PREDICCIÓN DE CLÚSTERS DE SERIES TEMPORALES DEMOGRÁFICAS.

Andrés M. Alonso[1], Daniel Peña[2], Julio Rodríguez[3].
[1]Instituto de Investigación Avanzada sobre Evaluación de la Ciencia y la Universidad y Departamento de Estadística. Universidad Carlos III de Madrid. 28903 Getafe (Madrid). [2]Departamento de Estadística. Universidad Carlos III de Madrid. 28903 Getafe (Madrid). [3]Dpto. Análisis Económico Economía Cuantitativa. Universidad Autónoma de Madrid. 28049 Madrid. andres.alonso@uc3m.es, daniel.pena@uc3m.es, jr.puerta@uam.es

Resumen

En el presente trabajo se propone la aplicación de técnicas clúster al modelado de las series temporales demográficas por edades. El objetivo es localizar la existencia de grupos de edades con una dinámica temporal similar, para posteriormente realizar una estimación conjunta del modelo que describa mejor las características comunes de los grupos de series. El método clúster utilizado permite la comparación de sus modelos generadores sin imponer la independencia de las series. Obtenidos los grupos de series con modelos generadores equivalentes y estimado el modelo generador común se realiza las predicciones a diferentes horizontes. Presentamos los resultados para las series de tasas brutas de mortalidad por grupos de edades simples de ambos sexos. Los predicciones obtenidas a partir de los clúster de series presentan un error cuadrático medio menor que las predicciones mediante modelos univariantes para cada una de las series. La principal ventaja de este método es que permite estimar los parámetros con mayor precisión y esto implica una reducción de la incertidumbre en los pronósticos.

Palabras clave: Clúster de series, series temporales, demografía, tasas brutas de mortalidad.

Abstract

Clusters prediction of demographic time series.

This paper proposes the application of cluster technique to the modeling of demographical by age time series. The objective is to find the existence of age groups with a similar time dynamic, to posteriorly have a full estimation of the model that describes better the common characteristics of the group series. After the groups series with equivalent models generators and estimated the common generator model, the predictions are made at different horizons. We present the results for the gross mortality rates by simple age groups in both sexes. The predicitions obtained from the clusters of sseries present a quadratic mean error lower than the predictions by univariant models for each series. The main advantage of this method is that it lets estimate the parameters with higher precision and this means a reduction of the incertainty in the pronostics.

Keywords: Clusters of time series, demografía, gross mortality rates.

INTRODUCCIÓN.

El cambio demográfico que se ha observado en el último cuarto del siglo pasado se prevé que tenga numerosas consecuencias socioeconómicas. En este periodo España ha experimentado importantes modificaciones en la dinámica de los componentes demográficos básicos: fecundidad, mortalidad e inmigración.

Este trabajo es parte de un proyecto que evalúa los cambios en las series demográficas y sus consecuencias en la predicción de la demanda de educación en España y Europa que está financiado por la Fundación BBVA. Un paso previo a la predicción de esta demanda es el análisis y predicción de las principales series temporales demográficas que describen tanto la población total como las tasas brutas de mortalidad, fecundidad, escolaridad e inmigración.

El análisis de las series demográficas se enfrenta a las siguientes condiciones:

- Longitud reducida de las series, debido a que el cambio se observa en los últimos 25 años.
- Estructura de dependencia entre las series contiguas en edades, lo que implica que en el análisis de las sexxistence eries marginalizadas existiría una gran pérdida de información.

La figura 1 muestra los logaritmos de las tasas brutas de mortalidad entre 1970 y 2001 en ambos sexos para edades seleccionadas. En ellos se aprecia la existencia de posibles componentes comunes tanto en la tendencia como en la estructura de dependencia dinámica.

La búsqueda de estructuras comunes en estas series permitirá proponer modelos más parsimoniosos e incrementar la precisión de las estimaciones de sus parámetros.

CONTRASTE DE IGUALDAD DE MODELOS

Sean $\{X_t\}_{t\in Z}$ e $\{Y_t\}_{t\in Z}$ dos procesos estacionarios que siguen los modelos P_X y P_Y, respectivamente. Sean $X=(X_1,..,X_n)$ e $Y=(Y_1,..,Y_n)$ vectores de observaciones de los procesos $\{X_t\}_{t\in Z}$ e $\{Y_t\}_{t\in Z}$ no necesariamente independientes.

