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, Luis Horna, 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 Clúster 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: HHHH

Capítulo 1

Introducción

Brasil tiene uno 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. Problema provocado por la dificultad que supone lidiar con la enorme cantidad de datos (accesibles desde la web de instituciones como ANA, ONS, NOAA, CPTEC, etc.) asociados a mediciones 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 características 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 través 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 únicamente 1 estación por cada clúster.

Ya que el problema se basa en identificar grupos de ríos cuyos Caudales se com-

portan 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 permitiría 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 deciencia estructural de la disponibilidad de energía, que a la larga tiene impactos económicos y sociales mayores que los cortes de energía.

Capítulo 2

Marco Teórico

2.1. Descomposición STL - Loess

STL es un procedimiento de filtrado que permite descomponer una serie de tiempo en sus componentes estacional, tendencia y Residuo. STL tiene un diseño simple que consiste en una secuencia de aplicaciones del Loess Smoother; la simplicidad permite el análisis de las propiedades del procedimiento y permite un cálculo rápido, incluso para series de tiempo muy largas y grandes cantidades de tendencia y suavizado estacional. Otras características de STL son la especificación de cantidades de suavizado estacional y de tendencias que varían, de manera casi continua, desde una cantidad muy pequeña de suavizado hasta una cantidad muy grande; estimaciones robustas de la tendencia y los componentes estacionales que no están distorsionados por un comportamiento aberrante en los datos; especificación del período de componente estacional a cualquier múltiplo entero del intervalo de muestreo de tiempo mayor que uno; y la capacidad de descomponer series de tiempo con valores perdidos.

2.2. Tratamiento de Valores perdidos

En este capítulo ilustraremos una propuesta para el tratamiento de valores perdidos en series de tiempo estacionales. Para ello consideremos una serie de tiempo $(X_t)_{t\in T}$, de la que conocemos las observaciones

$$x_1, x_2, ..., x_{j-1}, x_j, x_k, x_{k+1}, ..., x_n$$

donde $1 < j < k < n$. Recordemos que la descomposición STL-Loess permité descomponer aditivamente la serie en sus componentes de tendencia y estacionalidad inclusive en aquellos valores de t para los que no conocemos x_t , es decir, para $t = j+1, j+2, ..., k-1$.

Luego de descomoponer X_t obtenemos su tendencia D_t, estacionalidad S_t (para $t = 1, 2, ..., n$), y el residuo U_t (para $t = 1, 2, ..., j, k, ..., n$). Además estás series cumplen la relación siguiente.

$$X_t = D_t + S_t + U_t \tag{2.1}$$

Ilustramos lo antes mencionado en el siguiente gráfico (2.1) , que corresponde a la descomposición STL-Loess de una serie de datos mensuales de precipitación (lluvias) medidos en cierta zona de Brasil, está serie tiene estacionalidad anual (12 meses); notemos que es necesario fijar los parámetros de la descomposición adecuadamente ya que están asociados al número de retardos considerados al estimar tanto la componente estacional como la tendencia.

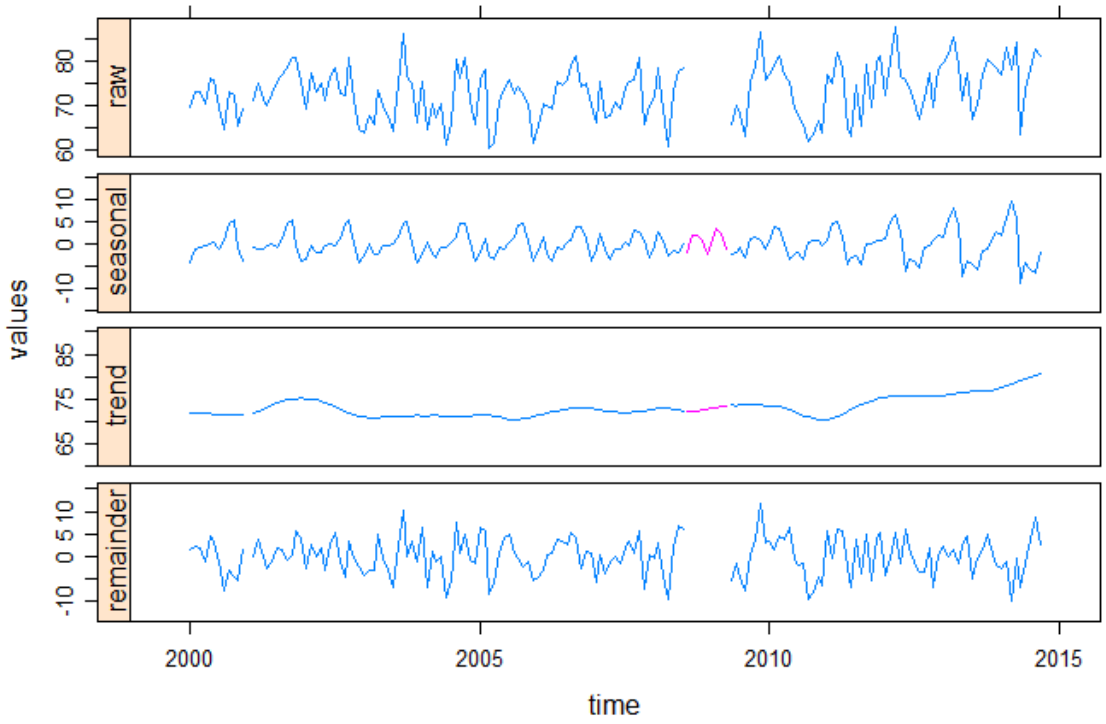
El gráfico muestra en la primera fila la serie climática, en segundo lugar muestra su componente estacional, en tercera fila encontramos su componente de tendencia, y finalmente el residuo. Como podemos notar las componentes de tendencia y estacionalidad están definidas en todo el dominio de tiempo.

