To cite this paper / Pentru a cita lucrarea:
Petrisor AI, Drane JW, Jackson K, Dragomirescu L (2002), Spatial Statistics: The DAC Statistics, 2002 Proceedings of the Section on Statistics and the Environment, American Statistical Association, pp. 2683-2686 (ISSN 1093-1880)

SPATIAL STATISTICS: THE DAC STATISTIC

Alexandru I. Petrişor, MSPH[1]; **J. Wanzer Drane**, PE, PhD[1]; **Kirby L. Jackson**, AB[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: Spatial statistics; DAC statistic; cluster

ABSTRACT

Spatial distributions find applications in various fields, such as public health, biology, ecology, geography, economics or sociology. Particular attention was given to their sensitivity to the location of origin and orientation of axes. 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. The DAC (Drane- Aldrich- Creangă) statistic is the difference between the empirical cumulative distribution of cases and that of the population at a particular point. For any size of a random sample of locations taken from the 6434 live births there is a noticeable variation of the location of the DAC statistic with random rotations within a given sample, when transformed back to original longitude and latitude. Simulations 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. Therefore, the DAC statistic should be used with caution, but it might be a good instrument to detect spatial clusters.

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). In Ecology, the same analyses are used to generate individual-based models (Huston, 1988). The area of application may be even wider, expanding to socio-economic problems, biology, or to geographical sciences.

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 that will be presented later on; 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. Another study used the measures of aggregation based on the counts of individuals in randomly sampled quadrates, and indices based on the spacing of the individuals, calculated from either nearest neighbor or "point to plant" distances (Paloheimo, 1976). Creangă (1998) created models of attenuations of point sources and their effects on the surrounding population. The Ederer-Myers-Mantel procedure is used to detect temporal-spatial clustering of events (Stark, 1967). The procedure is using a cell-occupancy approach and consists of dividing the time period into disjoint subintervals; under the null hypothesis of no clustering, cases are multinomially distributed among the subintervals, and the test statistic is the maximum number of cases occurring in a subinterval. Other authors propose a simpler test to detect within-family clustering of infected individuals, derived as the locally most powerful test for several parametric models designed to allow an increased within-family infectivity (Britton, 1995). Still others have recommended as a better approach to be vigilant for unusual environmental exposures, and to evaluate their possible impact, suggesting that cluster techniques might represent a part of a larger investigation including other epidemiological approaches (Smith, 1993). The theory of clustering processes and doubly stochastic processes were considered along with the dependence on the size of the quadrat to study the spatial pattern or distribution of a population dispersed over a continuous surface (Paloheimo, 1976).

In this investigation, the approach is to examine the empirical cumulative distributions. Probability is accumulated in a southwest - northeast direction as though we were working only in the first quadrant. It is well known that the origin of the axes will not affect the distribution of cases other than shifting the measures of location. In order to investigate the effect of a change in the origin and orientation of axes, the DAC statistic shall be used. It 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 (Petrişor, 2000).

The purpose of this paper is to investigate the sensitivity of the DAC statistic to the orientation of the axes of the cumulative distribution. Empirical distributions will be examined against arbitrary choices of orientations of axes.

THE DAC STATISTIC

The DAC statistic was introduced for the first time in the statistical literature through a study by Drane, Creangă, Aldrich, and Hudson (Drane, 1995; Petrişor 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.

The empirical cumulative distribution function is:

$F_n(x_1, x_2) = \frac{m(x_1, x_2)}{n}$, where $m(x_1, x_2)$ is the number of points of the sample of size *n* such that $x_{1_i} \le x_1$ and $x_{2_j} \le x_2$ (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),

$DAC(x_1, x_2) = F_m(x_1, x_2) - F_n(x_1, x_2)$.

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; Petrişor, 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.

The maximum absolute value of the DAC statistic represents the Kolmogorov-Smirnov statistic for two samples (Hollander, 1973).

BIRTH DATA 1998-1990, SPARTANBURG COUNTY, SC

The data came from a demonstration project sponsored by the Robert Woods Johnson Foundation. The object of the effort was to demonstrate the usefulness of geographically coded health events.

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. Each line contains, in order, the following variables: a line (1-6434), the actual latitude and longitude, and the infant's birth weight (Petrişor et al., 2000).

RESULTS

1. RANDOM LOCATIONS OF ORIGIN

The translation of origin is equivalent to adding constants to the coordinates of each data point. That is,

$T(X_1, X_2) = (X_1 + \alpha, X_2 + \beta)$ for all (x_1, x_2),

where $-\infty < \alpha, \beta < \infty$.

This change does not affect the order relationship between any possible set of data pairs. Only the measures of location, which change with a constant amount, are affected. As the cumulative distribution function is a step function and depends only on the order relationship between any possible set of data pairs, its shape is not influence by the change of the location of origin.

2. RANDOM ORIENTATIONS OF AXES

For these simulations, a special program, called "DAC.EXE", was created in Microsoft Q-Basic®. In order to increase the efficiency of this program (in terms of memory usage and speed), it was converted to an executable program using Quick Basic®. The program reads the initial data in comma-delimited format from a file titled *inp-data.txt*, prompts for the number of samples to be selected and for the size of each sample. It produces an output file in the same format called *out-data.txt*, containing as many lines as the number of samples indicates. Each line contains, in order:

- Maximum DAC statistic for respective sample (MaxDAC);
- The X value at which MaxDAC occurred;
- The Y value at which MaxDAC occurred;
- Maximum DAC statistic for rotated sample (Max DACr);
- The X value at which Max DAC occurred (in terms of original coordinates);
- The Y value at which Max DAC occurred (in terms of original coordinates) (Petrişor et al., 2000).

