CEAI, Vol. 4, No. 1, pp. 55-64, 2002 Printed in Romania

USING A QUICK-BASIC APPLICATION BASED ON THE DAC STATISTIC TO DETECT SPATIAL CLUSTERS

Alexandru I. PETRISOR[1], Liviu DRAGOMIRESCU[2], J. Wanzer DRANE [3], Kirby L. JACKSON[3] and David J. COWEN[4]

1 - Department of Environmental Health Sciences, Norman J. Arnold School of Public Health, University of South Carolina

2 - Department of Systems Ecology, Faculty of Biology, University of Bucharest

3 - Department of Epidemiology and Biostatistics, Norman J. Arnold School of Public Health, University of South Carolina

4 - Department of Geography, College of Liberal Arts, University of South Carolina

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. 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. A priori, clusters are expected in areas of high population densities, especially considering risk factors for low birthweight. The chosen axes are east-west (x) and north-south (y). From a mathematical perspective the choice is arbitrary. 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 its usefulness as a set of spatial descriptive statistic is not diminished in the least.*

Keywords: *Spatial statistics; DAC statistic; cluster; birth certificates; geocoding; low birth weight*

1. INTRODUCTION

Every diagnosis has both location and date associated with it. Location could be a home or place of employment according to expected etiology. In public health, location and time are used conjointly to detect space by time disease clusters [1-3, 14, 15] to increase the efficiency of health department's activity [3], or just to study the spatial pattern of a population dispersed over a continuous surface [9]. In Ecology, the same analyses are used to generate individual-based models [8]. Application may be even more comprehensive, expanding to include spatially indexed socio-economic, biology, or geographical data in general.

Different studies have indicated various approaches to space-time analyses over wide and expanding venues of applications. One approach is to work on disease risk from environmental hazards at three levels: distributional patterns relative to the locations of hazards, sentinel events in time and place, and case cluster strategies [2]. The DAC statistic is part of analyses of distributions that will be presented later on. The analysis of sentinel events recognizes that some events are more important than others when used to attract attention, and case-cluster strategies permit the identification of disease clusters without initially suggesting. 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 are also used in space-time analysis [9]. Models of attenuations of point sources were created to assess their effects on the surrounding population [5]. The Ederer-Myers-Mantel procedure is used to detect temporal clustering of events [14]. That procedure uses 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 [3]. 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 [13]. 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 [9]. In this investigation, the approach is to examine the empirical cumulative distributions of cases and non-cases. Probability is accumulated in a southwest - northeast direction arbitrarily placing all locations in the first quadrant. In order to suggest geographic clusters, 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 [2]. It is well known that the origin of the axes will not affect the distribution of cases other than shifting the measures of location [10]. The location of the maximum DAC statistic is not unique, moreover there can be a geometrical locus of it, and that locus varies as the orientation of the axes changes [10-12].

The purpose of this paper is to investigate whether the DAC statistic is a reliable instrument to detect spatial clusters. Layers of spatially indexed data will be projected over and onto a base map of the county from which the data were collected.

2. MATERIALS AND METHODS

2.1. *Birth data 1998-1990, Spartanburg County*

The data came from a demonstration project sponsored by the Robert Woods Johnson Foundation [11, 12]. The object of the effort was to demonstrate the usefulness of geographically coded health events. The one legal document that has a great promise of nearly a 100 percent response rate is 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. Altogether 6434 records were created. Out of these, 591 were cases. Cases were low birthweight babies. Low birthweights were defined as those less than or equal to 2500 grams. In the present study only longitude, latitude and the infant's birthweight were used [11, 12]. In 1990 the population of Spartanburg County was 226,800. There were 3,762 live births in Spartanburg County in 1990, out of which 302 were low birth weights [16].

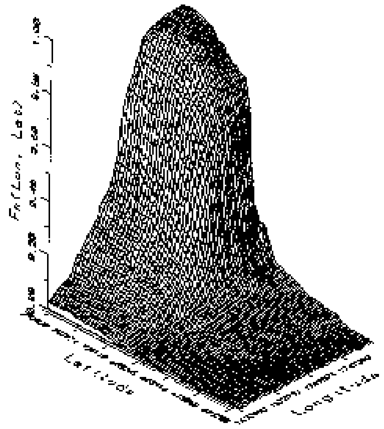


Fig. 1. Empirical Distribution of Live Births, N = 6434.

The results of a previous study [6] are presented below. Low birthweights were defined as those less than or equal to 2500 grams. Even if the two distributions presented in figures 1 and 2 appear similar to the naked eye, their differences, however small, are displayed in Figure 3. The graphs displayed in the following were produced using a Turbo-Pascal® application.

