Comparing three statistical techniques for Space-time clusters with county-level fertility data from Costa Rica.

Authors:
Gilbert Brenes Camacho, Mario Gutiérrez Leiva, Andrés Ramírez Gutiérrez

Centro Centroamericano de Población

Abstract
Population projections at a small area level are sensitive to random departures in temporal trends of population change components. In order to have robust trends for the county level projections in Costa Rica, the team in charge of forecasting decided to determine county clusters based on the past trends. In this paper, we compare three descriptive techniques used to build space-time clusters for the Crude Birth Rate CBR: Functional Data Analysis (FDA), the scan statistic, and Geographically Weighted Regression GWR. GWR was the technique that yielded clusters related to the diffusion paradigm for fertility decline. However, the scan statistic provided a more parsimonious set of clusters that are more tractable for population projections.

Introduction.

After every decennial census, in collaboration with the Central American Center for Population (CCP) from the University of Costa Rica, the Costa Rican National Institute of Census and Statistics (INEC) aims to update the population projections for the whole country and for its geographical subdivisions: provinces, counties (“cantones”), and districts. Costa Rica has a relatively small population: around 4.6 million people for 2011 (CCP, 2013). Besides, it is a country that has achieved very low mortality (life expectancy is 79.2 years in 2011) and very low fertility (TFR=1.9 children per woman) (INEC-CCP, 2013). These conditions produce a high volatility in annual estimates of demographic statistics by geographical unit (Crude Birth Rates CBR, Crude Death Rates CDR, Total Fertility Rates TFR) because there are some counties and districts with very small populations, and very small figures in the number of births and deaths.

In the previous population projections at the district level (Rosero-Bixby, 2008), CCP and INEC made the assumption that, for each demographic rate (fertility rates, survival ratios), the trend for each district will converge into a single national-level value. This
assumption proved wrong in the short run, affecting projections for young children who go to elementary school. The institutions involved in projections have decided to explore the possibility of making groups of counties with similar trends in fertility, mortality, and internal migration. We decided to try three different techniques: Functional data analysis FDA (Ramsay and Silverman, 2006), Kulldorf’s scan statistic for space-time data (Kulldorf, 2001), and Geographically Weighted Regression (GWR) (Fotheringham, Brundson and Charlton, 2002).

**Heterogeneity in Fertility Decline in Costa Rica**

In this article, the process is described for the Crude Birth Rate CBR. Therefore, we include a brief description of the historical process of fertility decline in Costa Rica. Fertility levels in Costa Rica started to decline in the 1960s decade, after a temporary increment in fertility after World War II during the 1950s. The reduction in the Total Fertility Rate (TFR) was very fast from a peak level of 7 children per woman during the early 1960s to 4 children per woman 15 years later. In 2002, the TFR reached the replacement level (2.1 children per woman) and in 2011 the TFR is 1.9 children per woman (CCP, 2013).

The first researchers that studied the process in the late 1960s and early 1970s noticed that there was geographical heterogeneity in CBR and TFR levels, and that this variability was associated with inter-county differentials in the literacy rate. The lowest TFRs were found in the counties located in Costa Rica’s Central Valley where San José—the capital— is, and also in counties located near the two main sea ports (Puntarenas and Limón) (Jimenez, 1969). On the other hand, higher fertility levels were observed in the counties located near the borders with Nicaragua and Panama. The counties with lower fertility levels were characterized by higher schooling levels than the rest of the country. Quantitative evidence from surveys in Costa Rica started to show as early as the 1970s that the prevalence of contraceptive methods use was increasing at a faster rate in the the Central Valley than in rural zones (Gomez, 1970).

Rosero-Bixby and Casterline (1994) tested the diffusion hypothesis of fertility decline for Costa Rica using spatial statistical methods. The “interaction diffusion” hypothesis
argues that technological innovations and values are spread in a society through social interaction between its members. It implies that fertility decline may spread across socioeconomic borderlines thanks to the communication of new ideas. In the developing world, the diffusion argument is also used to show that part of the reduction in birth rates can be explained by other mechanisms than just family planning programs or structural socio-economic growth (Rosero-Bixby and Casterline, 1994).