Estamos interesados en el siguiente contraste:

$$\begin{cases} H_0: & P_X = P_Y \\ H_1: & P_X \neq P_Y \end{cases}.$$

Para ello, comparamos los modelos generadores (AR ó ARMA):

$$\begin{cases} H_0: & \phi_X = (\phi_{X,1},..,\phi_{X,p})' = \phi_Y = (\phi_{Y,1},..,\phi_{Y,p})' \\ H_1: & \phi_X = (\phi_{X,1},..,\phi_{X,p})' \neq \phi_Y = (\phi_{Y,1},..,\phi_{Y,p})' \end{cases}.$$

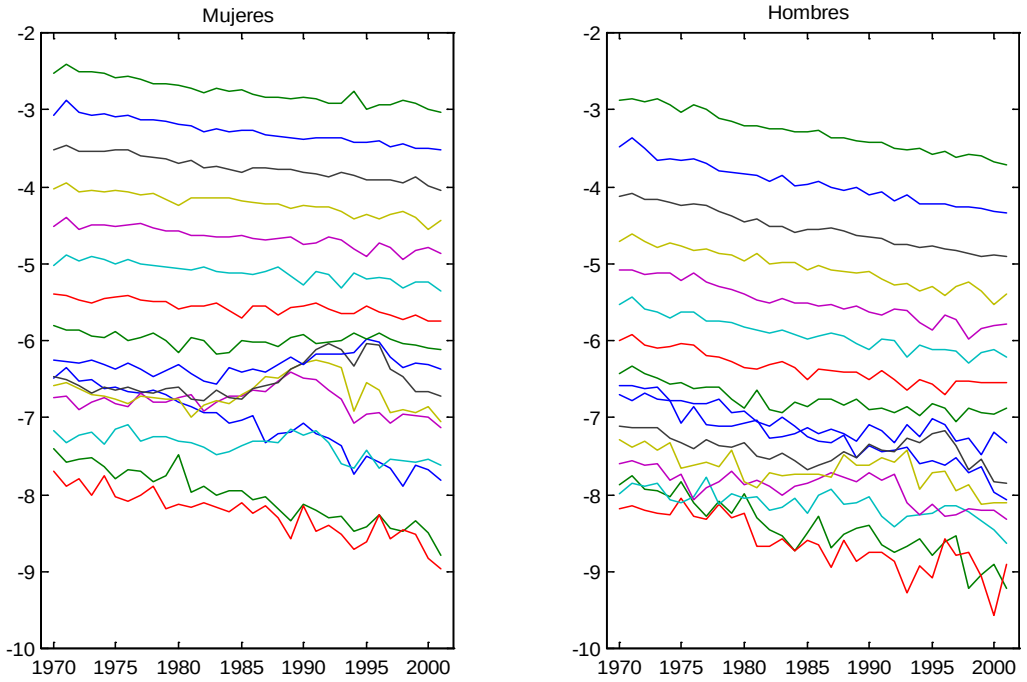

Fig. 1. Tasas de mortalidad por edad y sexo. España 1970 a 2001.

Suponemos que $\{X_t\}_{t\in Z}$ e $\{Y_t\}_{t\in Z}$ admiten una representación ARMA. Sea $k=\max(k_1,k_2)$ con k_1 y k_2 los órdenes de los modelos autorregresivos que aproximan a $\{X_t\}_{t\in Z}$ e $\{Y_t\}_{t\in Z}$ seleccionados a partir de una muestra de tamaño *n*. Podemos escribir un modelo conjunto para ambas series:

$$Z=\begin{bmatrix} X \\ Y \end{bmatrix} = W\pi + \varepsilon,$$

donde $W=\begin{bmatrix} W_X & 0 \\ 0 & W_Y \end{bmatrix}$, W_X y W_Y son las matrices $T-k\times k$ de observaciones retardadas, $\pi=[\pi_X \pi_Y]'$ y $\varepsilon=[\varepsilon'_X \varepsilon'_Y]'$. Suponemos que

$$E[\varepsilon]=0, E[\varepsilon\varepsilon']=V=\Sigma\otimes I_{n-k}, \text{y } \Sigma=\begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{yx} & \sigma_y^2 \end{bmatrix}$$

Bajo $H_0:\pi_X=\pi_Y$, el siguiente estadístico se distribuye asintóticamente como una χ_k^2 (ver, Maharaj (2000)):

$$D=(R\hat{\pi})'[R(W'\hat{V}W)^{-1}R']^{-1}(R\hat{\pi}),$$

donde $\hat{V}$ es el estimador por mínimos cuadrados de V, $\hat{\pi}$ es el estimador por mínimos cuadrados generalizados de π y $R=[I_p \vdots -I_p]$.

En Alonso y Maharaj (2005) se propone un procedimiento basado en técnicas de computación intensiva para un contraste de las estructuras de autocorrelación que es equivalente al desarrollado esta sección y que es independiente de modelo.