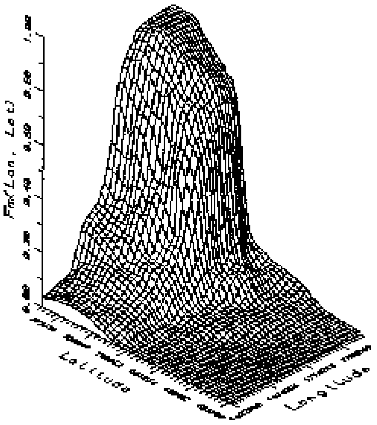


Fig. 2. Empirical Distribution of Low Birth Weights, N = 591.

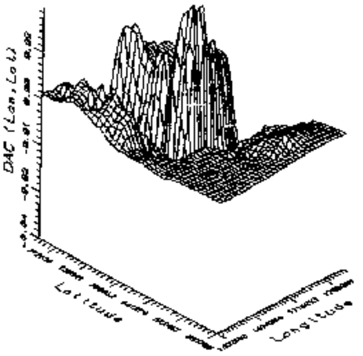


Fig. 3. Empirical Distribution of DAC Statistic.

When F_m grows to surpass F_n, $DAC>0$ and ≤ 0 elsewhere.

2.2. *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 [6]. 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 computation of the DAC statistic is difference between two empirical cumulative distribution functions. The empirical cumulative distribution function:

$F_n(x_1,x_2)=m(x_1, x_2)/n$, 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 $[0, n]$.

The DAC statistic spans the interval $[0, 1]$. 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 [6]. Since within the sample of size n there are m cases and n-m non-cases.

$$F_n(.)=(m/n)xF_m(.)+[(n-m)/n]xF_{n-m}(.)$$

$$DAC=[(n-m)/n]x[F_m(.)-F_{n-m}(.)]$$

Therefore, F_{n-m} may be substituted for F_n. The maximum absolute value of the DAC statistic represents the Kolmogorov-Smirnov statistic for two dimensions [7].

2.3. *Geographical information systems*

A GIS is a decision support system involving the integration of spatially referenced data in a problem-solving environment [4]. ArcView GIS© was used to convert the file produced in the previous step in a layer of information added to the other layers of geographic information about Spartanburg County, SC, its city limits and locations of other cities. Geographic information is accessible from the Department of Geography at the University of South Carolina, Columbia, SC (found on the Internet at: http://www.cla.sc.edu/gis/dataindex.html.

Special attention has to be paid to the coordinate systems. The original coordinates of the data were in state plane feet, whereas Spartanburg County geographic information were in meters, using the 1983 International coordinate systems. Coordinate transformations were performed using the projection utility extension associated with ArcView GIS©.

3. RESULTS AND DISCUSSION

Our research involved several steps. Initially, simulations were performed to investigate the sensitivity of the maximum DAC statistic to the location of origin and orientation of axes [[10]].

3.1. *Random locations of origin*

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

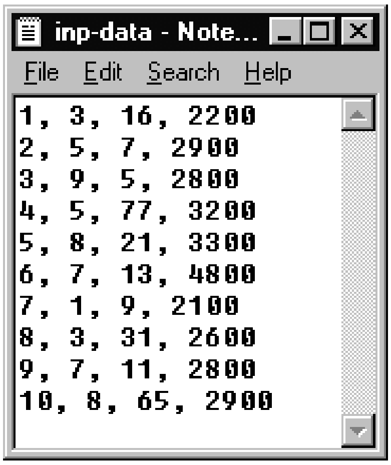
$T(x_1, x_2)=(x_1+\alpha, x_2+\beta)$ for all (x_1, x_2),

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.

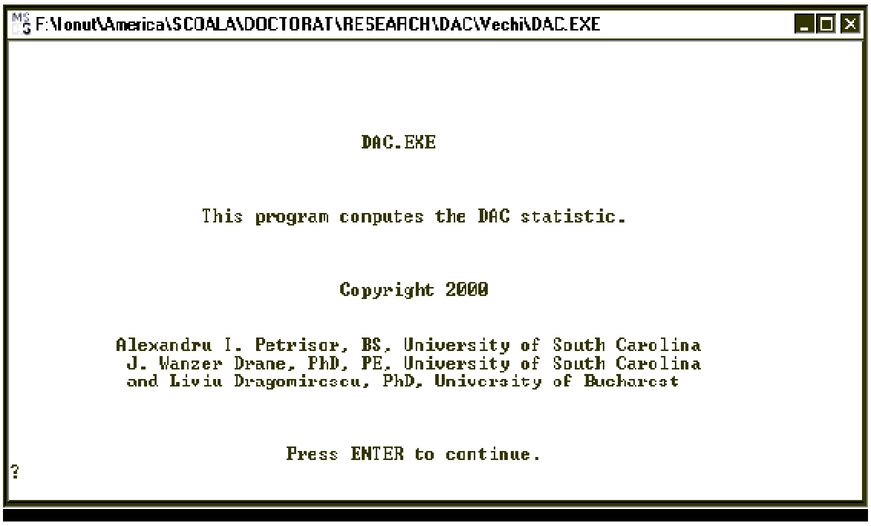
3.2. *Random orientations of axes*

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 <u>*inp-data.txt*</u> (Figure 1), prompts for the number of samples to be selected and for the size of each sample. The interface is simple, as presented in Figures 4 and 5.



