UN MÉTODO DE ALTERNATIVO PARA LA VIGILANCIA EPIDEMIOLÓGICA

Ramón Alvarez[1]; Andrés Castrillejo[1]

RESUMEN

En vigilancia epidemiológica se usan algunas herramientas gráficas para el estudio de fenómenos con *distribución tempo-espacial* como son algunas enfermedades transmisibles. Una de las herramientas que se usan, es el corredor endémico (CE) que permite establecer zonas de riesgo, de alerta y de epidemia para cierta enfermedad.
El CE permite clasificar a una serie $r_{i,j}$ del año i y período j intraanual en 4 grupos , luego de ser comparada con la media geométrica móvil $MG^{v}_{i,j}$ de orden v para el período j desde $(i-1-v, i-1)$, con una varianza $S^2_{(i-1-v;i-1,j)} = \tau_{i,j}$ De esa manera la serie r_{ij} puede ser clasificada como

- $G_{i,j}1$ sin riesgo
- $G_{i,j}2$ riesgo bajo
- $G_{i,j}3$ zona de alerta
- $G_{i,j}4$ zona de epidemia

Un uso habitual de esta herramienta es para la comparación de situaciones de riesgo para distintos años procurando identificar años similares.

Se propone una alternativa sencilla que posee las ventajas de:

- tener en cuenta la posible autocorrelación intra-anual de los datos
- tener cierta capacidad predictiva.

Este enfoque pretende emular el (CE) pero utilizando la información intra-anual de una manera elemental. Se construye una forma anual de CE y se buscan alternativas para medir proximidades respecto de él.

Se configuran n series $r_{(i,.)}$ (n años) que describen las trayectorias anuales. Se calcula la disimilaridad de cada serie respecto a al CE y se agrupan los años similares. Una medida de disimilaridad entre series bastante utilizada es la *Dynamic Time Warping* (DTW) que es una alternativa a la Distancia Euclídea para "alinear" y medir la proximidad de dos series. A partir de esta y otras formas de disimilaridad se hace clustering.

Adicionalmente se toman los posibles estados ($G_{i,j}1, \ldots, G_{i,j}4$) obtenidos por el (CE) y se comparan con el agrupamiento en esos grupos de riesgo obtenidos con las distintas medidas de disimilaridad.

Se comparan los resultados de aplicar esta aproximación con el (CE), sobre datos mensuales del período 1980-2009 de casos notificados de Hepatitis A para Uruguay.

Palabras clave: *Vigilancia Epidemiológica; Corredor Endémico; Clustering ; Dynamic Time Warping.*

[1]INSTITUTO DE ESTADISTICA

1. Introducción

En vigilancia epidemiológica se usan algunas herramientas gráficas para el estudio de fenómenos con *distribución tempo-espacial* como son algunas enfermedades transmisibles. Una de las herramientas que se usan, es el corredor endémico (CE) que permite establecer zonas de riesgo, de alerta y de epidemia para cierta enfermedad. Estos se basan en comparar la incidencia actual con la de períodos anteriores. La incidencia es una medida de frecuencia de eventos morbosos que se usa en salud pública, expresada generalmente como tasa cada 100.000 habitantes y que toma en cuenta los casos nuevos de una enfermedad en una población en un período determinado. Esta medida de frecuencia expresada para los casos nuevos da una idea de cual es la dinámica de la enfermedad bajo estudio. Si el número de casos observados es mayor que el de los esperados se confirma la existencia de lo que se peude llamar brote epidemico.

Concretamente el método de (CE) consiste en

$$
\begin{aligned}
&G_{i,j}1(\text{sin riesgo}) = r_{i,j} \leq MG^{v}_{i,j} - 2 * \phi_{1-\alpha/2} * \sqrt{\tau_{i,j}} \\
&G_{i,j}2(\text{riesgo bajo}) = r_{i,j} \leq M^{v}_{i,j} \\
&G_{i,j}3(\text{alerta}) = r_{i,j} \leq MG^{v}_{i,j} + 2 * \phi_{1-\alpha/2} * \sqrt{\tau_{i,j}} \\
&G_{i,j}4(\text{epidemia}) = r_{i,j} > MG^{v}_{i,j} + 2 * \phi_{1-\alpha/2} * \sqrt{\tau_{i,j}}
\end{aligned}
$$

El método utilizado para hallar el corredor endémico consiste en considerar la media de la incidencia semanal o mensual de los casos para un momento dado y el intervalo de confianza para la misma. Si se supone que se está trabajando con datos mensuales el (CE) permite clasificar a una serie $r_{i,j}$ del año i y período j intraanual en 4 grupos, luego de ser comparada con la media. Como habitualmente los datos pueden presentar una gran asimetría se aplica una transformación de los datos mediante Box Cox, aplicando logaritmo a las tasas calculadas y de esta manera lo que se considera para comparar es una media geométrica móvil $MG^{v}_{i,j}$ de orden v para el período j desde $(i-1-v, i-1)$, con una varianza $S^2_{(i-1-v;i-1,j)} = \tau_{i,j}$.(Bortman 1999) Se crea por tanto un clasificador (cCE) que para unaserie r_{ij} asigna a:

$$
\begin{aligned}
&G_{i,j}1(\text{sin riesgo}) = r_{i,j} \leq MG^{v}_{i,j} - 2 * \phi_{1-\alpha/2} * \sqrt{\tau_{i,j}} \\
&G_{i,j}2(\text{riesgo bajo}) = r_{i,j} \leq M^{v}_{i,j} \\
&G_{i,j}3(\text{alerta}) = r_{i,j} \leq MG^{v}_{i,j} + 2 * \phi_{1-\alpha/2} * \sqrt{\tau_{i,j}} \\
&G_{i,j}4(\text{epidemia}) = r_{i,j} > MG^{v}_{i,j} + 2 * \phi_{1-\alpha/2} * \sqrt{\tau_{i,j}}
\end{aligned}
$$

De esta manera si se esta trabajando con una serie mensual se tienen 12 tasa de incidencia promedio para los que se establece las 4 zonas antes mencionadas. Esta aproximación metodológica tiene algunas limitaciones que son importantes destacar como que es una herramienta fácil de implementar pero solo sirve a los efectos descriptivos y no se puede hacer predicción.

El supuesto que hay para la construcción de la tasa media de incidencia condicional , es que el proceso $r_{i,j}$ de generación de datos sigue una lógica de corte transversal, teniendo de esa manera 12 series de tamaño $n = 5$, una para cada mes. Si se piensa la cantidad de casos ocurridos o notificados sigue una evolución de un período a otro, con lo cual se deja de lado la dinámica que sigue el proceso.

