Survey on cluster tests for spatial area data

Friedrich Gebhardt
Stenzelbergweg 8, D-53229 Bonn*
friedrich.gebhardt@gmd.de

Updated version of GMD Report 7 (April 1998).
Work finished July 1999 with minor later amendments.

Table of contents

Abstract

Geographical data usually exhibit some amount of spatial dependency, a correlation between the values of neighbouring districts. Thus one wants to have measures for the strength of this dependency and tests for the deviation from randomly distributed values. There exist several tests. In this survey, they are collected and compared. This is done separately for binary variables assuming only two values and for real-valued variables. Among the tests are the black-black count, the black-white count, Moran's *I*, Geary's *c* and the Getis-Ord statistics. Some new statistics are proposed: a variant to the black-black count and statistics based on clusters composed of triplets of districts. Included are also new results on the distribution of Moran's *I* and its local version, based on simulations using several areas with 37 to 327 districts. Roughly speaking, the distribution of *I* is fairly close to a normal distribution and surprisingly independent of the underlying distribution of the district values while the local *I*'s are extremely far from normal and highly dependent on the underlying distribution.

1 Introduction

1.1 Goal

An important goal of data mining is to extract hidden relationships between objects, in particular relationships between some variables, possibly conditional on the values of other variables. However, looking unspecifically for possibly interesting properties of a data set

* The work was mostly performed when the author was with GMD – German Research Center for Information Technology, System Design Technology Institute, SET (now Autonomous intelligent Systems Institute, AiS).

involves a big danger: every meaningful data set, even random data, exhibits some peculiarities just by chance. As a rough guide to judge the importance of relationships in the data, statistical test procedures are used, but formally significant results do not prove the existence of deviations from randomness, they only give hints that have to be confirmed (or disproved) in further steps. This is, however, not the place to discuss the role of statistics in data mining in general; some specific caveats will be mentioned later (Section 6.2).

Geographical data show a peculiarity: in addition to the conventional variables, the relative geographical position of two objects is an important feature. Thus one may ask whether a variable shows a correlation between the values of neighbouring objects. Such a situation is called a spatial autocorrelation. 'Autocorrelation' refers to a correlation within one variable between the objects in analogy to the autocorrelation of time series. A different, though related, question is whether objects with some property are more or less evenly distributed in the space or exhibit a spatial clustering. This question is the main theme of this article.

We will give a survey on the existing methods of determining spatial clustering with emphasis on those that are useful for data mining. Essential characteristics of the exploratory analysis are that hundreds or thousands of tests are performed on a data set; thus the overall error probability cannot reasonably be kept at a predefined level. Consequently, the *exact* error probability of a single test is not important (an approximate value suffices), but it should be rather small (under 1% or even 0.1%) so that the distribution of the test statistic in the tails is needed where the normal approximation is often insufficient. In addition, since the multitude of tests necessarily leads to formally significant results, emphasis is on results that can easily be interpreted.

We will concentrate on area data – to be defined later – and mention some related work on point data and continuous data. The survey includes some new results by the author. These are expanded to some detail in Gebhardt (1998b, 2000). This report does not follow the style of a textbook – proofs are mostly omitted.

1.2 Data types

Geographic data can roughly be classified into three types.

First, there are *spatially continuous data* such as the elevation. This particular variable is virtually known for all points of earth with sufficient accuracy. Another example is the air pressure. It is known only for selected points and must be interpolated in between, but nevertheless it exists everywhere and thus is a spatially continuous variable. In addition, it is time-dependent.

The second type are *point pattern data* or *point data:* data that exist for some points only. An example is the epicenters of earthquakes; the location is the data element. Additional variables may be attached to the location: magnitude of the quake, time, duration. The data may be three-dimensional, for instance the center of the earthquake including also the depth. Another example is pollutant concentration. This is actually a continuous variable but sometimes measured only at so few points that interpolation is infeasible. Therefore it may have to be treated as point data.