One of the arguments that Rosero-Bixby and Casterline (1994) use to support their claim is the “ordered spatial pattern of decline” (p.439). They show a similar pattern than the one showed by Jimenez (1969) and described above: fertility decline started in the Central Valley and in the Northern Caribbean counties, then it was followed by counties in the North and Pacific Coast, and it was achieved more slowly by the counties near the borders with Panama and Nicaragua. It is worth clarifying that the places that are farthest away from the capital are those located near the borders.

The state of below-replacement fertility that Costa Rica has experienced during the 21st century has coincided with a reduction in the heterogeneity of fertility due to different characteristics. Vital statistics and survey evidence have shown that there are small differences in observed and expected fertility levels as well as in contraceptive use across place of residence and socioeconomic status (measured by educational levels), (Ministerio de Salud de Costa Rica et al., 2002).

This historical description is useful because it can be used to validate the clusters found in this analysis. Recall that the main purpose of the analysis is not to test a model of fertility decline, but to create uniform groups of geographical units that can be used to estimate trends in CBRs that can be used for population projections at a county level.

**Description of Statistical Techniques.**

**Functional Data Analysis.**

FDA is a set of statistical procedures that have been used before to establish criteria for population projections (Hyndman & Shang, 2009; Hyndman & Ulla, 2007). Functional data analysis refers to the analysis of information about curves or functions. More
specifically, it comprises multivariate techniques to analyze and reduce the dimensionality of data in which each “observation” is a temporal or spatial sequence of data that can be summarized as a function (Ramsay & Silverman, 2006). These techniques are ideal for describing and clustering a set of time series.

In FDA, each statistical unit is denoted with the subindex $i$. The sequence of demographic indicators through time for each statistical unit is denoted as $Y_i(t)$. In FDA, each $Y_i(t)$ is considered a function, and the purpose is to summarize, describe or analyze a sample of functions $Y_i(t)$. In this article, each county $i$ is the statistical unit, and each time series of the chosen demographic rate for the period 1975-2010 for each county is the function $Y_i(t)$.

FDA assumes that the data $x_{i1}, \ldots, x_{in}$, measured as discrete values, actually represent continuous functions. It also assumes that the data have measurement error.

Therefore, a common step is to smooth the time series so it better represents the continuous function. The smoothing process generally implies that the functions are linear combinations of a “base function”. Among the smoothing methods commonly used, there are k-order polynomials, Fourier series, and B-splines. According to Ramsay & Silverman, the polynomials are utilized for simple functions, while Fourier series require several assumptions, among which one of the most important is that the time series sequence is equally spaced. B-splines are more flexible because their implementation does not require too many assumptions: it needs shorter computing time, and the parameters can be estimated even if the measurements are not equally spaced (Ramsay & Silverman, 2006).

We use a series of techniques in order to establish the basis for the demographic indicators’ projections. In the first place, we use k-means clustering of functional data analysis. This functional data are used as raw data (the CBR and GFR) and spatially smoothed rates using functional data linear models.

In the case of the present paper, we have a set of 81 Costa Rican counties. Each county is FDA aims to summarize, describe or analyze the sample of functions $Y_i(t)$. It is common to summarize functional data with a Functional Data Linear Models, where
each of the functions \( Y_i(t) \) are response variables, which depend on a set of covariates \( Z_i(t) \) that are also functions of time. In the equation:

\[
Y_i(t) = \alpha(t) + Z_i(t)\beta(t) + V_i\gamma + \varepsilon_i(t)
\]

the regression coefficients can also vary over time \(-\alpha(t), \beta(t)\) in order to better describe the relationship between functions (Ramsay & Silverman, 2006), while other might be fixed \( V_i \gamma \). However, this analysis discards the utilization of a linear model, in favor of commonly used clustering algorithms. Specifically in k-means cluster analysis, the traditional algorithms that compute the distances between observations (Pitagorean, Mahalanobis, etc.) are used to assess the distances between the \( n \) set of \( Y_i(t) \) functions. Box-and-Jenkins models can be used to summarize the average time-series that summarize each cluster.