Para eso existen ya trabajos que proponen una modificación como los de (Bortman 1999) que combinan el método de CE con modelos de crecimiento, que se detallan en 2, asćomo

los trabajos de(Orellano & Reynoso 2011) y (Alvarez, Dibarboure & F. 2010) (este último presentado en las jornadas academicas de facultad) que comparan la metodología de (CE) con modelos autoregresivos de tipo SARIMA . Otra posibilidad seria la de considerar para describir y también hacer predición, el uso de modelos de conteo de tipo MLG, que se detallan también en 2

2. Metodología

Además de los enfoques mixtos que se mencionan en (Orellano & Reynoso 2011) y en (Dibarboure, Alvarez, Quian & Mazza 2009) que combinan la metdología tradicional de (CE) y que permiten hacer predicción estan los modelos puramente predictivos como los que consideran modelos de conteo, donde la serie en casos absolutos puede estimarse mediante modelos de la forma

$$E(y_t) = \mu_t, log(\mu_t) = \alpha + \beta ?t \tag{1}$$

En general estos modelos pueden ser de tipo Poisson, quasi Poisson cuando existe sobre dispersión , o de tipo Binomial Negativo , todos estos siendo ejemplos de los modelos lineales generalizados (MLG) que se pueden factorizar como $\log(\mu) = x^\top \beta$. Cuando la serie que se desea monitorear es expresada en tasas, mediante una reparametrización adecuada se puede estimar mediante un modelo de regresión beta

$$f(y;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \quad 0 < y < 1, \tag{2}$$

Estos modelos son los más adecuados ya que permiten decidir en base a modelos, que toman en cuenta el tiempo, si los casos observados o la proporción de los mismos, presentan valores muy diferentes a los esperados bajo el supuesto del modelo planteado. Sin embargo en este trabajo la idea que se busca es poder usar y contrastar diferentes metodos de tipo descriptivos como son los de clustering que se detallan en la sección 2.1.

2.1. Métodos de Clustering

Existen muy variados metodos de clustering,dentro de los cuales se encuentran los de tipo jerárquico y no jerárquico, sobre los cuales se pueden aplicar diferentes tipos de distancia, que toman en cuenta métricas diferentes que se adaptan al tipo de variable considerada. En estos métodos que se reseñan la lgica considerada es de la información vista a través de un corte transversal, donde se ignora el paso del tiempo, considerando la dinámica intraanual para datos mensuales, teniendo en cuenta solamente lo que pasa en los mismos meses de diferentes años tal como lo hace la metodología de (CE). Los métodos de clustering que si toman en cuenta el tiempo se presentan en la sección 2.2.

2.1.1. Metodo de WARD

Los métodos jerárquicos se caracterizan por generar una serie de particiones encajadas y requieren la definición de una distancia. Inicialmente, cada objeto se le asigna a su propio grupo, y entonces los algoritmos proceden iterativamente, en cada etapa unen los dos grupos más similares, continuando hasta que sólo quede un solo grupo. En cada etapa las distancias definidas entre las agrupaciones se recalculan por la fórmula disimilitud de Lance-Williams actualizádose de acuerdo con el método de agrupación particular que se utilice. Dentro de los métodos jerárquicos se considera el de **WARD**, que consiste en descomponer la variación total en variación en los grupos (*within*) y variación entre los grupos (*between*) y al estar frente a una partición dada, el método unirá aquellos grupos que produzcan el efecto de hacer mínima la variación *within* en la nueva partición.

$$T = W + B \tag{3}$$

Donde T es la matriz de varianzas y covarianzas del total, W la matriz de varianzas y covarianzas dentro de los grupos y B la matriz de varianzas y covarianzas entre grupos.

2.1.2. Metodo de k-MEANS

Con respecto a los métodos no jerárquicos se encuentra el de *k-means*, que es uno de los más utilizados, en función de su simplicidad y velocidad de convergencia de que es de orden($n * p$), donde n es la cantidad de observaciones y p el número de variables . A partir de un conjunto de de n observaciones ($x1$, $x2$, ..., xn), que se puede considerar un vector p, el método de k-means buscar encontrar una partición de los n individuos en k subconjuntos con $k \leq n$,de manera de minimizar la suma de cuadrados intraclase (SCIC):

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in S_i} \left\| \mathbf{x}_j - \mu_i \right\|^2 \tag{4}$$

considerando μ como el centroide de los puntos en el grupo S_i

Al inicio, todos los centros de los conglomerados están en la media de las celdas de Voronoi (que se puede interpretar como el conjunto de puntos de los datos que están más cerca del centro de ese grupo que de cualquier otro grupo).

El algoritmo funciona de la siguiente manera:

1. Se eligen en forma aleatoria los centros iniciales. Queda entonces la siguiente secuencia $m_1.m_2, ..., m_k$ de $kcentros$

2. Se asigna cada observación al cluster con la media mas próxima, es decir que la partición queda determinada por el diagrama de Voronoi que se generó con las medias iniciales

3. Se calcula los S_i de la siguiente manera

$$S_i^{(t)} = \left\{ x_p : \left\| x_p - m_i^{(t)} \right\| \leq \left\| x_p - m_j^{(t)} \right\| \; \forall \; 1 \leq j \leq k \right\} \tag{5}$$

donde cada uno de los x_p queda asignado a uno de los $S_i^{(t)}$.

4. El algoritmo se actualiza calculando las nuevas medias del grupo

$$\mathbf{m}_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{\mathbf{x}_j \in S_i^{(t)}} \mathbf{x}_j \tag{6}$$

5. El algoritmo se detiene luego que al reasignar alguna observación a otro grupo no hay cambios menores a una tolerancia prefijada en la (SCIC)

Los algoritmos habitualmente usados en los paquetes estadísticos están basados en los que plantearon MacQueen (1967),Lloyd (1957) y Forgy (1965). El algoritmo de Hartigan-Wong es el que se usa habitualmente y el que está implementado en R (R Development Core Team 2010), al trabajar con centros iniciales que se eligen en forma aleatoria, se recomienda ejecutar el algoritmo varias veces ($n = 10$)de manera de ver la estabilidad de los resultados.

2.1.3. Método PAM (partition around medoids)

El algoritmo (PAM),que se basa en la búsqueda de k objetos representativos o medoides entre las observaciones del conjunto de datos, de manera que representen adecuadamente la estructura de los datos. Un *medoide* se podría definir como el objeto perteneciente a un cluster o conglomerado, cuyo promedio de disimilaridad a todos los objetos en el conglomerado es mínima, es decir, que se puede considerar como el punto más céntrico de la agrupación considerada.

La implementación más frecuente de *k-medoides* es el agrupamiento alrededor de medoides (PAM) que tiene el siguiente algoritmo: [2]

1. Inicialización: Selección al azar de k de los n puntos de datos como los candidatos a medoides *fase de construcción*).

2. Se asigna cada observación al cluster con el medoide mas próximo, dependiendo de la distancia elegida (euclidiana, Manhattan o Minkowski). Luego se encuentra un mínimo local para la función objetivo, es decir, una solución de tal manera el cambio de observación con un medoide haga que la función objetivo decrezca (esto se denomina la fase de intercambio).

