To cite this paper / Pentru a cita lucrarea:
Petrisor AI (2000), Empirical Spatial Distributions, Master thesis, Department of Epidemiology and Biostatistics, School of Public Health, University of South Carolina, Columbia, SC, USA, 52 pp.

EMPIRICAL SPATIAL DISTRIBUTIONS AND THE DAC STATISTIC

by

Alexandru-Ionuț Petrişor

Bachelor of Science

University of Bucharest, 1997

Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Science in Public Health in the

Department of Epidemiology and Biostatistics

School of Public Health

University of South Carolina

2000

Department of Epidemiology
and Biostatistics
Director of Thesis

Department of Epidemiology
and Biostatistics
Second Reader

Department of Health Promotion
and Education
Third Reader

Dean of the Graduate School

CONTENTS

ACKNOWLEDGEMENTS

This thesis would have been impossible without the help, guidance and assistance from my thesis director, Prof. J. Wanzer Drane. I want to express my deepest appreciation for his permanent encouragement, and patience, as long as for his whole support during my program of study.

Special thanks are also due to Prof. Kirby Jackson for his help with the SAS® programming, and to Prof. Robert Valois for serving on my thesis committee and providing valuable suggestions.

I want to thank Prof. Liviu Dragomirescu for his assistance with programming and continued support during my program of study.

I remain grateful to all the friends I made during the studies for their support. Among them, I would reserve a special place for Col. John Means Harden III.

Last but not least, I am thankful for the caring love and support of my family, friends, professors and colleagues at the University of Bucharest.

This paper is dedicated to my mother and to the memory of my godfather.

ABSTRACT

Empirical distributions are not strangers to the biostatistician, but binary spatial distributions constructed from random locations indexed by longitude and latitude might be. This study uses the longitude and latitude as the (X_1, X_2) coordinates of the homes of mothers in Spartanburg, SC who gave birth to their babies in either 1989 or 1990. Mathematically, the above coordinates have an arbitrary origin as well as arbitrary orientations of their axes. What difference does it make to change either the location of the origin or the orientation of axes? This question is herein addressed. Since empirical and theoretical cumulative distributions are unaffected in shape by translations, this study addresses the effects of those distributions by reorienting the axes.

The DAC statistic is the difference between the distribution of cases and that of population at a particular point (x_1, x_2), and its maximum value was chosen as the measure of effect of different orientation of axes. For any size of a random sample of locations taken from those of the 6434 plus live births there is a noticeable variation of the location of MaxDAC with rotations from 30 through 360 degrees within a given sample, when transformed back to original longitude and latitude. The orientations were incremented at intervals of 30 degrees. Furthermore, a simulation exercise 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. Therefore, MaxDAC might

not be the measure of choice when using empirical spatial distributions. Obviously, more and deeper investigations are the order of the day.

CHAPTER 1

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 departments 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 present more importance than others, 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). 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).

In a study of the potential clusters of brain cancer cases in Rowan County, NC, the use of programs in CLUSTER did not support a generalized increase for brain cancer, but there were some peculiar geographic and temporal patterns. The authors recommend the selection of strategic, rare health events as sentinels of potential exposure to hazardous substances in the environment (Britton, 1995). In a study using 202 patients diagnosed with Burkitt's lymphoma in the West Nile District of Uganda between 1961-1975, the authors concluded that the patients involved in temporal clusters were significantly older than other patients were. They also found significant changes in incidence in different counties, which could be explained as case-ascertainment artifacts (Williams, 1978). Due to the fact that boundaries of a single cluster are often ill defined, and the geographic,

occupational or iatrogenic dimension is rarely obvious. In some cases using cluster tests to routinely scan disease registries to look for clusters is undesirable, and a better approach may be to be vigilant for unusual environmental exposures, and to evaluate their possible impact (Smith, 1993). In Ecology, case-based models incorporate the small-scale variability found in every ecological system, as well as the rules of mechanisms that govern interactions among individuals, allowing integration across many different levels in the traditional hierarchical organization of ecological processes (Huston, 1988).

An approach that has been applied is to assume that in a fairly large population the exact distribution of the final outcome of infected individuals is asymptotically multivariate normal and to derive a locally most powerful test for some parametric models (Britton, 1995).

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 look at the empirical cumulative distributions. The probability is accumulated basically on a southwest - northeast direction. It is well known that the origin of the axes will not affect the distribution of cases other than shifting the data points (i.e., adding or subtracting the same value from the coordinates). 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, 1997).

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.

2. BIVARIATE CUMULATIVE DISTRIBUTIONS AND SURVIVORSHIP FUNCTIONS

"*But since all our measurements and observations are nothing more than approximations to the truth, the same must be true of all calculations resting upon them, and the highest aim of all computations made concerning concrete phenomena must be to approximate as nearly as practicable, to the truth.*" (Gauss, 1963)

This quotation, lasting since 1809, synthesizes some of the principles that constitute the foundation of statistics. Theory is not part of the real world, but motivation for theory is. Once theory is in place, data can be used to mime theory. In that regard, there must be properties in the theoretical realm shared between them. That when constructing empirical functions, some of their properties, insofar as possible must also be properties of the theoretical counterparts.

Following are some statistical definitions and properties of the tools that are used for the purposes of this thesis.

Definition 1. **X** *is called a* **continuous random variable** *if its cumulative distribution function can be represented as:*

$$F_X(x_1, x_2) = P[X_1 \le x_1,\ X_2 \le x_2] = \int_{-\infty}^{x_2}\int_{-\infty}^{x_1} f_X(y_1, y_2)\,dy_1\,dy_2,$$

for $-\infty \le x_i \le \infty$ $(i = 1,2)$, where

$f_X(x_1, x_2)$ *is a* **probability density function**, *for which, that is, by definition, the following is true*:

$$\begin{cases} f_X(y_1, y_2) \geq 0, \text{ for all } y_1, y_2 \in (-\infty, \infty), \text{ and} \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_X(y_1, y_2)\, dy_1\, dy_2 = 1. \end{cases}$$

In this case $F_X(x_1,x_2)$ *is called a* **continuous** *distribution function* (Dudewicz, 1976).

Definition 2. *A* **bivariate cumulative distribution function** *is any function with the following properties* (Mendenhall, 1973):

(1) $F(-\infty, x_2) = \lim_{x_1 \to -\infty} F(x_1, x_2) = 0$ *for all* x_2

$F(x_1, -\infty) = \lim_{x_2 \to -\infty} F(x_1, x_2) = 0$ *for all* x_1

$F(-\infty, -\infty) = \lim_{x_1, x_2 \to -\infty} F(x_1, x_2) = 0$

(2) *If* $x_{1_1} < x_{1_2}$ *and* $x_{2_1} < x_{2_2}$, *then*

$$P[x_{1_1} < X_1 \leq x_{1_2}, x_{2_1} < X_2 \leq x_{2_2}] = F(x_{1_2}, x_{2_2}) - F(x_{1_2}, x_{2_1}) - F(x_{1_1}, x_{2_2}) + F(x_{1_1}, x_{2_1}) \geq 0$$

This is the *monotonicity property of sorts*, and it is not equivalent to:

$F(x_{1_1}, x_{2_1}) \leq F(x_{1_2}, x_{2_2})$ *for* $x_{1_1} \leq x_{1_2}$ *and* $x_{2_1} \leq x_{2_2}$)

(3) $F(x_1, x_2)$ *is continuous in each argument, i.e.*

$$\lim_{0<h\to 0} F(x_1 + h, x_2) = \lim_{0<h\to 0} F(x_1, x_2 + h) = F(x_1, x_2)$$

(Mood, 1974; and Wilks, 1946)

Theorem 1. *If* $\mathbf{X} = (X_1, X_2)$ *is a continuous bidimensional random variable, then*:

(1) $\lim_{\min(x_1, x_2) \to +\infty} F_{X_1, X_2}(x_1, x_2) = 1$

(2) *For each i (i = 1, 2)* $\lim_{x_i \to -\infty} F_{X_1,X_2}(x_1, x_2) = 0$

(3) $F_X(x_1, x_2)$ *is continuous from above in each argument* (Dudewicz, 1976).

The proof of this theorem is beyond the purposes of this thesis.

Definition 3. *If* $F_{X_1,X_2}(.,.)$ *is the joint cumulative distribution function of* x_1 *and* x_2*, then the cumulative distribution functions* $F_{x_1}(.)$ *and* $F_{x_2}(.)$ *are called* **marginal cumulative distribution functions** (Dudewicz, 1976) *and calculated as*:

$$F_{(x_1,+\infty)} = \int_{-\infty}^{x_1}\int_{-\infty}^{\infty} f_{x_1,x_2}\, dx_2\, dx_1 \text{ and } F_{(+\infty,x_2)} = \int_{-\infty}^{x_2}\int_{-\infty}^{\infty} f_{x_1,x_2}\, dx_1 dx_2 .$$

(Wilks, 1946).

Given that: $dF(x_1, x_2) = f(x_1, x_2)\, dx_1\, dx_2$, $dF(x_2) = f(x_2)\, dx_2$, and

$dF(x_1|x_2) = \frac{f(x_1, x_2)}{f(x_2)} dx_1$ (Mendenhall, 1973; and Wilks, 1962), we proved that the marginal distribution function is a distribution function, with the same properties as in the univariate case (Freund, 1980; Mendenhall, 1973; and Wilks, 1962):

1. $\lim_{x \to 0^+} F(x) = 0$

2. If $b > a$, then $F(b) \geq F(a)$

3. $F(\infty) = 1$

Remark 1. *Knowledge of the joint cumulative distribution function of* x_1 *and* x_2 *implies knowledge of the two marginal cumulative distribution functions* (Mood, 1974):

$F_{X_1}(x_1) = F_{X_1, X_2}(x_1, \infty)$ *and* $F_{X_2}(x_2) = F_{x_1, x_2}(\infty, X_2)$

Remark 2. $F_{X_1}(x_1) + F_{X_2}(x_2) - 1 \leq F_{X_1, X_2}(x_1, x_2) \leq \sqrt{F_{X_1}(x_1) F_{X_2}(x_2)}$ *for all for all* x_1 *and* x_2 (Mood, 1974).

