**F\_AA\_225**

**FACULTAD DE CIENCIAS**

**CARRERA/PROGRAMA DE INGENIERÍA MATEMÁTICA**

**PLAN DE TRABAJO DE TITULACIÓN/TESIS**

**TIPO DE TRABAJO DE TITULACIÓN: Proyecto de Investigación**

|  |  |
| --- | --- |
| **I.- INFORMACIÓN BÁSICA** | |
| **PROPUESTO POR:**  Cristian David Pachacama Simbaña | **LÍNEA DE INVESTIGACIÓN:**  Estadística Aplicada |
| **AUSPICIADO POR:** | **FECHA:**  *02/05/2018* |
| **RELACIÓN:**  **Nombre del Proyecto de Investigación:** *Forecast and Impact of extreme low levels of streamflow in hydropower plants.*  **Director del Proyecto de Investigación:** PhD. Adriana Uquillas | |
| **II.- INFORMACIÓN DEL TRABAJO DE TITULACIÓN** | |
| 1. **Título del Trabajo de Titulación**   *ANÁLISIS CLUSTER DE SERIES DE TIEMPO CON COVARIANTES* | |
| 1. **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.* | |
| 1. **Justificación**   *2.1 Justificación Teórica*  *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.*  *2.2 Justificación Metodológica.*  *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: Kovac̃ić 1998; 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.*  *2.3 Justificación Práctica*  *La justificación práctica establece el ¿para qué o para quién se resuelve el problema? Se debe indicar, con base en fuentes bibliográficas, que la investigación que se propone tiene una aplicación concreta, relacionada con el planteamiento del problema o los beneficiarios directos del problema.* | |
| 1. **Hipótesis (Si aplica)**   *Indicar que trata de probar y definir como explicaciones tentativas del fenómeno investigado.* | |
| 1. **Objetivo General** | |
| 1. **Objetivos Específicos** | |
| 1. **Metodología** | |
| 1. **Plan de Trabajo** | |
| 1. **Bibliografía**  * *Caiado, J., Crato, N., & Peña, D. (2006). A periodogram-based metric for time series classification. Computational Statistics & Data Analysis, 50(10), 2668-2684.* * *Montero, P., & Vilar, J. A. (2014). TSclust: An R package for time series clustering. Journal of Statistical Software, 62(1), 1-43.* * *Liao TW (2005). “Clustering of Time Series Data : A Survey.” Pattern Recognition.* * *Fu TC (2011). “A Review on Time Series Data Mining.” Engineering Applications of Artificial Intelligence.* * *Kovac̃ić ZJ (1998). “Classification of Time Series with Applications to the Leading Indicator Selection.” In Data Science, Classification, and Related Methods – Proceedings of the Fifth Conference of the International Federation of Classification Societies (IFCS-96), Kobe, Japan, March 27–30, 1996, pp. 204–207. Springer-Verlag.* * *Kaufman, L. and P.J. Rousseeuw, Finding Groups in Data (John Wiley & Sons, New York, 1990).* * *Struzik ZR, Siebes A (1999). “The Haar Wavelet in the Time Series Similarity Paradigm.” In Principles of Data Mining and Knowledge Discovery – Proceedings of the Third Euro-pean Conference, PKDD-99, Prague, Czech Republic, September 15–18, 1999, pp. 12–22. Springer-Verlag.* * *Galeano P, Peña D (2000). “Multivariate Analysis in Vector Time Series.” Resenhas do Instituto de Matemática e Estatı́stica da Universidade de São Paulo.* * *Caiado J, Crato N, Peña D (2006). “A Periodogram-Based Metric for Time Series Classification.” Computational Statistics & Data Analysis.* * *Kaufman, L. and P.J. Rousseeuw, Clustering large data sets (with discussion), in: E.S. Gelsema and L.N. Kanal (Eds.), Pattern Recognition in Practice II (North-Holland, Amsterdam, 1986).* * *Kerr MK, Churchill GA (2001). “Bootstrapping Cluster Analysis: Assessing the Reliability of Conclusions from Microarray Experiments.” Proceedings of the National Academy of Sciences of the United States of Americ.* * *Yeung KY, Haynor DR, Ruzzo WL (2001). “Validating Clustering for Gene Expression Data.” Bioinformatics.* * *Datta S, Datta S (2003). “Comparisons and Validation of Statistical Clustering Techniques for Microarray Gene Expression Data.”* | |
| 1. **Cronograma** | |

Firma Firma

|  |  |  |
| --- | --- | --- |
| Cristian David Pachacama Simbaña |  |  |
| Proponente 1 |  |  |
| Email: cristian.pachacama01@gmail.com  Telf.: 0958873211 |  |  |

Firma Firma

|  |  |  |
| --- | --- | --- |
| Adriana Uquillas Andrade |  | Nombre completo del Profesor |
| DIRECTOR |  | CODIRECTOR (*Si aplica*) |
| Email: adriana.uquillas@epn.edu.ec  Telf.: |  | Email:  Telf.: |