ELSEVIER

Computational Statistics & Data Analysis 32 (2000) 259–271

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

A new geometric approach to data analysis using the Minkowski polytope ☆

Marco Bonetti

Department of Biostatistics, Harvard School of Public Health and Dana-Farber Cancer Institute, 655 Huntington Avenue, Boston, MA 02115, USA

Abstract

We introduce a new approach to the analysis of random samples in $\Re^d$, based on a geometric transformation called the "Minkowski polytope" (MP). We describe how a theorem by Minkowski guarantees the existence and uniqueness of such a transformation, discuss its construction, and state a result about the almost sure convergence of a scaled version of the MP in $\Re^2$. We show how the shape of the MP is sensitive to the presence of outliers and correlation in the sample. Finally, we use the MP to develop a new Monte Carlo test for spatial randomness over non-uniform populations, and illustrate its application on a well-known dataset of leukemia cases in the state of New York. © 2000 Published by Elsevier Science B.V. All rights reserved.

Keywords: Exploratory data analysis; Clustering of disease; Flatness; Leukemia

1. Introduction

A traditional way of analyzing a numerical sample is to form functions of the sample, or *statistics*, and to subject these to study. Usual statistics are either single values such as the sample mean, or vectors such as the order statistics. Motivated by recent work in the theory of random (geometric) sets, we extend this approach to

☆ This work was supported in part by ONR Grant N00014-90-J-1641 at the Department of Statistics of the University of Connecticut. The author acknowledges his former academic advisor, Professor Richard Vitale, for his support, encouragment and guidance.

E-mail address: bonetti@jimmy.harvard.edu (M. Bonetti)

0167-9473/00/$ - see front matter © 2000 Published by Elsevier Science B.V. All rights reserved.
PII: S 0167-9473(99)00080-8

set-valued statistics, that is, to functions of the sample that are sets. Of course, the sample mean can be thought of as a (singleton) set, but our aim here is to consider richer possibilities.

1.1. The Minkowski polytope

We are interested in geometric representations of the sample itself from which other statistics can be derived as functionals. Among these, we focus on the Minkowski polytope (MP) constructed from the sample. A *polytope* is defined as the intersection of a finite number of closed half spaces. In the form we consider it, the mapping from a sample in general position to its MP entails only loss of location information, and even that can be avoided with a somewhat more involved version.

It is instructive to consider the planar case (indeed much of our analysis has focused on this case for its tractability). Consider a sample $\{x_1, x_2, \ldots, x_n\}$ in $\Re^2$ that is subjected to centering $\{x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_n - \bar{x}\}$. The latter comprises vectors that sum to zero. Suppose that these are ordered according to their angle of inclination to the (positive) x-axis, and a polygon is formed by starting with one vertex at the origin and succeeding vertices given at the partial sums of the vectors. If the polygon is rotated through $\pi/2$, then it has the following property: it consists precisely of sides that have lengths $||x_i - \bar{x}||$ and normals $x_i - \bar{x}$. This polygon is the MP. Apart from location, the sample (assumed once again in general position) can be retrieved from the polygon (indeed if we center the polygon at $\bar{x}$, then there is no loss of information).

The existence of the Minkowski polytope in arbitrary dimensions is given by the following:

Theorem 1 (Minkowski, 1903). *Let $v_1, v_2, \ldots, v_n$ be pairwise different zero-sum vectors in $\Re^d$ that span $\Re^d$. Then there exists a convex polytope $P \subset \Re^d$ with facet normals $v_1, v_2, \ldots, v_n$ and corresponding facet volumes $||v_1||, ||v_2||, \ldots, ||v_n||$. P is unique up to translation.*

Here, "facet" and "volumes" are the higher-dimensional extensions of the two-dimensional concepts of "sides" and "lengths". A proof of the theorem can be found in Minkowski (1903). Of major interest in the theory of convex bodies, Theorem 1 has been used in several applied areas, such as physics (Jerison, 1996), astronomy (Lamberg, 1993), and image processing (Little, 1983).

1.2. The reconstruction problem

We have seen in the previous section that the construction of the MP in $\Re^2$ is immediate. There is another case in which the construction is direct. In fact, it is quite easy to show that the Minkowski polytope corresponding to $d + 1$ normals in $\Re^d$ can be obtained in closed form. In general, however, the problem of the construction for $d \geq 3$ is non-trivial because of the unknown adjacency relationships among the facets.