Pues bien, notemos que bastaría conocer los valores de U_t en para todo t e inmediatamente conoceríamos los de X_t por la ecuación (2.1). Así, bastaría simular los valores perdidos de U_t que tiende a comportarse como un proceso estacionario, suponiendo que se especificaron bien los parámetros de la descomposición STL. Una forma simple de simular dichos datos, es partiendo de la función de distribución empírica de los Residuos

$$\hat{F}_0(u) := \frac{1}{n} \sum_{i=1}^{n} I_{(x_i \leq u)}$$

una vez calculada $\hat{F}_0(u)$ usamos el algoritmo siguiente.

Figura 2.1: Descomposición STL-Loess



2.3. Distancias y Funciones de Disimilitud entre Series de Tiempo

Desde un punto de vista general el término proximidad indica el concepto de cercanía en espacio, tiempo o cualquier otro contexto. Desde un punto de vista matemático, ese término hace referencia al concepto de disimilaridad o similaridad entre dos elementos. Sea O un conjunto finito o infinito de elementos (individuos, estímulos sujetos u objetos) sobre los que queremos definir una proximidad.

Dados dos puntos $o_i, o_j \in O$ y δ es una función real de $O \times O \to \mathbb{R}$, con $\delta_{ij} = \delta(o_i, o_j)$. Se diría que δ es una disimilaridad si verifica

- $\delta_{ij} = \delta_{ji}, \forall i, j.$
- $\delta_{ii} \leq \delta_{ij}, \forall i, j.$
- $\delta_{ii} = \delta_o, \forall i.$

La primera condición podría eliminarse, aunque resulta necesaria si se desea comparar con una distancia. No obstante, esa condición suele violarse cuando las disimilaridades provienen de juicios emitidos por sujetos, ya que éstos no siempre

califican igual al par (i, j) que al par (j, i). Las condiciones segunda y tercera suelen establecerse igualmente para $\delta_o = 0$, aunque también es conocido que cuando a un individuo le son presentados dos estímulos idénticos, éste tiende a asignarles algún valor de disimilaridad no nulo y generalmente positivo, y además no siempre se define $\delta_o \geq 0$ 0 ya que, si por ejemplo las disimilaridades provienen de una transformación, éstas podrían ser negativas. Una función real s de $O \times O \rightarrow \mathbb{R}$, se dirá que es una similaridad si verifica:

- $s_{ij} = s_{ji}, \forall i, j.$
- $s_{ij} \leq s_{ii}\ \forall i, j.$
- $s_{ii} = s_o, \forall i.$

Algunos autores consideran $s_o = 0$, y además suponen que $0 \leq s_{ij} \leq 1$ ya que una similaridad es un término opuesto al de disimilaridad por lo que deberá existir alguna transformación monótona t tal que $t(s) = \delta$. Una transformación de ese tipo podría ser $\delta = 1 - s$ si $0 \leq s \leq 1$, aunque otra utilizada en INDSCAL podría ser $\delta = -s$, sin que s deba estar acotada.
Puesto que la idea fundamental sobre la que se basa el MDS es la de asociar disimilaridades a distancias, hemos de verificar, entre otras, que se cumpla la desigualdad triangular. No obstante, si se cumplen los demás axiomas salvo éste, es posible transformar los datos para que ésta también sea verificada, tomando $c = \text{máx}_{i,j,h}(\delta_{hj} - \delta_{hi} - \delta_{ij})$ de forma que, $\gamma_{ii} = 0$; $\gamma_{ii} = \delta_{ij} + c$, $\forall i \neq j$.
Existen diferentes medidas para el cálculo de disimilaridades o similaridades entre un par de variables o individuos. Si consideramos una matriz de datos x_{ri}, obtenida de n objetos sobre p variables, algunos ejemplos de medidas son:

- *Distancia euclídea ponderada*

$$\delta_{rs} = \left(\sum_i w_i (x_{ri} - x_{si}) \right)^{1/2}$$

- *Métrica de Minkowski*

$$\delta_{rs} = \left(\sum_i x_{ri} - {x_{si}}^{\lambda} \right)^{1/\lambda}, \quad \lambda \geq 1$$

- *Separación angular*

$$\delta_{rs} = 1 - \frac{\sum_i x_{ri} x_{si}}{(\sum_i x_{ri}^2 \sum_i x_{si}^2)^{1/2}}$$

- Variograma

$$\delta_{rs} = \frac{\gamma_{rs}}{\sum_{r,s} \gamma_{rs}}$$

Donde:

$$\bar{X}_{r.} = \frac{1}{n} \sum_{i} X_{ri}$$

$$\gamma_{rs} = \frac{1}{(n-1)} \left[\sum_{i} (X_{ri} - X_{si})^2 - n(\bar{X}_{r.} - \bar{X}_{s.})^2 \right]$$

2.4. Algoritmos de Clusterización

2.5. Análisis de Componentes Principales Funcional

2.6. Modelo SARIMAX

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.

Capítulo 4

Conclusiones y Recomendaciones

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.