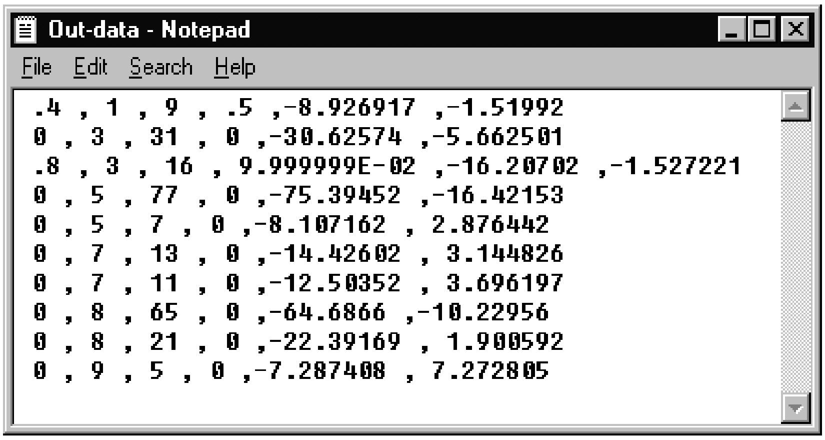
```
1, 3, 16, 2200
2, 5, 7, 2900
3, 9, 5, 2800
4, 5, 77, 3200
5, 8, 21, 3300
6, 7, 13, 4800
7, 1, 9, 2100
8, 3, 31, 2600
9, 7, 11, 2800
10, 8, 65, 2900
```

Fig. 4. Input Data.



```
                         DAC.EXE


              This program conputes the DAC statistic.


                      Copyright 2000


      Alexandru I. Petrisor, BS, University of South Carolina
       J. Wanzer Drane, PhD, PE, University of South Carolina
       and Liviu Dragomirescu, PhD, University of Bucharest


                   Press ENTER to continue.
?
```

Fig. 5. First Screen of the Simulation Program: Introduction to the Program.



```
.4 , 1 , 9 , .5 ,-8.926917 ,-1.51992
0 , 3 , 31 , 0 ,-30.62574 ,-5.662501
.8 , 3 , 16 , 9.999999E-02 ,-16.20702 ,-1.527221
0 , 5 , 77 , 0 ,-75.39452 ,-16.42153
0 , 5 , 7 , 0 ,-8.107162 , 2.876442
0 , 7 , 13 , 0 ,-14.42602 , 3.144826
0 , 7 , 11 , 0 ,-12.50352 , 3.696197
0 , 8 , 65 , 0 ,-64.6866 ,-10.22956
0 , 8 , 21 , 0 ,-22.39169 , 1.900592
0 , 9 , 5 , 0 ,-7.287408 , 7.272805
```

Fig. 6. Output Data.

The application produces an output file in the same format called out-data.txt, containing as many lines as the number of samples indicates (Figure 6). 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) [11].

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. Observations are selected based on the equality of the index with the number randomly generated. 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 7 and 8. The graphs were produced using SAS®.

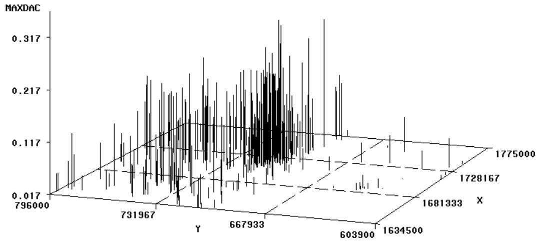


Fig. 7. Location of the Maximum DAC Statistic for 400 Random Samples.

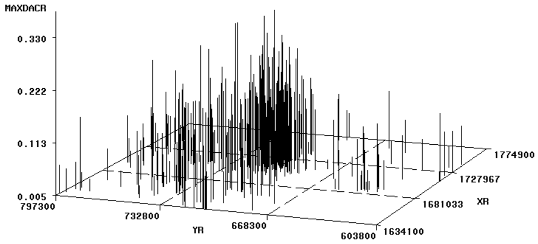


Fig. 8. Location of the Maximum DAC Statistic for 400 Random Samples Rotated with Random Angles.

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 9 and 10 using a Turbo-Pascal® plotting application.