The simplest available reconstruction algorithm known to us is given in Little (1983) for $\Re^3$. He developed an iterative algorithm that constructs the MP from the extended Gaussian image (EGI) of the object. The EGI representation is used in image processing to describe 3-D scenes, and is equivalent to the specification of the vectors with directions normal to each face and length equal to the area of each face. The polytope P that is sought minimizes

$$\Phi(P) = \Phi(P(\lambda)) = ||v_1||\lambda_1 + \cdots + ||v_n||\lambda_n$$

over all polytopes L with volume at least unity. The values v_i's are the areas of the faces, and the λ_i's are the corresponding distances of the faces from the origin.

Little's approach is to solve the constrained minimization problem with standard linear programming techniques using the reduced gradient method described in Gill et al. (1981). The region of the admissible vectors λ can be shown to be convex, and this allows one to consider the problem from the optimization point of view.

Also of interest is the way in which Little constructs the polytope as intersection of the specified half-spaces. He makes use of a "dual transform" (Huffman, 1977). This transform takes a plane with equation

$$Ax + By + Cz + 1 = 0$$

into the point $(A, B, C) \in \Re^3$. The n planes forming P correspond therefore to n points in $\Re^3$, and it can be shown that the convex hull of these points provides the adjacency information for P. In particular, any face of the hull corresponds to a vertex of P; and any two points incident on an edge of the hull correspond to a pair of faces of P that share an edge. In other words, this adjacency information allows determination of the vertices.

Once the (translated) P has been obtained, re-scaling by $V^{1/3}(P)$ is done to have unit volume. Then the value of $\Phi(P)$ is computed and the optimizing algorithm is invoked, that computes the next step through the gradient of $V(P)$. Iteration continues until the reduction in the value of $\Phi(P)$ is smaller than a pre-determined value $\delta > 0$. Little's paper also contains a detailed illustration of the construction of the MP from a set of vectors in $\Re^3$, with interesting stereo views of the polygons.

The iterative algorithm requires $O(n \log n)$ operations for each iteration, and the number of iterations depends on the algorithm used. The one proposed achieves linear rate of convergence, i.e. if one calls ε_i the error at step i, then the algorithm satisfies

$$\lim_{i \to \infty} \frac{|\varepsilon_{i+1}|}{|\varepsilon_i|} = \gamma < 1.$$

Gritzmann and Hufnagel (1995) point out that there is no polynomial bound on the running time of Little's algorithm. They suggest a more involved algorithm, and show that if the dimensionality d of the problem is known, then the reconstruction problem can be solved in polynomial time.

The use of the MP for data analysis is motivated by the following result, that deals with the large sample behavior of the MP.

2. A strong law of large numbers for the MP

In traditional statistics we are often interested in the convergence of a sequence of random variables to a particular limiting value. For example, one might consider the sequence $\{T_n = (1/n)\sum_{i=1}^{n} x_i\}$ of the sample means calculated from random samples of increasing sample size n and ask whether as $n \to \infty$ the sequence converges to the population mean μ. We are now interested in the limiting behavior of the Minkowski polygon constructed from a sample arising from a particular distribution in $\Re^2$. Scaling the MP to take into account the increasing sample size, there is convergence of the scaled Minkowski polygon to a convex body characteristic of the population from which the samples are obtained. We first need to introduce the concept of support function.

The support function $h_C(v)$ of a convex body C is defined as follows: $\forall v \in \Re^d, h_C(v) = \sup_{y \in C} \langle v, y \rangle$ (see Schneider, 1993). The support function uniquely identifies the body C and conveniently allows one to deal with translations, changes in scale, and rotations in the axes. We say that a sequence $\{C_n\}$ of convex bodies *converges* to a convex body C if the sequence of the corresponding support functions $\{h_{C_n}\}$ converges pointwise to h_C. The evident positive homeogeneity of support functions allows them to be restricted to the unit sphere S^{d-1}. Here we concentrate on the case of vectors $\{x_1, x_2, \ldots, x_n\}$ arising from a probability distribution in $\Re^2$, in which case $h_C(v)$ can be defined as a function of the angle $\theta \in [0, 2\pi)$:

$$h_C(\theta) = \sup_{x \in C} \langle e(\theta), x \rangle, \quad \text{where } e(\theta) = [\cos\theta, \sin\theta]'.$$

Theorem 2 (Bonetti, 1996). *If $Z_1, Z_2, \ldots, Z_n$ is a random sample obtained from an absolutely continuous distribution F_Z on the plane with zero mean and finite variances, then the scaled (i.e. divided by n) Minkowski polygon converges a.s. to a convex body associated with F_Z. Such convex body is described by the support function*

$$h\left(\theta - \frac{\pi}{2}\right) = E_Z 1(\Theta \leq \theta) R \cos(\Theta - \theta),$$

where Θ and R are the expression in polar coordinates of the random variable Z.