Due to the Quick Basic® processor, the maximum sizes allowed by the program ranged from either 20 samples of size 400 or 40 samples of size 200. This problem was overcome through a completely random device based on the computer clock. In successive steps, the program was able to draw 1,000 samples of size 400. The samples were rotated with random angles and the results are displayed below in figures 1 and 2.

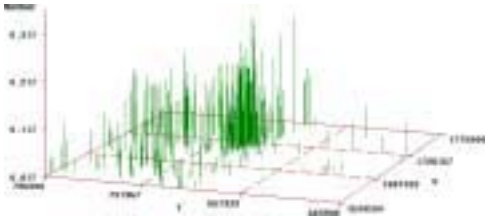

Figure 1. Location of the Maximum DAC Statistic for 400 Random Samples

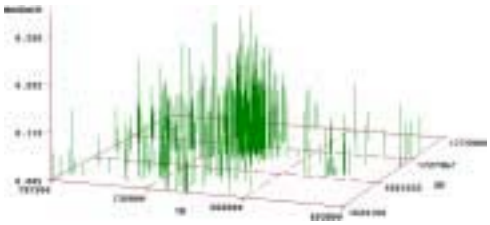

Figure 2. Location of the Maximum DAC Statistic for 400 Random Samples Rotated with Random Angles

In the next step, the DAC statistic was computed for all 6434 observations. Data were rotated arbitrarily and the DAC statistic was recomputed for the rotated data. The results are displayed in Figures 3 and 4 using a Turbo-Pascal® plotting application.

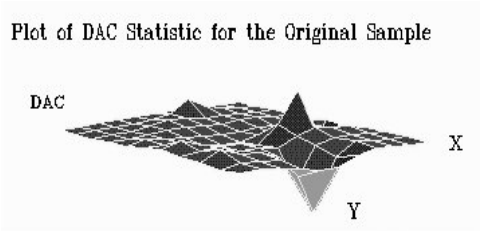

Figure 3. Location of the Maximum DAC Statistic for the Original Data

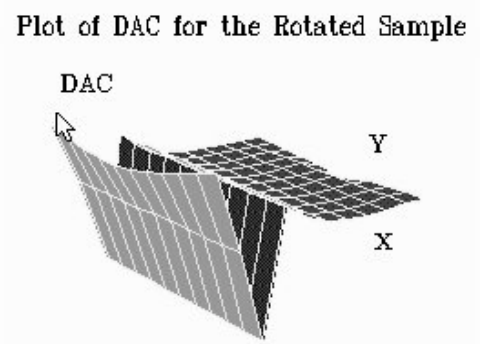

Figure 4. Location of the Maximum DAC Statistic for the Rotated Data

It may be noticed even with a naked eye that the maximum DAC statistic occurs at approximately the same location before and after rotating the samples arbitrarily. This may support the reliability of the maximum DAC statistic in terms of detecting spatial clusters.

DISCUSSION

The results indicated that the DAC statistic does not depend on the location of the origin. However, the dependence on the orientation of axes has an analytical expression that may not be easily detected. In this example, the maximum DAC statistic appears to be a reliable instrument in detecting spatial clusters independently of the orientation of axes.

In real life example, the maximum DAC statistic does not have necessarily an analytical expression, therefore it is almost impossible to find its geometrical locus. The question remains whether it will still remain a reliable instrument in detecting spatial or temporal clusters. Obviously, more and deeper research are the order of the day.

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 Tibără 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

Creangă D.L. 1998. Spatial Data Analysis for Distributions of Health Events. Doctoral Thesis. *University of South Carolina*. Columbia

Drane J.W., Creangă 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

Dudewicz E.J.. 1976. Introduction to Probability and Statistics. *Holt, Rinehart and Winston*: 59-64, 90-93

Freund J.E., and Walpole R.E.. 1980. Mathematical Statistics. *Prentice-Hall, Inc.* Englewood Cliffs: 78

Gauss (translator Davis C.H.). 1963. Theory of the Motion of the Heavenly Bodies Moving about the Sun in Conic Sections. *Dover Publications Inc.* New York

Hollander M., Wolfe D.A. 1973. Nonparametric Statistics Methods. *John Wiley & Sons*. New York

Huston M., DeAngelis D., and Post W. 1988. New Computer Models Unify Ecological Theory. *BioScience* **38**: 682-691

Mendenhall W., and Scheaffer R.L. 1973. Mathematical Statistics with Applications. *Duxburry Press*. Massachusetts: 33, 46-47, 60-64, 165

Mood, A.M., Graybill F.A., and Boes, D.C. 1974. Introduction to the Theory of Statistics. *McGraw-Hill*: 132-133

Paloheimo J.E., and Vokov A.M. 1976. On measures of Aggregation and Indices of Contagion. *Math. Biosci.* **30**: 69-97

Petrişor AI. 2000. Empirical Spatial Distributions and the DAC Statistic. Master Thesis. *University of South Carolina*. Columbia

Petrişor A.I., Drane J.W. et al. 2000. Review of the DAC Statistics. *Bulletin of South Carolina Academy of Science* **63**: 94-5

Smith D., and Neutra R. 1993. Approaches to Disease Cluster Investigations in a State Health Department. *Stat. Med.* **12**: 1757-1762

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

Wilks S.S. 1946. Mathematical Statistics. *Princeton University Press.* Princeton: 8-13

Wilks S.S. 1962. Mathematical Statistics. *John Wiley & Sons, Inc.* New York: 33, 46-47, 60-64

*** 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