3. Se repiten los pasos anteriores hasta que los medoides queden estables (es decir que haya cambios en los medoides).

2.1.4. FANNY

A diferencia de los métodos anteriores donde se obtiene una partición en k subconjuntos , es decir $N = \bigcup_1^n x_i = \bigcup_1^k \bigcup_{j=1}^{j=n_k} c_j$, donde c_j son los clusters determinados mediante algún algoritmo, en el agrupamiento *difuso* (fuzzy), cada observación se 'extiende' a través de los distintos grupos. En este método se puede establecer: m parámetro de incertidumbre (fuzziness parameter), v_k es el centro del cluster c y u_{ik} el grado de pertenencia del individuo i al grupo k. Si se tiene n el número de observaciones, k el número de clusters, r es el parámetro de ajuste del modelo y $d(i,j)$ la disimilaridad entre las observaciones i y j, el método *fanny* busca minimizar la función objetivo (Maechler, Rousseeuw, Struyf, Hubert & Hornik 2012)

$$\sum_{v=1}^{v=n} \frac{\sum_1^i \sum_1^j u_{i,v}^r u_{j,v}^r d(i,j)}{2\sum_1^n u_{j,}^r} \tag{7}$$

2.1.5. SILUETA

El método *Silueta* en realidad no es método de clustering en si mismo sino un método de interpretación y validación del número de conglomerados o cluster hallado(Kaufman & Rousseeuw 1990). Esta técnica permite obtener una representación gráfica de que tan bien esta cada observación dentro de un determinado grupo

.

Tiene la ventaja que puede ser utilizada para datos que hayan sido clasificados a través de cualquier método como por ejemplo *k-medias* 2.1.3 o k-medoides 2.1.3. Para cada observación i, sea $a(i)$ la disimilaridad promedio de i con todos los demás observaciones dentro del mismo grupo. Se puede interpretar que (i) esta bien emparejado con los restantes integrantes del grupo cuanto menor sea $a(i)$. Para los demás clusters C se define $d(i,C)$ como la disimilaridad promedio de i con los datos de C. Se repite el procedimiento para cada grupo del cual el i no es miembro y se determina $b(i) := min_C d(i,C)$, que representa la mínima disimilaridad promedio de i con cualquier otro grupo, lo que representa la disimilaridad entre i y los cluster vecinos. Se define entonces el estadístico $S(i)(silueta)$ como

$$S(i) = \frac{(b(i) - a(i))}{max[(b(i) - a(i))]} \tag{8}$$

De la definición anterior, se puede ver que $-1 \leq S(i) \leq 1$, en donde si $S(i)$ esta cerca de 1 significa que los individuos estan correctamente clasificados , valores cercanos a 0 que los grupos no están bien determinados y valores negativos der $S(i)$ que esa observación debería pertenecer a otro grupo.

2.2. Clustering en Series de Tiempo

El Análisis de Cluster (Conglomerados) o Clustering tiene por objetivo encontrar grupos(*clusters*) de objetos no etiquetados que sean"similares" entre si al interior de cada grupo y diferentes entre grupos, al igual que los metodos de clustering para datos trasnversales.

El clustering y la clasificación de series temporales o flujos de datos (*data stream*) ha tomado mucho interés estimulado por los progresos en las tecnologías y la recolección de datos. Las series de tiempo difieren de los datos "no-temporales" en la relación de interdependencia de

las mediciones. Los datos pueden tomar valores discretos o continuos, estar uniformemente muestreados o no, pueden se multivariados o univariados y tener igual o diferente longitud.

Las aproximaciones al clustering de series de tiempo pueden dividirse, según Liao (Liao 2005) en:

- basadas en datos crudos (*raw data*).
- basadas en atributos (*feature data*).
- basadas en modelos.

Cada una de estas aproximaciones se diferencian en el proceso que conduce de los datos al resultado del clustering.

Algunos algoritmos generales utilizados son presentados por Liao en (Liao 2005) en su revisión:

- *Relocation* clustering

  - Se comienza con un clustering inicial C con k clusters.
  - Se computa la matriz de disimilaridad $\forall t$ y se guardan todas las matrices para el cálculo de la similaridad de la trayectoria.
  - Se busca un C' mejor que C en términos del criterio de Ward generalizado. C' se obtiene reubicando 1 elemento de C_p en C_q o intercambiando dos elementos entre ellos ($C_p, C_q \in C; p, q = 1 \ldots k; p \neq q$).

  Si no existe C' el procedimiento se detiene, sino se reemplaza C con C' y se repite el último paso. Este algoritmo funciona solo con series de igual longitud.

- Jerárquico aglomerativo. Agrupa series de tiempo en un árbol(jerarquía). No se puede ajustar la pertenencia de cada serie a un grupo luego de que este está formado. No esta restringido a series de igual longitud, usando una distancia apropiada.

- *k-means*(*fuzzy c-means*) La idea es el minimizar una función objetivo (la distancia entre los objetos y sus respectivos centros). Si hay n objetos x_i y sus k centros v_c, $\text{mín} J(U,V) = \sum_{c=1}^{k} \sum_{i=1}^{n} u_{ci} ||x_i - v_c||^2$ sujeto a :$u_{ci} \in \{0,1\} \ \ \forall c,i$ y a: $\sum_{c=1}^{k} u_{ci} = 1 \ \ \forall i$

  - Elegir k ($2 \leq k \leq n$) ; ε ; iniciar el contador $l = 0$ y los centros iniciales $V^{(0)}$.
  - Distribuir $x_i \ \ \forall i$ para determinar $U^{(l)}$ tal que se minimize J.
  - Modificar los centros $V^{(l)}$.
  - Detener si el cambio en V es menor que ε o incrementar l y repetir los dos pasos anteriores.

  Este algoritmo funciona mejor con series de igual longitud porque en otro caso el concepto de centro del cluster no es claro.

2.2.1. Medidas de Disimilaridad

La diferencia principal del clustering de series de tiempo radica en como computar la disimilaridad entre dos objetos (Liao 2005). Además de algunas dificultades ya mencionadas, la elección de la métrica seleccionada para evaluar la disimilaridad es crítica y la literatura se ocupa de esto (Corduas 2007) o compara resultados del uso de diferentes métricas (Ding, Trajcevski, Scheuermann, Wang & Keogh 2008).

Se pueden distinguir al menos dos aproximaciones principales para evaluar la proximidad entre series de tiempo (Chouakria & Nagabhushan 2007), en primer lugar una aproximación paramétrica en donde la proximidad se mide entre los coeficientes ajustados a la serie a través de polinomios, modelos ARIMA o transformadas de Fourier ((Maharaj 2000),(Caiado, Crato & Peña 2006),(Gada & Puttagunta 2001))y en segundo lugar una aproximación no-paramétrica se basa en la descripción temporal de la serie((Vilar, Vilar & Pértega 2009),(Vilar, Alonso & Vilar 2010)).