A proof of this result can be obtained from the author. (Since submission of the present paper, a more general result has also been obtained in Bonetti and Vitale, 1999.) Examples of this form of convergence can be observed in Figs. 1–3 below, where the scaled MP is constructed from samples of size 1000 arising from different distributions.

Example 1. Consider the sample as coming from a bivariate normal random variable $Z \sim N_2[0, \sigma^2 I_2]$. For this simple case it is well known that the polar coordinates $(\theta(Z), \rho(Z))$ are independent, and Theorem 2 indicates that the limiting scaled MP is a circle of radius $r = E(\rho)/2\pi$ "sitting" on the origin.

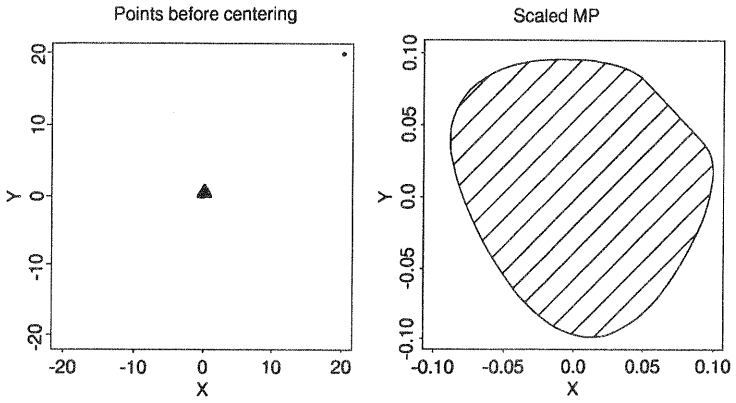

Fig. 1. Scaled MP constructed from a sample of size 1000 from a uniform distribution on a triangle, plus an extreme outlier at (20,20).

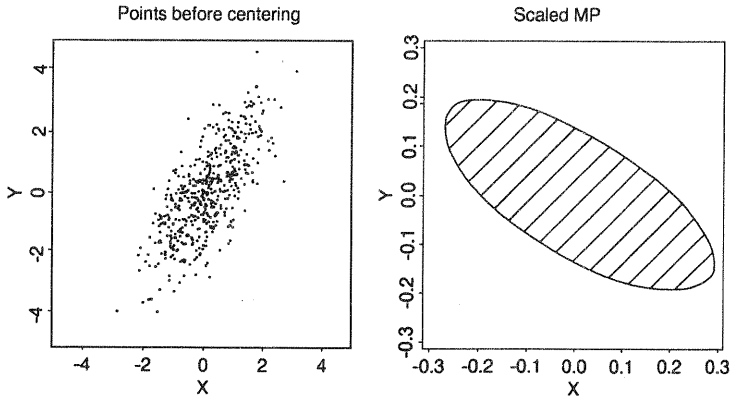

Fig. 2. MP constructed from a sample of size 1000 obtained from a normal distribution having correlated coordinates.

3. Exploratory data analysis using the MP

Exploratory data analysis is traditionally accomplished through numerical and graphical work, using techniques such as stem-and-leaf displays, numerical summaries, and q–q plots. Along these lines, we discuss some ways in which the MP can be used to develop new exploratory data analytical techniques. We concentrate on the $\Re^2$ case, and the three properties of outlying observations, collinearity, and clustering. The discussion that follows is informal, and aimed at exploring the effect of these properties on the MP.

3.1. Flatness

Recall from Section 1 the construction of the MP in $\Re^2$: given our sample, we create a polygon based on Minkowski's theorem. If the underlying distribution

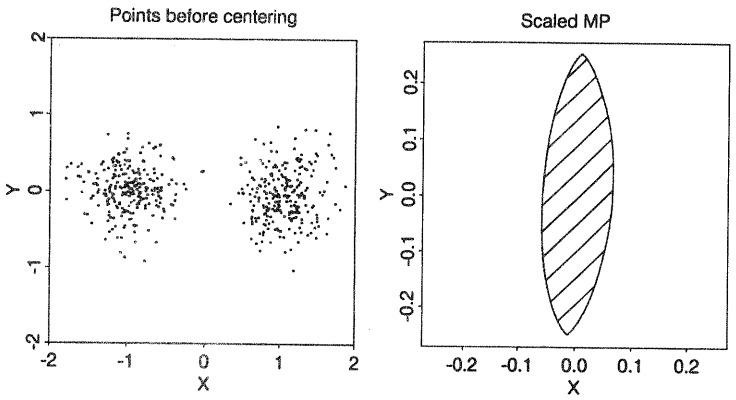

Fig. 3. MP constructed from a sample of size 1000 obtained from a mixture of two equally weighted circular normal distributions centered at $(-1,0)$ and $(1,0)$.