**Scan statistic**

The scan statistic was designed for purely spatial clustering processes (Kulldorf, 1997). It can estimate clusters based on different probability distributions: Bernoulli (for individual cases), Poisson (for aggregate rates), exponential and Gaussian (for continuous variables). Given that the project has CBR information, we use the Poisson models. The algorithm constructs different circles of radius \( r \) at all the coordinate sets (usually centroids), and it defines a cluster if the incidence rate of the cluster is significantly different to the average area rate, based on a likelihood ratio test. Let \( S \) be the purely spatial scan statistic, \( N \) the observed total number of cases (in the whole study area), \( n_Z \) the number of cases in circle \( Z \), \( \mu(Z) \) the expected number of cases in circle \( Z \), and \( L(Z) \) the maximum likelihood for circle \( Z \), then, as Kulldorf (1997, 2001) explains, under the Poisson model,

\[
S = \max \left[ \frac{L(Z)}{L_0} \right] = \max \left[ \frac{L(Z)}{L_0} \right] = \max \left[ \frac{n_Z}{\mu(Z)} \left( \frac{N - n_Z}{N - \mu(Z)} \right)^{N-n_Z} \right]
\]
assuming that $nZ > \mu(Z)$. P-values for each test are computed using Monte Carlo simulations. The algorithm accounts for the inflation or Type I error due to multiple testing.

For the space time clustering process, $S$ is no longer a two-dimensional circle, but is defined as a three-dimensional cylinder. The size of the base circle varies for detecting the spatial clusters, while the height of the cylinder also varies in order to detect different period clusters. Given these cylinders, the algorithm determines monotonic linear trends, rather than curvilinear.

**Geographically Weighted Regression (GWR)**

GWR is a Generalized Linear Model that includes different weights for each observation (Fotheringham et al., 2002). Let $y_i$ be a dependent variable, $x_{ik}$ a set of $k$ independent variables, $\beta_k(u_i, v_i)$ the regression model parameters that can vary geographically as defined by the $u_i$ and $v_i$ geographical coordinates, and $\varepsilon_i$ the error term.

$$y_i = \beta_0(u_i, v_i) + \sum_k \{\beta_k(u_i, v_i)x_{ik}\}$$

The $\beta_k(u_i, v_i)$ parameters are estimated using a weighted information matrix. The weights depend on the shape of the kernel and a bandwidth. Two types of bandwidths have been programmed in the algorithm: a fixed constant bandwidth across the study area, and a geographically variable bandwidth. For the purpose of this analysis, we use the constant bandwidth. Given this fixed bandwidth, the kernel can be defined based on a density function. The weight of the $j^{th}$ data point at the $i^{th}$ regression point is then estimated as (Fotheringham et al., 1997):

$$w_{ij} = \exp \left[-\left(\frac{d_{ij}^2}{b^2}\right)\right]$$

where $d_{ij}$ is the Euclidean distance between the regression point $i$ and the data point $j$, and $b$ is the bandwidth. If the bandwidth increases, the data point $j$ is weighted in such a way as to resemble its neighbor areas. Therefore, $b$ plays a smoothing role within the
model. For the purpose of this analysis, a larger \( b \) generates a more parsimonious clustering because the parameters for the area.

The GWR algorithm has not been programmed for space-time models. Given that the main purpose of the analysis is to define spatial-temporal clusters rather than to estimate an explanatory model, we computed the average geometric rate of change of the CBRs for each county and for the period under study, and then estimate a GWR model whose dependent variable is this geometric rate of change.

**Data sources**
The Crude Death Rates for each county were calculated based on information from Vital Statistics and population estimates. Birth data are processed by the National Institute of Statistics and Censuses (Instituto Nacional de Estadística y Censos INEC), the Costa Rican public institution in charge of official statistics. These data are available at the website of the Central American Center for Population (CCP): http://ccp.ucr.ac.cr/censos/. The size of each county’s population for every year of the 1975-201 period is obtained from the most recent population estimates and projections by district, computed by CCP and INEC (Rosero-Bixby, 2008). The information is available at the CCP website. In addition, General Fertility Rates (GFR) were also computed using the same data sources.