These properties of $F_{X_1, X_2}(x_1, x_2)$ are here stated because of their usefulness when one considers bivariate empirical distributions, which are at the foundation of the DAC statistics and this investigation. From this point forward theoretical properties shall be used by reference only as they apply to the empirical distributions and the statistics derived therefrom.

3. THE DAC STATISTIC

3.1 DEFINITION

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

1. 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 $x_{1i} \leq x_1$ and $x_{2j} \leq x_2$ [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].

2. The DAC statistic is:

$$DAC(x_1, x_2) = F_m(x_1, x_2) - F_n(x_1, x_2),$$ for all permissible values of (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). 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 because of the following theorem by Drane et al. (Drane, 1995)

Theorem 1. *Let* θ, $0 < \theta < 1$, *be the prevalence of a disease. Let* $F_{n-m}(x)$ *be the empirical distribution of non-cases. It follows that*:

$F_m - F_n = (1 - \theta) [F_m - F_{n-m}]$.

Proof. The cumulative distribution for the entire sample is expressible as a convex decomposition made up of the empirical distributions F_m and F_{n-m}, namely $F_n = \theta F_m + (1-\theta) F_{n-m}$.

Thus $F_m - F_n = F_m - [\theta F_m + (1-\theta) F_{n-m}] = (1 - \theta) [F_m - F_{n-m}]$.

Since $F_m - F_n$ and $F_m - F_{n-m}$ are everywhere proportional, it is arbitrary in the selection of $F_m - F_n$ or $F_m - F_{n-m}$ as the DAC statistic. F_m and F_n are both defined on the subset of the total sample occupied by the cases, whereas F_m and F_{n-m} could possibly have no points (X_1, X_2) in common. Therefore, it is prudential to chose DAC = $F_m - F_n$ because of the shared support.

Remarks

1. Since $F_m - F_n = (1 - \theta) [F_m - F_{n-m}]$ and both are step functions, it is sufficient to evaluate them at (x_1, x_2) of the case data.

2. $F_m(x) - F_n(x) \neq 0$ almost everywhere $0 < F_m, F_n < 1$

Not only are the isoboles different, but the contours of the equation:

$$F_m(x) - F_n(x) = \text{constant}$$

are not predictable á priori.

3. The maximum absolute value of the DAC statistic represents the Kolmogorov-Smirnov statistic for two samples. This is defined as follows (Hollander, 1973):

(1) Define the order statistics $Z_{(1)} \le Z_{(2)} \le ... \le Z_{(n)}$, as X_{1_1}, ..., X_{1_m}, and

X_{2_1}, ..., $X_{2_{n-m}}$, arranged in increasing order.

(2) Define $F_m(a) = \frac{\text{Number of } X_1\text{'s} \le a}{m}$ and $G_n(a) = \frac{\text{Number of } X_2\text{'s} \le a}{n}$

(3) Using Theorem 1, the test statistic is:

$$J_i = \frac{m(n-m)}{d} \max_{i=1,...,N} \left\{ \left| F_m\left(Z_{(i)}\right) - G_{n-m}\left(Z_{(i)}\right) \right| \right\} = \frac{m(n-m)}{d(1-\theta)} \max_{i=1,\,m} \left\{ \left| F_m\left(x_1,\ x_2\right) - F_n\left(x_1,\ x_2\right) \right| \right\},$$

where *d* is the greatest common divisor of *m* and *m-n*.

For one sample, the test statistic is:

$$J_0 = \sqrt{m} \max_{-\infty < a < \infty} \left\{ \left| F_m(a) - F_0(a) \right| \right\}$$

A two-sided Kolmogorov-Smirnov test uses the following couple of hypotheses:

H_0: $P(X \le a) \equiv P(Y \le a)$ vs. H_1: $P(X \le a)$ $P(Y \le a)$.

3.2 ISOBOLES

Isoboles are iso-probability curves in which the location of a data point is represented by a point on a graph the axes of which are the longitude (denoted by X_1) and latitude (X_2). They can be visualized using an isobologram such as the following one.

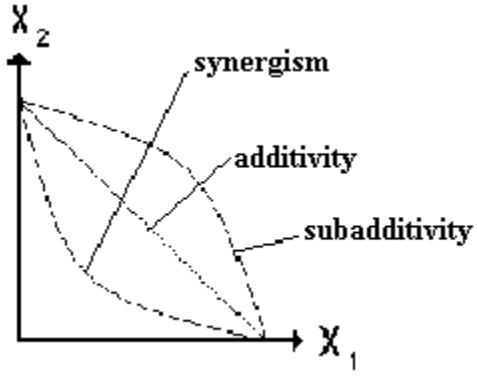

Figure 1. Isobologram

Isobolograms are a simple way of visualising interactions. Synergistic interaction is shown by a line of points curving underneath the diagonal additivity line to produce a concave-up isobole. This type of interaction indicates that the combined effect on two dimenssions is greater than the sum on each dimension individually. In contrast, a line of points curving above the zero interaction line, to produce a concave-down isobole, indicates that the two dimensions are less than additive but cannot be clasified as antagonistic. Finally, a line of points along the zero interaction line indicates that the combination is merely additive, or to put it another way, the combined effect on both

dimensions together is the sum on each one separately. The "effect" is the probability of a binary response indicated by a zero if absent and one if present.

Let H = 1 if an event occurs and zero otherwise. Then P (H = 1 | $X_1 = x_1$, $X_2 = x_2$) is the dose response function. Setting it equal to numbers between zero and one produces isoboles.

The behavior of the cumulative distribution function, F(x), for a random variable X, is given by its mathematical properties (Dudewicz, 1976; Huston, 1988; and Wilks, 1946):

1. F(0) = 0

2. F(b) = F(a) if b > a

3. F(∞) = 1

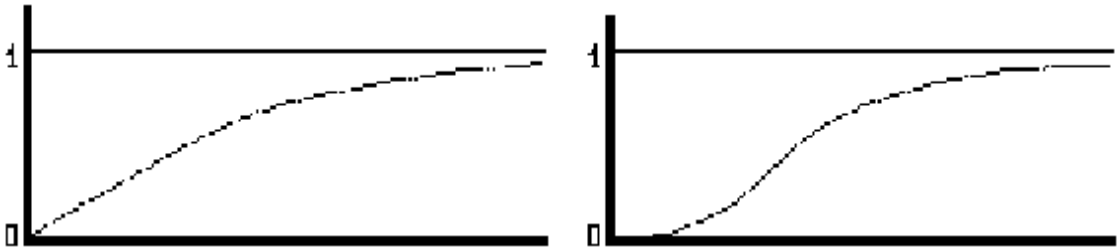

Figure 2. Prototypical Graphs of Univariate Cumulative Distribution Functions

As also covered in Chapter 2, the cumulative distribution function in two variables becomes a surface. We expand on the univariate properties by making reference to Theorem 1, that is:

1. F(0, 0) = 0

2. F(∞,∞) = 1

3. $F_n(\infty, x_2) = F_n(x_2)$, $F_n(x_1, \infty) = F_n(x_1)$

On any ray (0, ∞) through a point (x_1, x_2), for any $\delta, \eta > 0$, $F_n(x_1 + \delta, x_2) \geq F_n(x_1, x_2)$, $F_n(x_1, x_2 + \eta) \geq F_n(x_1, x_2)$, and from the properties of the bivariate cumulative,

4. $F_n(x_1 + \delta, x_2 + \eta) - F_n(x_1 + \delta, x_2) - F_n(x_1, x_2 + \eta) + F_n(x_1, x_2) \geq 0$.

3.3.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. During 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 birthweight.

The data set consisted of 6434 lines of observations, corresponding to 6434 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: an identification number (1-6434), the actual latitude and longitude, and the infant's birthweight.

The following graphic displays the geographical location of the study populations. The limits of Spartanburg County were displayed for a better visualization.