is sufficiently uniform, then an outlying observation has its corresponding vector $(x_i - \bar{x}, y_i - \bar{y})$ longer than the others; this means that the corresponding side in the MP will be longer than "expected". That is, because of the way in which the polygon is constructed, an outlier will cause the polygon's tending to be *flat* (see Fig. 1).

We can consider some measure of such flatness, for instance, the quantity

$$F(\{x_1, x_2, \ldots, x_n\}) = \frac{\text{area(MP)}}{\text{perimeter}^2\text{(MP)}}. \tag{1}$$

It is well known that this ratio ranges between 0 and $1/4\pi$, where 0 corresponds to a segment and $1/4\pi$ corresponds to the circle (in fact, this is the isoperimetric ratio, see Mitrinović et al., 1989, p. 443). This quantity, therefore, is a measure of how far the polygon is from being a circle.

How can we be sure that any flatness of the polygon is indeed due to the presence of one (or possibly more) outliers? In fact, multicollinearity and clustering also produce flatness: in a situation of collinearity, for example, the sample points tend to fall in a pattern that favors certain directions in the vectors $(x_i - \bar{x}, y_i - \bar{y})$. This, in turn, corresponds to favoring certain orientations in the sides of the polygons. (See Fig. 2 for an illustration of this.) Also, if for example the sample is formed by two clusters, there will also be certain preferred directions (see Fig. 3).

3.2. *ρ-histograms and θ-histograms*

We suggest plotting the two simple histograms corresponding to the lengths ρ_i's and to the angles θ_i's of the centered vectors $x_i - \bar{x}$. Careful observation of these two histograms in many different simulated cases (see for example Fig. 4) has illustrated its value in determining the contribution of individual observations to the flatness measured in the polygon. In particular, individual outlying observations can be detected by one (or more) spikes in the ρ-histogram. Also, both the presence of two clusters and correlation between the coordinates result in preferred directions, but

