# Análisis de Datos Funcionales para tendencias poblacionales de áreas pequeñas.


## Abstract

El Análisis de Datos Funcionales (FDA: Functional Data Analysis en ingles) es un conjunto de técnicas multivariadas para reducir la dimensionalidad de datos definidos temporalmente o secuencialmente. Estas técnicas estadísticas permiten el análisis de problemas espacio-temporales. El artículo utiliza análisis de conglomerados ("cluster analysis"), componentes principales, y modelos lineales a datos con estructura funcional para describir las tendencias en la Tasa Global de Fecundidad, la Tasa Bruta de Natalidad y la Tasa Bruta de Mortalidad, para poder definir proyecciones de los componentes del cambio poblacional para áreas pequeñas: cantones en Costa Rica. Estas técnicas son importantes para este país porque hay tamaños muy pequeños de población en este nivel administrativo, por lo que hace que las estimaciones anuales sean muy volátiles. El análisis de conglomerados de series de tiempo muestra heterogeneidad geográfica en las tendencias. Los 9 conglomerados obtenidos pueden ser usados para proyecciones a nivel cantonal. Desde una perspectiva sustantiva, los conglomerados concuerdan con los hallazgos clásicos de Rosero-Bixby y Casterline (1994) sobre la difusión de la caída de la fecundidad en Costa Rica.


## Introducción.

Después de cada censo decenal, en colaboración con el Centro Centroamericano de Población (CCP), el Instituto Nacional de Estadística y Censos (INEC) actualiza sus proyecciones de población para todo el país y por unidades geográfcas: provincias, cantones y distritos. Costa Rica tiene una población relativamente pequeña (4.3 millones según el último Censo de Población 2011). Adicionalmente, como nación, ha arribado a niveles muy bajos de mortalidad (esperanza de vida al nacer de 79.2 años en 2011) y de fecundidad (Tasa Global de Fecundidad de 1.85 hijos por mujer en 2011). Estas condiciones producen una volatilidad muy alta en las estimaciones demográficas (Tasas Brutas de Natalidad y Mortalidad, Tasas Globales de Fecundidad) porque hay algunos cantones o distritos que tienen poblaciones muy pequeñas, así como muy pocos nacimientos y defunciones.

En proyecciones de población anteriores (publicadas en 2002) al nivel de distrito, se realizaron supuestos de que los componentes del cambio poblacional de los cantones convergerían hacia los niveles promedio del país. Este supuesto demostró estar errado en el corto plazo, afectando proyecciones de menores de edad, particularmente las estimaciones usadas para medir la población en edad escolar. El siguiente artículo propone una metodología para establecer grupos con tendencias

similares en fecundidad, mortalidad y migración interna. La metodología se basa en las técnicas del Análisis Funcional de Datos AFD (FDA Functional Data Analysis), una serie de procedimientos estadísticos que ya han sido utilizados para proyecciones de población (Hyndman & Shang, 2009; Hyndman & Ulla, 2007). El AFD se refiere al análisis de información sobre curvas o funciones. Más específicamente, comprende un grupo de técnicas multivariadas para analizar y reducir la dimensionalidad de datos en que cada "observación" es una secuencia temporal, espacial ó gráfica que puede ser resumida en una función (Ramsay & Silverman, 2006). Estas técnicas son ideales para describir y agrupar un conjunto de series de tiempo.

En este artículo se usan varias técnicas con el fin de establecer la base para proyecciones de indicadores demográficos. En el primer lugar, se usan técnicas de conglomerados por K-medias ("k-means clustering") de AFD. Estos datos funcionales son usados como información "en bruto" (Tasa Bruta de Natalidad TBN y Tasa Global de Fecundidad TGF), que después se "suavizan" utilizando modelos lineales para datos funcionales. Con fines ilustrativos, se usarán las técnicas del FDA a las series de tiempo de Tasas Brutas de Natalidad y Global de Fecundidad publicadas por CELADE para los países de América Latina y el Caribe.

Métodos: Análisis Funcional de Datos AFD.

Cada unidad estadística por analizar se denota con el subíndice i. La secuencia de indicadores demográficos para cada unidad estadística a través del tiempo se denota como Yi(t). En AFD, cada Yi(t) es considerada una función, y se pretende resumir, describir o analizar una muestra de funciones Yi(t). En el análisis cantonal para Costa Rica, cada cantón es la unidad estadística, y la serie de tiempo entre 1975 y 2010 para cada cantón correspondería a la función Yi(t). En el caso del análisis por país, cada país es denotado por i, y la serie de tiempo entre 1950 y 2010 por Yi(t).

