

Clustering en series de tiempo

Juan Kalemkerian
jkalem@cmat.edu.uy

Facultad de Ciencias (Centro de Matemática),
Universidad de la República.

Seminario SIESTA, Montevideo, Uruguay.
4 de noviembre de 2020.

Contenidos

- 1 Introducción
- 2 Distance Time Warping (DTW)
 - Warping functions
 - Definición de $DTW(x,y)$
 - Cálculo de $DTW(x,y)$
 - Dos virtudes importantes de la metodología DTW
 - Restricciones
 - TADPole
- 3 Shape based distance (SBD)
 - Definición
 - Detección de los centroides
- 4 Aplicación: covid19
 - En América del Sur
 - En el mundo
- 5 Trabajo a futuro
- 6 Referencias

Contenidos

- 1 **Introducción**
- 2 **Distance Time Warping (DTW)**
 - Warping functions
 - Definición de $DTW(x,y)$
 - Cálculo de $DTW(x,y)$
 - Dos virtudes importantes de la metodología DTW
 - Restricciones
 - TADPole
- 3 **Shape based distance (SBD)**
 - Definición
 - Detección de los centroides
- 4 **Aplicación: covid19**
 - En América del Sur
 - En el mundo
- 5 **Trabajo a futuro**
- 6 **Referencias**



¿Por qué es necesario usar técnicas particulares de clustering cuando los datos son series de tiempo?

- Las series de tiempo son a menudo muy largas por lo que están en un espacio de alta dimensión lo que dificulta las técnicas habituales de clustering (por ejemplo, los datos generados por un electrocardiograma ocupan un giga de espacio, un blog habitual de la web ocupa alrededor de 5 gigas, etc).
- Las técnicas habituales de clustering toman cada objeto (o sea cada dato) como un único individuo de forma atómica (toman cada dato como un átomo) mientras que si cada dato es una serie de tiempo, dentro de cada dato tenemos mucha información, tenemos estructura en cada dato, lo cual nos permite por ejemplo tomar como criterio de homogeneidad alguna medida que observe la estructura interna de dicha serie.

FORMAS DE REALIZAR CLUSTERING EN SERIES DE TIEMPO

En general hay tres formas de realizar clustering en series de tiempo: las basadas en la forma de la curva, las basadas en alguna característica y las basadas en modelos.

- 1 Las basadas en la forma de la curva en general consideran como medida de similitud alguna característica que compare las formas de las curvas.
- 2 Las basadas en alguna característica, reducen la dimensión ya que de cada serie se toma en cuenta un vector que capte dichas características para luego aplicar técnicas convencionales de clustering.
- 3 Las basadas en modelos, también reducen la dimensión puesto que ajustan una familia de modelos a cada serie y luego se realizan clustering convencionales con los vectores de los parámetros estimados.

Contenidos

- 1 Introducción
- 2 Distance Time Warping (DTW)
 - Warping functions
 - Definición de $DTW(x,y)$
 - Cálculo de $DTW(x,y)$
 - Dos virtudes importantes de la metodología DTW
 - Restricciones
 - TADPole
- 3 Shape based distance (SBD)
 - Definición
 - Detección de los centroides
- 4 Aplicación: covid19
 - En América del Sur
 - En el mundo
- 5 Trabajo a futuro
- 6 Referencias

Supongamos que tenemos la observación de dos series de tiempo $x = (x_1, x_2, \dots, x_n)$, $y = (y_1, y_2, \dots, y_m)$ de longitudes n y m , una alineación π entre ambas series de tiempo será vincular los n valores observados de las x con los m valores observados de las y de manera creciente en el tiempo (subíndices) vinculando x_1 con y_1 (inicio de la curva x con inicio de la curva y), x_n con y_m (final de la curva x con final de la curva y), donde se permite que cada x_j esté vinculado con más de un valor y_j de modo no decreciente. Lo formalizamos en la siguiente definición.

Definition (Warping functions)

Fijados n, m y $p \leq n + m - 1$, un alineamiento entre el conjunto de índices $I_n = \{1, 2, \dots, n\}$ e $I_m = \{1, 2, \dots, m\}$ es una función $\pi = (\pi_1, \pi_2) : I_p \rightarrow I_n \times I_m$ tal que

- 1 $1 = \pi_1(1) \leq \pi_1(2) \leq \dots \leq \pi_1(p) = n$, donde $\pi_1(i+1) - \pi_1(i) \leq 1$ para todo $i = 1, 2, \dots, p$,
- 2 $1 = \pi_2(1) \leq \pi_2(2) \leq \dots \leq \pi_2(p) = m$, donde $\pi_2(i+1) - \pi_2(i) \leq 1$ para todo $i = 1, 2, \dots, p$.

Es decir que π_1 recorre de manera no decreciente los índices desde 1 hasta n , permitiendo repeticiones o aumentando de a una unidad y análogamente con π_2 .

Ejemplo.