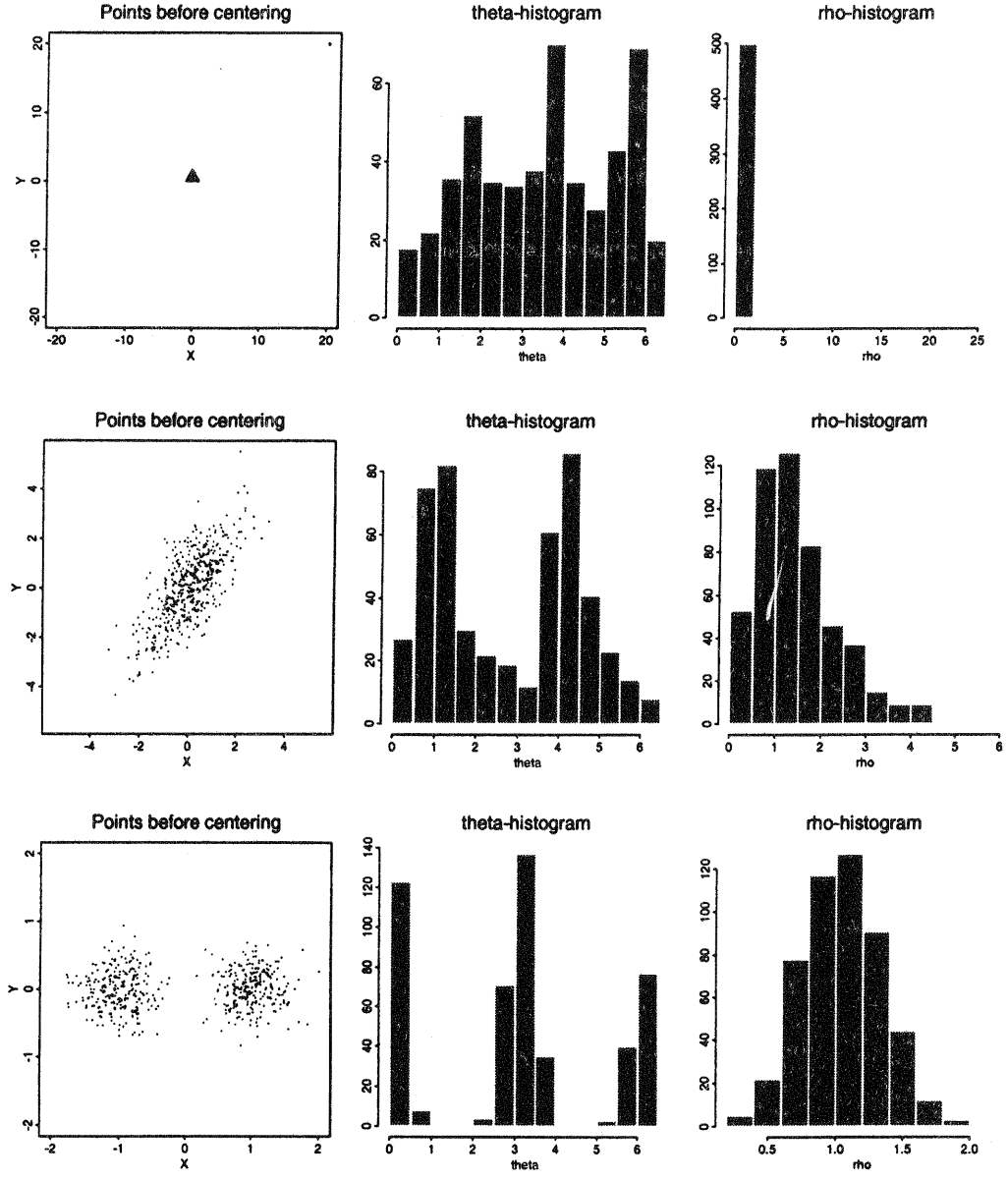

Fig. 4. Clustering vs. correlation.

the θ-histogram for the latter case has sharper gaps. Moreover, the distance between the clusters causes very few vectors to have small lengths (the sample mean tends to fall between the clusters, with few points close to it), and this is clearly shown by the ρ-histograms. Observe how rotation of the distribution leaves the ρ-histogram unchanged, while it shifts the θ-histogram along the x-axis (modulo 2π).

The definition of one more quantity can be suggested: let the "second-order" outlying measure be the area corresponding to the triangle obtained connecting the side

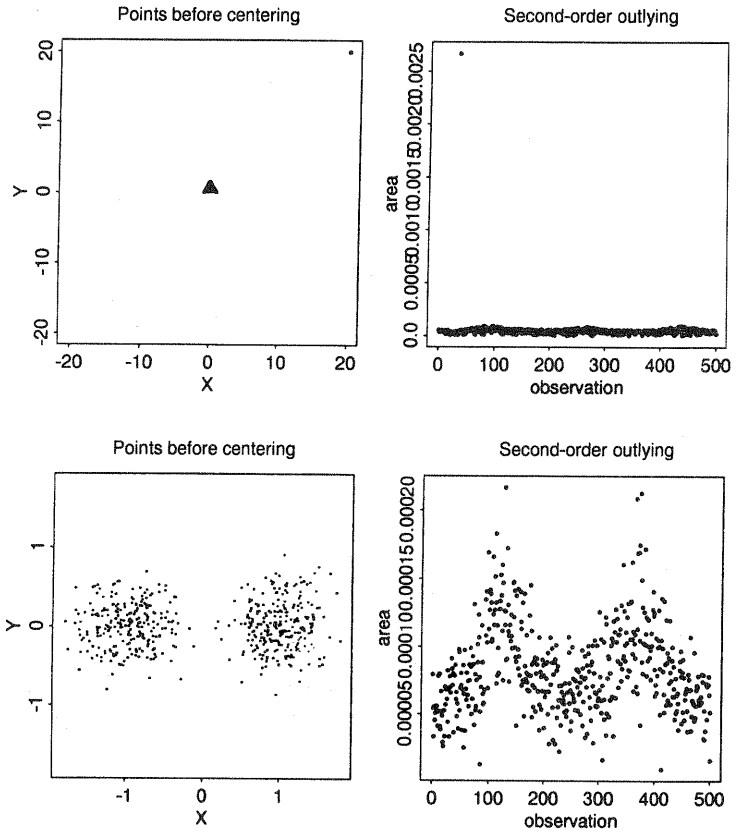

Fig. 5. Second-order outlying measure.

corresponding to each observation with the center of gravity of the MP. This new measure (M_i) depends not only on the length of the vector corresponding to the observation, but also on the overall influence of that observation on the final shape of the polygon. Simulated examples suggest (see Fig. 5) that observation of an index plot of these quantities M_i allows can be helpful in distinguishing the presence of clustering from correlation, even in the presence of outliers (which are also easily identified). The definition of this new measure M_i can easily be extended to general samples in $\Re^d$ if one thinks of (hyper-)volumes instead of areas, thus providing the necessary reduction in dimensionality.