The “functions” for FDA were smoothed with B-splines using R, estimating 6th order polynomials, with the years 1975, 1980, 1990, 2000 and 2010 as nodes. The 6th order polynomial was useful for modelling the slowdown in the decline that occurred during the 1980s.

**Results**
In the FDA analysis, we found 4 clusters of the trends in the county-level CBRs in Costa Rica (Figure 1). We chose the final number of clusters by computing the eta-statistic for the rate of CBR decline between 1975 and 2010 for different total of clusters; we had a
sequence of eta-statistics for a defined number of clusters and we computed the difference between the eta-statistic with \( k \) number of clusters minus the eta-statistic with \( k-1 \) number of clusters; finally we chose the number of clusters such that the difference in these sequential eta-statistics is close to zero. Figure 1 maps the clusters. The yellow counties represent the clusters that had barely a change during the period, while the dark green counties had a sharp decline in their fertility. The map shows that there has been a strong convergence in inter-county birth rates, because most of the counties are clustered into a group with an intermediate rate of change (-0.01226 per year, on average), and this cluster is composed of counties spread all over the Costa Rican territory, from the Pacific coast to the Caribbean coast, and from the border with Nicaragua to the border with Panama. The three counties with the fastest decline rate are located in the country’s Northwest region, while the counties with the slowest decline rate are in the Central Valley and near the Pacific Ocean. Given that the two clusters with the slowest rates are not contiguous, in addition to the large cluster with most of counties, there is little evidence for a spatial clustering that can be used for defining trends for the county-level projections.

The scan statistic has been primarily used for detecting clusters of disease outbreaks. Therefore, the scan statistic procedure did not produce “exhaustive” clusters, that is, it did not classify all the counties into separate clusters, but leaves out all the counties with average trends. Therefore, the gray area in the map represents the counties with average trends (Figure 2). We decided to consider those gray zones as clusters of “average” fertility trends. Neither are they contiguous to each other, as some of the clusters defined by the FDA. There are three clear areas of “average” trends: the counties near the border with Panama, the four counties near the Western part of the border with Nicaragua, and the counties in Heredia province plus the county of San Jose (the capital). The latter area represents the precursors of fertility decline in Costa Rica, while the other two areas are typically classified as laggards, given the early descriptions by Jimenez (1969), Gomez (1970), and Rosero-Bixby and Casterline (1994). The counties in the Central Valley cluster had very low CBRs, while the other two have relatively high CBRs, but they appear to have a similar pace of decline.
The other clusters are the ones detected as such by the scan statistic. Cluster 1 (shaded in light blue) and Cluster 2 (shaded in yellow) agglomerate counties that started the 1970s with low fertility—although not as low as the precursors—and their decline was below average. The average annual decline in CBRs in cluster 1 has been 2.02%; in cluster 2, it has been 2.3%. Both clusters have relative risks that are under one (0.92 and 0.95, respectively) and both have counties located in the Central Valley. While cluster 1 basically comprises the Eastern side of the Central Valley and the El General Valley, cluster 2 is composed of the Western part of the Central Valley, as well as most of the Pacific Coast: from the Peninsula and the Gulf of Nicoya to Garabito and Quepos. These two clusters are what Rosero-Bixby and Casterline (1994) describe as the followers that were located in an axis parallel to the Pacific Coast. In Figure 1, these counties are in clusters that have low or intermediate levels of CBR decline. Finally, the third cluster has the sharpest decline, and it is composed of two neighboring counties: Pococi and Sarapiqui, in the Eastern part of the border with Nicaragua. These two counties are important areas of banana plantations, but their secondary cities (particularly, Guapiles and Cariari) have experienced a fast degree of urban growth linked to the services for the banana industry. Their CBRs fell 2.6% on average, and they still have a 7% higher incidence ratio than the average. Both counties started with very high CBRs in the 1970s and still have high levels.

The clustering achieved with the scan statistic has the advantage that, after mapping the results, the clusters are composed of contiguous counties, although it was necessary to create separate groups with sets of counties classified under “average change”. This clustering has the advantage that it agrees with the theoretical framework used by other authors (Jimenez, 1969; Gomez, 1970; Rosero-Bixby and Casterline, 1994).