Si $n = 2$ $m = 3$, tenemos 5 posibles alineamientos que son los siguientes:

$$(1) : \begin{array}{ccc} x_1 & x_2 & \\ \downarrow & \downarrow & \searrow \\ y_1 & y_2 & y_3 \end{array} \quad \pi : (1, 1) \rightarrow (2, 2) \rightarrow (2, 3).$$

$$(2) : \begin{array}{ccc} x_1 & x_2 & \\ \downarrow & \swarrow & \downarrow & \searrow \\ y_1 & & y_2 & y_3 \end{array} \quad \pi : (1, 1) \rightarrow (2, 1) \rightarrow (2, 2) \rightarrow (2, 3).$$

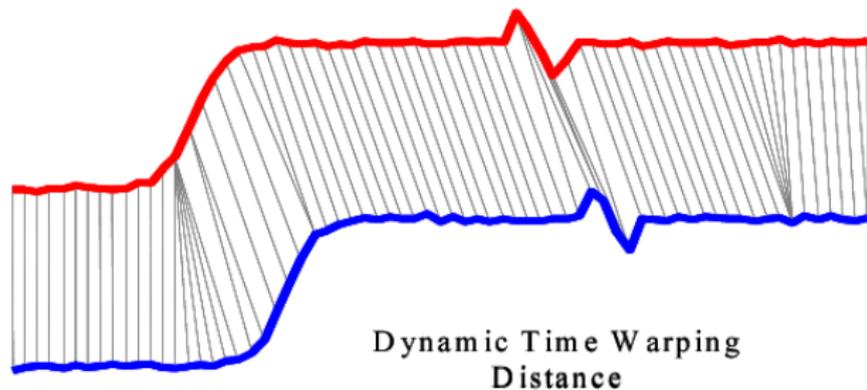
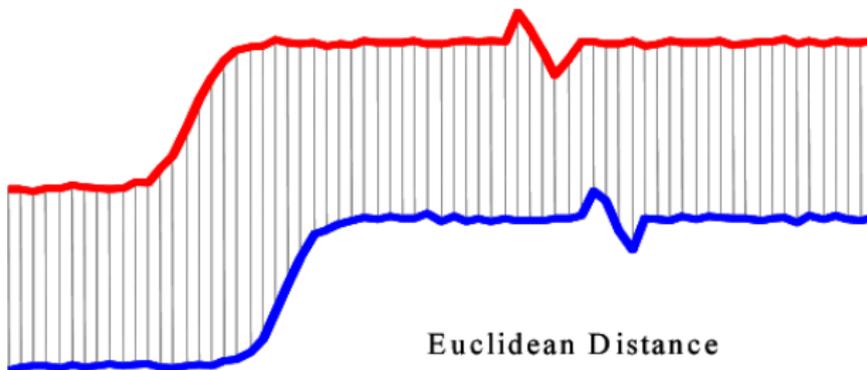
$$(3) : \begin{array}{ccc} x_1 & x_2 & \\ \downarrow & \searrow & \searrow \\ y_1 & y_2 & y_3 \end{array} \quad \pi : (1, 1) \rightarrow (1, 2) \rightarrow (2, 3).$$

En el caso (5) del ejemplo anterior, podemos dar una medida de la “discrepancia entre las curvas x e y ” calculando la cantidad

$$|x_1 - y_1| + |x_1 - y_2| + |x_1 - y_3| + |x_2 - y_3|$$

y en general dado el alineamiento π tomamos como medida de la discrepancia

$$D_{x,y}(\pi) = \sum_{i=1}^p |x_{\pi_1(i)} - y_{\pi_2(i)}|.$$



Definición (Distancia DTW entre las series x e y)

Dadas las series $x = (x_1, x_2, \dots, x_n)$, $y = (y_1, y_2, \dots, y_m)$ la distancia DTW entre ellas se define mediante

$$DTW(x, y) = \min_{\pi \in A(n, m)} D_{x, y}(\pi) = \min_{\pi \in A(n, m)} \sum_{i=1}^p |x_{\pi_1(i)} - y_{\pi_2(i)}|$$

siendo $A(n, m)$ el conjunto de todos los posibles alineamientos entre I_n e I_m .

- De esta forma, consideramos todas las posibles “deformaciones en el tiempo” que vinculan x con y , y tomamos el alineamiento que minimiza la diferencia entre ambas series medidas a través de $D_{x,y}(\pi)$.
- DTW no es una función de distancia puesto que no cumple la desigualdad triangular.
- La cantidad de alineamientos posibles es $D(n-1, m-1)$ siendo

$$D(n, m) = \sum_{k=0}^{m \wedge n} \binom{m}{k} \binom{n}{k} 2^k \text{ (números de Delannoy).}$$

- $D(n, m)$ calcula la cantidad de trayectorias tomando valores en el cuadrículado $\{0, 1, 2, \dots, n\} \times \{0, 1, 2, \dots, m\}$ que unen el punto $(0, 0)$ con el punto (n, m) donde del punto (i, j) puede pasarse al $(i + 1, j)$, o al $(i, j + 1)$ o al $(i + 1, j + 1)$ (es decir que los posibles movimientos son en una unidad en dirección horizontal a la derecha, o en una unidad en dirección vertical para arriba, o en una unidad en ambas coordenadas en dirección diagonal hacia arriba).
- En la próxima figura vemos los $25(= D(3, 2))$ caminos de Delannoy posibles que unen $(0, 0)$ con $(3, 2)$

DOS VIRTUDES IMPORTANTES DE ESTA METODOLOGÍA

- 1 Puede ser utilizado para series de distinto tamaño.
- 2 Puede ser utilizado para el caso de series multivariadas.