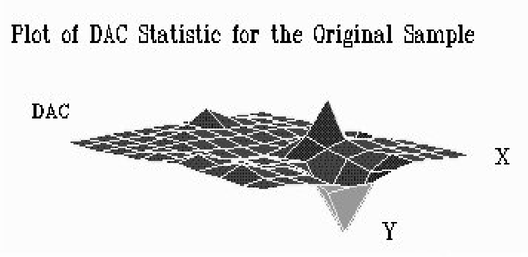


Fig. 9. Location of the Maximum DAC Statistic for the Original Data.

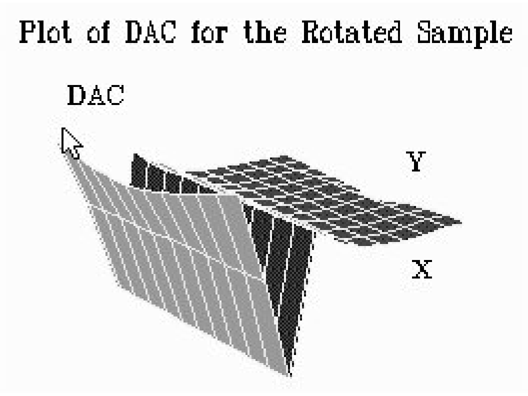


Fig. 10. Location of the Maximum DAC Statistic for the Rotated Data.

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.

At this stage, the question remained whether the DAC statistic is a reliable instrument to detect spatial clusters [[11]]. A new application was created to use the DAC statistic with the Spartanburg data to detect clusters of low birthweight [[11], [12]]. The program read the initial data in comma-delimited format from an input file, prompted for the weight limit for normal births, and produced an output file in the same format, containing as many lines as the number of observations indicated. Each line contained, in order, the location (latitude and longitude) and the value of the DAC statistic, as well as the values of the cumulative distributions for the cases and for the entire sample. The results were used to create the map displayed in Figure 11 using ArcView GIS©. This figure presents a chloropleth map of the positive values of the DAC statistic in Spartanburg County, SC. The shading intensity is directly proportional to the density of positive values in the area. Cities are displayed as black dots. It may be easily noticed that the peaks of the DAC statistic concentrate around the cities. Positive values occur in the northeast part of Spartanburg and around Cowpens, Chesnee, Landrum, Campobello, and Inman. The highest values can be found around Spartanburg.

Figure 12 is a three-dimensional representation of the positive values of the DAC statistic in Spartanburg County, SC, in relationship to the position of the cities. The height of each peak and the shading intensity are directly proportional to the density of positive values in the area. Cities are displayed as black dots. The area of the county appears as a semitransparent gray shape.

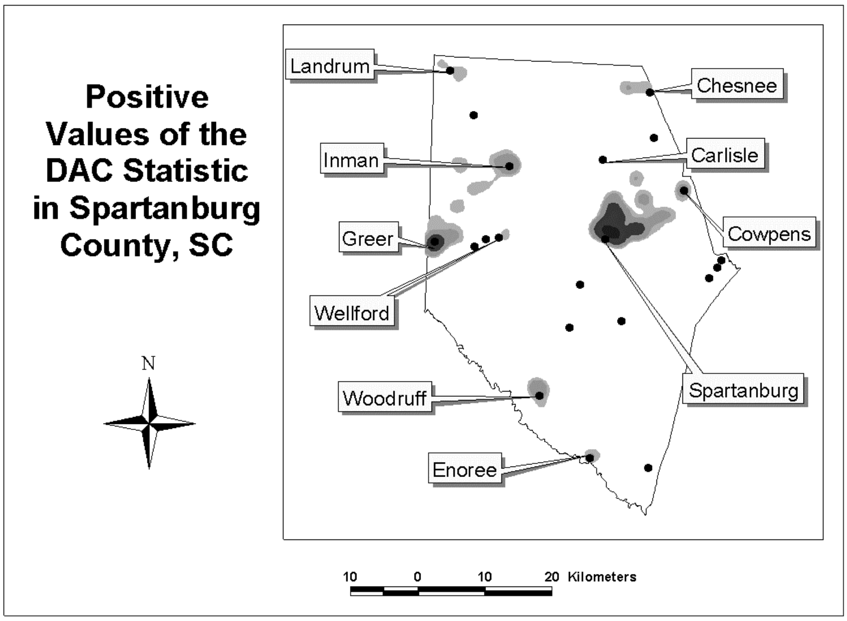


Fig. 11. Map of the positive DAC statistic values in Spartanburg County, SC in relationship with the position of the main cities. The shading intensity is directly proportional to the density of positive values in the area. Main cities are displayed as black dots.