En el análisis de conglomerados vía K-medias, los algoritmos tradicionales que calculan la distancia entre observaciones (Pitágoras, Mahalanobis, etc.) son usados para calcular las distancias entre el conjunto n de Yi(t) funciones. Sin embargo, también se pueden usar medidas de distancia especiales para algunas funciones (véase Delicado et al., 2010, que usan la distancia de Kullback-Leibler para comparar pirámides de población). Modelos de Box-Jenkins son usados para resumir la serie de tiempo promedio que describe cada conglomerado.

En AFD, es posible definir cada función como variables respuesta que dependen de un conjunto de covariables Zi(t) que también son funciones del tiempo, como en la ecuación:

$$Yi(t)= \alpha(t)+ Zi(t)\beta(t)+ Vi\gamma+ \varepsilon i(t)$$

Algunos de los coeficientes de regresión pueden variar a través del tiempo –$\alpha(t)$, $\beta(t)$– con el fin de describir mejor la relación entre funciones (Ramsay & Silverman, 2006), mientras que otros coeficientes pueden ser fijos en el tiempo, como $\gamma$. En un Modelo Lineal Funcional determinado espacialmente,

algunas de las covariables Vi pueden ser las coordenadas geográficas de las unidades estadísticas (cantones, países), para tomar en cuenta la heterogeneidad espacial.

Fuentes de Datos y aplicación específica del método a los datos

Para el análisis de los cantones (municipios) costarricenses, se calcularon Tasas Brutas de Natalidad (TBN) y Tasas Globales de Fecundidad (TGF) para el período 1975-2010. Los datos de nacimientos por cantón son procesados por el Instituto Nacional de Estadística y Censos INEC de Costa Rica. Las poblaciones que sirven de denominadores se tomaron de las más recientes Estimaciones y Proyecciones de población por cantón (Rosero-Bixby, 2008), realizadas conjuntamente ente el INEC y el CCP. Ambos conjuntos de datos pueden ser accesado , y están disponibles en el sitio web del CCP (http://censos.ccp.ucr.ac.cr/). Cabe aclarar que no se hicieron ajustes por subregistro de nacimientos.

Para definirlas como "funciones de datos", las series cantonales de tiempo fueron suavizadas con B-splines usando el software R. Se usó un polinomio de orden 6 y los años 1975, 1980, 1990, 2000 y 2010 como nodos o pivotes. El polinomio de orden 6 permitió modelar aceptablemente el receso en el descenso de la fecundidad ocurrido durante inicios de la década de los ochentas.

Las TBN y TGF para América Latina y el Caribe se tomaron de las publicaciones de Observatorio Demográfico de CELADE (2011). Se tomó el período 1950-2010. Dado que los indicadores están dados por quinquenios, la serie temporal fue asociada a los años intermedios de los quinquenios. Se decidió juntar tanto a los países continentales como a las repúblicas insulares no-hispanohablantes del Caribe, pues se considera que algunos de los patrones observados en los países continentales son similares a los patrones caribeños. Los indicadores ya son calculados por CELADE, por lo que no se realiza ningún ajuste para su cálculo. El suavizamiento también se hizo con B-splines, tomando como pivotes a los años: 1952, 1972, 1992, 2002, y 2007. Se empleó un polinomio de orden 3 (splines cúbicos). Se decidió tomar un orden menor al usado para los cantones costarricenses, porque las series temporales están compuestas apenas por 12 puntos, en lugar de 36.

Resultados.

Discusión

Bibliografía

CELADE (2011). Observatorio Demográfico No. 7: Proyección de Población. Santiago de Chile

Delicado P, Giraldo R, Comas C, Mateu J (2010). “Statistics for spatial functional data: Some recent contributions”. Environmetrics 21:224–239

Hyndman RJ, Booth H, Yasmeen F (2011). “Coherent mortality forecasting: the product-ratio method with functional time series models”. Monash University, Department of Econometrics and Business Statistics. Working Paper 01/11.

Hyndman RJ, Shang HL(2009). “Forecasting functional time series”. Journal of the Korean Statistical Society (2009), 38(3), 199–221.

