USING KRIGING AND THE DAC STATISTIC TO PREDICT LOW BIRTHWEIGHT CLUSTERS IN SPARTANBURG COUNTY, SC

Alexandru I. Petrisor, MSPH[1]; **J. Wanzer Drane**, PE, PhD[1]; and **Liviu Dragomirescu**, PhD[2]
1 - School of Public Health, University of South Carolina, Columbia, SC, USA
2 - Department of Systems Ecology, University of Bucharest, Bucharest, Romania

Key Words: DAC statistic; Spatial statistics; Low birthweight clusters; Empirical cumulative distribution; Kriging

Abstract

Spatial distributions find applications in various fields, such as public health, biology, ecology, geography, economics or sociology. The DAC statistic is the difference between the empirical cumulative distribution of cases and that of non-cases at a particular point. Simulations indicated that the location of the maximum DAC statistic is not unique. Previous studies attempting to use the DAC statistic to suggest or even predict occurrences of low birthweight clusters faced various problems, such as the discrete nature of sample data or software limitations. This study uses the longitude and latitude as the coordinates of the homes of mothers in Spartanburg County, SC who gave birth to their babies in 1989 or 1990, and the DAC statistic in conjunction with various kriging approaches. For the particular data set used in this study, SAS yielded inconclusive results. The results obtained using ArcGIS confirm previous findings suggesting low birthweight clusters located in Spartanburg, Chesnee, and possibly Inman and Greer. Therefore, the DAC statistic should be used with caution, but its usefulness as a set of spatial descriptive statistic is not diminished in the least.

Introduction

Space-time analyses represent important issues, due to their wide area of application. In public health, they are used to detect disease space and time clusters (Aldrich, 1993; Aldrich, 1997; Britton, 1995; Stark, 1967; and Williams, 1978), to increase the efficiency of health department's activity (Britton, 1995), or just to study the spatial pattern or distribution of a population dispersed over a continuous surface (Paloheimo, 1976). Different studies have indicated various approaches to space-time analyses over wide and expanding venues of applications. One approach was to work on disease risk from environmental hazard at three levels: analyses of distribution, analyses of sentinel events, and case cluster strategies (Aldrich, 1997). The analysis of distribution refers to the DAC statistic; the analysis of sentinel events recognizes that some events are more important than others when used to draw attention, and case-cluster strategies permit the identification of disease clusters. The DAC statistic is defined as the difference between the empirical distribution for the cases and that of the total sample (Aldrich et al., 1997). A simulation indicated that the location of the maximum DAC statistic is not unique, moreover there is a geometrical locus of it, and this varies as the orientation of the axes changes (Petrisor, 2000). Other studies investigated the usefulness of the DAC statistic in suggesting spatial clusters. Sampling provided discrete data and the analysis could not point directly to potential clusters. SAS yielded inconclusive results (Petrisor et al. (2), 2001), whereas the location of clusters depended on classification if using ArcView GIS (Petrisor et al. (1), 2001).

Spatial prediction, referred as kriging (Johnston et al., 2001; Piegorsch et al., 2001) may provide a way to generate smooth continuous surfaces and predict the behavior of the DAC statistic at each location within the investigated area, suggesting also location of potential clusters. Various kriging procedures are available in SAS (Piegorsch et al., 2001) and ArcGIS (Johnston et al., 2001).

The purpose of this study is to use the DAC statistic in conjunction with various approaches to kriging to suggest potential low birthweight clusters in Spartanburg county, SC.

The DAC Statistic

The DAC statistic was introduced for the first time in the statistical literature through a study by Drane, Creanga, Aldrich, and Hudson (Drane, 1995; Petrisor et al., 2000). The purpose of introducing the DAC statistic was to provide an instrument to detect spatial clusters, or, more generally, areas with health problems. The computation of the DAC statistic is based on the empirical cumulative distribution function, defined as:

$$F_n(x_1, x_2) = m(x_1, x_2) / n \qquad (1)$$

where $m(x_1, x_2)$ is the number of points of the sample of size *n* such that $x_{1_i} \leq x_1$ and $x_{2_j} \leq x_2$ (4). As (x_1, x_2) covers the entire sample from (0, 0) to (max x_1, max x_2), $m(x_1, x_2)$ spans the interval [0, n].

The DAC statistic is, for all permissible values of (x_1, x_2),

$$DAC\,(x_1, x_2) = F_m\,(x_1, x_2) - F_n\,(x_1, x_2) \qquad (2)$$

F_m is the empirical cumulative distribution function of all cases, and F_n is the empirical cumulative distribution function of the total population (Drane, 1995; Petrisor, 2000). If within the sample of size n there are m cases and n-m non-cases, F_{n-m} may be substituted for F_n.