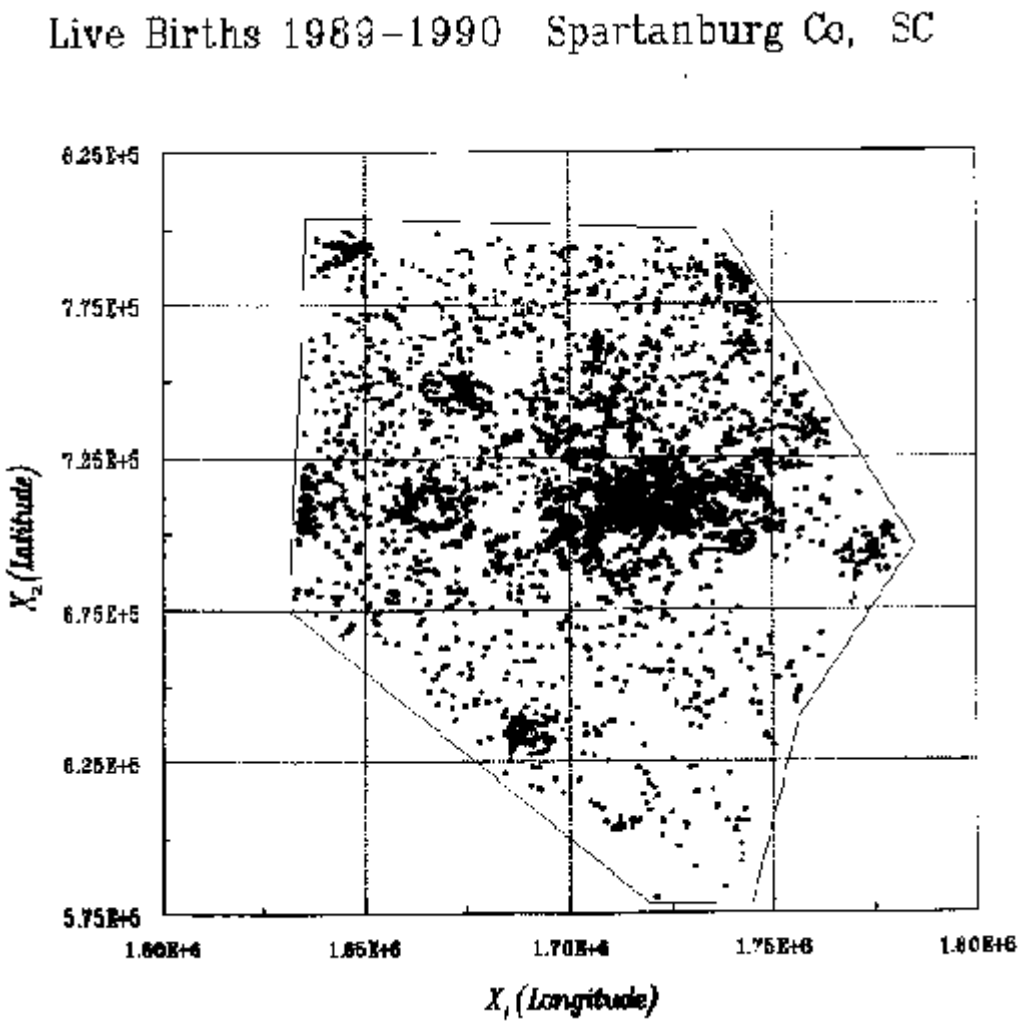

Figure 3. Study Population

The following graphic displays the geographical location of the cases. The limits of Spartanburg County were displayed for a better visualization.

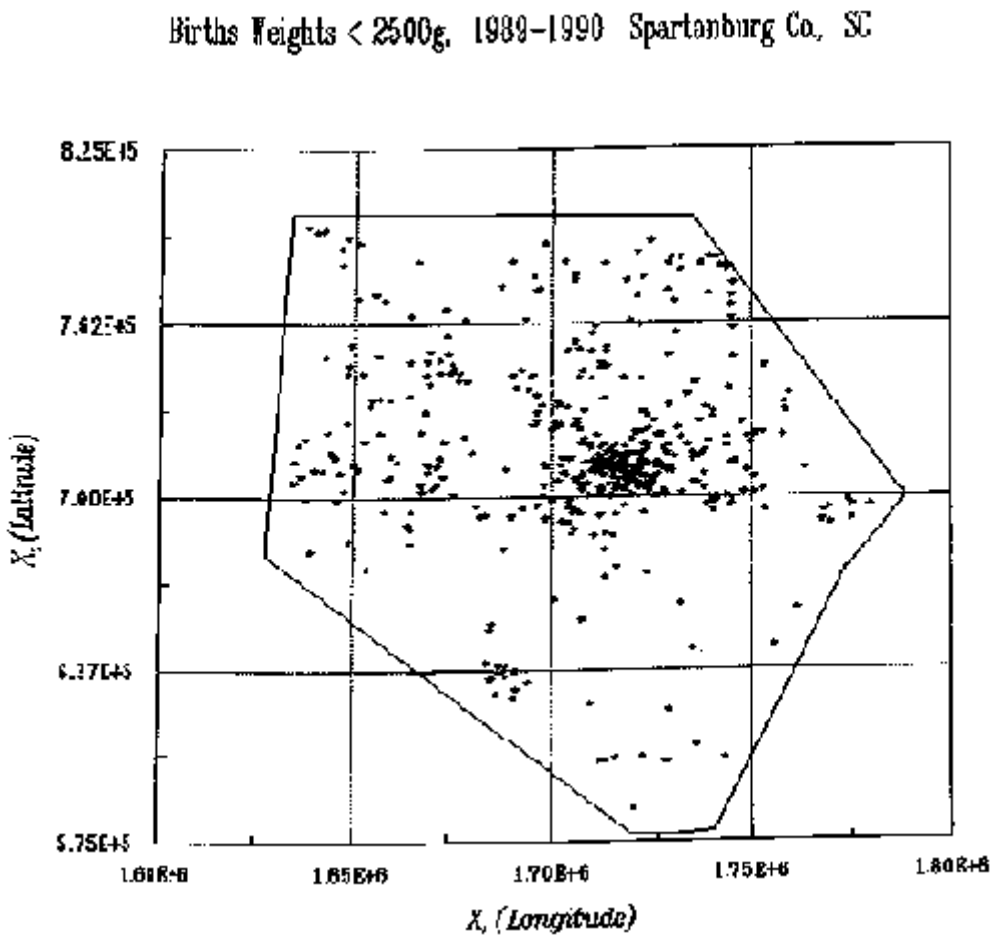

Figure 4. Cases

Spartanburg County has now a population of 247,548. In 1990, its population was 226,800. There were 3,762 live births in Spartanburg County in 1990, out of which 302 were low birth weights (SCDHEC, 1993). The map of Spartanburg County is displayed below.

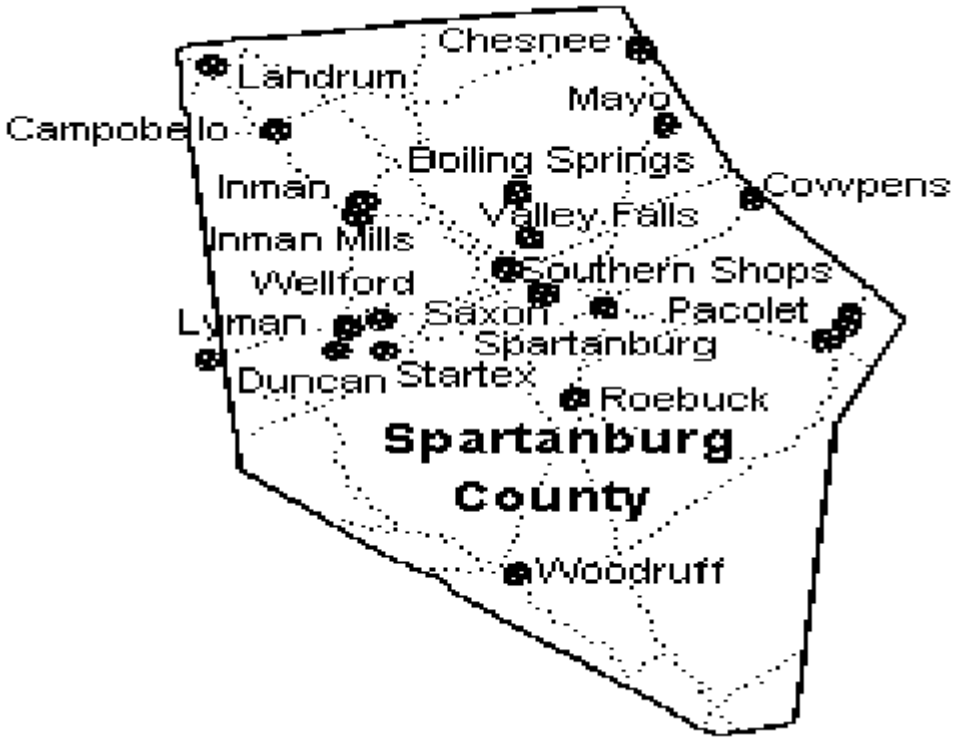

Figure 5. Map of Spartanburg County, SC

If the figure 5 and either 3 or 4 are overlaid, it can easily be noticed that areas with high density of cases and population correspond to the center of the county, more exactly to the cities Spartanburg, Saxon, Valley Falls, Boiling Springs, and Pacolet. The question remains whether cases accumulate at a greater rate than base population in these areas, i.e. there is a cluster of low birthweights.

CHAPTER 4

BIRTH DATA 1989-1990, SPARTANBURG COUNTY, SC

The data consists of all live births in Spartanburg County, South Carolina for the two years 1989 and 1990. Cases were low birthweight babies. The results of a previous study (Drane, 1995) are presented below. Low birthweights were defined as those less than or equal to 2500 grams. Even if the two distributions presented in figures 4 and 5 appear similar to the naked eye, their differences, however small, are displayed in Figure 8.