The characterization given above is not general, and we cannot expect it to be so. This is particularly so when the distribution is highly non-homogeneous (for example, in the case of a human population), in which case making any statement about the characteristics of the distribution is going to be very challenging.

There are two particular situations that require comment: (a) two or more of the vectors actually are positively proportional; and (b) two or more vectors are not only positively proportional, but they are replications of the same vector. Both cases introduce a problem since Minkowski's Theorem does not apply. Positive propor-

tionality of one or more vectors means that the MP would have a facet to which would correspond more than one vector, thus making it impossible to identify the effect of each contributing vector. In other words, we would not be able to transform the MP back to the vectors, since a large facet could correspond to many "short" vectors, to fewer "longer" vectors, or to some identical vectors.

Both situations arise with probability zero if the sample is obtained from a continuous probability distribution, but they might be observed in practical applications, due to discretization. An example of situation (2) above will be analyzed in the next section, where population counts will be given for cells corresponding to a sub-division of a region of interest (in $\Re^2$). It is therefore important to understand the consequences of these situations. In fact, they do not necessarily constitute a particular problem for the practical construction of the MP, but they require more care in its interpretation.

The flatness of the MP can be thought of as being a numerical summary of the shape of the sample. We now discuss as an example of application of this fact to a new test for the spatial randomness of disease cases.

4. A test for spatial randomness

Waller et al. (1994) study leukemia cases recorded in an eight-county region in upstate New York during the five-year period 1978–1982. The data, obtained from the state Cancer Registry, report 591 cases of leukemia over a population of little over 1 million individuals, and the interest is in the detection of possible spatial clusters in the cases. The detection of clustering in the leukemia cases could provide some hints about factors that might be related to the incidence of the disease in the population. In this section we make use of the MP to construct a new test for spatial randomness over non-homogeneous populations.

The eight-county region was divided by the Census bureau into 790 subregions or cells with population counts and leukemia incidence counts (see Fig. 6 for a map of the region). The cells were defined by using US Census block groups (for all but one county), and for each cell there are spatial coordinates of the centroid of the cell and the pair of counts for the population and the leukemia cases. Because of confidentiality issues, the exact geographic locations are not available.

A test for the randomness of the geographical distribution of the leukemia cases must consider the underlying distribution of the population of the region. The authors of the study applied several different techniques to this data set. The main three are:

- A method proposed in Whittemore et al. (1987) based on the mean distance between all pairs of cases, which has been shown to be asymptotically normally distributed under the null hypothesis of randomness.
- A graphical procedure called "geographical analysis machine" or GAM (Openshaw et al., 1988) based on the number of cases observed within a circle of fixed radius and scanning the region. This method picks the circles that