Algunas de las medidas de disimilaridad utilizadas son la Euclidea, la de Minkowski, el coeficiente de correlación de Pearson. Otras mas específicas para el caso de las series de tiempo como por ejemplo:

- *Dinamic Time Warping* ((Sakoe & Chiba 1978),(Ratanamahatana & Keogh 2004),(Roche n.d.)(Keogh & Ratanamahatana 2004))
- *Short Time Series*
- *Probability-based distance function for data with errors* (Liao 2005)
- Kullback-Liebler
- *J divergence and symmetric Chernoff information*
- Fréchet (Chouakria & Nagabhushan 2007)
- la métrica AR

De todas esas medidas de disimilaridad solo se considera y se presenta la (DTW), vista como una forma de alinear series y evaluar que tanto se deforman.

2.3. Métodos de Alineación

En los métodos de alineación (dtw) se busca una curva de deformación que relaciones 2 series $x(k)$,$x(k)$ (Giorgino 2009),(Tormene, Giorgino, Quaglini & Stefanelli 2008) (*warping curve*) $\phi(k), k = 1 \dots T$:

$$\begin{aligned} \phi(k) &= \Big(\phi_x(k), \phi_y(k)\Big) \quad \text{con} \\ \phi_x(k) &\in \{1 \dots N\}, \\ \phi_y(k) &\in \{1 \dots M\} \end{aligned}$$

las funciones de deformación ϕ_x y ϕ_y se asocian a X y Y respectivmente. Dado ϕ, se calula la distorsion promedio acumulada entre las series deformadas X y Y :

$$d_\phi(X,Y) = \sum_{k=1}^{T} d(\phi_x(k), \phi_y(k))\, m_\phi(k)/M_\phi$$

Se imponen algunas restricciones como

$$\phi_x(k+1) \geq \phi_x(k)$$
$$\phi_y(k+1) \geq \phi_y(k)$$

la idea que se busca con (DTW) es encontrar la alineacion ϕ optima que minimice

$$D(X,Y) = \min_{\phi} d_{\phi}(X,Y) \tag{9}$$

3. Aplicación

Se comparan los resultados de aplicar esta aproximación con el (CE), sobre datos mensuales del período 1980-2009 de casos notificados de Hepatitis A para Uruguay.

- Se trabaja con los datos en tasas para la aplicacion de clustering sobre las 30 series
- Se aplica el clasificador (cCE), se crean 25 series com media geométrica (MG) ,25 series ordinales en zonas (SOZ)
- Se calculan distancias entre una serie y la correspondiente (MG)

4. Resultados

En el grafico 6 se puede ver como es la evolucion de las series mensuales en grupos de 5 años

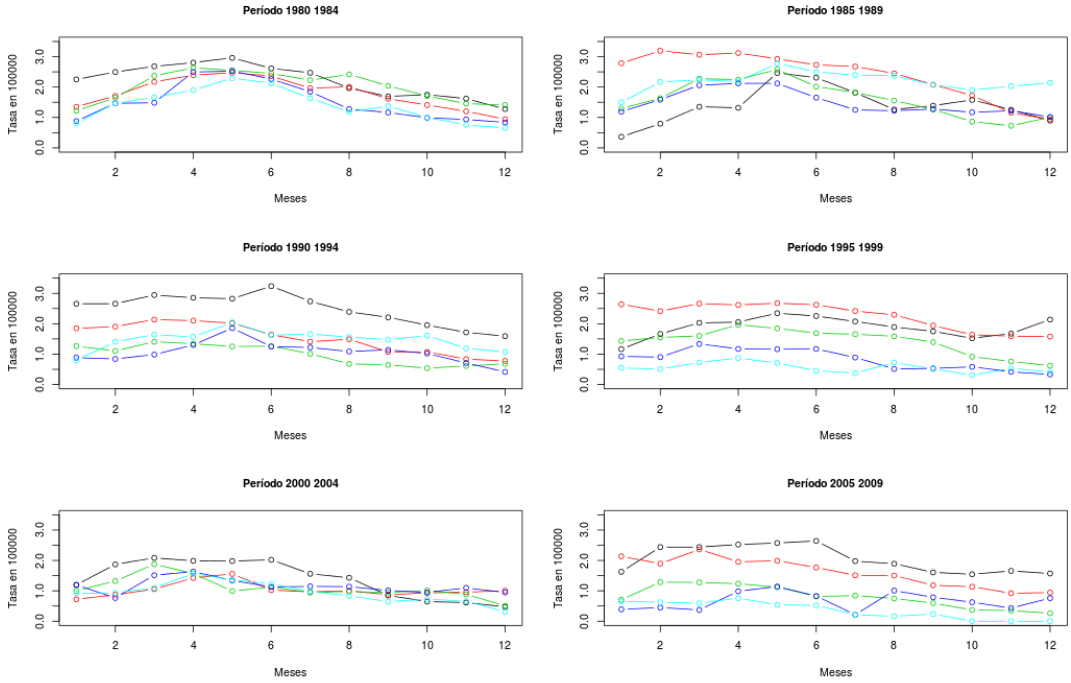

Figura 1: Series

Las 30 series mensuales tiene la siguiente distribucion vista en cortes transversales por mes

```
> summary(tasa_log1[,3:14])
      mes1             mes2             mes3             mes4
 Min.   :0.3646   Min.   :0.4510   Min.   :0.3715   Min.   :0.759
 1st Qu.:0.8239   1st Qu.:0.9105   1st Qu.:1.3420   1st Qu.:1.362
 Median :1.1933   Median :1.5102   Median :1.7681   Median :1.961
 Mean   :1.2786   Mean   :1.5196   Mean   :1.7836   Mean   :1.891
 3rd Qu.:1.4820   3rd Qu.:1.8874   3rd Qu.:2.2659   3rd Qu.:2.358
 Max.   :2.7855   Max.   :3.1953   Max.   :3.0645   Max.   :3.121
      mes5             mes6             mes7             mes8
 Min.   :0.5409   Min.   :0.456   Min.   :0.2144   Min.   :0.165
 1st Qu.:1.3450   1st Qu.:1.187   1st Qu.:0.9888   1st Qu.:1.003
 Median :2.0320   Median :1.729   Median :1.5963   Median :1.354
 Mean   :1.9675   Mean   :1.778   Mean   :1.5312   Mean   :1.422
 3rd Qu.:2.5466   3rd Qu.:2.349   3rd Qu.:1.9750   3rd Qu.:1.891
 Max.   :2.9605   Max.   :3.236   Max.   :2.7369   Max.   :2.451
      mes9            mes10            mes11            mes12
 Min.   :0.2383   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
 1st Qu.:0.8532   1st Qu.:0.7579   1st Qu.:0.6742   1st Qu.:0.5375
 Median :1.2237   Median :1.0173   Median :0.9284   Median :0.9167
 Mean   :1.2468   Mean   :1.1071   Mean   :0.9996   Mean   :0.9177
 3rd Qu.:1.6152   3rd Qu.:1.5718   3rd Qu.:1.2460   3rd Qu.:1.0601
 Max.   :2.2109   Max.   :1.9533   Max.   :2.0302   Max.   :2.1411
```

Una vez que se aplican los corredores (CE), tomando como primer serie sobre la que se puede aplicar la ventana temporal de tamaño 5 es 1985, se puede ver graficamente como queda

en su totalidad la serie observada, vista como una sola serie, a diferencia de la metodología de (CE) que siempre la ve dentro del año y en referencia a los 5 años que la preceden.