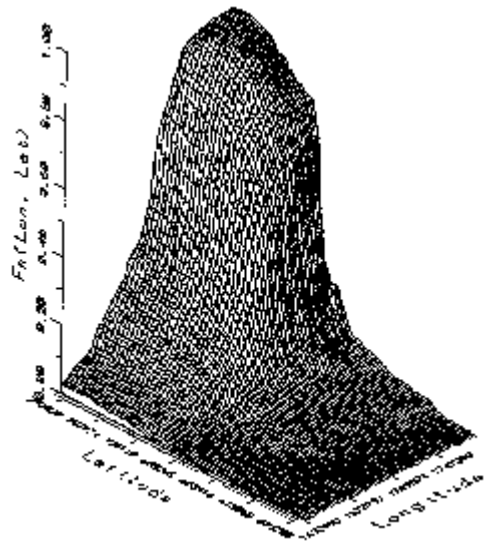

Figure 6. Empirical Distribution of Live Births, N = 6434

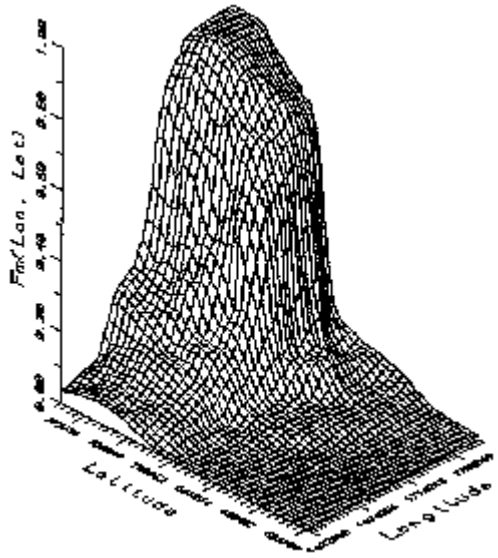

Figure 7. Empirical Distribution of Low Birth Weights, N = 591

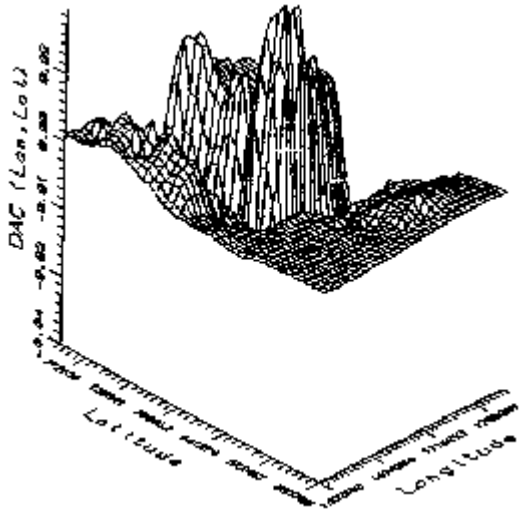

Figure 8. Empirical Distribution of DAC Statistic

When F_m grows to surpass F_n, DAC > 0 and ≤ 0 elsewhere. Visualizing the projection of DAC(x_1, x_2) onto the map of Spartanburg county, the question remains: "What are the effects of random orientation of x(X_1, X_2) on DAC(x_1, x_2)?" The object of this study is to, at least partially, answer this question.

5. RESULTS

5.1 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 magnitudes, 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.

For a more rigorous proof, consider the definition given in Drane, 1995:

$$F_n(X_1, X_2) = \frac{m(X_1, X_2)}{n},$$

where $m(x_1, x_2)$ is the number of points such that $X_1 \leq x_1$ and $X_2 \leq x_2$.

If we translate the origin, the cumulative distribution function becomes:

$$F_n(X_{1_i} + \alpha, X_{2_j} + \beta) = \frac{m(X_{1_i} + \alpha, X_{2_j} + \beta)}{n},$$

where m ($x_{1_i}+\alpha$, $x_{2_j}+\beta$) is the number of points such that $X_1+\alpha \leq x_{1_i}+\alpha$ and $X_2+\beta \leq x_{2_j}+\beta$, that is equivalent to $X_1 \leq x_{1_i}$ and $X_2 \leq x_{2_j}$. Thus, m ($x_{1_i}+\alpha$, $x_{2_j}+\beta$) = m (x_{1_i}, x_{2_j}), and $F_n(x_{1_i}+\alpha, x_{2_j}+\beta) = F_n(x_{1_i}, x_{2_j})$. Q.E.D.

5.2 RANDOM ORIENTATIONS OF AXES

For these simulations, a special program, called "FINAL.EXE", was created in Microsoft Q-Basic™. The original program is presented in Appendix 1. 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)

Due to the Quick Basic™ processor, the maximum sizes allowed by the program range from either 20 samples of size 400 or 40 samples of size 200.

In the following graphical representation, the centroids within each sample size were plotted using actual latitudes and longitudes. Sample sizes ranged from 30 to 750 in increments of 30. For each sample size, 10 samples were selected.

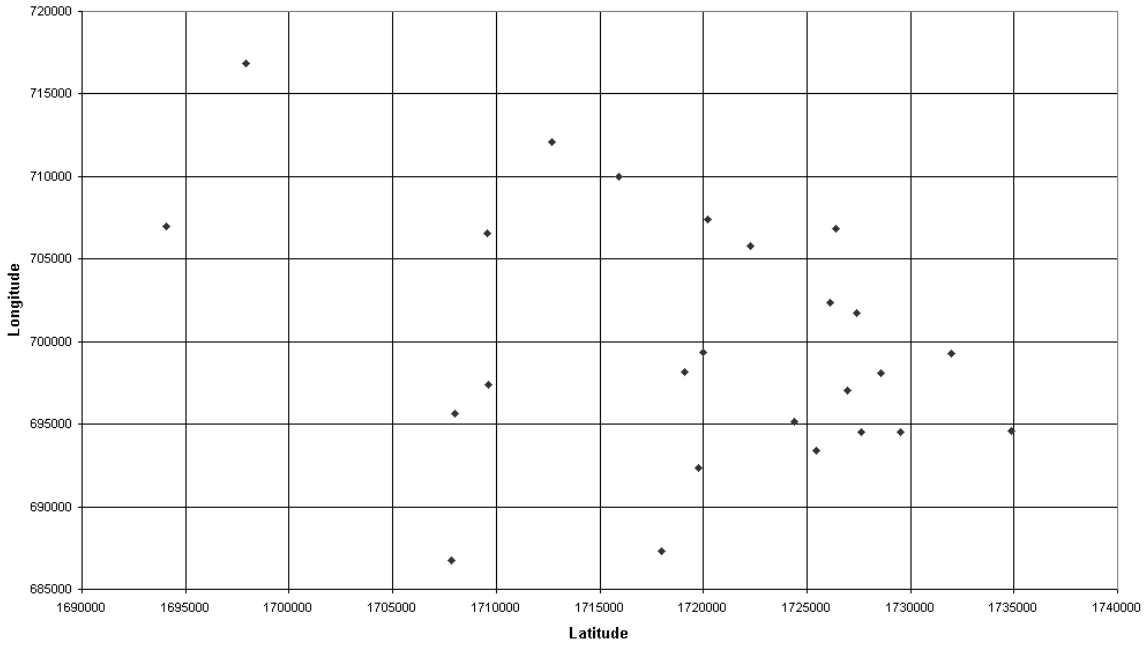

Figure 9. A Plot of Average Max DAC Location for Various Sample Sizes

In this phase, it could be noticed that the location of the maximum DAC statistic for the original and for the rotated sample differed one from each other and between the samples. Thus, the next step was to assess the dependence of this variation of the sample size.

In order to perform this analysis, the results of the first phase were used by a SAS™ application, presented in Appendix 2, in order to calculate a measure of variability between samples, and test its dependence on the sample size. This measure (referred as deviation and denoted by D) was analogous to the sample standard deviation and computed using the formula:

$$D = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2 + (y_i - \bar{y})^2}{n - 1}}$$

The results are presented graphically below.

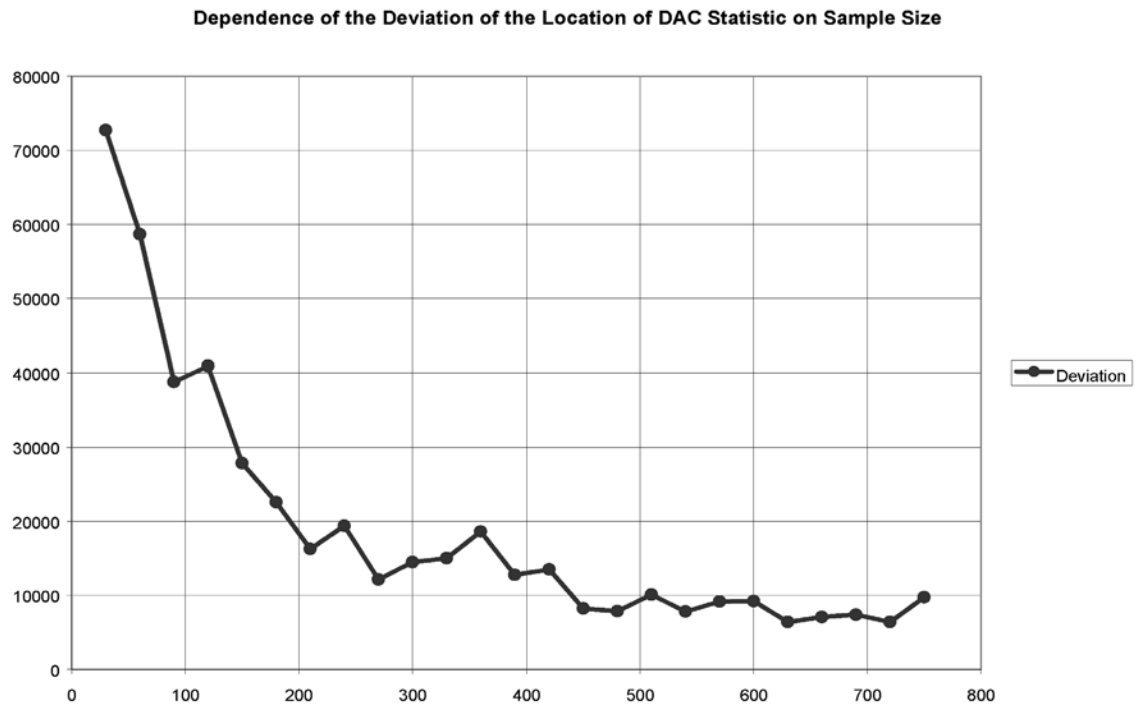

Figure 10. Dependence of the Location of the Maximum DAC Statistic on Sample Size

Even somebody without mathematical inclination can easily notice that the so-called deviation decreases with the increase of the sample size. Large variability is observed for small samples (less than 200 observations). If the sample size increases to several hundreds, the variability reduces and tends to remain stable. Even for large samples, the deviation is less than 8000 m (i.e., 8 km) or 5 miles.