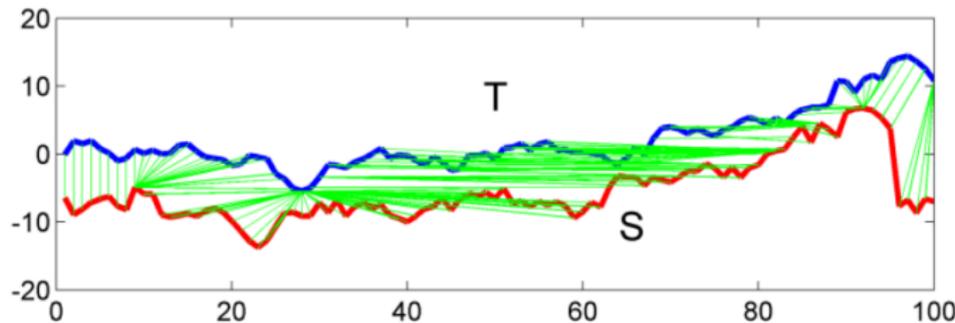
RESTRICCIONES

- Se llaman también restricción por ventanas (window constraints) y tienen dos fines, por un lado minimizar el tiempo de ejecución del programa y por otro eliminar posibles patologías en la deformación.
- Se suele utilizar un 10% de la longitud de la serie como ventana, pero pueden (a veces) obtenerse mejores resultados con un porcentaje menor. La más utilizada de las ventanas es la propuesta por Sakoe & Chiba.
- Cuando se usan ventanas inclinadas y el tamaño de las mismas es w , se consideran los puntos en el rango $[(i, j - w); (i, j + w)]$. De esta forma, cuando la ventana es de tamaño w , en cada paso $2w + 1$ puntos son usados.



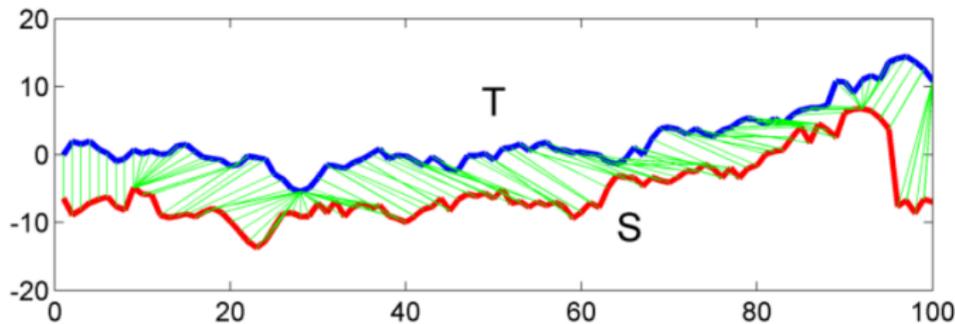
a

Classic DTW

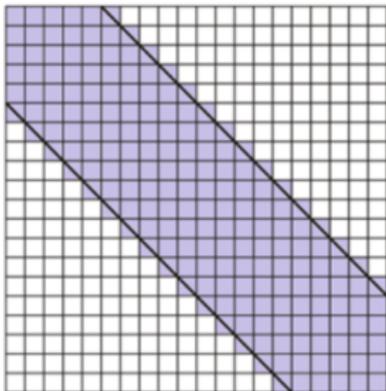


b

Restricted DTW

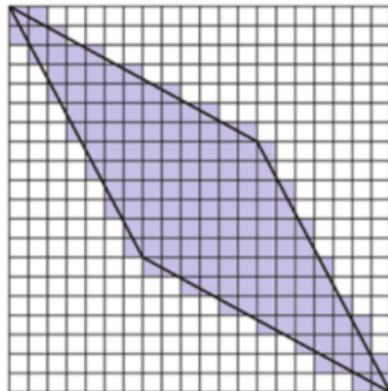


a



Sakoe-Chiba band

b



Itakura parallelogram

GLOBAL ALIGNMENT KERNEL DISTANCE (Cuturi, 2011)

Definición

*Dadas las series de tiempo $x = (x_1, x_2, \dots, x_n)$,
 $y = (y_1, y_2, \dots, y_m)$*

$$k_{GA}(x, y) = \sum_{\pi \in A(n, m)} e^{-|x_{\pi_1(i)} - y_{\pi_2(i)}|}$$

En esta definición, en lugar de quedarse con sólo la alineación óptima, se hace uso de todos los posibles alineamientos entre las series x e y .

TADPole (Begum, Ulanova, Wang & Keogh, 2015)

- Se usa la distancia DTW con un algoritmo más eficiente de recorte. Puede ser considerado una clase de algoritmos PAM dado que el centroide de cada grupo, es un elemento perteneciente al mismo. Los autores proponen un algoritmo inspirado en el algoritmo DP (density peaks) propuesto por Rodríguez & Laio en 2014.
- Trabajando con series reales, se ha observado que los mejores clustering en series de tiempo se obtienen cuando se eliminan algunos datos. El algoritmo DP se ha mostrado que es capaz de ignorar datos anómalos. DP también tiene la habilidad de funcionar con series cuyo gráfico tenga cualquier tipo de forma.
- El algoritmo DP asume que los centros de los clusters están rodeados por vecinos de menor densidad local y tienen una distancia mayor a todo punto en una zona de mayor densidad local.



- El algoritmo DP tiene dos parámetros que son la densidad local (ρ_i) y la distancia desde puntos con mayor densidad local (δ_i).

Definición

Dados x_1, x_2, \dots, x_m m series de tiempo a agrupar el clusters definimos

$$\rho_i = \#\{x_j : d(x_i, x_j) \leq d_c\}$$

$$\delta_i = \min\{d(x_i, x_j) : d(x_i, x_j) > d_c\}$$

La densidad local del punto i (ρ_i) se define como el número de puntos que están a una distancia menor o igual a un valor de corte d_c .

La distancia desde puntos de mayor densidad (δ_i) es la mínima distancia entre el punto i y los puntos en zonas de mayor densidad.