Birth Data 1998-1990, Spartanburg County, SC

The data came from a demonstration project sponsored by the Robert Woods Johnson Foundation. The one legal paper, which had a great promise of nearly a 100% response rate, was the birth certificate. It was chosen. For the period 1989-1992 nearly all of the live births in Spartanburg County SC were geocoded. The longitude and latitude of the mother's home was affixed to the birth certificate data of the baby. For this particular biostatistical methodological investigation the only data used were the longitude, latitude and the baby's birth weight.

The data set consisted of 6434 lines of observations, corresponding to 6434 live births. Out of these, 591 were cases. Cases were low birthweight babies. Low birthweights were defined as those less than or equal to 2500 grams (Petrisor et al., 2000).

Results

Results of the previous studies are presented below. Figure 1 displays a map of the DAC statistic in Spartanburg County using SAS. Each location is represented through a vertical line of height proportional to the value of the DAC statistic at that location. The results are inconclusive first due to the quality of the representation and second due to the lack of a base map unto which values could be projected to identify the location of potential clusters.

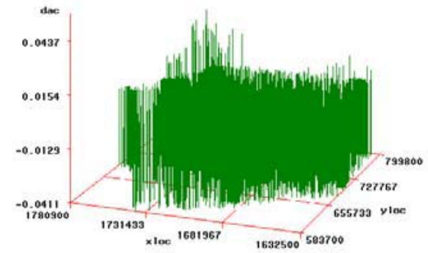

Figure 1. Values of the DAC Statistic for the Spartanburg Data

A Turbo-Pascal application was used to produce smoother surfaces through an interpolation algorithm. The results are displayed in Figure 2. It can easily be noticed that beside the quality of the representation, the lack of a base map remains a problem.

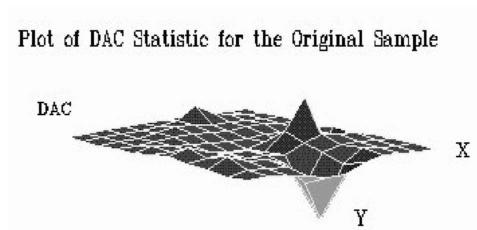

Figure 2. Location of the Maximum DAC Statistic for the Spartanburg Data

Figure 3 presents a density map of positive DAC values in Spartanburg County, SC. The shading intensity is directly proportional to the density of positive values in the area. Main cities are displayed as black dots. Even if in this case the location of clusters becomes possible, various classification criteria point to different areas of high density of positive DAC values indicating potential clusters. The extreme cases are "too many areas", i.e. very small areas centered at each positive DAC value, and "one area", i.e. Spartanburg city.

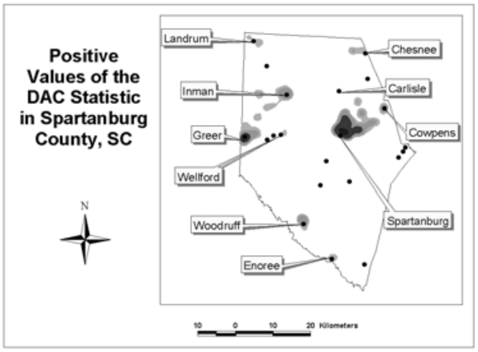

Figure 3. Map of the positive DAC statistic values in Spartanburg County in relationship with the position of the main cities

Our project used three approaches to kriging available in ArcGIS. Ordinary kriging uses semivariogram or covariance models relying on spatial relationships among the data, assuming intrinsic stationarity and that the true mean of the data (i.e. mean DAC value) is constant but unknown (Johnston et al., 2001). Figure 4 displays the semivariogram corresponding to using ordinary kriging for the DAC data. It may be argued that the assumption of a constant mean does not hold in this case.

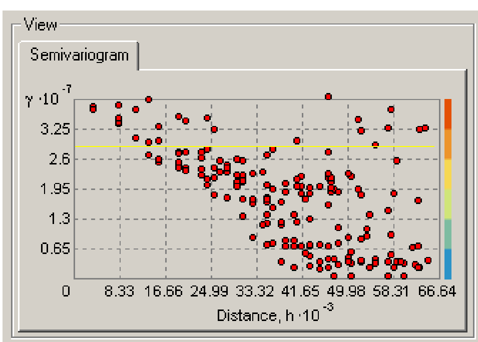

Figure 4. Ordinary Kriging: Semivariogram

Fitting this model provides the map displayed in Figure 5. Grey shades indicate negative values of the DAC statistic. Of interest for our study are black shades suggesting clusters of positive values indicating low birthweights.