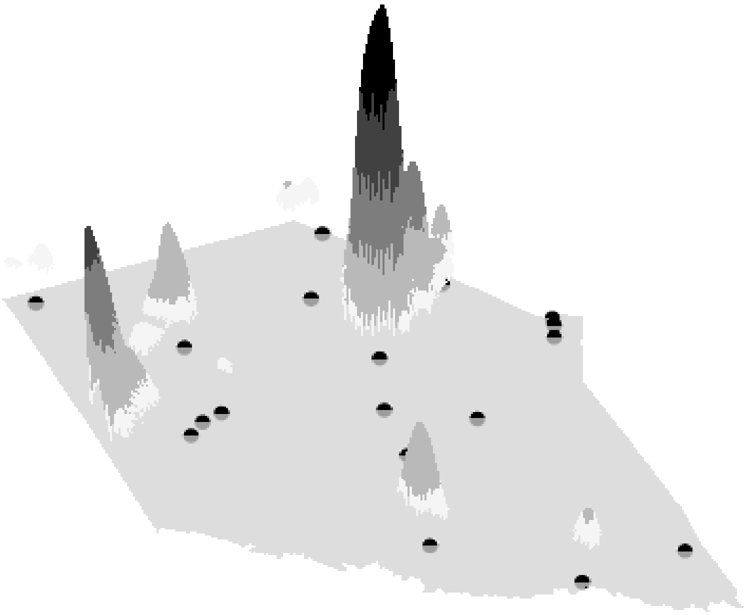


Fig. 12. Three-dimensional representation of the positive DAC statistic values in Spartanburg County, SC in relationship to the position of the main cities. The height of each peak and the shading intensity is directly proportional to the density of positive values in the area. Main cities are displayed as black full dots. The area of the county appears as a semitransparent gray shape.

It is expectable to find more DAC values around the large cities, and even more expectable for the peaks of the DAC statistic, to occur around these places. Our results show that maximum values tend to occur mostly in the Northwestern part of the county. This may be an indication of clustering. Furthermore, the peaks detected around cities, especially the larger ones, Spartanburg and Greer, may indicate problems in these areas. Epidemiological studies conducted in these areas might explain the causes of these clusters. Obviously, more and deeper investigations are the order of the day. The main limitation to the usage of this application for simulation purposes was represented by the size of the Quick Basic® processor. This limited sampling to a formula where the product between the number of samples to be selected and the sample size could not exceed 8000, therefore the program is not suitable for large data sets. To overcome this limitation, other applications should be developed using different software.

2.4. *Generalization*

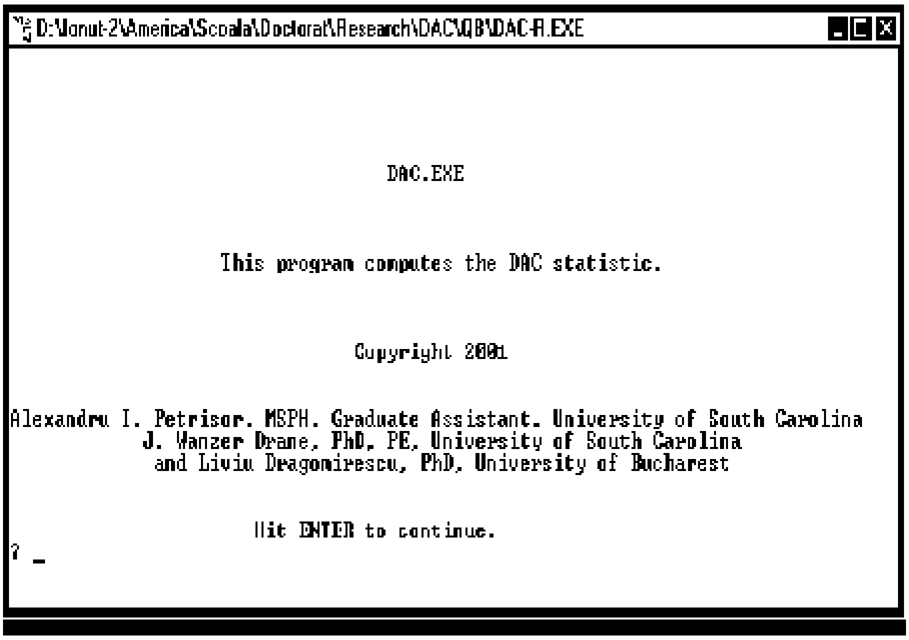
After the Medical Infobahn for Europe Special Topic Conference "Healthcare Telematics Support in Transition Countries", held in Bucharest, Romania, in June 2001, several epidemiologists expressed their interest in the program. As a result, a version of this program is available in English or Romanian free of charge for academic and research purposes. A copy of the program may be sent via e-mail upon request by the first author (Alexandru Petrisor, e-mail: aipetri@mailbox.sc.edu). The final version of this program allows the user specify the input file and the discrimination limit (in our study, the birth weight at which a child is considered normal). The application provides also explanations related to the statistical theory behind the program, as well as copyright issues. The interface is presented below in figures 13-16.