Contenidos

- 1 Introducción
- 2 Distance Time Warping (DTW)
 - Warping functions
 - Definición de $DTW(x,y)$
 - Cálculo de $DTW(x,y)$
 - Dos virtudes importantes de la metodología DTW
 - Restricciones
 - TADPole
- 3 Shape based distance (SBD)
 - Definición
 - Detección de los centroides
- 4 Aplicación: covid19
 - En América del Sur
 - En el mundo
- 5 Trabajo a futuro
- 6 Referencias

- Propuesto por Paparrizos & Gravano en 2016, dan un procedimiento para realizar clustering de acuerdo a la forma de las curvas a agrupar basadas en la correlación cruzada entre las series. En particular definen una distancia para tener en cuenta en este método de clusterización y dan un algoritmo para calcular los centroides de los grupos.
- La medida está basada en el cálculo de las correlaciones cruzadas entre x e y para todos los posibles “delays” que existan entre ambas series.

Definición (shift de s términos para una serie $x = (x_1, \dots, x_n)$)

El vector shift de tamaño s de la serie x es el vector $x_{(s)} \in \mathbb{R}^n$ definido mediante

$$x_{(s)} = \begin{cases} (0, \dots, 0, x_1, x_2, \dots, x_{n-s}) & \text{si } 0 \leq s \leq n-1 \\ (x_{n+s}, \dots, x_{n-1}, x_n, 0, \dots, 0) & \text{si } -n+1 \leq s < 0 \end{cases} .$$

Los casos $s = n$ y $s = -n$, no los consideramos puesto que queda el vector nulo. Por lo tanto tenemos $2n - 1$ posibles vectores shift para x .

Definición (Correlación cruzada)

Dadas las series $x = (x_1, x_2, \dots, x_n)$ e $y = (y_1, y_2, \dots, y_n)$ definimos la sucesión de correlación cruzada entre ambas series a

$$CC_w(x, y) = R_{w-n}(x, y) \text{ para } w = 1, 2, 3, \dots, 2n - 1$$

siendo

$$R_k(x, y) = \begin{cases} \sum_{i=1}^{n-k} x_{i+k} y_i & \text{si } 0 \leq k \leq n-1 \\ \sum_{i=1}^{n+k} y_{i-k} x_i & \text{si } -n+1 \leq k \leq 0 \end{cases}$$

y la correlación cruzada normalizada como

$$NCC_w(x, y) = \frac{CC_w(x, y)}{\sqrt{R_0(x, x) R_0(y, y)}}$$

Definición (Shape-based distance (SBD))

Dadas las series $x = (x_1, x_2, \dots, x_n)$ e $y = (y_1, y_2, \dots, y_n)$ definimos la distancia basada en la forma (shape-based distance) entre ambas como

$$SBD(x, y) = 1 - \max_{w \in \{1, 2, \dots, 2n-1\}} NCC_w(x, y).$$

- $SBD(x, y)$ toma valores entre 0 y 2 y da cero cuando hay “perfecta similaridad” entre las series ya que para el w donde se obtiene $\max_{w \in \{1, 2, \dots, 2n-1\}} NCC_w(x, y)$, este máximo es 1 si y sólo si $x_{i+k} = \lambda y_i$ para determinado $\lambda > 0$, para todo i es decir que ambas curvas tendrán la misma curva con un delay de $|k|$ (k puede ser positivo o negativo).
- Observamos que esta distancia es invariante por shifts y para que sea invariante por cambios de localización y escala basta aplicarla a los datos normalizados restando su media y dividiendo entre su desviación

- Los autores proponen utilizar la distancia *SBD* para la extracción del centroide como sigue: si le llamamos μ_k a un posible centroide del grupo P_k , entonces

$$\mu_k^* = \arg \max_{\mu_k} \sum_{x^{(i)} \in P_k} NCC(x^{(i)}, \mu_k).$$

- La complejidad del algoritmo de clusters basado en esta distancia es

$$O\left(\max\left\{mkn \log n, mn^2, kn^3\right\}\right)$$

siendo k la cantidad de clusters, m la cantidad de series y n la longitud de cada serie.

- Observar que esta metodología de clusters soporta únicamente series univariadas (a diferencia de DTW por ejemplo) y además es fácilmente adaptable para series de distinto tamaño a partir de la definición de CC_w .

Contenidos

- 1 Introducción
- 2 Distance Time Warping (DTW)
 - Warping functions
 - Definición de $DTW(x,y)$
 - Cálculo de $DTW(x,y)$
 - Dos virtudes importantes de la metodología DTW
 - Restricciones
 - TADPole
- 3 Shape based distance (SBD)
 - Definición
 - Detección de los centroides
- 4 **Aplicación: covid19**
 - En América del Sur
 - En el mundo
- 5 Trabajo a futuro
- 6 Referencias

- Aplicaremos algunas de estas técnicas para separar clusters de países de Sudamérica (continental) observando las series de tiempo del acumulado de muertos por covid19.
- Para cada país, se usaron los datos correspondientes al total de muertos por covid 19 desde el 22 de enero de 2020 al 27 de setiembre de 2020, formando una serie de 250 datos.
- Se utilizó el paquete “dtwclust” en R que tiene implementadas todas las metodologías aquí expuestas más otras.



