ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE CIENCIAS

ANÁLISIS CLÚSTER PARA SERIES DE TIEMPO ESTACIONALES Y MODELIZACIÓN DE CAUDALES DE RÍOS DEL BRASIL.

TRABAJO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERÍA MATEMÁTICA

PROYECTO DE INVESTIGACIÓN

CRISTIAN DAVID PACHACAMA SIMBAÑA
cristian.pachacama01@epn.edu.ec

Director: UQUILLAS ANDRADE ADRIANA, PH.D.
adriana.uquillas@epn.edu.ec

OCTUBRE 2018

DECLARACIÓN

Yo, CRISTIAN DAVID PACHACAMA SIMBAÑA, declaro bajo juramento que el trabajo aquí descrito es de mi autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y que he consultado las referencias bibliográficas que se incluyen en este documento.

A través de la presente declaración cedo mis derechos de propiedad intelectual, correspondientes a este trabajo, a la Escuela Politécnica Nacional, según lo establecido por la Ley de Propiedad Intelectual, por su Reglamento y por la normatividad institucional vigente.

Cristian David Pachacama Simbaña

CERTIFICACIÓN

Certifico que el presente trabajo fue desarrollado por CRISTIAN DAVID PACHACAMA SIMBAÑA, bajo mi supervisión.

Uquillas Andrade Adriana, Ph.D.
Directora del Proyecto

AGRADECIMIENTOS

A mi familia, ya que su cariño y apoyo me llevaron a donde ahora estoy. A mis mejores amigos Miguel, Rubi, Luis y Pablo, por brindarme su amistad y tan gratos momentos.

A mi tutora Adriana por ser una guía y apoyarme desde el primer momento a alcanzar esta meta, gracias por depositar su confianza en mi. A los profesores Erwin Jimenez, y de manera especial a Juan Carlos Trujillo quienes hicieron nacer en mi la pasión por la Matemática, pasión que espero inspirar a más generaciones de estudiantes.

Finalmente, a grandes matemáticos de la historia como George Cantor, Simeón Poisson , Abraham Wald, y Karl Pearson, cuyo trabajo me inspiró a profundizar en el conocimiento de esta bella ciencia.

DEDICATORIA

A mis padres Magdalena y Lucio, por su incondicional amor, sus sabios consejos y su paciencia, siempre lo tendré presente. A Isabel por su apoyo incondicional, y por aparecer en el momento exacto en mi vida para llenarla de felicidad. A Miguel, Rubi, Pablo y Luis por brindarme su sincera amistad.

Índice general

Índice de figuras

Índice de tablas

Resumen

En el presente trabajo se utilizará el Análisis Clústser para agrupar estaciones (asociadas a ríos en Brasil), basándonos en el comportamiento temporal del caudal de los ríos que se mide en dichas estaciones, y posteriormente modelar los caudales (1 modelo por clúster) usando variables micro y macro climáticas.

Palabras clave: Análisis Clúster para Series de Tiempo, descomposición STL, SARIMAX, Análisis de Componentes Principales Funcional.

Abstract

Keywords: HH

Capítulo 1

Introducción

Capítulo 2

Marco Teórico

Capítulo 3

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 [14] 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. [9] 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 aso-

ciados 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: [16]; [10]; [3]; [4]). 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 [14] son variaciones de procedimientos generales como por ejemplo: K-Means, K-Medoids, PAM, CLARA [11] 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 [8]; [7] ; [15] ; [12]. 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 [2] 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.

Apéndice

Bibliografía

[1] K. ALLIGOOD, T. SAUER, AND J. YORKE, *CHAOS: An Introduction to Dynamical Systems*, Springer-Verlag New York, New York, 1996.

[2] G. E. BOX, G. M. JENKINS, G. C. REINSEL, AND G. M. LJUNG, *Time series analysis: forecasting and control*, John Wiley & Sons, 2015.

[3] J. CAIADO, N. CRATO, AND D. PEÑA, *A periodogram-based metric for time series classification*, Computational Statistics & Data Analysis, 50 (2006), pp. 2668–2684.

[4] A. D. CHOUAKRIA AND P. N. NAGABHUSHAN, *Adaptive dissimilarity index for measuring time series proximity*, Advances in Data Analysis and Classification, 1 (2007), pp. 5–21.

[5] K. CHUNG AND R. WILLIAMS, *Introduction to Stochastic Integrtion*, Birkhäuser, New York, 1990.

[6] S. DATTA AND S. DATTA, *Comparisons and validation of statistical clustering techniques for microarray gene expression data*, Bioinformatics, 19 (2003), pp. 459–466.

[7] R. O. DUDA, P. E. HART, D. G. STORK, ET AL., *Pattern classification*, International Journal of Computational Intelligence and Applications, 1 (2001), pp. 335–339.

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

[9] T.-C. FU, *A review on time series data mining*, Engineering Applications of Artificial Intelligence, 24 (2011), pp. 164–181.

[10] P. GALEANO AND D. P. PEÑA, *Multivariate analysis in vector time series*, Resenhas do Instituto de Matemática e Estatística da Universidade de São Paulo, 4 (2000), pp. 383–403.

[11] L. KAUFMAN AND P. J. ROUSSEEUW, *Clustering large data sets*, in Pattern Recognition in Practice, Volume II, Elsevier, 1986, pp. 425–437.

[12] M. K. KERR AND G. A. CHURCHILL, *Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments*, Proceedings of the National Academy of Sciences, 98 (2001), pp. 8961–8965.

[13] M. LAKSHMANAN AND S. RAJASEKAR, *Nonlinear Dynamics*, Springer-Verlag Berlin Heidelberg, Berlin, 2003.

[14] T. W. LIAO, *Clustering of time series data?a survey*, Pattern recognition, 38 (2005), pp. 1857–1874.

[15] S. SALVADOR AND P. CHAN, *Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms*, in Tools with Artificial Intelligence, 2004. ICTAI 2004. 16th IEEE International Conference on, IEEE, 2004, pp. 576–584.

[16] Z. R. STRUZIK AND A. SIEBES, *The haar wavelet transform in the time series similarity paradigm*, in European Conference on Principles of Data Mining and Knowledge Discovery, Springer, 1999, pp. 12–22.

[17] R. TIBSHIRANI AND G. WALTHER, *Cluster validation by prediction strength*, Journal of Computational and Graphical Statistics, 14 (2005), pp. 511–528.

[18] R. TIBSHIRANI, G. WALTHER, AND T. HASTIE, *Estimating the number of clusters in a data set via the gap statistic*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63 (2001), pp. 411–423.

[19] K. Y. YEUNG, D. R. HAYNOR, AND W. L. RUZZO, *Validating clustering for gene expression data*, Bioinformatics, 17 (2001), pp. 309–318.