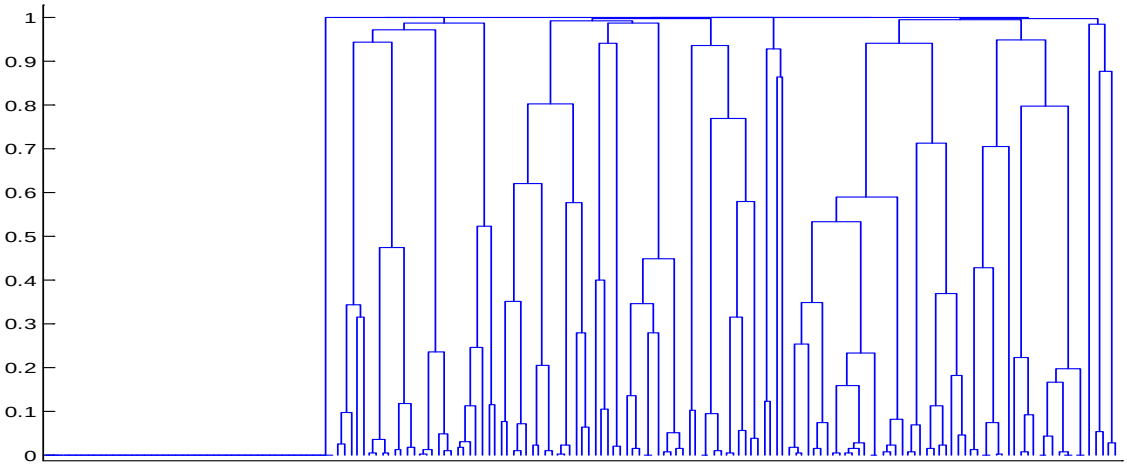

Fig. 2. Cluster jerárquico mediante comparación de modelos

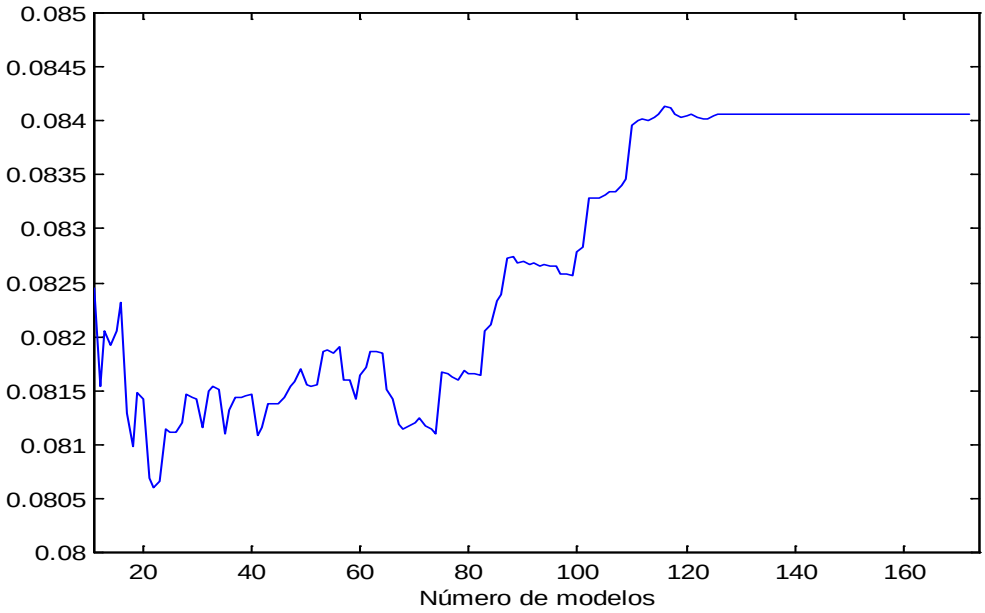

Fig. 3. Error absoluto medio de predicción según el número de modelos considerado

RESULTADOS DEL PROCEDIMIENTO CLUSTER.

Basándose en el anterior contraste de igualdad de modelos en series temporales, Maharaj (2000) propone un procedimiento de clusters jerárquicos que permite agrupar las series de forma que cada uno de los grupos sea homogéneo respecto al modelo generador.

En la figura 2 se muestra el resultado de este procedimiento aplicado a la diferencias de las 172 series de tasas brutas de mortalidad por edades y sexo (edades simples de 0 años a 84 años y grupo de 85 y más años) para el período 1970 – 2000. En el eje de ordenadas se representa el 1-*p-valor* del contraste de igualdad de modelos.

Estableciendo como punto de corte, por ejemplo, el valor 0.95 obtendríamos en torno a 15 grupos de series. Al estar basado el dendrograma en el encadenamiento del vecino más lejano se asegura que todas las comparaciones dos a dos, dentro un cluster, establecen que el modelo en las series es igual.

La estructura jerárquica obtenida mediante este procedimiento nos permite definir desde 1 a 172 modelos en las series de tasas.

En cada cluster de series se estima el modelo común mediante mínimos cuadrados generalizados.