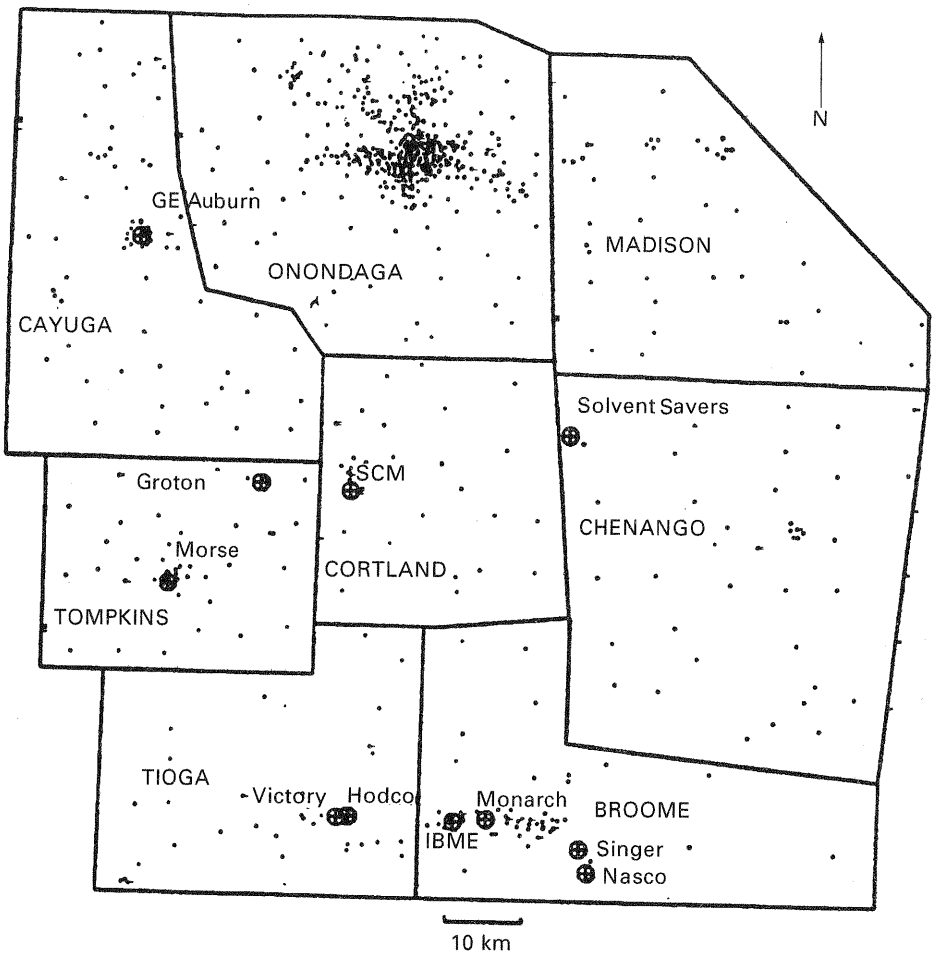

Fig. 6. Map of New York study area with location of cell centroids indicated by · and hazardous waste sites containing trichloroethylene indicated by ⊕ (reprinted with permission from Waller et al. (1994)).

contain an "excessively high" (with respect to a Poisson distribution) number of cases, and is looked at as a descriptive tool.

- Two Monte Carlo testing procedures, described in Besag and Newell (1991) and in Turnbull et al. (1990), that evaluate different test statistics based on the GAM idea, but in which the circles are constrained to contain the same population and the same number of cases, respectively.

We do not discuss the details of each method here. Rather, we quote the overall conclusion of the authors of Waller et al. (1994) that "the evidence of clustering of the cases is rather weak, although there is some suggestion that there may be a mild effect when one considers larger radii in the method of Turnbull et al. (1990)". The authors indicate that this may be due to the fact that "by trying to maintain a type I error rate in the face of so many multiple tests [...] the power of the procedures must necessarily be very low". (Waller et al., 1994, p. 16).

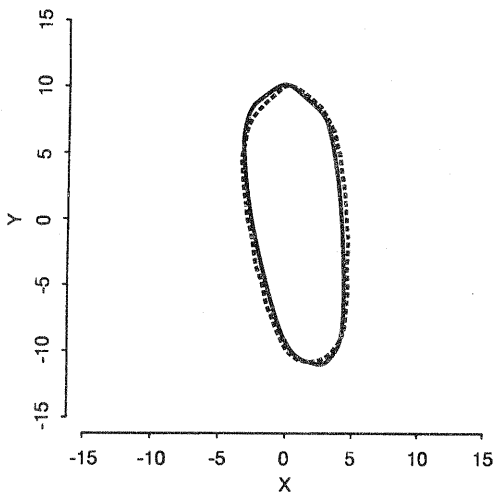

Fig. 7. MP constructed from the leukemia cases (dashed line) and MP from the whole population (solid line).

Table 1
Summary of sample of flatness values

Min	Q_1	Median	Q_3	Max
0.04968	0.05477	0.05601	0.05719	0.06186

Let us consider the sampling distribution of the flatness measure F=Area/Perimeter^2 of the MP constructed from the centered sample consisting of the coordinates of the leukemia cases. Comparison of the value corresponding to the actual sample of leukemia cases with such sampling distribution allows one to construct a procedure for testing the hypothesis that such sample be indeed a *random* sample from the population. An estimate of the sampling distribution of the flatness of the MP is obtainable through Monte Carlo simulation of the sampling process, when we sample by giving to each individual in the population the same probability of becoming one of the cases.