The final result is an output file called "RESULTS.txt", containing the coordinates, and the values of the DAC statistic and the cumulative distribution function for the whole sample and for cases only evaluated at each location.

4. CONCLUSIONS

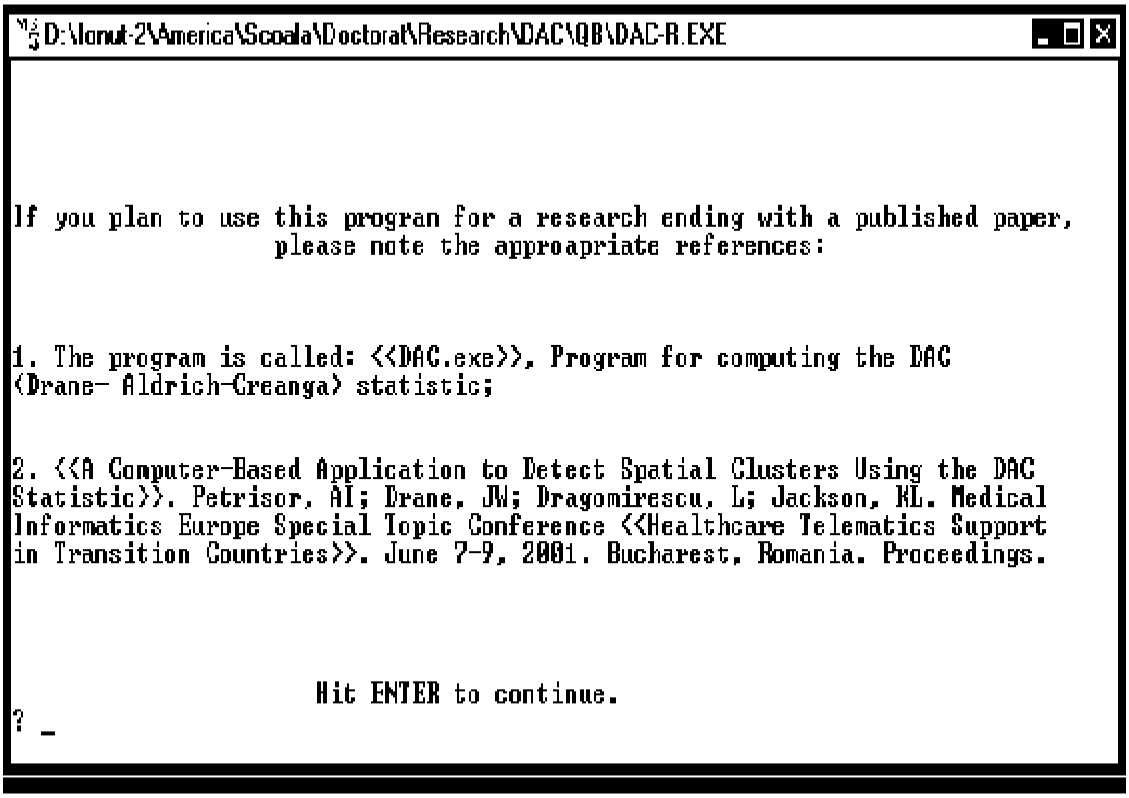
The results of the simulations indicated that the DAC statistic does not depend on the location of the origin. However, the dependence on the orientation of axes has an analytical expression that may not be easily detected. In real life example, the maximum DAC statistic does not have necessarily an analytical expression, therefore it is almost impossible to find its geometrical locus.

In our example, the DAC statistic appeared to be a possible instrument to detect spatial or temporal clusters. The question remains whether the reliability of this instrument expands over other real life examples. Obviously, more and deeper research are the order of the day.



