Examining the Reliability of the DAC Statistic in Detecting Spatial Clusters and its Relationship with the Survivorship Function

To cite this paper / Pentru a cita lucrarea:
Drane JW, Petrisor AI, Dragomirescu L (2003), Examining the Reliability of the DAC Statistic in Detecting Spatial Clusters and its Relationship with the Survivorship Function, International Conference on Reliability and Survival Analysis, Columbia, SC, USA, May 23, 2003

J. Wanzer Drane, PhD [1]; **Alexandru I. Petrisor**, PhD Candidate [2], and **Liviu Dragomirescu**, PhD[3]

(1) Department of Epidemiology and Biostatistics, Arnold School of Public Health, University of South Carolina, Columbia, SC 29208

(2) Department of Environmental Health Sciences, Arnold School of Public Health, University of South Carolina, Columbia, SC 29208, e-mail: alex@sc.edu

(3) Department of Ecology, Faculty of Biology, University of Bucharest, Romania

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. The purpose of introducing the DAC statistic was to provide an instrument to suggest spatial clusters, or, more generally, areas with possible health problems. The DAC statistic is the difference between two empirical cumulative distribution functions. The empirical cumulative distribution function is:

$F_n(x,y)=m(x, y)/n$, where $m(x, y)$ is the number of points of the sample of size n such that $x_i \le x$ and $y_j \le y$.

As (x, y) covers the entire sample from $(0, 0)$ to $(\max x, \max y)$, $m(x, y)$ spans the interval $[0, n]$.

The DAC statistic spans the interval $[0, 1]$. For all permissible values of (x, y),

$DAC(x, y)=F_m(x, y)-F_n(x, y)$.

F_m is the empirical cumulative distribution function of all cases, and F_n is the empirical cumulative distribution function of the total population.

To discuss the dependence of the DAC statistic on the location of origin, consider that the translation of origin is equivalent to adding constants to the coordinates of each data point. That is,

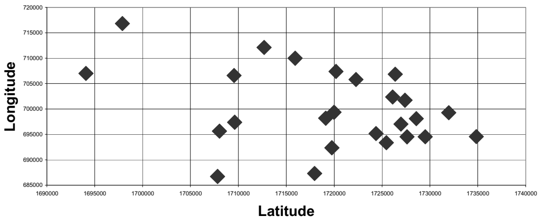
$T(x, y)=(x+\alpha, y+\beta)$ for all (x, y), where $-\infty<\alpha, \beta<\infty$.

Therefore, $F_n(x+\alpha,y+\beta)=m(x+\alpha, y+\beta)/n$, where $m(x_{1i}+\alpha, x_{2j}+\beta)$ is the number of points such that $x+\alpha \le x_i+\alpha$ and $y+\beta \le y_j+\beta$, that is equivalent to $x_i \le x$ and $y_j \le y$. Thus, $m(x_{1i}+\alpha, x_{2j}+\beta) = m(x_{1i}, x_{2j})$, and

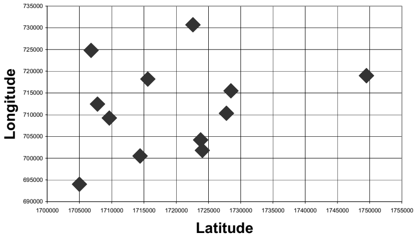
$F_n(x_{1i}+\alpha, x_{2j}+\beta) = F_n(x_{1i}, x_{2j})$. Q.E.D.

The dependence on the orientation of axes is presented below.

Location of average DAC Statistic across samples

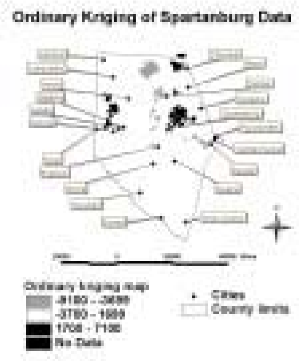


Location of MAX DAC for rotations of data with different angles



Simulations indicated that the location of the Max DAC statistic is not unique, moreover there is a geometrical locus of it, and this varies as the orientation of the axes changes.

Nevertheless, the question remains whether these properties affect the ability of the DAC statistic to detect spatial clusters. Since our concerns are rather theoretical, the DAC statistic remains an useful instrument for its purpose, especially when used in conjunction with kriging or other spatial prediction techniques. The image on the right side illustrates the detection of low birth clusters in Spartanburg County, SC, using 1988-1990 birth certificate data consisting of 6434 observations.



The Survivorship Function

Recall that: $F(x, y) = P(X \le x$ and $Y \le y)$, whereas $S(x, y) = P(X > x$ and $Y > y)$, while its compliment is $P(X \le x$ or $Y \le y)$:

$P(X \le x$ or $Y \le y) = P(X \le x) + P(Y \le y) - P(X \le x$ and $Y \le y) = F(x) + F(y) - F(x, y)$

$S(x, y)$ is on an intersection of $\{X > x\}$ and $\{Y > y\}$ with **strict** inequalities while $F(x, y)$ is on $\{X \le x\}$ and $\{Y \le y\}$ with non-strict inequalities. Since $F(x, \infty) = F(x)$ and $F(\infty, y) = F(y)$,

$S(x, y) + [F(x) - F(x, y)] + [F(y) - F(x, y)] + F(x, y) = 1$, and

$S(x, y) = 1 - F(x) - F(y) + F(x, y)$

Implications

The statistical methodology built around the survivorship function has grown considerably, and there are important theory pieces for discrete and continuous data as well. The instantaneous log-odds ratio (ILOR) proved to be a remarkable instrument in the detection of spatial clusters. However, the DAC statistic represents a new addition. Our research concerning the ability of the DAC statistic to detect spatial clusters and to it sensitivity to the orientation of axes could be extended to the survivorship function, given the mathematical developments associated with the latter.