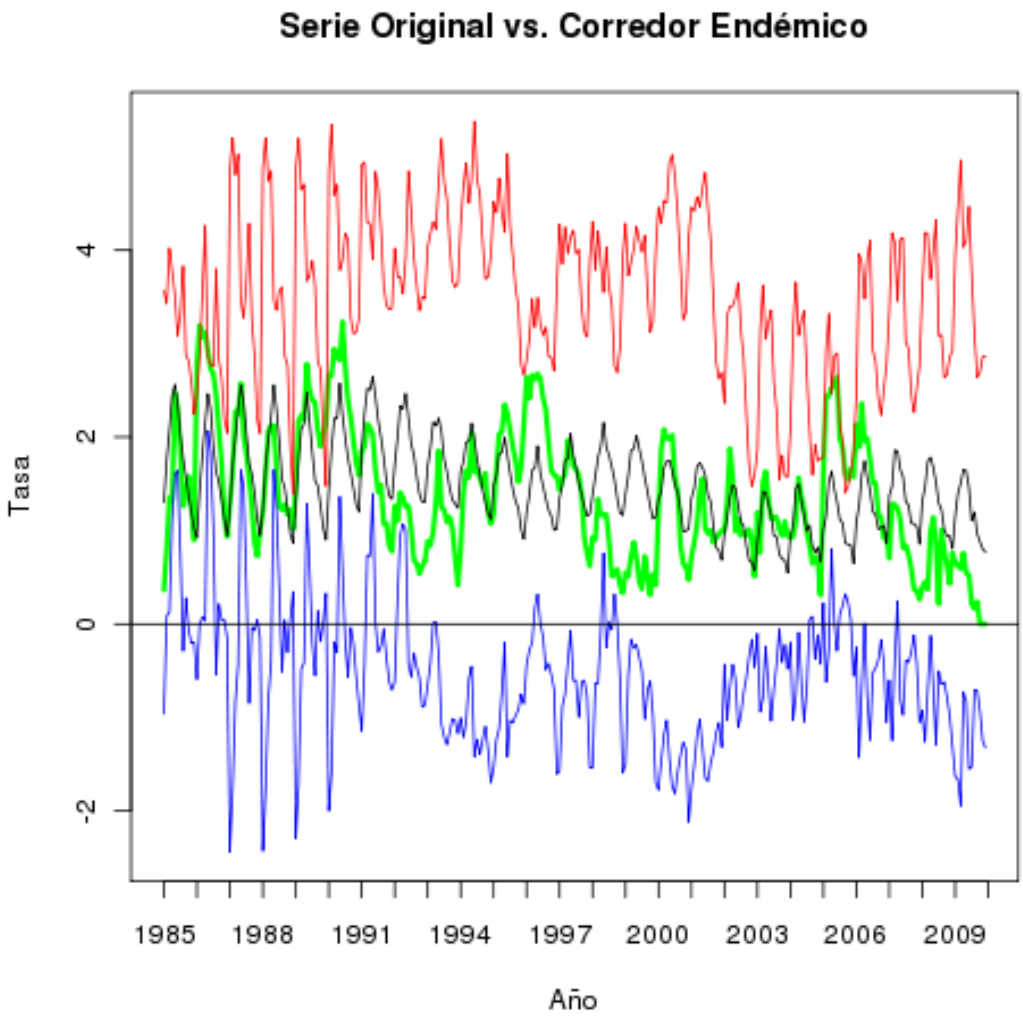

Figura 2: Series observadas período 1980-2009

La figura 3 permite ver mas en detalle que es lo que realmente se ve a través de (CE) y como se pierde parte de la dinamica de mas largo plazo al observar el año solamente

Teniendo en cuenta graficamente como es el fenomeno a lo largo de los 30 años se aplican en forma secuencial los diferentes metodos de clustering vistos en 2.1, con lo cual a instancias del metodo de Ward se evalua conservar 3 grupos, ya que los diferentes indicadores de ajuste (R^2) y demas estadísticos asi lo indican

Para comparar los resultados del metodo de WARD se consideran 3 grupos tambien para el método de *k-means* y el de los *medoides*, viendo que el método de *k-means* diferencia los 3 centroides a traves de una diferencia de nivel, mas que de forma

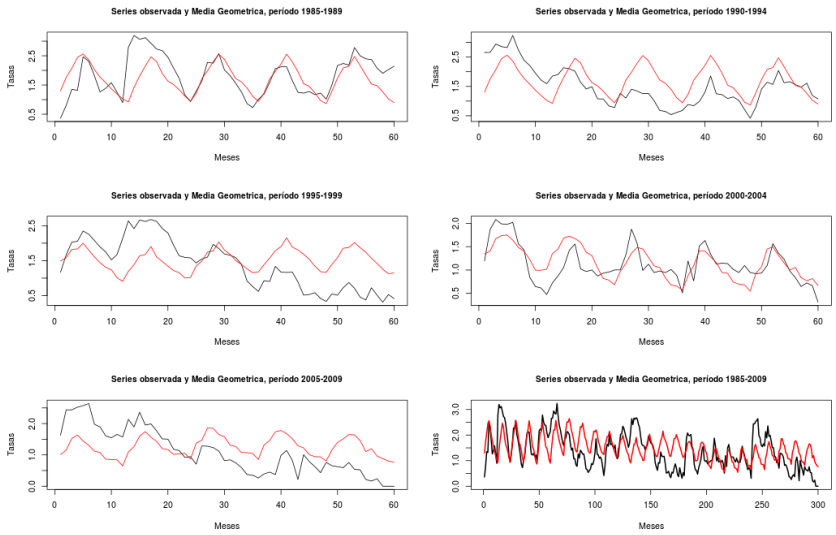

Figura 3: Series obsrevadas y su correspondiente centro del (CE)