Finally, the whole data set was rotated with angles between 30^0 and 360^0, in increments of 30^0. A plot of the maximum DAC statistic corresponding to each rotation is presented below.

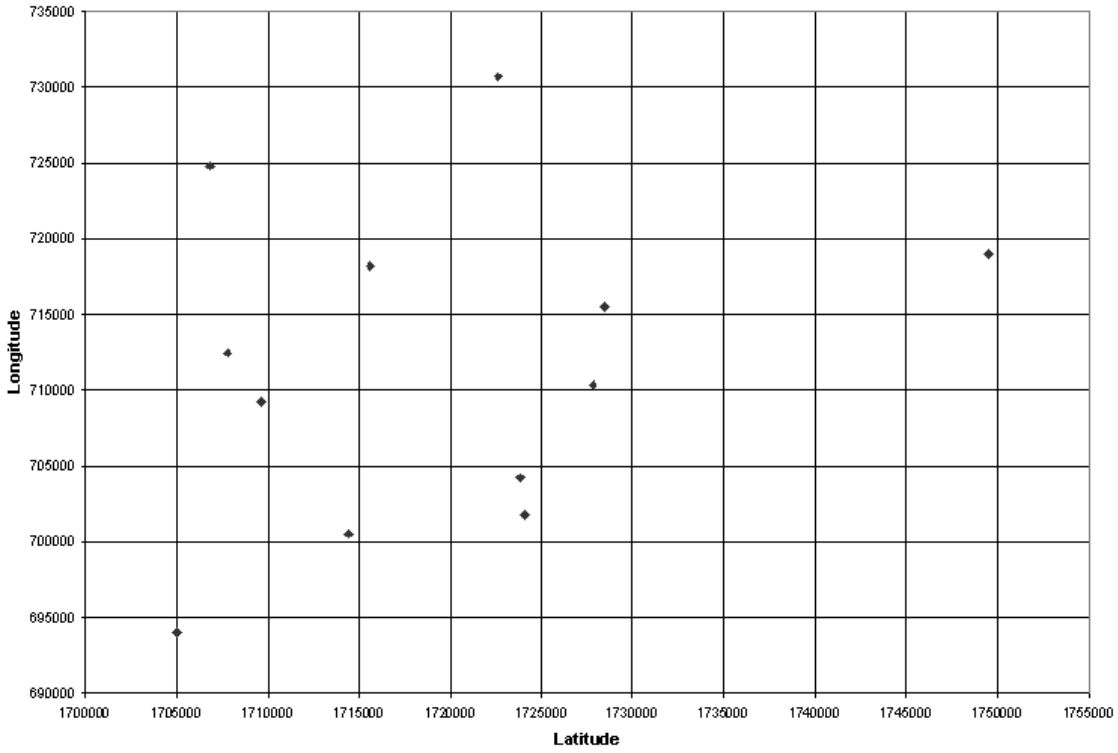

Figure 11. A Plot of Max DAC Location for 12 Rotations from 30^0 to 360^0 of the Whole Data Set

There appears even to the naked eye to be a large variation of the Max DAC statistic location for rotations of the whole data set with different angles. However, the computation of D indicated a value of 31,720.70, i.e. approximately 32 km or 20 miles.

CHAPTER 6

DISCUSSION

The results indicated that the DAC statistic does not depend on the location of the origin. The sensitivity of the DAC statistic to the orientation of axis depends on the sample size, and it decreases as the sample size increases.

The first statement is the result of a mathematical proof. In section 5.1 it was shown that the DAC statistic does not depend on the location of the origin, based on the definitions of $F_n (X_1, X_2)$ and $F_m (X_1, X_2)$.

The second statement resulted from the simulations presented in section 5.2 and is subject to several limitations. The most important limitation is related to the size of the bytes of Quick Basic®. Because of the 8 or 16 bit processor of Quick Basic® indexing is limited and acts to constrain our present efforts at large samples. It is possible that using another programming language this problem will disappear. For the moment we are limited to the product of sample size and number of samples not to exceed 8000.

The investigated problem should studied in greater depth, using better technology for the simulations in order to permit the selection of more samples and increasing the sample size such as converting the Quick Basic® program for simulations to a SAS® application.

SAS programs do not suffer the same limitations of dimension. SAS will probably permit more valid results without any other special technical requirements.

Using another programming language such as Turbo Pascal® or SAS® the problem of index limits should be overcome. This will allow sampling of arbitrary sizes giving rise to an attainable limit of D. Hopefully, that limit is zero. That is, as $n \rightarrow \infty$ Max $D \rightarrow 0$ with probability one.

POSTSCRIPT

Generally such simulation exercises as these give one the comparison from the outset that mathematical approaches are intractable. These negative results from the point of view of our expectations caused us to reflect on whether an example could be easily constructed showing the characteristics of the DAC statistic. The following is such an example.

Consider the two cumulative distributions on the unit square and the resultant DAC (x,y).

$F_1 = xy$, $F_2 = x^2y^2$ and DAC (x, y) = $F_1 - F_2$

DAC = $xy - x^2y^2 = xy(1 - xy)$, $0 \le x, y \le 1$

Southwest Accumulation.

Here accumulation is from the southwest most point (0, 0).

$F_1 = xy$, $F_2 = x^2y^2$

The limits of accumulation are:

$F_1(0, 0) = 0$; $F_1(0, 1) = 0$; $F_1(1, 0) = 0$; $F_1(1, 1) = 1$, and

$F_2(0, 0) = 0$; $F_2(0, 1) = 0$; $F_2(1, 0) = 0$; $F_2(1, 1) = 1$.

If DAC has a maximum, its first partial derivatives must simultaneously be equal to zero.

$$\frac{\partial DAC}{\partial x} = y - 2xy^2 = y(1 - 2xy)$$

$$\frac{\partial DAC}{\partial y} = x - 2x^2y = x(1 - 2xy), \text{ and}$$

$$\frac{\partial^2 DAC}{\partial x^2} = -2y^2 \le 0, \quad \frac{\partial^2 DAC}{\partial y^2} = -2x^2 \le 0, \text{ indicating a local maximum.}$$

These two partial derivatives are zero when x = y = 0 and wherever 1 - 2xy = 0 or $xy = \frac{1}{2}$ which is an rectangular hyperbola intersecting the unit square at $(1, \frac{1}{2})$ and $(\frac{1}{2}, 1)$ and passing through $(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$. Thus DAC (x, y) is maximum along the locus $xy = \frac{1}{2}$ inside the unit square and at its boundaries where they intersect, and its value is 0.25.

A plot of the DAC statistic, created using the SAS® program presented in the appendices, against x and y is presented below.

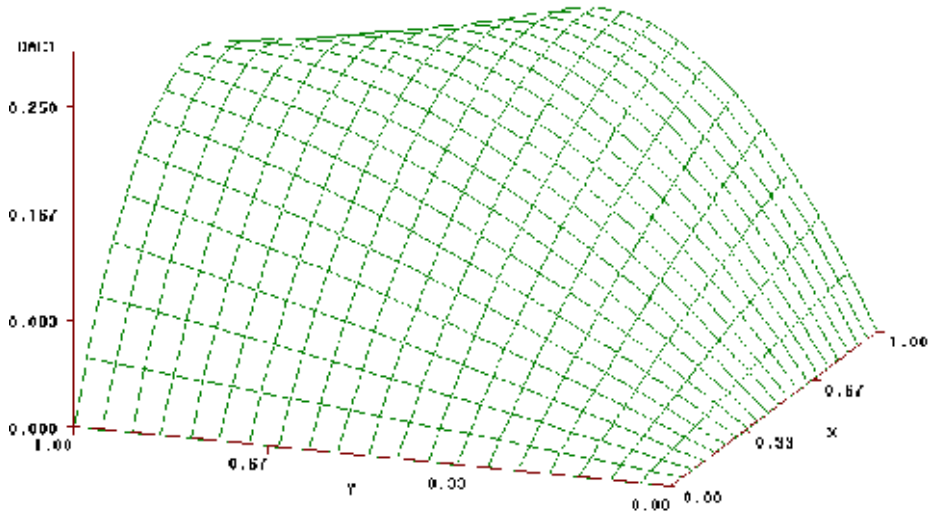

Figure 12. A Plot of the DAC Statistic against the Coordinates, for Southwest Accumulation

For any point (x, y) one has the option to accumulate points from any direction. If the conjecture that Max DAC (x, y) is invariant under the rotation of the axes, then the hyperbola just derived will be duplicated regardless of the direction from which we accumulate x and y.

From F_1 (x, y) and F_2 (x, y) we obtain:

f_1 (x, y) = 1 and f_2 (x, y) = 4xy on $0 \leq x, y \leq 1$.

Northwest Accumulation.

