Plan de Proyecto de Titulación

Cristian Pachacama

11 de mayo de 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 CON COVARIANTES.

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 esta enorme cantidad de datos asociados al sector es un problema prioritario. Dentro de ellos, un problema específico es la identificación de regiones (no necesariamente geográficas) en las que los Caudales de los ríos poseen un comportamiento similar en el tiempo, esto debido a que de hacerlo es posible simplificar el modelamiento de Caudales.

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 simplificar el modelamiento, pasando de un modelo por río, a un modelo por grupo (Clúster).

3.4. Hipótesis

3.5. Objetivo General

El principal objetivo del proyecto es el de identificar y agrupar estaciones (asociadas a ríos en Brasil) en zonas, basandonos en el comportamiento temporal del caudal de los rios (medido en dichas estaciones), a fin de facilitar el modelamiento de los cuadales de al rededor de 150 estaciones repartidas en todo Brasil, que como sabemos posee uno de los sistemas hidricos más grandes y comlejos del mundo, por lo que al estudiar su comportamiento es posible escalarlo y adaptarlo a sistemas menos complejos como el de nuestro país.

3.6. Objetivos Específicos

- 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 al Caudal de los ríos, que tienen la peculiaridad de que las series estan conformadas por observaciones diarias de los caudales y además suelen ser series estacionales, y este tipo de series no han sido analizadas con esta técnica previamente.

- Comparar el modelamiento de estas series de tiempo usando el análisis cluster versus el modelamiento sin un previo análisis, a fin de mostrar las ventajas en cuanto a tiempo y eficiencia de la metodología planteada.

3.7. 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: Kovac1998; Struzik and Siebes 1999; Galeano and Peña 2000; Caiado, Crato, and Peña 2006). 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 desemejanza, se puede obtener una matriz de desemejanza inicial por pares y luego se usa un algoritmo de agrupamiento convencional para formar grupos de objetos. De hecho, la mayoría de los enfoques de agrupamiento de series de tiempo revisados por Liao (2005) son variaciones de procedimientos generales, por ejemplo: K-Means, K-Medoids, PAM, CLARA (Kaufman & Rousseeuw (2005)) o de Clúster jerárquico que utilizan una gama de disimilitudes específicamente diseñadas para tratar con series de tiempo. Una etapa adicional dentro del análisis 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 para validar los resultados de un análisis de agrupamiento y determinar qué algoritmo de agrupamiento ofrece el mejor rendimiento (Kerr y Churchill 2001; Yeung y otros 2001; Datta y Datta 2003). Esta validación puede basarse únicamente en las propiedades internas de los datos o en alguna referencia externa.

3.8. Plan de Trabajo