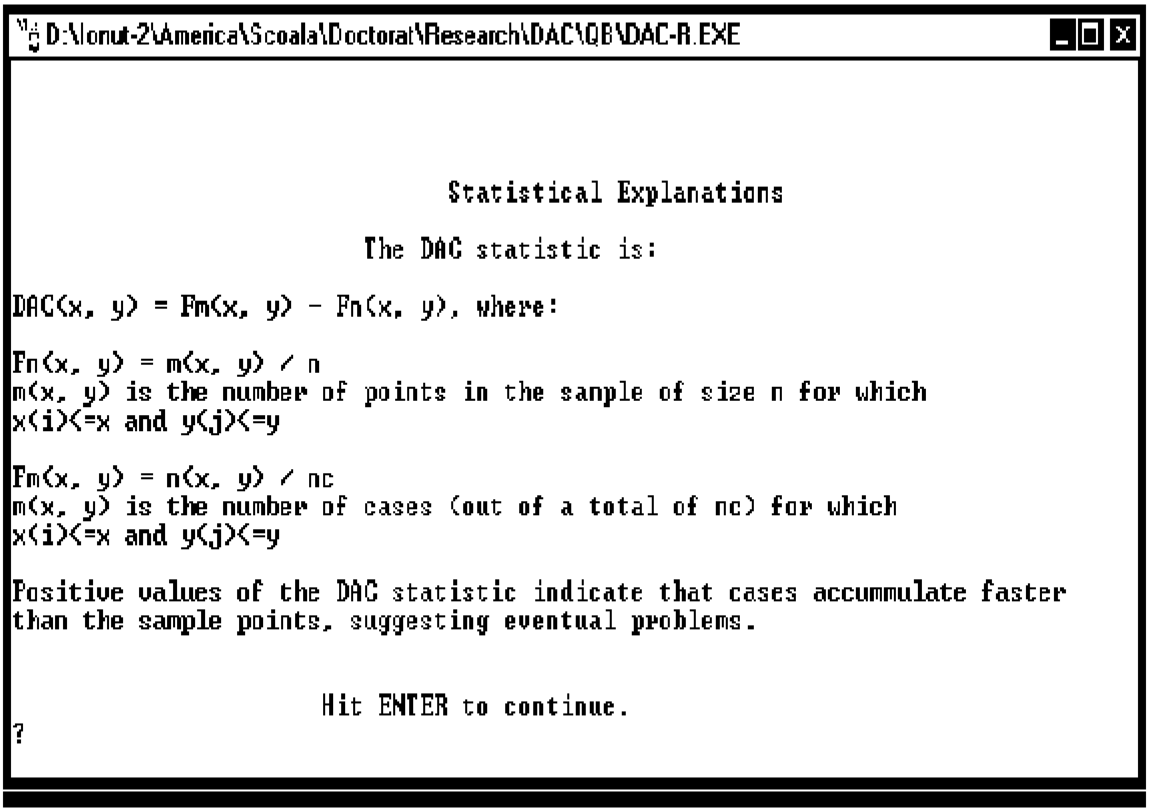
```
D:\Ionut-2\America\Scoala\Doctorat\Research\DAC\QB\DAC-A.EXE

                                    DAC.EXE


                    This program computes the DAC statistic.


                                 Copyright 2001


Alexandru I. Petrisor. MSPH. Graduate Assistant. University of South Carolina
            J. Wanzer Drane, PhD, PE, University of South Carolina
             and Liviu Dragomirescu, PhD, University of Bucharest


                           Hit ENTER to continue.
? _
```

Fig. 13. Interface of the Final Version of DAC.EXE: Authorship.


D:\Ionut-2\America\Scoala\Doctorat\Research\DAC\QB\DAC-R.EXE

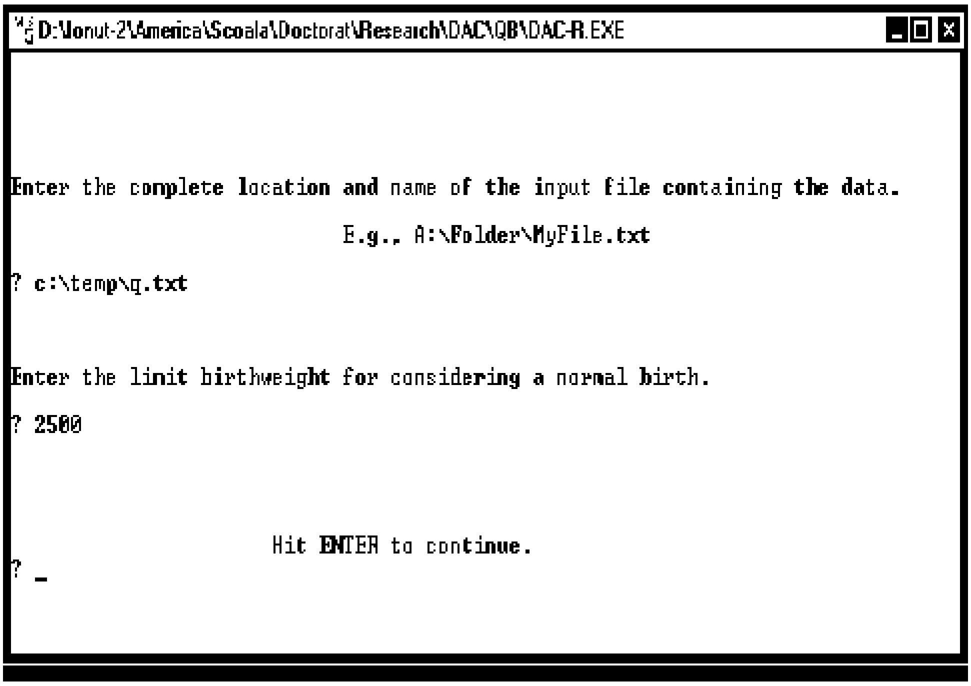
```
If you plan to use this program for a research ending with a published paper,
                     please note the approapriate references:


1. The program is called: <<DAC.exe>>, Program for computing the DAC
(Drane- Aldrich-Creanga) statistic;


2. <<A Computer-Based Application to Detect Spatial Clusters Using the DAC
Statistic>>. Petrisor, AI; Drane, JW; Dragomirescu, L; Jackson, KL. Medical
Informatics Europe Special Topic Conference <<Healthcare Telematics Support
in Transition Countries>>. June 7-9, 2001. Bucharest, Romania. Proceedings.



                         Hit ENTER to continue.
? _
```