If accumulation is from the northwest most point (0, 1), the analytical expressions change to:

$F_1 = \int_y^1 dy \int_0^x dx = x(1-y)$ and $F_2 = \int_y^1 dy \int_0^x 4xy\, dx = x^2(1-y^2)$, and

$DAC = x\,(1-y) - x^2\,(1 - y^2)$

And the limits of accumulation become:

$F_1 = x(1 - y)$, $F_2 = x^2(1 - y^2)$

$F_1(0, 1) = 0$; $F_1(0, 0) = 0$; $F_1(1, 0) = 1$; $F_1(1, 1) = 0$

$F_2(0, 1) = 0$; $F_2(0, 0) = 0$; $F_2(1, 0) = 1$; $F_2(1, 1) = 0$

Furthermore

$\frac{\partial DAC}{\partial x} = (1-y) - 2x(1-y^2) = (1-y)\big(1-2x(1+y)\big)$, and

$\frac{\partial^2 DAC}{\partial x^2} = -2(1-y^2) \le 0$, indicating a local maximum.

This partial derivative is zero when $y = 1$ or when $1 - 2x\,(1 + y) = 0$, i.e. $x = \frac{1}{2(y+1)}$

The last is the equation of a hyperbola intersecting the unit square at $(\frac{1}{4},1)$ and passing through $(\frac{1}{2},0)$ and $(\frac{1}{3},\frac{1}{2})$. Thus DAC (x, y) is maximum along the locus $x = \frac{1}{2(y+1)}$ inside the unit square and its boundaries where they intersect.

$\frac{\partial DAC}{\partial y} = -x + 2x^2(1-y) = x\big(1-2x(1-y)\big)$, and

$\frac{\partial^2 DAC}{\partial y^2} = -2x^2 \le 0$, indicating a local maximum.

This partial derivative is zero when x = 0 or when 1 - 2x (1 - y) = 0, i.e. $x = \frac{1}{2(1-y)}$

The last is the equation of a hyperbola intersecting the unit square at $(1, \frac{1}{2})$ and passing through $(\frac{1}{2}, 0)$ and $(\frac{3}{4}, \frac{1}{3})$. Thus DAC (x, y) is maximum along the locus $x = \frac{1}{2(1-y)}$ inside the unit square and its boundaries where they intersect.

The maximum value of the DAC statistic is again 0.25.

A plot of the DAC statistic against x and y is presented below.

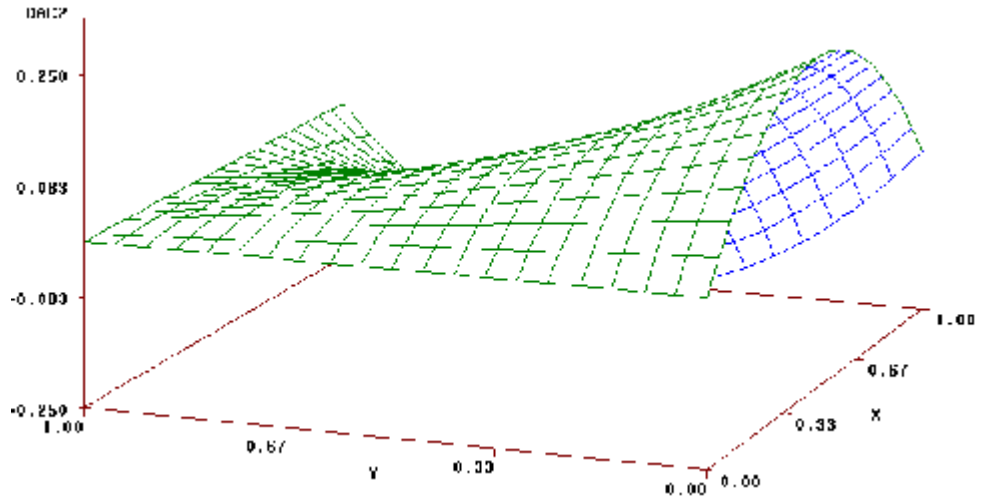

Figure 13. A Plot of the DAC Statistic against the Coordinates, for Northwest Accumulation

Northeast Accumulation.

If accumulation changes from the northeast most point (1, 1), the analytical expressions change again:

$$F_1 = \int_y^1 dy \int_x^1 dx = (1-x)(1-y) \text{ and } F_2 = \int_y^1 dy \int_x^1 4xy\, dx = (1-x^2)(1-y^2)$$

$$DAC = (1 - x)(1 - y) - (1 - x^2)(1 - y^2)$$

and the limits of accumulation become

$F_1 = (1 - x)(1 - y)$, $F_2 = (1 - x^2)(1 - y^2)$

$F_1(1, 1) = 0$; $F_1(0, 1) = 0$; $F_1(0, 0) = 1$; $F_1(1, 0) = 0$

$F_2(1, 1) = 0$; $F_2(0, 1) = 0$; $F_2(0, 0) = 1$; $F_2(1, 0) = 0$

Furthermore

$\frac{\partial DAC}{\partial x} = (y - 1) + 2x(1 - y^2) = (y - 1)\left(1 - 2x(1 + y)\right)$, and

$\frac{\partial^2 DAC}{\partial x^2} = 2(1 - y^2) \geq 0$, indicating a local minimum.

This is equal to zero when $y = 1$ or $x = \frac{1}{2(y + 1)}$, and is the equation of a hyperbola

intersecting the unit square at $(\frac{1}{4}, 1)$ and passing through $(\frac{1}{2}, 0)$ and $(\frac{1}{3}, \frac{1}{2})$. Thus DAC

(x, y) is minimum along the locus $x = \frac{1}{2(y + 1)}$ inside the unit square and its boundaries

where they intersect.

$\frac{\partial DAC}{\partial y} = (x - 1) + 2y(1 - x^2) = (x - 1)\left(1 - 2y(1 + x)\right)$, and

$\frac{\partial^2 DAC}{\partial y^2} = 2(1 - x^2) \geq 0$, indicating a local minimum.

This equals zero when $x = 1$ or when $x = \frac{1}{2y} - 1$, and is the equation of a hyperbola

intersecting the unit square at $(1, \frac{1}{4})$ and passing through $(0, \frac{1}{2})$ and $(\frac{1}{2}, \frac{1}{3})$. Thus DAC

(x, y) is minimum at $x = y = \frac{\sqrt{3} - 1}{2} = 0.3660254$.

The minimum value of the DAC statistic is in this case $\frac{9-6\sqrt{3}}{4}$ = - 0.3480762.

A plot of the DAC statistic against x and y is presented below.

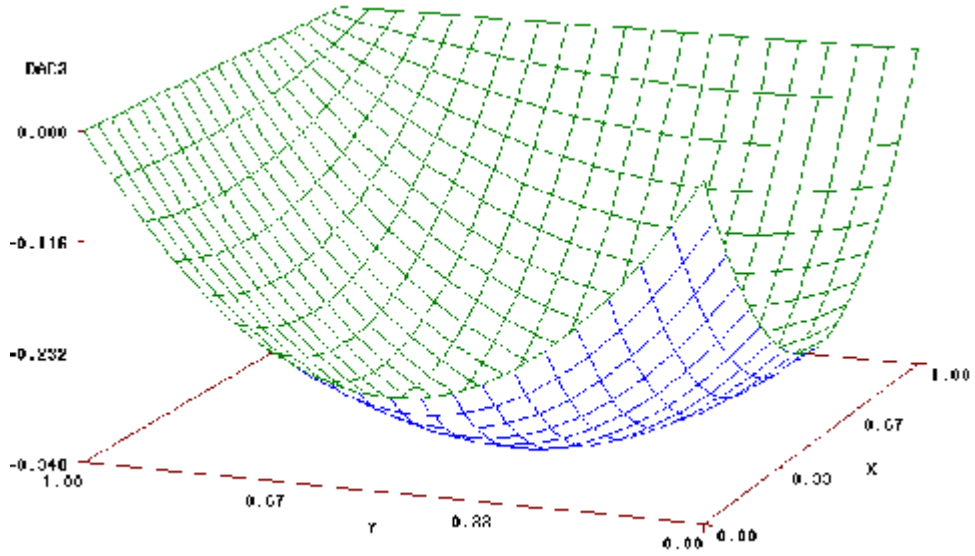

Figure 14. A Plot of the DAC Statistic against the Coordinates, for Northeast Accumulation

Southeast Accumulation.

Finally, if accumulation changes from the southeast most point (1, 0), the analytical expressions become:

$$F_1 = \int_0^y dy \int_x^1 dx = y(x-1) \text{ and } F_2 = \int_0^y dy \int_x^1 4xy\, dx = y^2(1-x^2)$$

DAC = $y(1 - x) - y^2(1 - x^2)$ and the limits of accumulation are

$F_1 = y(1 - x)$, $F_2 = y^2(1 - x^2)$

$F_1(0, 0) = 0$; $F_1(0, 1) = 1$; $F_1(1, 0) = 0$; $F_1(1, 1) = 0$

$F_2(0, 0) = 0$; $F_2(0, 1) = 1$; $F_2(1, 0) = 0$; $F_2(1, 1) = 0$

Furthermore

$\frac{\partial DAC}{\partial x} = -y + 2y^2(1-x) = y\left(-1+2y(1-x)\right)$, and

$\frac{\partial^2 DAC}{\partial x^2} = -2y^2 \le 0$, indicating a local maximum.

This equals zero when y = 0 or when $x = 1-\frac{1}{2y}$, and is the equation of a hyperbola intersecting the unit square at $(\frac{1}{2},1)$ and passing through $(0,\frac{1}{2})$ and $(\frac{1}{4},\frac{1}{3})$. Thus DAC (x, y) is maximum along the locus $x = 1-\frac{1}{2y}$ inside the unit square and its boundaries where they intersect.