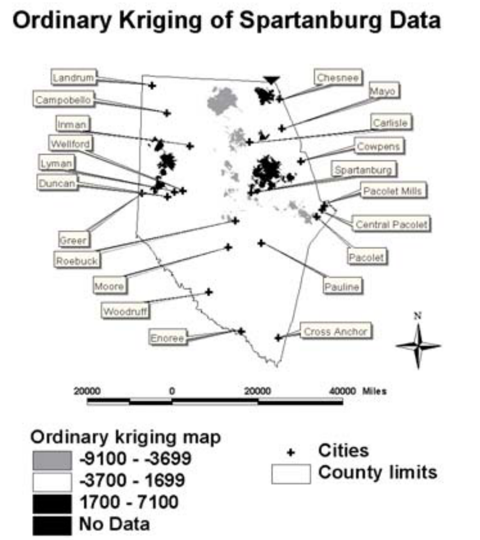

Figure 5. Ordinary Kriging: Prediction Map

Simple kriging, as defined in ArcGIS, uses semivariogram or covariance models relying on spatial relationships among the data, assuming intrinsic stationarity and that the true mean of the data (i.e. mean DAC value) is constant and known (Johnston et al., 2001). Figure 6 displays the semivariogram corresponding to using simple kriging for the DAC data. Apparently, the semivariogram fits the data pretty well.

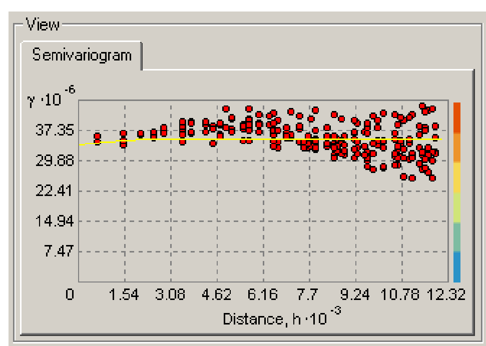

Figure 6. Simple Kriging: Semivariogram

Fitting this model provides the map displayed in Figure 7. Grey shades indicate negative values of the DAC statistic. Of interest for our study are black shades suggesting clusters of positive values indicating low birthweights.

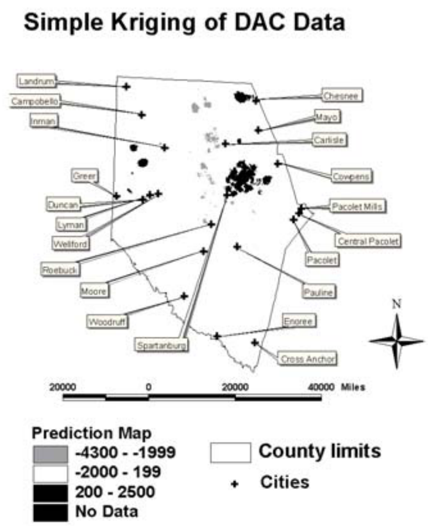

Figure 7. Simple Kriging: Prediction Map

Universal kriging, also available in ArcGIS, uses semivariogram or covariance models relying on spatial relationships among the data, assuming that the true mean of the data (i.e. mean DAC value) is some deterministic function (Johnston et al., 2001). Figure 8 displays the semivariogram corresponding to using universal kriging for the DAC data. It may be argued that the semivariogram does not fit the model.

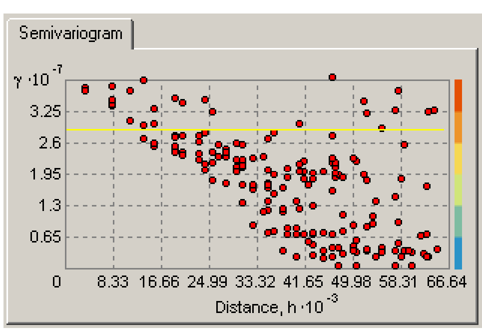

Figure 8. Universal Kriging: Semivariogram

Fitting this model provides the map displayed in Figure 9. Grey shades indicate negative values of the DAC statistic. Of interest for our study are black shades suggesting clusters of positive values indicating low birthweights.

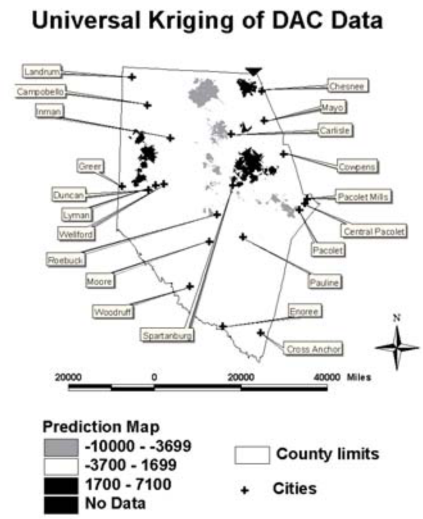

Figure 9. Universal Kriging: Prediction Map

There were two attempts to kriging with SAS. The first one was designed to perform an ordinary kriging. The first problem was to generate a prediction grid. Given the size of the original data set (6434 observations), generating a fine resolution prediction grid resulted into exceeding the allocated memory for a proper running of the program. Therefore, no results were obtained before limiting the prediction grid to a low-resolution one.