Acumulado de muertes por millón de habitantes desde el 22 de enero al 27 de setiembre de 2020 (13 países de América del Sur)

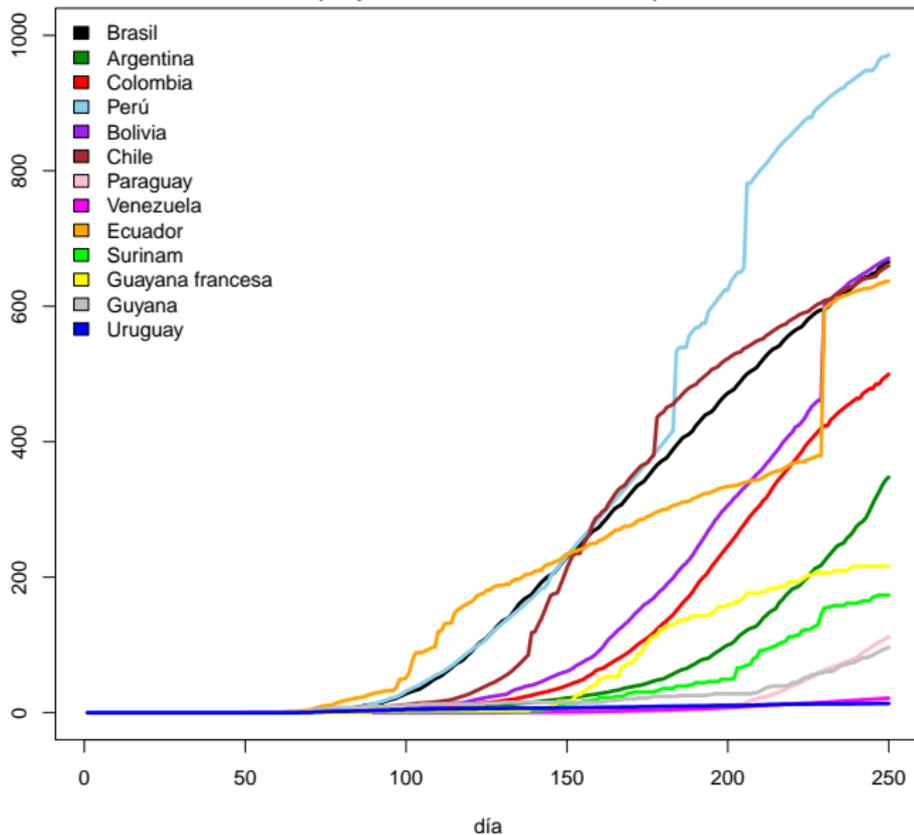


Table : Medidas que sugieren la cantidad óptima de clusters (cuando la distancia utilizada es DTW): Silhouette coeficiente (SIL), Calinski Harabasz (CH), Davis Boudin (DB), Dunn (D), COP

<i>DTW</i>	$k = 2$	$k = 3$	$k = 4$	$k = 5$
SIL ↑	0.789	0.578	0.559	0.378
CH ↑	25.46	16.50	10.02	9.56
DB ↓	0.242	0.327	0.504	2.158
D ↑	0.151	0.009	0.132	0.066
COP ↓	0.099	0.077	0.055	0.058

- La metodología GAK también sugiere $k = 2$ clusters (según SIL, CH, DB y D) mientras que COP sugiere $k = 5$ grupos.

Definición de las medidas de pureza de clusters utilizadas.

En todos los casos se parte de $x_1, x_2, \dots, x_n \in \mathbb{R}^p$ separados en k grupos disjuntos y que contienen a las n observaciones que llamamos C_1, C_2, \dots, C_k y se utilizan las siguientes notaciones:

$d(x, y)$ indicará la distancia euclídeana entre x e y ($\in \mathbb{R}^p$); \bar{x} el centroide de las n observaciones, \bar{C}_j el centroide del grupo C_j

(para cada j); dado un conjunto A , $|A|$ indica la cantidad de

elementos del conjunto A , $\text{diam}(A) \stackrel{\text{def}}{=} \sup_{a, a' \in A} d(a, a')$

(diámetro del conjunto A); dados dos conjuntos A y B , la

distancia entre ambos la definimos mediante

$d(A, B) \stackrel{\text{def}}{=} \inf_{a \in A, b \in B} d(a, b)$.

- Silhouette (\nearrow)(Rousseauw 1987)

$$\text{SIL} = \frac{1}{n} \sum_{j=1}^k \sum_{x \in C_j} \frac{a(x, C_j) - b(x, C_j)}{\max \{a(x, C_j), b(x, C_j)\}}$$

siendo $a(x, C_j) = \frac{1}{|C_j|} \sum_{y \in C_j} d(x, y)$,

$b(x, C_j) = \min_{C_h \neq C_j} \frac{1}{|C_h|} \sum_{y \in C_h} d(x, y)$.

- Calinski–Harabasz (\nearrow)(1974)

$$\text{CH} = \frac{(n - k) \sum_{j=1}^k |C_j| d(\bar{C}_j, \bar{x})}{(k - 1) \sum_{j=1}^k \sum_{x_i \in C_j} d(x_i, \bar{C}_j)}$$

- Dunn (\nearrow)(1973)

$$D = \frac{\min_{i,j=1,2,\dots,k} d(C_i, C_j)}{\max_{i,j=1,2,\dots,k} \text{diam}(C_i)}$$



- Davis–Bouldin (↘) (1979)

$$DB = \frac{1}{k} \sum_{j=1}^k \max_{i \neq j} \frac{S(C_j) + S(C_i)}{d(\bar{C}_j, \bar{C}_i)}$$

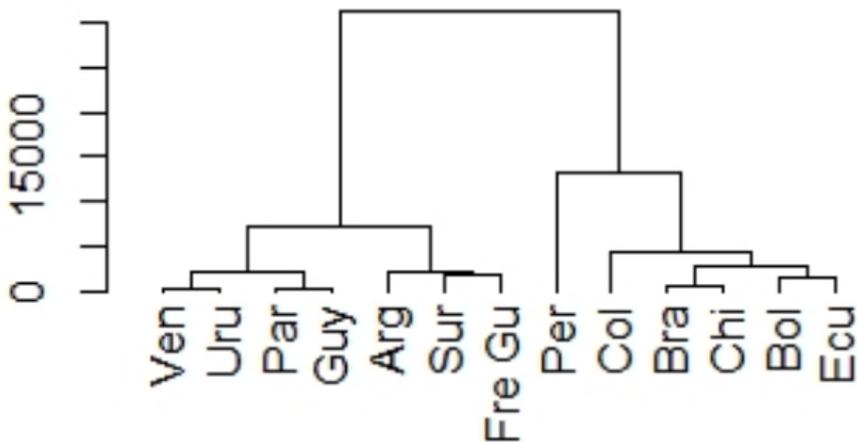
siendo $S(C_j) = \frac{1}{|C_j|} \sum_{x \in C_j} d(x, \bar{C}_j)$.