Hyndman RJ, Ullah MS (2007). “Robust forecasting of mortality and fertility rates: a functional data approach”. Computational Statistics & Data Analysis (2007), 51, 4942–4956.

Ramsay JO, Silverman BW (2006), Functional Data Analysis, 2nd ed., Springer, New York.

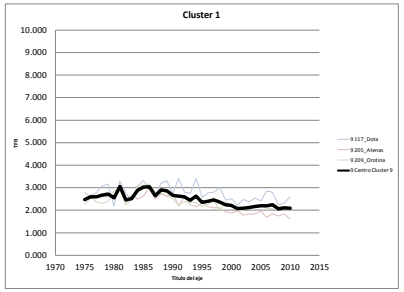
Rosero Bixby L (2008). “ Estimaciones y proyecciones de población por distrito y otras áreas geográficas: Costa Rica 1970-2030. (Actualizadas en 2008). San Jose, C.R.: CCP / INEC.

Rosero-Bixby L, Casterline J (1994). Interaction diffusion and fertility transition in Costa Rica. Social Forces, 73(2): 435-462.

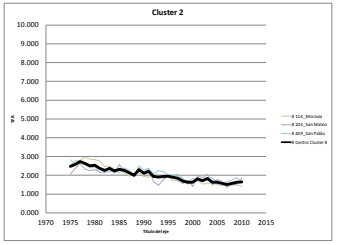
Figures.

Figure 1. Nine clusters of trends in county level Total Fertility Rates (TFRs) in Costa Rica, 1975-2010.

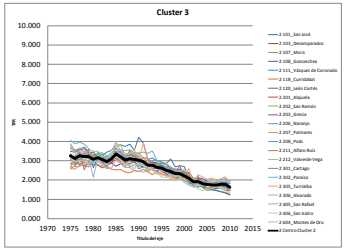
Cluster 1



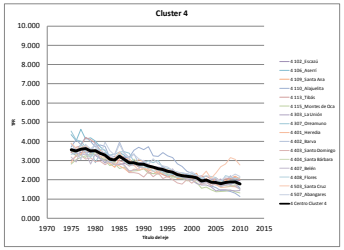

Cluster 2



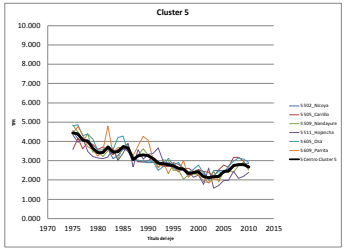

Cluster 3



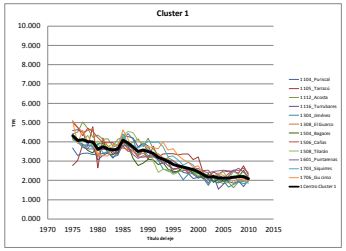

Cluster 4



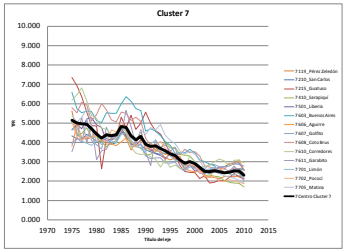

Cluster 5



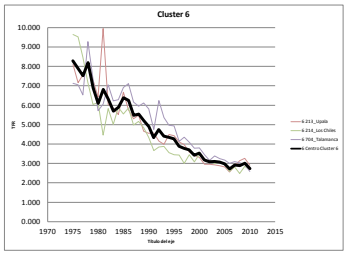

Cluster 6




Cluster 7




Cluster 8




Cluster 9

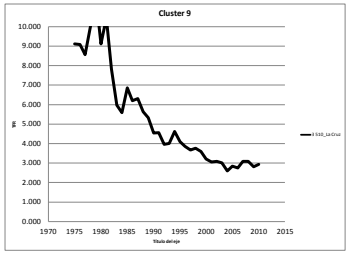


Figure 2. Costa Rican counties: Cluster map of trends in TFR decline 1975-2010.