Fig. 14. Interface of the Final Version of DAC.EXE: Copyright and Credits.


D:\Ionut-2\America\Scoala\Doctorat\Research\DAC\QB\DAC-R.EXE

```
                                   Statistical Explanations

                              The DAC statistic is:

DAC(x, y) = Fm(x, y) - Fn(x, y), where:

Fn(x, y) = m(x, y) / n
m(x, y) is the number of points in the sanple of size n for which
x(i)<=x and y(j)<=y

Fm(x, y) = n(x, y) / nc
m(x, y) is the number of cases (out of a total of nc) for which
x(i)<=x and y(j)<=y

Positive values of the DAC statistic indicate that cases accummulate faster
than the sample points, suggesting eventual problems.


                         Hit ENTER to continue.
?
```

Fig. 15. Interface of the Final Version of DAC.EXE: Statistical Explanations.


D:\Ionut-2\America\Scoala\Doctorat\Research\DAC\QB\DAC-R.EXE

```
Enter the complete location and name of the input file containing the data.
                                        E.g., A:\Folder\MyFile.txt
? c:\temp\q.txt

Enter the limit birthweight for considering a normal birth.
? 2500


                              Hit ENTER to continue.
? _
```

Fig. 16. Interface of the Final Version of DAC.EXE: User Specifies Input File and the Discrimination Limit.

5. REFERENCES

[1] Aldrich, T.E., and Drane, J.W. - "Cluster 3.1: Software System for Epidemiologic Analysis". Atlanta, GA: USDHHS, ATSDR, 1993.

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

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

[4] Cowen, D.J. – "GIS versus CAD versus DBMS: What are the Differences?", *Photogrammetric Engineering and Remote Sensing*, 54, pp. **1551-1555**, 1988.

[5] Creanga, D.L. - "Spatial Data Analysis for Distributions of Health Events". *Doctoral Thesis*: University of South Carolina, Columbia, 1998.

[6] Drane, J.W., Creanga, D.L., Aldrich, T.E., and Hudson, M.B. – "Detecting Adverse Health Events via Empirical Spatial Distributions", Abstract, *Symposium on Statistical Methods*. Atlanta, GA: USDHHS, PHS, CDC, 1995.

[7] Hollander, M., and Wolfe, D.A. – "Nonparametric Statistics Methods", John Wiley & Sons. New York, NY, 1973.

[8] Huston, M., DeAngelis, D., and Post, W. – "New Computer Models Unify Ecological Theory", *BioScience*, 38, pp. **682-691**, 1988.

[9] Paloheimo, J.E., and Vokov, A.M. – "On measures of Aggregation and Indices of Contagion", *Math. Biosci.*, 30, pp. **69-97**, 1976.

[10] Petrisor, A.I. – "Empirical Spatial Distributions and the DAC Statistic", Master Thesis, University of South Carolina, Columbia, 2000.

[11] Petrisor, A.I., Drane, J.W., Jackson, K.L., and Dragomirescu, L. – "Review of the DAC Statistics", *Bulletin of South Carolina Academy of Sciences*, 63, **94-95**, 2000.

[12] Petrisor, A.I., Drane, J.W., Jackson, K.L., and Dragomirescu, L. – "Spatial Statistics: The DAC Statistics", *Proceedings of the*

Joint Statistical Meetings, Atlanta, GA, 2001.

[13] Petrisor A.I., Drane, J.W., Dragomirescu, L., and Jackson, K.L. – "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, 2001.

[14] Smith, D., and Neutra, R. – "Approaches to Disease Cluster Investigations in a State Health Department", *Stat. Med.*, 12, pp. **1757-1762**, 1993.

[15] Stark, C.R., and Mantel, N. – "Temporal-Spatial Distribution of Birth Dates for Michigan Children with Leukemia", *Cancer Res.*, 27, **1749-75**, 1967.

[16] 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", *Br. J. Cancer*, 37, pp. **109-122**, 1978.

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