Finally, we estimated clusters based on the smoothing procedure used by Geographically Weighted Regression—GWR smoothing from now on—(Figure 3). We found 5 clusters. The yellow and light green clusters had a fast decline in CBRs. Again, they are located near the border with Nicaragua (Los Chiles, Upala, Guatuso, La Cruz, Sarapiquí), on the Caribbean Coast (Limón Province), and some Eastern counties of San Jose’s Metropolitan Area. On the contrary, the dark green clusters are located
on the Pacific Coast and part of the Alajuela province. They had a small CBR decline. The average decline was observed in the rest of San Jose’s Great Metropolitan Area (from Alajuela to Cartago). The kernel used by the GWR algorithm allowed to have groups composed by contiguous counties, although the group with the fastest decline was spatially separated into at least 5 sub-clusters: two in the North West zone near the border with Nicaragua, three in the Caribbean coast (including Sarapiqui), and another in the North Eastern part of the Central Valley (in the right part of the segregated box at the bottom). Nonetheless, the clustering also agrees with the observations of the authors commented above.

When the three clustering techniques are compared, the data show that there is not an exact agreement across the three procedures. All the clustering techniques highlight the slow pace of fertility decline in some areas of the Central Valley, mainly in the Western part of the central Valley. However, the scan statistic grouped all the Eastern counties of the Central Valley under this label, the GWR smoothing determined that only part of this area had low fertility decline—while the rest of the Eastern counties had average decline— and the FDA technique grouped most of the Central Valley counties under the slow pace label.

Additionally, the scan statistic classified the patterns in the North West and in the South East as an “average” decline, the FDA and the GWR smoothing results suggest that the CBRs fell rapidly during the period. The scan statistic and the GWR smoothing seem to identify low or high levels of fertility throughout the period, while FDA differentiates the trends.

We could also highlight that the clustering obtained by the scan statistic is more parsimonious than the one obtained by GWR, because the 5 groups defined by the pace of decline must be subdivided in order to have certain contiguity in the clusters. In this sense, the scan statistic has the advantage that most of the clusters are composed of counties that are in a “spatial continuous”, except for the counties that have an “average” decline. Given the advantages of the scan statistic, we reproduced the same procedure using the General Fertility Rate (GFR). The results obtained are basically the same, although the clustering based on the GFR classified some few counties into
the average decline group, rather than on the clusters detected by the scan statistic algorithm.

**Conclusions**
The main goal of this analysis was to establish groups with similar trends in the decline of the CBR over the period 1975-2010, in order to use these groups to estimate trends for projecting births at a county and district level. The best grouping should have spatial contiguity (clusters of neighboring clusters), agreement with historical analysis of the fertility decline in Costa Rica, and parsimony in the number of clusters. The three techniques rendered clusters that agree with historical and empirical analyses. The result generated with FDA is parsimonious, but there is a large cluster that consists of most of the counties; therefore, its explanatory power is limited. The result from the scan statistic algorithm is more parsimonious than the result from GWR, and most of the clusters are depicted as continuous areas. However, the counties that are not clustered because they have average rates of decline have to be grouped in a separate stage after the algorithm has finished. This can be viewed as a limitation if the users of the clustering procedure would like to program it as an automatic process.

For the purpose of the county-level projections in Costa Rica, the clustering based on the scan statistic seems to be the best choice because it achieves the three conditions described at the beginning of these conclusions.

**References.**


Figure 1. Costa Rican counties: Map of clusters of trends in CBR decline 1975-2010, determined by FDA procedures. (1=CBR already low, 9=Sharpest decline from high CBR)
Figure 2. Costa Rican counties: Map of clusters of trends in CBR decline 1975-2010, determined by the scan statistic procedure.

(Cluster 1: small decline, Cluster 2: Second smallest decline, Cluster 3: Sharpest decline, Cluster “Ausencia de cluster”: Average decline)
Figure 3. Costa Rican counties: Map of clusters of trends in CBR decline 1975-2010, determined by a Geographically Weighted Regression smoothing of intercepts of the CBR decline.