$\frac{\partial DAC}{\partial y} = 1-x-2y(1-x^2) = (x-1)\left(2y(x+1)-1\right)$, and

$\frac{\partial^2 DAC}{\partial y^2} = 2x^2 \ge 0$, indicating a local minimum.

This equals zero again when x = 1 or when $x = \frac{1}{2y}-1$, and is the equation of a hyperbola intersecting the unit square at $(1,\frac{1}{4})$ and passing through $(0,\frac{1}{2})$ and $(\frac{1}{2},\frac{1}{3})$. Thus DAC (x, y) is minimum along the locus $x = \frac{1}{2y}-1$ inside the unit square and its boundaries where they intersect.

The maximum value of the DAC statistic is again 0.25.

A plot of the DAC statistic against x and y is presented below.

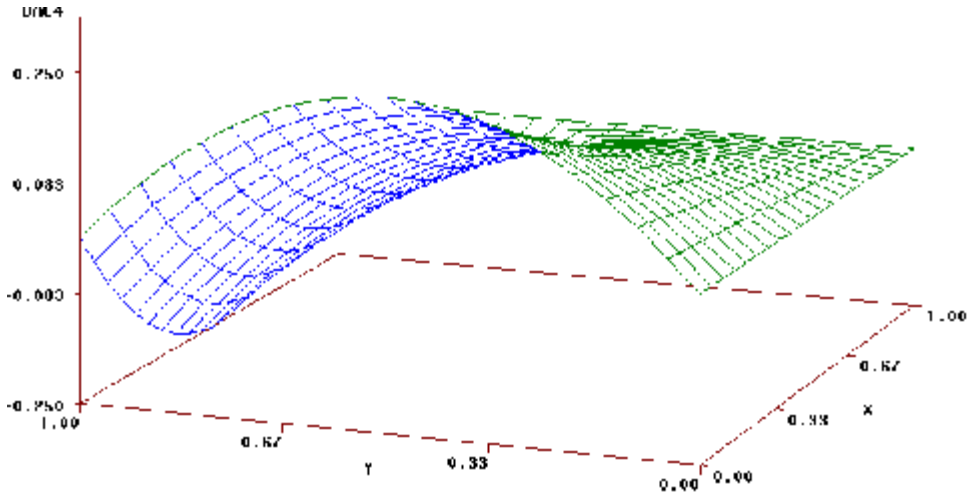

Figure 15. A Plot of the DAC Statistic against the Coordinates, for Southeast Accumulation

As this exercise indicated, 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. At this point, another limitation of this study may become evident. As in the initial approach we were looking for a unique Max DAC statistic, the number of samples was not a determining factor in assessing the dependence of DAC statistic on the orientation of axes. Given this new theoretical support, the importance of this issue must be underlined. It is likely and expected that a simulation using a large number of samples would have suggested that, in fact, we are dealing with a locus and not a unique value. Instead, large sample sizes were used to the detriment of our cause, to discover the truth. This may indicate a possible direction of study, i.e. performing the same simulations, but selecting more samples of an appropriate sample size. Obviously, the improvements already suggested, referring to the usage of some more favorable software, such as Fortran®, Turbo Pascal® or SAS® remain also valid.

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

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

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

APPENDICES

1. Following is the Q-Basic® code main program used for the simulations. The results were presented in Chapter 5.2.

```
CLS : PRINT : PRINT : PRINT : PRINT : PRINT
PRINT "                    Final.EXE" : PRINT : PRINT : PRINT
PRINT "          This program computes the DAC statistic." : PRINT : PRINT
PRINT : PRINT "                    Copyright 2000" : PRINT : PRINT
PRINT "     Alexandru I. Petrisor, BS, University of South Carolina"
PRINT "      J. Wanzer Drane, PhD, PE, University of South Carolina, and"
PRINT "      Liviu Dragomirescu, PhD, University of Bucharest" : PRINT : PRINT
PRINT : PRINT "                    Press ENTER to continue."
n$ = "c:\temp\inp-data.txt" : n2$ = "c:\temp\out-data.txt"
INPUT a$: CLS
OPEN n$ FOR INPUT AS #1
n = 0
DO UNTIL EOF(1)
INPUT #1, id, x, y, bw : n = n + 1
LOOP
CLOSE #1
```

```
OPEN n$ FOR INPUT AS #2
DIM id(n): DIM x(n): DIM y(n): DIM bw(n)
FOR i = 1 TO n
INPUT #2, id(i), x(i), y(i), bw(i)
NEXT i
CLOSE #2
OPEN n$ FOR INPUT AS #3 : OPEN n2$ FOR OUTPUT AS #4
teta = RND * 2 * 3.141593 : PRINT : PRINT : PRINT
PRINT "                    The number of samples is: ": INPUT nos
PRINT : PRINT : PRINT : PRINT "                         The sample size is: ": INPUT nobs
CLS : DIM sid(nos, nobs): DIM sx(nos, nobs): DIM sy(nos, nobs) : DIM sxr(nos, nobs):
DIM syr(nos, nobs): DIM sbw(nos, nobs): DIM ncs(nos) : DIM Ffnxy(nos, nobs)
DIM Ffmxy(nos, nobs): DIM DAC(nos, nobs) : DIM Ffnxyr(nos, nobs)
DIM Ffmxyr(nos, nobs): DIM DACr(nos, nobs) : DIM na(nos, nobs)
DIM MaxDAC(nos): DIM XMax(nos): DIM YMax(nos) : DIM MaxDACr(nos)
DIM XMaxr(nos): DIM YMaxr(nos)
FOR i = 1 TO nos
FOR j = 1 TO nobs
na(i, j) = INT(RND * n) + 1
FOR z = 1 TO n
IF na(i, j) = id(z) THEN sid(i, j) = id(z): sx(i, j) = x(z): sy(i, j) = y(z): sbw(i, j) = bw(z)
NEXT z
IF sbw(i, j) <= 2500 THEN ncs(i) = ncs(i) + 1
```

```
NEXT j

NEXT i

FOR i = 1 TO nos

FOR j = 1 TO nobs

FOR k = 1 TO nobs

IF sx(i, j) < sx(i, k) THEN GOTO 102

102 SWAP sid(i, j), sid(i, k) : SWAP sx(i, j), sx(i, k) : SWAP sy(i, j), sy(i, k)

SWAP sbw(i, j), sbw(i, k)

NEXT k

NEXT j

NEXT i

FOR i = 1 TO nos

FOR j = 1 TO nobs

FOR l = 1 TO nobs

IF sx(i, l) = sx(i, j) AND sy(i, l) < sy(i, j) THEN GOTO 103

103 SWAP sid(i, l), sid(i, j) : SWAP sx(i, l), sx(i, j) : SWAP sy(i, l), sy(i, j)

SWAP sbw(i, l), sbw(i, j)

NEXT l

NEXT j

NEXT i

FOR i = 1 TO nos

FOR j = 1 TO nobs

FOR n = 1 TO nobs
```

```
IF sy(i, n) >= sy(i, j) AND sx(i, n) >= sx(i, j) THEN Ffnxy(i, n) = Ffnxy(i, n) + 1: GOTO
20
20 REM "Calcul Fm(X, Y)"
IF sbw(i, n) <= 2500 AND sbw(i, j) <= 2500 AND sy(i, n) >= sy(i, j) AND sx(i, n) >=
sx(i, j) THEN Ffmxy(i, n) = Ffmxy(i, n) + 1
NEXT n
NEXT j
NEXT i
FOR i = 1 TO nos
FOR j = 1 TO nobs
IF ncs(i) <> 0 THEN GOTO 23 ELSE GOTO 30
23 Ffnxy(i, j) = Ffnxy(i, j) / nobs : Ffmxy(i, j) = Ffmxy(i, j) / ncs(i)
24 IF sbw(i, j) <= 2500 THEN GOTO 25 ELSE GOTO 30
25 DAC(i, j) = Ffmxy(i, j) - Ffnxy(i, j): GOTO 33
30 DAC(i, j) = 0
33 CLS
NEXT j
NEXT i
FOR i = 1 TO nos
MaxDAC(i) = DAC(i, 1)
FOR j = 1 TO nobs
IF DAC(i, j) >= MaxDAC(i) THEN GOTO 3365 ELSE GOTO 3368
3365 MaxDAC(i) = DAC(i, j) : XMax(i) = sx(i, j) : YMax(i) = sy(i, j)
```

```
3368 CLS
NEXT j
NEXT i
FOR i = 1 TO nos
FOR j = 1 TO nobs
sxr(i, j) = sx(i, j) * COS(teta) + sy(i, j) * SIN(teta)
syr(i, j) = sy(i, j) * COS(teta) - sx(i, j) * SIN(teta)
NEXT j
NEXT i
FOR i = 1 TO nos
FOR j = 1 TO nobs
FOR k = 1 TO nobs
IF sxr(i, j) < sxr(i, k) THEN GOTO 100
100 SWAP sid(i, j), sid(i, k) : SWAP sx(i, j), sx(i, k) : SWAP sxr(i, j), sxr(i, k)
SWAP sy(i, j), sy(i, k) : SWAP syr(i, j), syr(i, k) : SWAP sbw(i, j), sbw(i, k)
NEXT k
NEXT j
NEXT i
FOR i = 1 TO nos
FOR j = 1 TO nobs
FOR k = 1 TO nobs
IF sxr(i, j) = sxr(i, k) AND syr(i, j) < syr(i, k) THEN GOTO 101
101 SWAP sid(i, j), sid(i, k) : SWAP sx(i, j), sx(i, k) : SWAP sxr(i, j), sxr(i, k)
```

```
SWAP sy(i, j), sy(i, k) : SWAP syr(i, j), syr(i, k) : SWAP sbw(i, j), sbw(i, k)
NEXT k
NEXT j
NEXT i
FOR i = 1 TO nos
FOR j = 1 TO nobs
FOR m = 1 TO nobs
IF syr(i, j) >= syr(i, m) AND sxr(i, j) >= sxr(i, m) THEN Ffnxyr(i, j) = Ffnxyr(i, j) + 1:
GOTO 40
40 CLS
IF sbw(i, j) <= 2500 AND sbw(i, m) <= 2500 AND syr(i, j) >= syr(i, m) AND sxr(i, j) >=
sxr(i, m) THEN Ffmxyr(i, j) = Ffmxyr(i, j) + 1
NEXT m
NEXT j
NEXT i
FOR i = 1 TO nos
FOR j = 1 TO nobs
IF ncs(i) <> 0 THEN GOTO 2223 ELSE GOTO 2230
2223 Ffnxyr(i, j) = Ffnxyr(i, j) / nobs : Ffmxyr(i, j) = Ffmxyr(i, j) / ncs(i)
2224 IF sbw(i, j) <= 2500 THEN GOTO 2225 ELSE GOTO 2230
2225 DACr(i, j) = Ffmxyr(i, j) - Ffnxyr(i, j): GOTO 2233
2230 DACr(i, j) = 0
2233 CLS
```

```
NEXT j
NEXT i
FOR i = 1 TO nos
MaxDACr(i) = DACr(i, 1)
FOR j = 1 TO nobs
IF DACr(i, j) >= MaxDACr(i) THEN GOTO 65 ELSE GOTO 68
65 MaxDACr(i) = DACr(i, j) : XMaxr(i) = sx(i, j) : YMaxr(i) = sy(i, j)
68 CLS
NEXT j
NEXT i
CLOSE #3
FOR i = 1 TO nos
PRINT #4, MaxDAC(i); ","; XMax(i); ","; YMax(i); ","; MaxDACr(i); ","; XMaxr(i); ",";
YMaxr(i)
NEXT i
CLOSE #4
PRINT : PRINT : PRINT : PRINT "                    Press ENTER to end the program.":
INPUT b$: CLS : PRINT : PRINT : PRINT : PRINT : PRINT : PRINT : PRINT : PRINT
PRINT : PRINT "                    Thank you for using Final.EXE."
PRINT : PRINT : PRINT : PRINT : PRINT : PRINT : PRINT : PRINT : PRINT
PRINT "                    Press ENTER to exit the program.": INPUT b1$: CLS
END
```