The MP obtained from the leukemia cases (shown in Fig. 7) has a flatness measure of 0.0518034. The results of our simulations (based on 5000 samples of size 591) are shown in Table 1. Observation of the sorted values shows that the 60th value is 0.05180081 and the 61st is 0.05183548, so that our (2-sided) testing procedure gives an estimated p-value of about 0.024. We thus reject the null hypothesis of spatial randomness of the leukemia cases.

Observe that we are given the population and leukemia cases counts corresponding to the centroids of the cells, and we do not know the specific location of each individual. Our analysis is therefore "discretized" to the centroids. We can try to artifically reconstruct a possible distribution of the individuals within each cell, for

example uniformly within the cell. This might be important since the dependence of the result on the size and shape of each cell might modify the outcome of the testing procedure. We have experimented in this direction, but the limited information available to us about the boundaries of the cells forced us to introduce an approximation of such boundaries through the construction of the Voronoi diagram (see O'Rourke, 1993) corresponding to the centroids. Uniform spreading of the population (and of the sample) within the cells has lead to results extremely similar to the ones obtained without the introduction of this reconstruction.

5. Discussion

Summarizing, we have introduced a set-valued data-analytical tool that presents some interesting characteristics. The development of a distributional theory for the MP and for statistics obtained from it appears rather difficult, but the derivation of the SLLN is encouraging. Ongoing work is aimed at studying more of the behavior of the MP, and will appear elsewhere.

Our results from the application discussed in Section 4 are different from the ones reported by Waller et al. (1990) in that our results show evidence of non-randomness in the leukemia cases. The analysis that we have described here, however, is meant as an illustration of the method, and it should not be considered as being conclusive for this delicate problem. The method that we have introduced consists of a single overall test over the region, and as such it does not present the problem of repeated testing. Also, the method does not require any distributional assumption, nor estimation of parameters.

References

Besag, J., Newell, J., 1991. The detection of clusters in rare diseases. J. Roy. Statist. Soc. Ser. A 154, 143–155.

Bonetti, M., 1996. Geometric methods in data analysis. Ph.D. Dissertation, Department of Statistics, University of Connecticut, Storrs, CT.

Bonetti, M., Vitale, R.A., 1999. Asymptotic behavior of a Set. Statistic. In press, Discrete Comput. Geom.

Gill, P.E., Murray, W., Wright, M.H., 1981. Practical Optimization. Academic Press, New York.

Gritzmann, P., Hufnagel, A., 1995. A polynomial time algorithm for Minkowski reconstruction. Proceedings of the 11th Annual Symposium on Computational Geometry, pp. 1–9.

Huffman, D.A., 1977. A duality concept for the analysis of polyhedral scenes. In: Elcock, E.W. and Michie, D. (Eds.), Machine Intelligence. 8, pp. 475–492.

Jerison, D., 1996. A Minkowski problem for electrostatic capacity. Acta Math. 176 (1), 1–47.

Lamberg, L., 1993. On the Minkowski problem and the lightcurve operator. Ann. Acad. Sci. Fenn. Ser. A I Math. Dissertationes 87, 1–107.

Little, J.J., 1983. An iterative method for reconstructing convex polyhedra from extended Gaussian images. Proc. AAAI, National Conference on Artificial Intelligence, Washington D.C., pp. 247–250.

Minkowski, H., 1903. Volumen und oberfläche. Math. Ann. 57, 447–495.

Mitrinović, D.S., Pečarić, J.E., Volenec, V., 1989. Recent Advances in Geometric Inequalities. Kluwer Academic Publishers, Boston, MA.

Openshaw, S., Craft, A.W., Charlton, M., Birch, J.M., 1988. Investigation of leukemia clusters by use of a geographical analysis machine. "Lancet 1 (8580): 272–3".

O'Rourke, J., 1993. Computational Geometry in C. Cambridge University Press, New York.

Schneider, R., 1993. Convex Bodies: the Brunn–Minkowski Theory. Cambridge University Press, New York.

Turnbull, B.W., Iwano, E.J., Burnett, W.S., Howe, H.L., Clark, L.C., 1990. Monitoring for clustering of disease: application to leukemia incidence in upstate New York. Amer. J. Epidem. 132, S136–S143.

Waller, L.A., Turnbull, B.W., Clark, L.C., Nasca, P., 1994. Spatial pattern analyses to detect rare disease clusters. In: Lange, N. et al. (Eds.), Case Studies in Biometry. Wiley, New York, pp. 3–23.

Whittemore, A.S., Friend, N., Brown, B.W., Holly, E.A., 1987. A test to detect clusters of disease. Biometrika 74, 631–635.