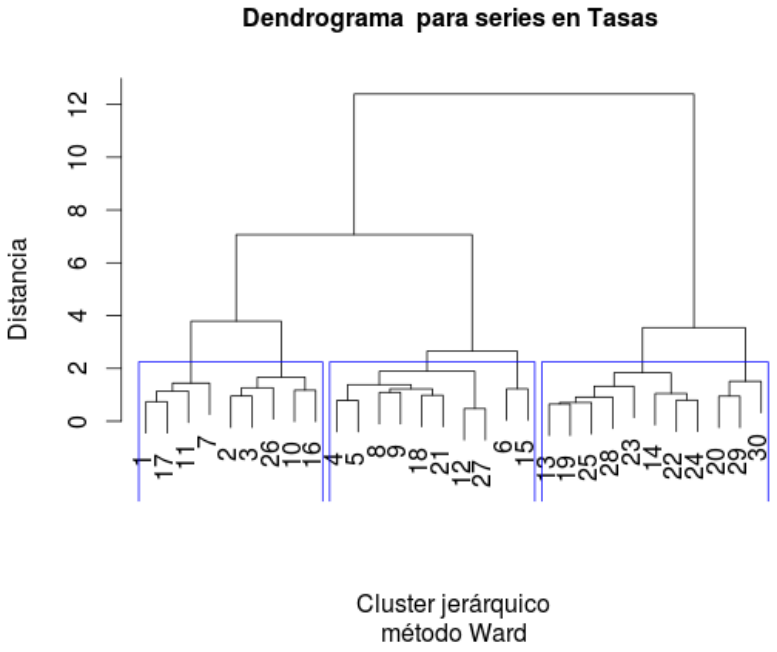

Figura 4: Dendrograma para método de WARD para series 1980-2009

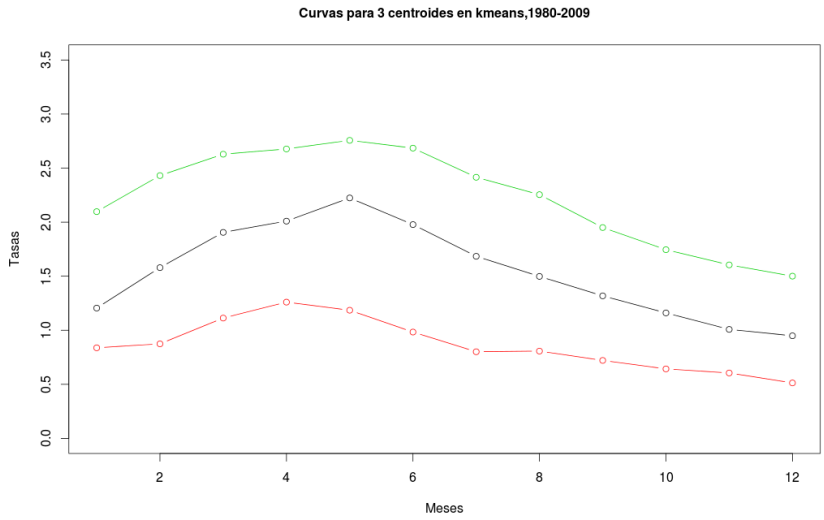

Figura 5: Centroides para método *k-means* en 3 grupos

Para el caso de la aplicación de PAM se puede ver que el considerar 3 grupos muestra que algunas de las series, con el caso extremo de 1995 tienen un valor de silueta negativa, indicando que esa serie, si bien pertenece al cluster 2 no está correctamente asignada. Para ver que tan estable es este resultado en 3 grupos, se aplica el algoritmo PAM cambiando el número de grupos, donde se resume a través del gráfico 7 que lo correcto es considerar 2 grupos.

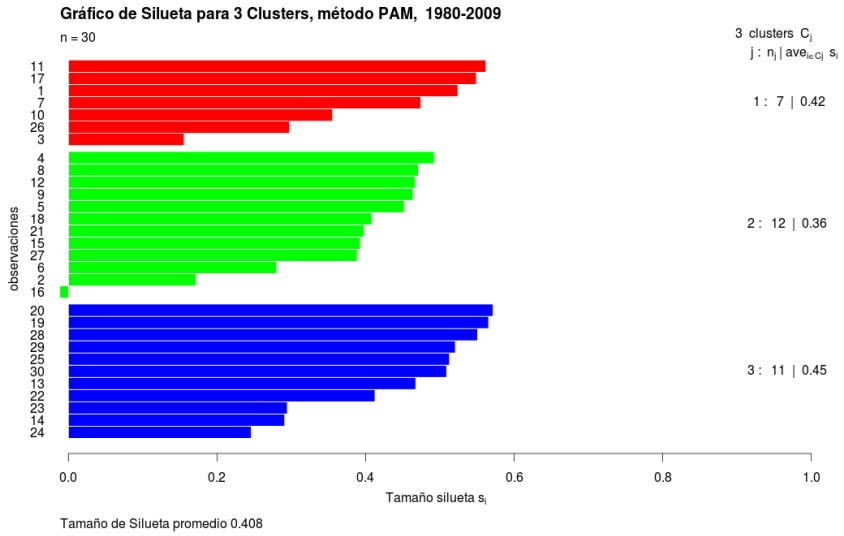

Figura 6: Gráfico de silueta para método PAM con 3 grupos

También el algoritmo de cluster difuso, luego de ser aplicado , muestra resultados que indican que lo óptimo es considerar un total de 2 grupos, con la diferencia de que se sabe el grado de pertenencia de la serie al grupo en la que esta se clasificó

Luego de evaluar la clasificación hecha por los diferentes metodos de clustering se considera las series clasificadas combinando el (CE) con un metodo de clustering de tipo PAM,aplicando una disimilaridad de gower, que permite trabajar con variables ordinales. Esta nuevo clasificación se aplica entonces sobre el clasificador (cCE) presentado en 1. La nueva serie es entonces una serie de valores ordinales, mostrando las 4 zonas y es por eso que se debe aplicar uan

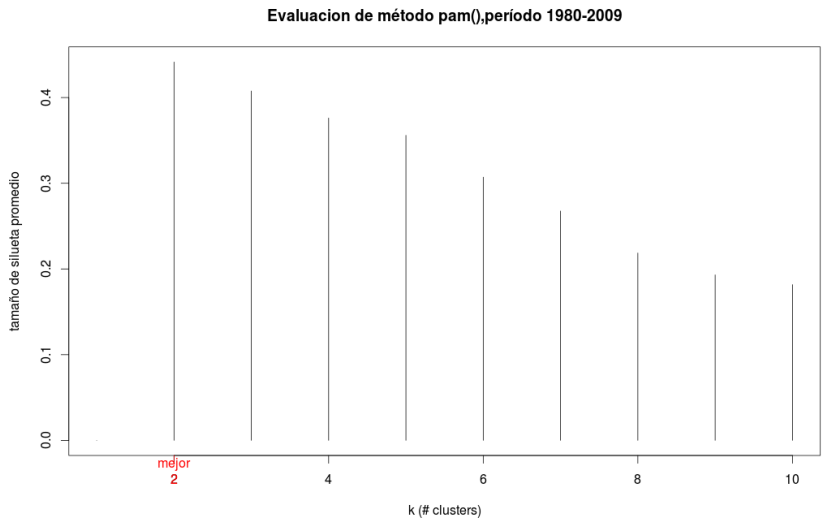

Figura 7: Número óptimo de clusters para método PAM

distancia no euclidea.