(1=TFR already low, 9=Sharpest decline from high TFR)

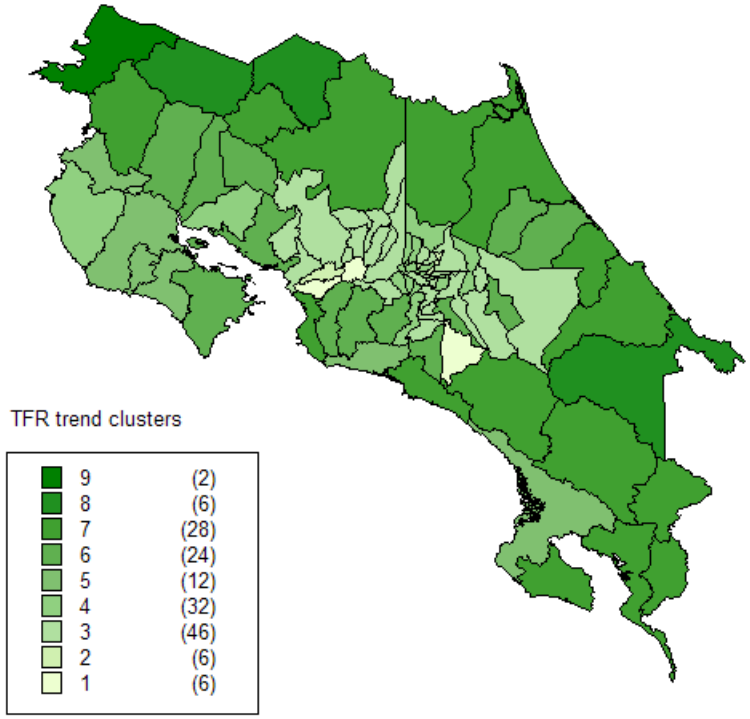


Figura X. Costa Rica: Curva suavizada de la serie de tiempo de Tasas Brutas de Natalidad, por conglomerados de cantones

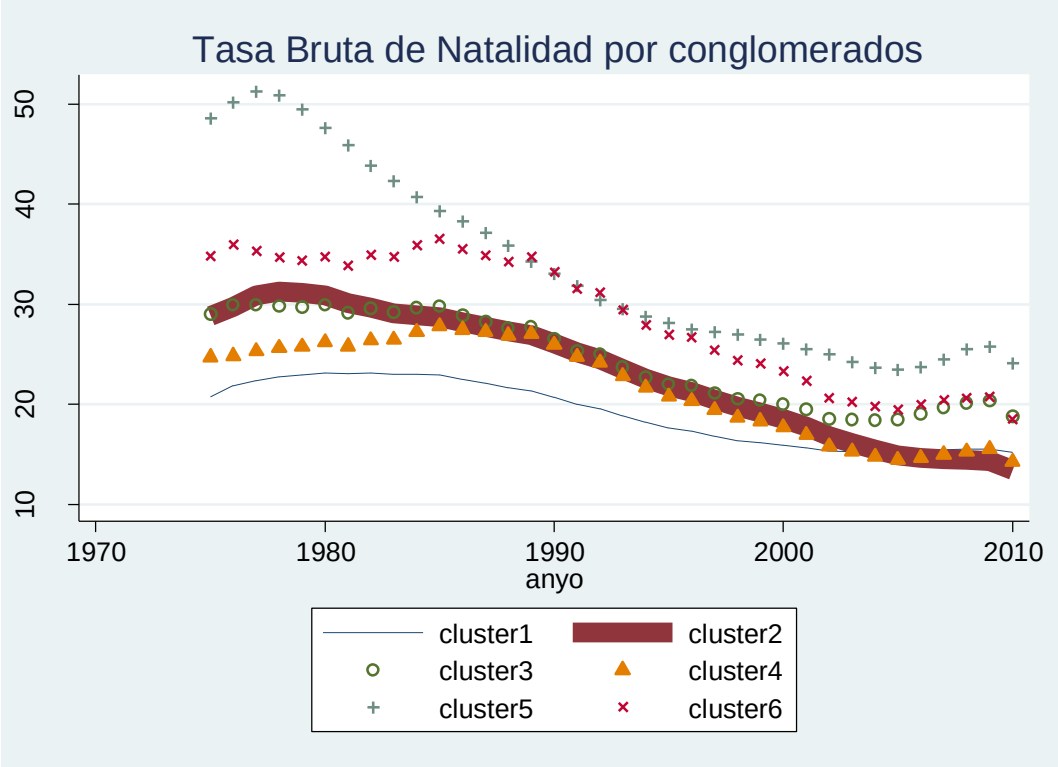


Figura X. Costa Rica: Curva no suavizada de la serie de tiempo de Tasas Brutas de Natalidad, por conglomerados de cantones.