- COP (↘) (2010)

$$COP = \frac{1}{n} \sum_{j=1}^k |C_j| \frac{\frac{1}{|C_j|} \sum_{x \in C_j} d(x, \bar{C}_j)}{\min_{x \notin C_j} \max_{y \in C_j} d(x, y)}$$

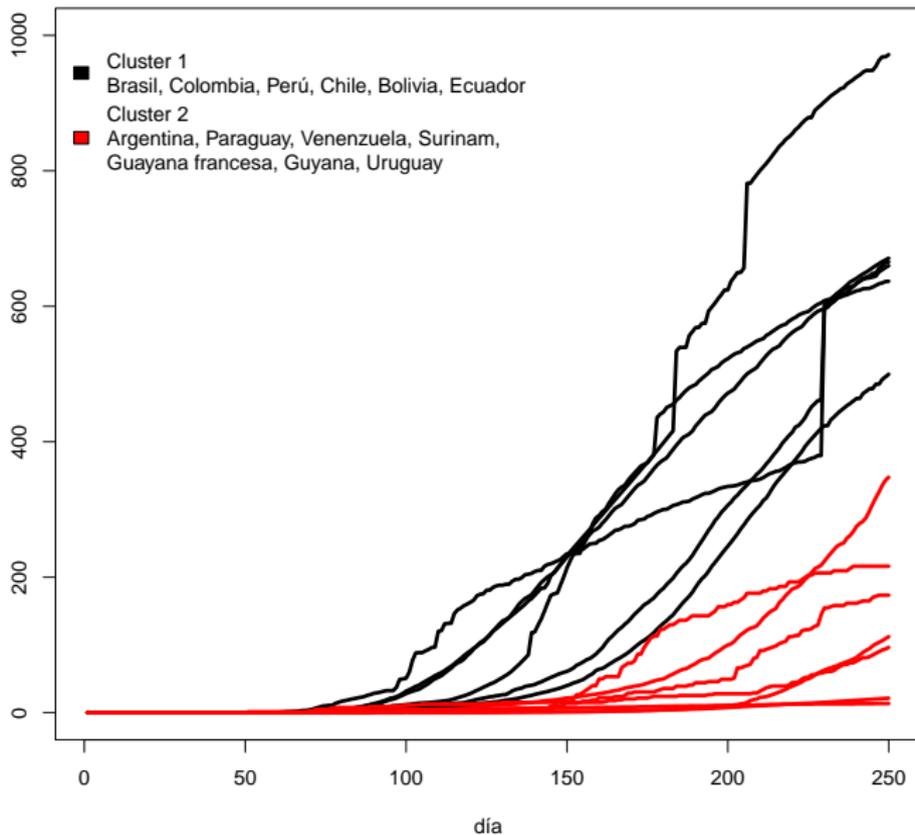


Agrupamiento de países (según DTW) de América del Sur.



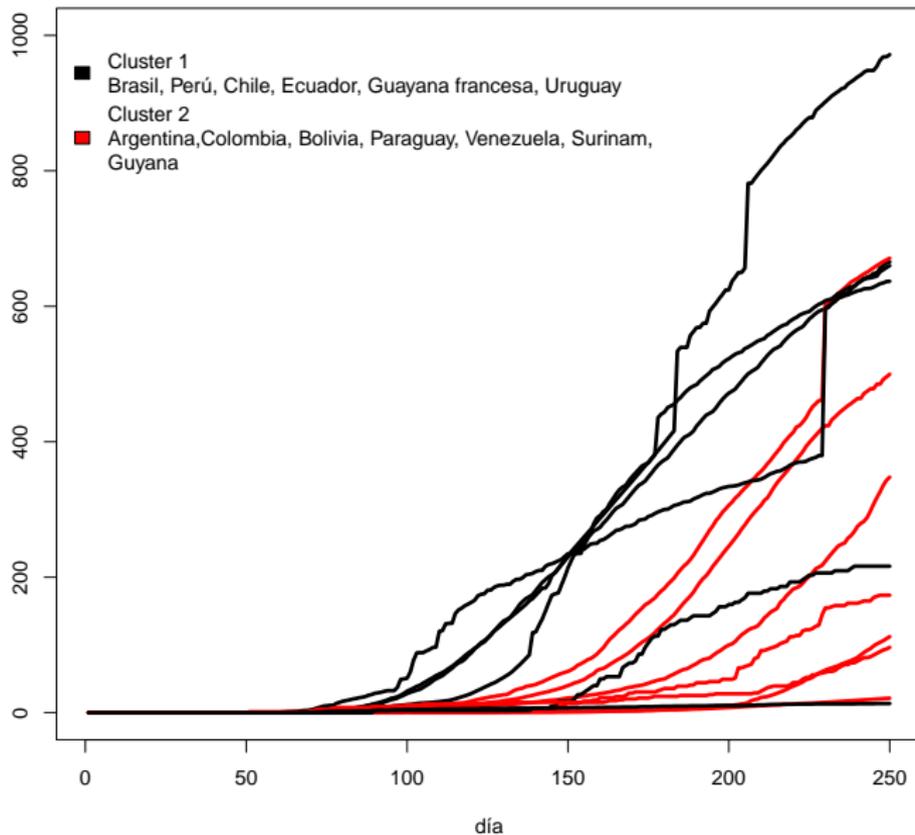


Clusters 1 y 2 (metodología DTW)