2. Following is the Q-Basic® code main program used to rotate the whole data set. The results were presented in Chapter 5.2.

```
CLS : PRINT : PRINT : PRINT : PRINT : PRINT
PRINT "                    Rotation.EXE" : PRINT : PRINT : PRINT
PRINT "This program rotates a data set and computes the DAC statistic."
PRINT : PRINT : PRINT : PRINT "                    Copyright 1999"
PRINT : PRINT : PRINT "      Alexandru I. Petrisor, BS, University of South
Carolina"
PRINT "      J. Wanzer Drane, PhD, PE, University of South Carolina"
PRINT "      and Liviu Dragomirescu, PhD, University of Bucharest"
PRINT : PRINT : PRINT: PRINT "                    Press ENTER to continue."
n$ = "c:\temp\inp-data.txt" : n2$ = "c:\temp\rot-data.txt" : INPUT a$: CLS
1011 OPEN n$ FOR INPUT AS #1
CLS : PRINT : PRINT : PRINT : PRINT : PRINT : n = 0
DO UNTIL EOF(1)
INPUT #1, id, x, y, bw
n = n + 1
IF bw <= 2500 THEN nc = nc + 1
LOOP
CLOSE #1
OPEN n$ FOR INPUT AS #2
DIM id(n): DIM x(n): DIM y(n): DIM bw(n): DIM xr(n): DIM yr(n)
DIM Ffnxyr(n): DIM Ffmxyr(n): DIM DACr(n)
```

```
FOR i = 1 TO n
INPUT #2, id(i), x(i), y(i), bw(i)
NEXT i
CLOSE #2
OPEN n$ FOR INPUT AS #3
OPEN n2$ FOR OUTPUT AS #4
nobs = n : CLS : PRINT : PRINT : PRINT : PRINT : PRINT
PRINT "        Enter the rotation angle, in multiples of 30 degrees (1-12):"
INPUT i: CLS : teta = (i / 12) * 2 * 3.141593
FOR j = 1 TO nobs
xr(j) = x(j) * COS(teta) + y(j) * SIN(teta)
yr(j) = y(j) * COS(teta) - x(j) * SIN(teta)
NEXT j
FOR j = 1 TO nobs
FOR k = 1 TO nobs
IF xr(j) < xr(k) THEN GOTO 100
100 SWAP id(j), id(k) : SWAP x(j), x(k) : SWAP xr(j), xr(k) : SWAP y(j), y(k)
SWAP yr(j), yr(k) : SWAP bw(j), bw(k)
NEXT k
NEXT j
FOR j = 1 TO nobs
FOR k = 1 TO nobs
IF xr(j) = xr(k) AND yr(j) < yr(k) THEN GOTO 101
```

```
101 SWAP id(j), id(k) : SWAP x(j), x(k) : SWAP xr(j), xr(k) : SWAP y(j), y(k)
SWAP yr(j), yr(k) : SWAP bw(j), bw(k)
NEXT k
NEXT j
FOR j = 1 TO nobs
FOR m = 1 TO nobs
IF yr(j) >= yr(m) AND xr(j) >= xr(m) THEN Ffnxyr(j) = Ffnxyr(j) + 1: GOTO 40
40 CLS
IF bw(j) <= 2500 AND bw(m) <= 2500 AND yr(j) >= yr(m) AND xr(j) >= xr(m) THEN
Ffmxyr(j) = Ffmxyr(j) + 1
NEXT m
NEXT j
FOR j = 1 TO nobs
IF nc <> 0 THEN GOTO 2223 ELSE GOTO 2230
2223 Ffnxyr(j) = Ffnxyr(j) / nobs : Ffmxyr(j) = Ffmxyr(j) / nc
2224 IF bw(j) <= 2500 THEN GOTO 2225 ELSE GOTO 2230
2225 DACr(j) = Ffmxyr(j) - Ffnxyr(j): GOTO 2233
2230 DACr(j) = 0
2233 CLS
NEXT j
MaxDACr = DACr(1)
FOR j = 1 TO nobs
IF DACr(j) >= MaxDACr THEN GOTO 65 ELSE GOTO 68
```

```
65 MaxDACr = DACr(j) : XMaxr = x(j) : YMaxr = y(j) : Fffnxyr = Ffnxyr(j)
Fffmxyr = Ffmxyr(j)
68 CLS
NEXT j
CLOSE #3
CLS
PRINT #4, MaxDACr; ","; XMaxr; ","; ; YMaxr; ","; Fffnxyr; ","; Fffmxyr
CLOSE #4
PRINT : PRINT : PRINT : PRINT "                    Press ENTER to continue."
INPUT t$ : CLS : PRINT : PRINT : PRINT : PRINT : PRINT : PRINT
PRINT "                    Press ENTER to end the program.": INPUT b$: CLS
PRINT : PRINT : PRINT : PRINT : PRINT : PRINT : PRINT : PRINT : PRINT
PRINT "                    Thank you for using Rotation.EXE."
PRINT : PRINT : PRINT : PRINT : PRINT : PRINT : PRINT : PRINT : PRINT
PRINT "                    Press ENTER to exit the program.": INPUT b1$: CLS
END
```

3. Following is the SAS codes used to analyze the results produced by the previous program. The results were presented in Chapter 5.2.

a) To determine mean coordinates of the DAC statistic, after the rotation of the samples, within a set of samples of different sizes selected in the same number:

```
data one;
input n mdac xmdac ymdac mdacr xmdacr ymdacr @@;
cards;
(data obtained using the previous program (1) are placed here)
;
proc sort;
by n;
proc means;
var xmdacr ymdacr;
by n;
run;
```

b) To compute the deviation (D), as defined in Chapter 5.2

```
data one;
input n mdac xmdac ymdac mdacr xmdacr ymdacr @@;
if n=30  then avx=...[1];
else if n=60  then avx=...[1];
.
.
```

```
.
if n=30  then avy=...[1];
else if n=60  then avy=...[1];
.
.
.
d=sqrt(((xmdacr-avx)**2+(ymdacr-avy)**2)/(n-1));
cards;
(data obtained using the previous program (1) are placed here)
;
proc sort;
by n;
proc means sum;
var d;
by n;
run;
```

[1] Actual values are yielded by the first program (2 a)

4. Following is the SAS® code I used to produce the figures presented in the Postscript.

```
data one;
input x0 y0;
do i=0 to 20;
do j=0 to 20;
x=.05*i; y=.05*j;
dac1=x*y*(1-x*y);
dac2=x*(1-y)-x**2*(1-y**2);
dac3=(x-1)*(y-1)-(1-x**2)*(1-y**2);
dac4=y*(1-x)-y**2*(1-x**2);
output; end; end;
cards;
0
0
;
proc print;
proc univariate plot; var dac1 dac2 dac3 dac4;
proc g3d;
plot y*x=dac1; plot y*x=dac2; plot y*x=dac3; plot y*x=dac4;
run;
```