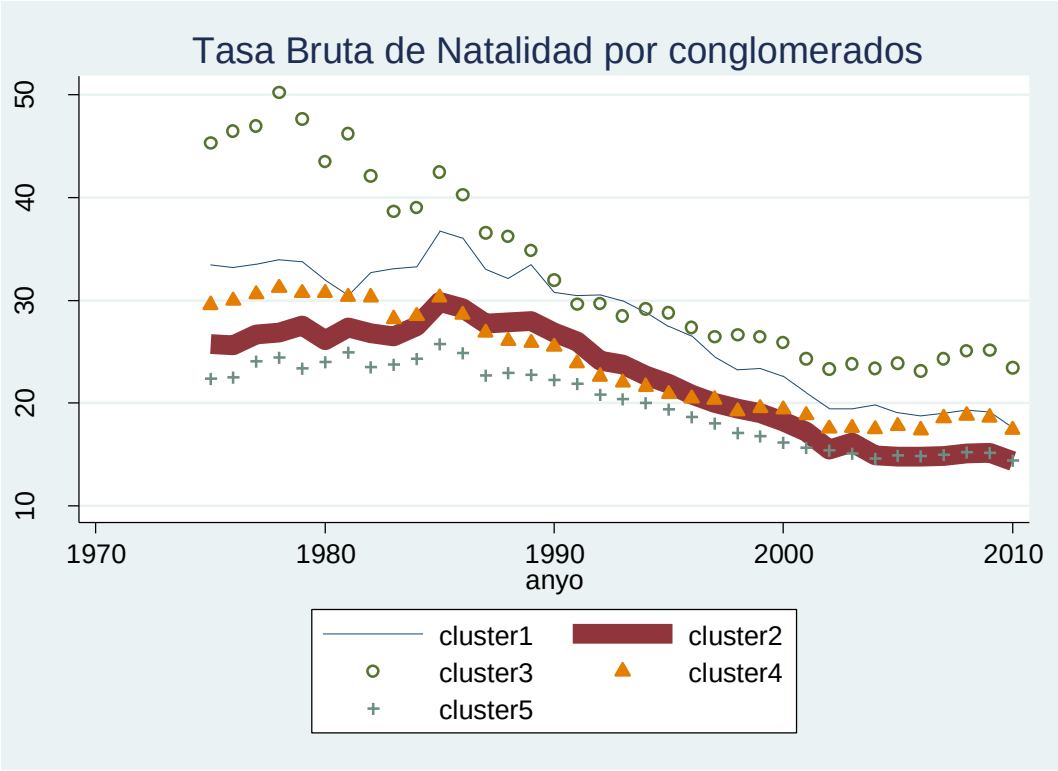


Figura X. Costa Rica: Curva suavizada de Tasas Globales de Fecundidad, por conglomerados de cantones.