Clusters 1 y 2 (metodología SBD)

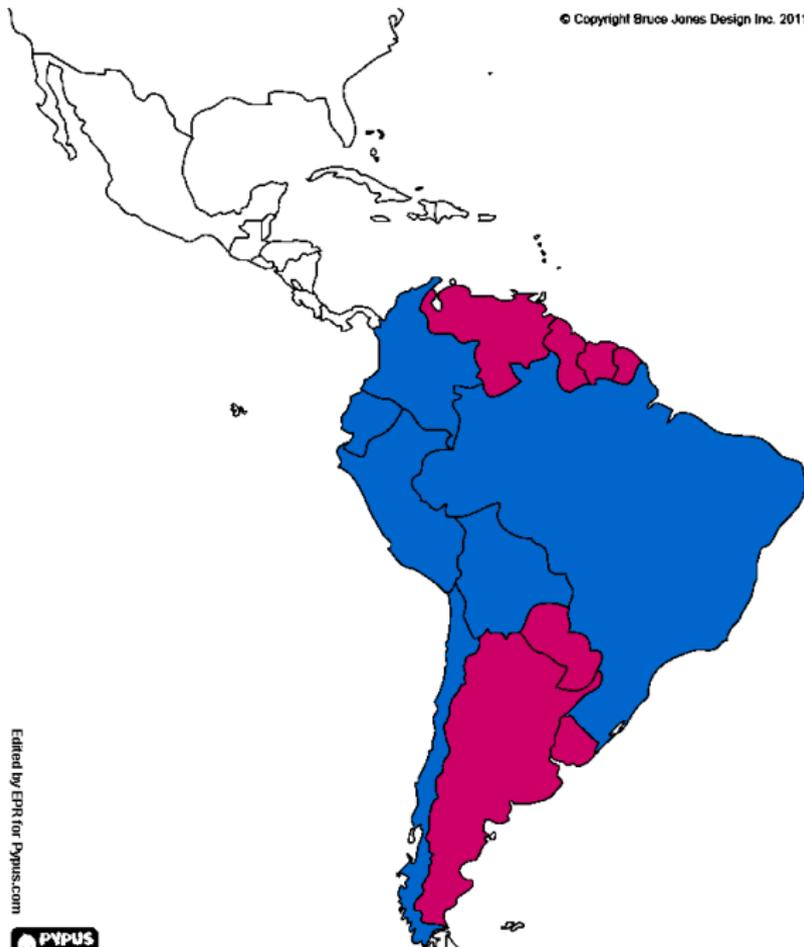


oooooooooooooooooooo

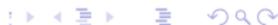
oooo

oooooooooooo●oooo

© Copyright Bruce Jones Design Inc. 2011



Edited by EPR for Pypus.com



- Considerando 175 países del mundo, usando DTW, al igual que en América del Sur, salvo COP, todas las medidas sugieren separar en dos clusters.
- COP sugiere separar en 11 clusters (aunque los valores son todos pequeños y similares para esta medida).
- Separando en 11 clusters, en los siguientes gráficos, vemos la distribución geográfica (por continente) de aquellos países que se encuentran en el mismo cluster que Uruguay.
- No aparecen América del Norte (por no tener ningún país en el mismo cluster que Uruguay), Oceanía (puesto que están todos) ni América del Sur (sólo está Venezuela).