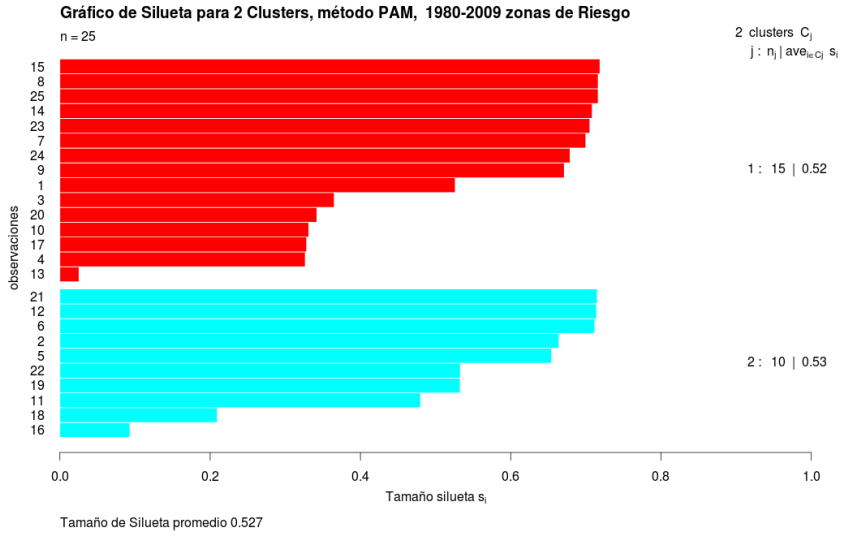

Figura 8: Gráfico de silueta para método PAM con 2 grupos sobre series clasificadas en zonas de riesgo

Hasta el momento los resultados presentados solo son los que dan cuenta de las 30 series clasificadas de diferente manera o de las zonas de riesgo creadas con los el (cCE). Sin embargo en estos resultados no se sigue la lógica del (CE) ya que en los grupos creados a partir de las zonas se consideraron todas, mientras que el (CE) solo mira la series y la banda de confianza que se crean a partir de esta. Por lo tanto en los resultados que siguen, se muestra que resulta de solo considerar 2 series (la observada y el centro de su correspondiente (CE), que no es mas que la media geométrica $MG^{v}_{i,j}$ de orden 5, aplicando la metodología de *dtw* que muestra como se alinean y /o se deforman 2 series.

Para eso a las series de $1985 - 2009$ expresadas en zonas de riesgo se les calcula la disimilaridad de Gower y se evalua como queda las 2 series que estan mas proximas en terminos de

Año	clu.ward	clu.kmeans	clu.fan	clu.pam
1980	1	3	1	1
1981	1	1	2	1
1982	1	3	1	1
1983	2	1	2	1
1984	2	1	2	1
1985	2	1	2	1
1986	1	3	1	1
1987	2	1	2	1
1988	2	1	2	1
1989	1	3	1	1
1990	1	3	1	1
1991	2	1	2	2
1992	3	2	3	2
1993	3	2	3	2
1994	2	1	2	2
1995	1	1	2	1
1996	1	3	1	1
1997	2	1	2	1
1998	3	2	3	2
1999	3	2	3	2
2000	2	1	2	1
2001	3	2	3	2
2002	3	2	3	2
2003	3	2	3	2
2004	3	2	3	2
2005	1	3	1	1
2006	2	1	2	1
2007	3	2	3	2
2008	3	2	3	2
2009	3	2	3	2

la distancia de *dtw* que evalua todo el año, pero expresadas en sus valores originales en tasas, tratando de ver que cual es el efecto artificial que produce el clasificador (cCE) que no considera de forma diferente aquellos valores que apenas estan en un zona, muy cerca de la frontera o por el contrario lo que estan lo mas lejos posible sn dejar de pertenecer a esa zona. Las series que

son mas distantes con la *disimilaridad de gower* son $1999, 2005$

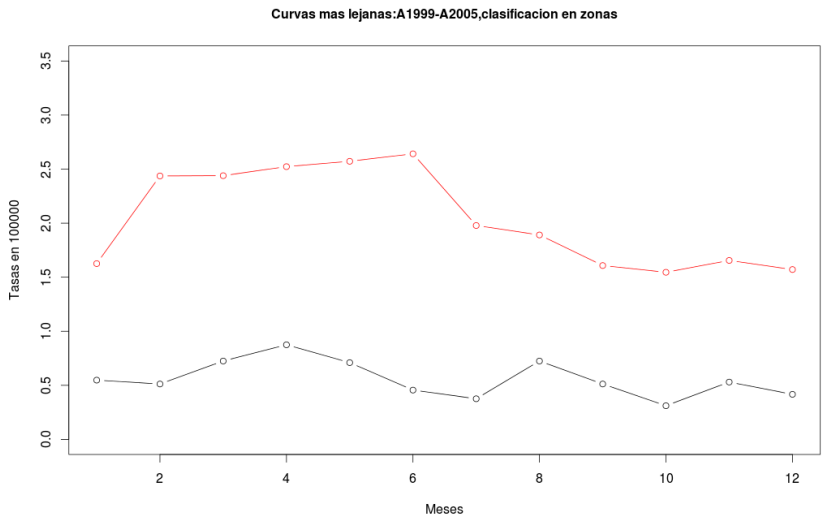

Figura 9: Series mas distantes clasificadas en zonas de resgo

A estas 2 series se las alinea, mostrando el siguiente comportamiento:

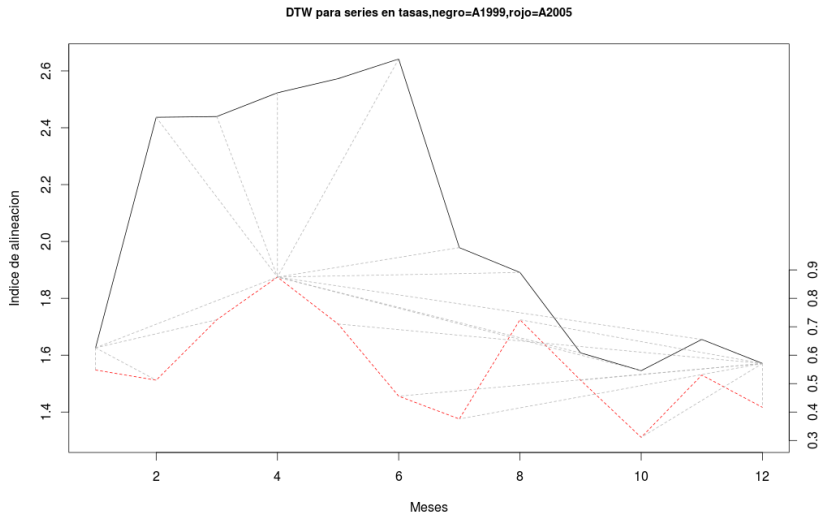

Figura 10: Alineación según *dtw* de las series mas distantes clasificadas según zonas de riesgo