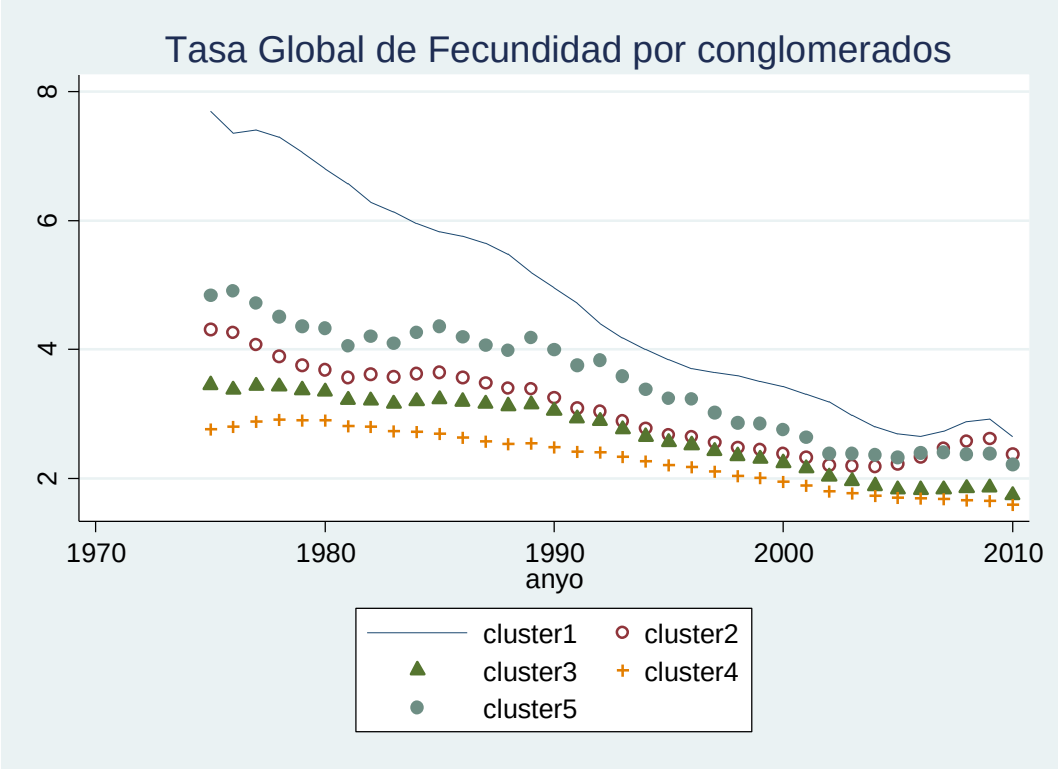
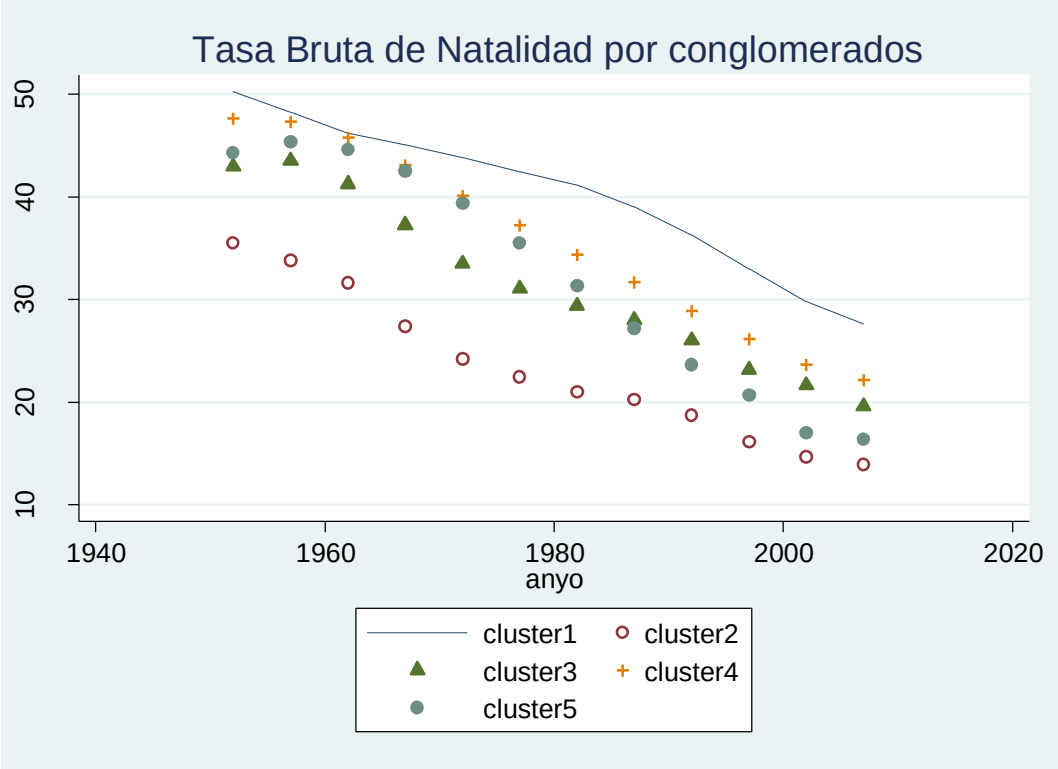


Figura X.  América Latina y el Caribe: Curva suavizada de Tasas Brutas de Natalidad, por conglomerados.




Fuente:

Figura X. América Latina y el Caribe: Curva suavizada de Tasas Globales de Fecundidad, por conglomerados.