© Copyright Bruce Jones Design Inc. 2011



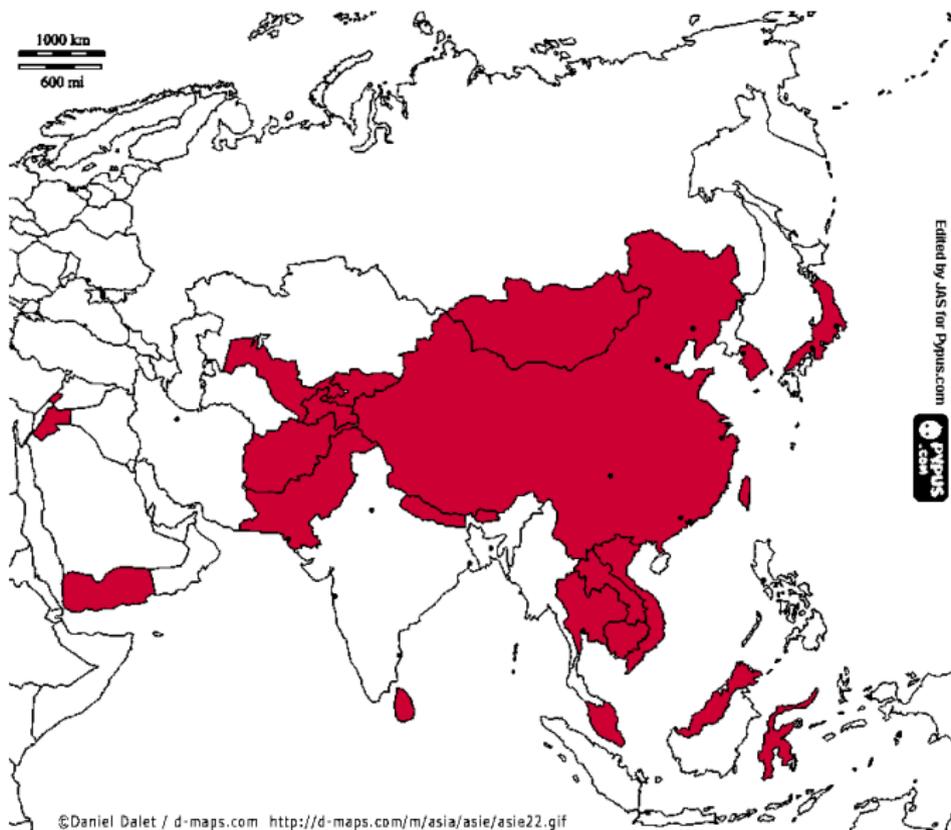
Edited by EPR for Papyrus.com



oooooooooooooooooooo

oooo

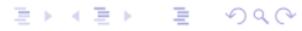
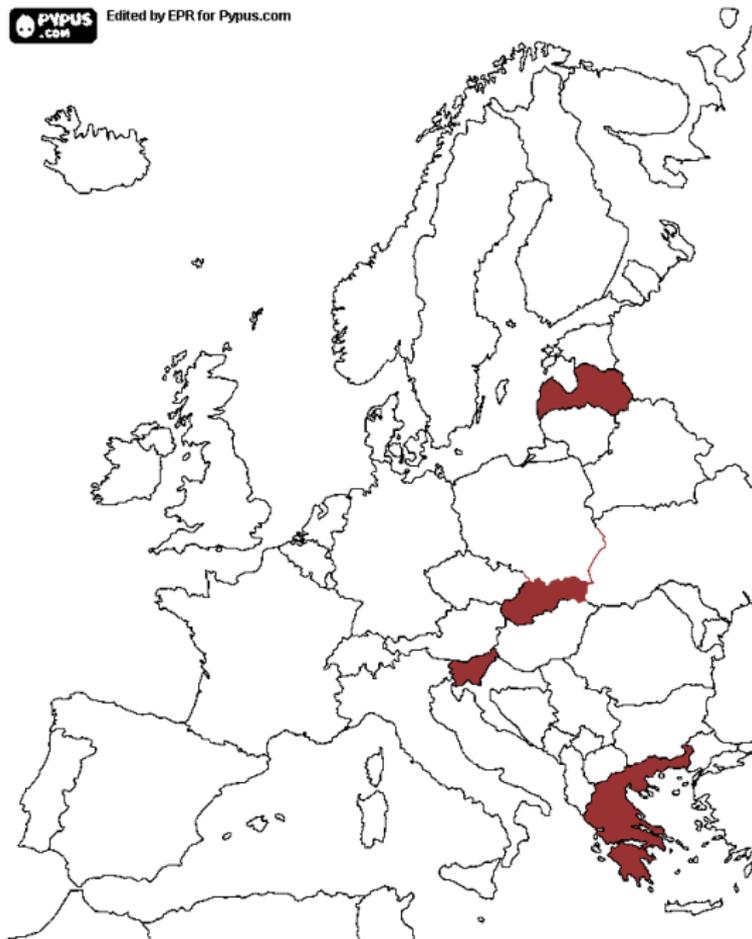
oooooooooooooooo●●



©Daniel Dalet / d-maps.com <http://d-maps.com/m/asia/asia/asia22.gif>



Edited by EPR for Pypus.com



Contenidos

- 1 Introducción
- 2 Distance Time Warping (DTW)
 - Warping functions
 - Definición de $DTW(x,y)$
 - Cálculo de $DTW(x,y)$
 - Dos virtudes importantes de la metodología DTW
 - Restricciones
 - TADPole
- 3 Shape based distance (SBD)
 - Definición
 - Detección de los centroides
- 4 Aplicación: covid19
 - En América del Sur
 - En el mundo
- 5 Trabajo a futuro
- 6 Referencias

- Realizar un estudio comparativo de los índices que miden la pureza de los clusters aplicados a series de tiempo (SIL, CH, etc).
- Investigar la clusterización mediante una nueva medida de distancia basadas en porcentajes de recurrencias.
- Aplicación de estas técnicas para series de tiempo que tengan interés para Uruguay: series económicas, meteorológicas, médicas, etc.

Contenidos

- 1 Introducción
- 2 Distance Time Warping (DTW)
 - Warping functions
 - Definición de $DTW(x,y)$
 - Cálculo de $DTW(x,y)$
 - Dos virtudes importantes de la metodología DTW
 - Restricciones
 - TADPole
- 3 Shape based distance (SBD)
 - Definición
 - Detección de los centroides
- 4 Aplicación: covid19
 - En América del Sur
 - En el mundo
- 5 Trabajo a futuro
- 6 Referencias



- Begum N, Ulanova L, Wang J, Keogh E. “Accelerating Dynamic Time Warping Clustering with a Novel Admissible Pruning Strategy”. Proceedings of 21 th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. (2015).
- Cuturi, M. “Fast Global Alignment Kernels”, Proceedings 28º International Conference on Machine Learning, https://icml.cc/Conferences/2011/papers/489_icmlpaper.pdf (2011).
- Montalto, P., Aliotta, M., Cannata, A., Cassisi, C., Pulvirenti, A., “Similarity measures and dimensionality reduction techniques for time series data mining”. Chapter in Advances in Data Mining Knowledge Discovery and Applications (2012).
- Sakoe, Chiba (1978). “Dynamic programming algorithm optimization for spoken word recognition”. IEEE Trans Acoustics Speech Signal Process. ASSP 26:43–49.

- Paparrizos, J., Gravano, L., “*k*-Shape: Efficient and Accurate Clustering of Time Series”. ACM SIGMOD Record <https://doi.org/10.1145/2949741.2949758> (2016).