatura pero nos centraremos en aquellas medidas asociadas a la autocorrelación (simple, e inversa), correlación cruzada y periodograma de las series (Ver: [Struzik and Siebes, 1999]; [Galeano and Peña, 2000]; [Caiado et al., 2006]; [Chouakria and Nagabhushan, 2007]). Estos enfoques basados en características tienen como objetivo representar la estructura dinámica de cada serie mediante un vector de características de menor dimensión, lo que permite una reducción de dimensionalidad (las series temporales son esencialmente datos de alta dimensionalidad) y un ahorro significativo en el tiempo de cálculo, además de que nos ayudan a alcanzar el objetivo central por el que usaremos el Análisis Clúster que es el de la modelización de series de tiempo.

Una vez que se determina la medida de disimilitud, se obtiene una matriz de disimilitud inicial (que contiene la disimilitud entre parejas de series), y luego se usa un algoritmo de agrupamiento convencional para formar los clústers (grupos) con las series. De hecho, la mayoría de los enfoques de agrupamiento de series de tiempo revisados por [Liao, 2005] son variaciones de procedimientos generales como por ejemplo: K-Means, K-Medoids, PAM, CLARA [Kaufman and Rousseeuw, 1986] o de Clúster jerárquico que utilizan una gama de disimilitudes específicamente diseñadas para tratar con series de tiempo y algunas de sus características.

Una etapa adicional dentro del análisis clúster consiste en determinar la cantidad de clústers que es más apropiada para los datos. Idealmente, los clústers resultantes no solo deberían tener buenas propiedades estadísticas (compactas, bien separadas, conectadas y estables), sino también resultados relevantes. Se han propuesto una variedad de medidas y métodos para validar los resultados de un análisis clúster y determinar tanto el número de clústers, así como identificar qué algoritmo de agrupamiento ofrece el mejor rendimiento, algunas de estas ellas pueden encontrarse en [Fraley and Raftery, 1998]; [Duda et al., 2001] ; [Salvador and Chan, 2004] ; [Kerr and Churchill, 2001]. Esta validación puede basarse únicamente en las propiedades internas de los datos o en alguna referencia externa.

Posteriormente se procede al modelamiento SARIMAX de los caudales usando además variables micro y macro climáticas que podrían explicar de mejor manera el comportamiento de estos caudales, cabe mencionar que se modelara únicamente un caudal por cada clúster (grupo). El modelo SARIMAX propuesto en [Box et al., 2015] es un modelo SARIMA (Seasonal Autoregressive Integrated Moving Average) que incluye variables exógenas. Es decir, compone un modelo de regresión ordinario que usa variables exógenas en el modelo

SARIMA que se usa para estudiar series de tiempo estacionales.

3.7 Plan de Trabajo

El plan de trabajo se resume en la Tabla 3.7.

Actividad	Descripción
Obtención de Datos	Descarga de Bases de Datos Brasil de las fuentes: • Datos Hídricos Operador Nacional do Sistema Elétrico Brasil (ONS). • Datos Meteorológicos Globales: National Weather Service. (NOAA), Global Climate Observing System (GCOS)y CPTEC. • Datos de Clima: Climatic Reserch Unit, Departamento de Ciências Atmosféricas (U. São Paulo), e Instituto Nacional de Meteorología de Brasil (INMET)
Análisis Preliminar	Estudio teórico de modelos a fin de identificar mejores técnicas que puedan usarse dentro de la metodología planteada.
Tratamiento de Datos	• Depuración de Datos. • Estandarización de Datos. • Unificación de Datos para el modelamiento.
Análisis y Modelamiento	• Análsis y comparación de métricas y funciones de disimilitud. • Comparación de distintos algoritmos de Clusterización. • Validación del Análisis. • Modelamiento SARIMAX del caudal a partir de variables micro y macro climáticas (un modelo por clúster).
Automatización	Identificación de parámetros escenciales del análisis clúster de las series de tiempo de caudales, y luego automatizar dicho análisis mediante su programación en el software estadístico R.
Conclusiones y Recomendaciones	

Table 1: Plan de Trabajo

3.8 Cronograma

A continuación se muestra el cronograma de actividades planificadas resumidas en la Tabla 3.8 .

	Abril	Mayo	Junio	Julio	Agosto
Obtención de Datos	X	X			
Tratamiento de Datos		X			
Análisis y Modelamiento		X	X	X	
Automatización			X	X	X
Conclusiones y Recomendaciones					X

Table 2: Cronograma de Trabajo

References

[Box et al., 2015] Box, G. E., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.

[Caiado et al., 2006] Caiado, J., Crato, N., and Peña, D. (2006). A periodogram-based metric for time series classification. *Computational Statistics & Data Analysis*, 50(10):2668–2684.

[Chouakria and Nagabhushan, 2007] Chouakria, A. D. and Nagabhushan, P. N. (2007). Adaptive dissimilarity index for measuring time series proximity. *Advances in Data Analysis and Classification*, 1(1):5–21.

[Duda et al., 2001] Duda, R. O., Hart, P. E., Stork, D. G., et al. (2001). Pattern classification. *International Journal of Computational Intelligence and Applications*, 1:335–339.

[Fraley and Raftery, 1998] Fraley, C. and Raftery, A. E. (1998). How many clusters? which clustering method? answers via model-based cluster analysis. *The computer journal*, 41(8):578–588.

[Fu, 2011] Fu, T.-c. (2011). A review on time series data mining. *Engineering Applications of Artificial Intelligence*, 24(1):164–181.

[Galeano and Peña, 2000] Galeano, P. and Peña, D. P. (2000). Multivariate analysis in vector time series. *Resenhas do Instituto de Matemática e Estatística da Universidade de São Paulo*, 4(4):383–403.

[Kaufman and Rousseeuw, 1986] Kaufman, L. and Rousseeuw, P. J. (1986). Clustering large data sets. In *Pattern Recognition in Practice, Volume II*, pages 425–437. Elsevier.

[Kerr and Churchill, 2001] Kerr, M. K. and Churchill, G. A. (2001). Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. *Proceedings of the National Academy of Sciences*, 98(16):8961–8965.

[Liao, 2005] Liao, T. W. (2005). Clustering of time series data—a survey. *Pattern recognition*, 38(11):1857–1874.

[Salvador and Chan, 2004] Salvador, S. and Chan, P. (2004). Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. In *Tools with Artificial Intelligence, 2004. ICTAI 2004. 16th IEEE International Conference on*, pages 576–584. IEEE.

[Struzik and Siebes, 1999] Struzik, Z. R. and Siebes, A. (1999). The haar wavelet transform in the time series similarity paradigm. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 12–22. Springer.