At this stage, SAS provided a unique estimate with the same standard error for all the locations within the predicted grid. It could be argued that this value represents an estimate of the true mean of the data (Johnston, 2001; Piegorsch, 2001).

The next step involved an attempt to universal kriging. Again, the allocated memory for a proper running of the program was exceeded and no results were obtained.

In summary, all the attempts to kriging with SAS provided inconclusive results for the Spartanburg data.

Discussion

Despite of the kriging method used in ArcGIS, all the results indicated that the predicted clusters of low birthweights occurred close to Spartanburg, Chesnee, and possibly Inman and Greer. In this example, the DAC statistic appears to be a useful instrument in suggesting spatial clusters if used in conjunction with spatial prediction methods. Nevertheless, the DAC statistic should be used with caution, but its usefulness as a set of spatial descriptive statistic is not diminished in the least.

In this study, even if easier to control from a statistical viewpoint, kriging with SAS was limited by the size of the data set resulting into exceeding the allocated memory for a proper running of the program. Therefore, given these limitations, the results provided by SAS are inconclusive. At the same time, limitations to kriging with ArcGIS refer to the ability to control the modeling process statistically.

Acknowledgement

We are thankful to Dr. John Grego from the Department of Statistics at the University of South Carolina for all the time and efforts he invested in exploring various kriging approaches with SAS using our data.

References

Aldrich T.E., and Drane J.W. 1993. Cluster 3.1: Software System for Epidemiologic Analysis. USDHHS, ATSDR, Atlanta, GA 30333

Aldrich T.E., Krautheim K., Kinee E., Drane J.W., and Tibara D. 1997. Statistical Methods for Space-Time Cluster Analysis. *Proceedings of the International Symposium on Computer Mapping in Epidemiology and Environmental Health*. 226-236

Britton T. 1995. Tests to Detect Clustering of Infected Individuals within Families – Research Report. *Institute of Actuarial Mathematics and Mathematical Statistics, Stockholm University*: 1-18

Drane J.W., Creanga D.L., Aldrich T.E., and Hudson M.B. 1995. Detecting Adverse Health Events via Empirical Spatial Distributions (Abstract). Symposium on Statistical Methods 1995, U.S.D.H.H.S., P.H.S., C.D.C., Atlanta, GA, January 24-26, 1995

Johnston K., Ver Hoef J.M., Krivoruchko, K., and Lucas, N. 2001. Using ArcGIS Geostatistical Analyst. *ESRI*, Redlands, CA

Piegorsch W.W., and Edwards, D. 2001. Notes on Temporal and Spatial Analysis for a Course in Environmetrics. *Department of Statistics. University of South Carolina*. Columbia, SC

Petrisor A.I. 2000. Empirical Spatial Distributions and the DAC Statistic. Master Thesis. *University of South Carolina*. Columbia

Petrisor A.I., Drane J.W., Jackson K.L., and Dragomirescu, L. 2000. Review of the DAC Statistics. *Bulletin of South Carolina Academy of Science* **63**: 94-5

Petrisor A.I., Drane J.W., Jackson K.L., and Dragomirescu, L. 2001. Spatial Statistics: The DAC Statistics, *Proceedings of the Joint Statistical Meetings*, Atlanta, GA

Petrisor A.I., Drane J.W., Dragomirescu, L. and Jackson, K.L. 2001. A Computer-Based Application to Detect Spatial Clusters Using the DAC Statistic, *Proceedings of the Medical Infobahn for Europe Special Topic Conference "Healthcare Telematics Support in Transition Countries"*, Bucharest, Romania

Stark C.R., and Mantel N. 1967. Temporal-Spatial Distribution of Birth Dates for Michigan Children with Leukemia. *Cancer Res.* **27**: 1749-1775

Williams E.H., Smith P.G., Day N.E., Geser A., Ellice A., and Tukei, P. Space-Time Clustering of Burkitt's Lymphoma in the West Nile District of Uganda: 1961-1975. 1978. *Br. J. Cancer* **37**: 109-122

*** The Division of Biostatistics, Office of Vital Records and Public Health Statistic, South Carolina Department and Environmental Control. 1993. South Carolina Vital and Morbidity Statistics 1990. Volume I: Annual Vital Statistics Series