ANÁLISIS COMPARATIVO DE PREDICCIONES

Para analizar la influencia del número de modelos considerados (uno por cluster) en la capacidad predictiva global del procedimiento propuesto analizamos el error absoluto relativo medio en la predicción de las tasas para el año 2001.

En la figura 3 se observa que los errores de predicción son muy similares cuando se consideran más de diez grupos, e incluso para valores menores que 100 grupos son ligeramente inferiores a considerar un modelo para cada serie temporal.
Al no existir una pérdida en los errores medios de predicción se justifica la selección de un esquema más parsimonioso, en cuanto al número de modelos diferentes, para la predicción a corto plazo. La selección de un esquema de modelos más parsimonioso se corrobora si se aplican criterios de información como el AIC o el BIC que tienen en cuenta, simultáneamente, tanto el error de predicción como el número de parámetros a estimarse.
La estimación conjunta de los parámetros de un modelo para un grupo de series conduce a la reducción de la incertidumbre asociada a la estimación siempre que los grupos provengan del mismo modelo generador.
Esta mejora implicará la obtención de intervalos de predicción más precisos que si utilizamos un modelo distinto para cada serie.

.
En la figura 4 presentamos los incrementos porcentuales, respecto a los obtenidos con la estimación individual de cada una de las series, en las longitudes de los intervalos de predicción del 95% con horizonte desde 1 a 25 años. Se observa una reducción media de la longitudes en torno al 5% cuando se opta por un número reducido de modelos.

CONCLUSIÓN.
En este trabajo hemos propuesto un procedimiento para la modelación y predicción de un alto número de series temporales que se basa en la agrupación por igualdad de modelos generadores. Se ha aplicado el procedimiento a un conjunto de datos reales demográficos y se obtienen resultados satisfactorios en cuanto al error de predicción y a la incertidumbre asociada a la estimación de los modelos.

Agradecimientos.
Este trabajo ha sido financiado por el proyecto #2233 “Previsión del efecto de los cambios de la natalidad en la demanda de educación en España y en la Unión Europea” de la Fundación BBVA.

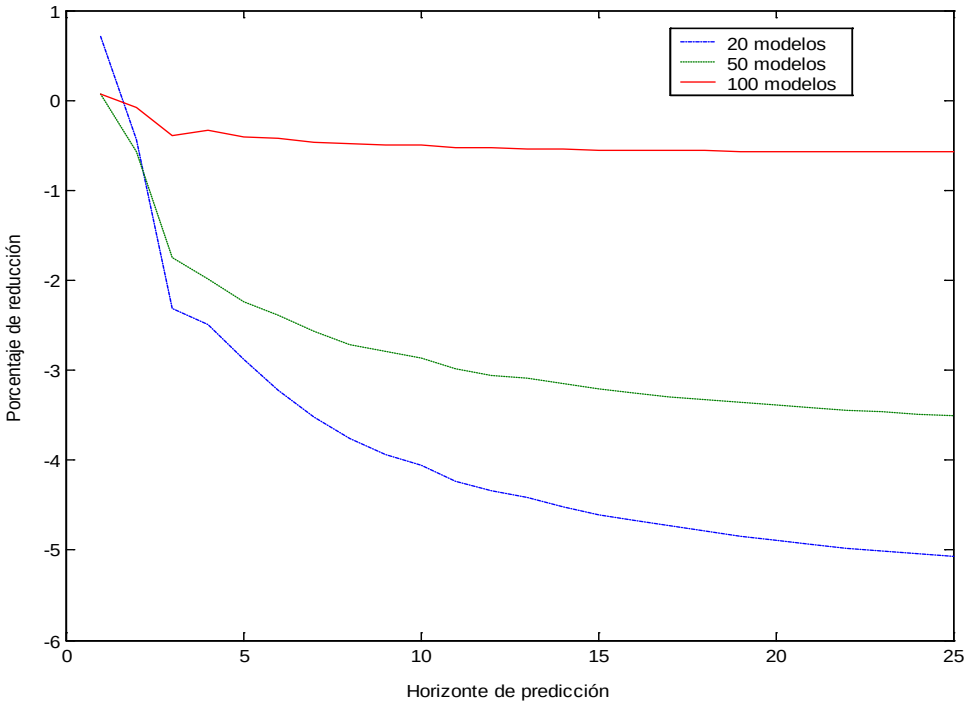

Fig. 4. Reducción de la incertidumbre de predicción según el número de modelos considerado

REFERENCIAS.
Alonso, A.M. and Maharaj, E.A. (2005) Comparison of time series using subsampling, *Computational Statistics and Data Analysis*, 50 (10), 2589-2599.
Maharaj, E.A. (2000) Clusters of time series, *Journal of Classification*, 17 (2), 297-314.