5. Conclusiones y futuros paso

Los resultados que se encontraron hasta ahora muestran que los diferente metodos de clustering condicionan los grupos, ya que estos sean de tipo transversal o de tipo longitudinales

Por un lado los métodos de tipo tranversal ponen de manifiesto una estructura de datos que varía según sea el algoritmo usado. En cualquiera de ellas el comportamento no es monótono ya que nunca quedan clusters formados por series consecutivas en todo el período. En particular los clusters que se forman con el metodo *k-means* muestran esencialmente una diferencia de nivel en los centroides. Al trabajar con el método de los medoides la división en 2 grupos de series se produce a partir de la serie 1981 y la serie 2004, que son los medoides seleccionados La ventaja por otro lado del metodo de PAM es que los centros originales son realmente series observadas a diferencia de los centroides del metodo *k-means*.

Por otra lado el método *dtw* visto para tener en cuenta el tiempo permite ver como se deforma y se desalinean 2 series; por ejemplo

```
15      A1999 1999    1    1    1    1    1    1    1    1    1    1    1    1
21      A2005 2005    4    4    4    4    4    4    4    4    4    4    4    4
```

Esas 2 series son las que luego de pasar por el clasificador (cCE) muestran estar mas alejadas, sin embargo ese clasificador las ubica, sin considerar que en una parte del año los valores de las tasas son mas próximos (se julio a diciembre) (es lo que aparece en la figura 10).

Los futuros pasos son poder tener una clasificación de las series ya clasificadas en zonas de riesgo pero de las que se tenga en cuenta la dinámica intraanual (a través de DTW por ejemplo). Ese agrupamiento se comparará con el que se haga sobre las 25 series clasificados con DTW en sus valores originales pero comparandolas de a pares, es decir cada serie con su correspondiente (CE).

6. Bibliografía

Alvarez, R., Dibarboure, H. & F., M. (2010), Comparacion de 2 metodos de vigilancia epidemiologica para la hepatitis tipo a.

Bortman, M. (1999), 'Elaboración de corredores o canales endémicos mediante planillas de cálculo', *Revista Panamericana de Salud Pública* **5**, 1 – 8.

Caiado, J., Crato, N. & Peña, D. (2006), 'A periodogram-based metric for time series classification', *Computational Statistics & Data Analysis* **50**, 2668 – 2684.

Chouakria, A. & Nagabhushan, P. (2007), 'Adaptive dissimilarity index for measuring time series proximity', *Advances in Data Analysis and Classification* **1**, 5 – 21.

Corduas, M. (2007), 'Dissimilarity criteria for time series data mining', *Quaderni di Statistica* **9**, 107 – 129.

Dibarboure, H., Alvarez, R., Quian, J. & Mazza, F. (2009), Estudio epidemiológico de la hepatitis A en Uruguay. Una revisión de los últimos 30 años 1980 - 2009. Gran Premio Nacional de Medicina de la Academia Nacional de Medicina.

Ding, H., Trajcevski, G., Scheuermann, P., Wang, X. & Keogh, E. (2008), 'Querying and mining of time series data: experimental comparison of representations and distance measures', *PVLDB* **1**(2), 1542–1552.

Gada, K. K. D. & Puttagunta, V. (2001), Distance measures for effective clustering of arima time-series, *in* 'In proceedings of the 2001 IEEE International Conference on Data Mining', pp. 273–280.

Giorgino, T. (2009), 'Computing and visualizing dynamic time warping alignments in R: The dtw package', *Journal of Statistical Software* **31**(7), 1–24.
URL: *http://www.jstatsoft.org/v31/i07/*

Kaufman, L. & Rousseeuw, P. (1990), *Finding Groups in Data: An Introduction to Cluster Analysis*, John Wiley, New York.

Keogh, E. & Ratanamahatana, A. (2004), 'Everything you know about dynamic time warping is wrong', *3rd Workshop on Mining Temporal and Sequential Data, in conjunction with 10th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD-2004), Seattle, WA* .

Liao, T. W. (2005), 'Clustering of time series data - a survey', *Pattern Recognition* **38**(11), 1857–1874.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M. & Hornik, K. (2012), *cluster: Cluster Analysis Basics and Extensions*. R package version 1.14.2 — For new features, see the 'Changelog' file (in the package source).

Maharaj, E. (2000), 'Clusters of time series', *Journal of Classification* **17**, 297 – 314.

Orellano, P. W. & Reynoso, J. I. (2011), 'Nuevo método para elaborar corredores endémicos', *Revista Panamericana de Salud Pública* **29**, 309 – 314.

Paliwal, K., Agarwal, A. & Sinha, S. (1982), 'A modification over sakoe and chiba's dynamic time warping algorithm for isolated word recognition', *Signal Processing* **4**(4), 329 – 333.

R Development Core Team (2010), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Ratanamahatana, C. & Keogh, E. (2004), Making time-series classification more accurate using learned constraints, *in* M. W. Berry, U. Dayal, C. Kamath & D. B. Skillicorn, eds, 'Proceedings of the Fourth SIAM International Conference on Data Mining - SDM', SIAM.

Roche, A. (n.d.), 'Arboles de decisión y series de tiempo'. Tesis de Maestría en Ingeniería Matemática, Facultad de Ingeniería - UdelaR. Montevideo, 2009.

Sakoe, H. & Chiba, S. (1978), 'Dynamic programming algorithm optimization for spoken word recognition', *IEEE Transactions on Acoustics, Speech, and Signal Processing* **26**(1), 43–49.

Tormene, P., Giorgino, T., Quaglini, S. & Stefanelli, M. (2008), 'Matching incomplete time series with dynamic time warping: An algorithm and an application to post-stroke rehabilitation', *Artificial Intelligence in Medicine* **45**(1), 11–34.
URL: *http://dx.doi.org/10.1016/j.artmed.2008.11.007*

Vilar, J., Alonso, A. & Vilar, J. (2010), 'Non-linear time series clustering based on non-parametric forecast densities', *Computational Statistics and Data Analysis* **54**, 2850 – 2865.

Vilar, J., Vilar, J. & Pértega, S. (2009), 'Classifying time series data: A nonparametric approach', *Journal of Classification* **26**, 3 – 28.