Still another example is the residence of persons having a particular disease. In analyzing such data one has of course to take into account that people are not evenly spread over earth. If this fact is neglected, one might find locations with high population density rather than high disease risk.

The third type consists of *area data*. A connected *region* is subdivided into smaller parts (states, counties, statistical districts, fields, cells, pixels), here called *districts*. The districts are non-overlapping and their union is the whole region. Often it is assumed that each district is connected.

To each district, the value of one or more variables is attached. In this report we distinguish binary and real-valued variables. *Binary variables* take on two values only: 1 and 0, yes and no, black and white, marked and unmarked. The values may be inherently binary, for instance the election districts won by a particular party or the distribution of two fiber types in a cross section of a muscle, analyzed by Venema (1992). The binary variable may signify one class of a nominal variable. Often the binary data are derived from a real-valued variable, for instance whether that variable exceeds a natural or an arbitrarily chosen threshold.

In many cases the original variables have to be normalized somehow to make the values comparable. For instance, comparing the numbers of unemployed in a county makes little sense; the unemployment rate should be considered. Finding the proper normalization can be a problem. Thus the number of traffic accidents in a county depends certainly on the population (or the number of licensed cars) but also on the road net and the proportion of cars from other counties on the road.

Even with normalization, the areas should be comparable to achieve proper results. If for instance the number of inhabitants is the normalizing factor, the population of the areas should not differ too much. It makes no sense to compare Vatican City with Russia.

1.3 Example

As an illustration consider Figure 1 (on page 32). It shows 171 counties in north-west Germany, marked by the motor vehicle codes (if a city and the surrounding county have the same code, the latter one is distinguished here by an asterix). Some examples to help locate the region: HB Bremen, H Hannover, E Essen, K Köln (Cologne), AC Aachen, BN Bonn, F Frankfurt, SB Saarbrücken.

Basis for the variable under consideration is the number of persons working and counted for social security (sozialversicherungspflichtig Beschäftigte, henceforth called “workers” for short). Figure 1 (page 32) shows the counties with a high share of aliens among the workers. There seem to be two clusters, one in the industrial region between Essen and Bonn and the other one around Frankfurt. Is this a chance result? The average of the county values is 0.071 with a standard deviation of 0.034. The largest value, 0.183, pertains to Groß Gerau (GG) near Frankfurt, the smallest one to Dannenberg (DAN) in the north. As a binary variable, we will arbitrarily bisect the rate of aliens at 0.11, i.e., counties with a rate > 0.11 are considered black or marked. Incidentally, the results for bisecting at 0.10 or 0.12 are comparable. The data are taken from Statistisches Bundesamt (1994, 1995) and refer to 1993.

1.4 Overview

The main part deals with the analysis of area data: binary variables in chapter 3, real-valued variables in chapter 4. Common concepts are introduced before in chapter 2.

The treatment of point data and spatially continuous data is sketched in chapter 5.

The last chapter characterizes some textbooks, points to other related work and warns of some dangers and pitfalls particularly with spatial data; finally it offers some general conclusions.

We neglect here models in space *and* time, although they are also treated in several textbooks, for instance in Anselin (1988), Cressie (1993).

2 General model for area data

A goal of the analysis of spatial data is to find spatial correlations and spatial regularities or peculiarities. Due to the two-dimensional structure there is a high chance that counties with large (or small) values of the considered variable are neighbours; thus one has to be cautious in interpreting seemingly conspicuous concentrations. For a statistical test, one has to specify the null hypothesis (no peculiarity is present, usually independence of all districts is assumed) and a proper alternative or set of alternatives, the counter hypotheses. The concepts common to most tests of spatial area data are developed in this chapter.

2.1 Null hypothesis

The null hypothesis with area data is usually that all values are independent and distributed according to a known or unknown common distribution function.

Sometimes the assumption of a common distribution is grossly in error, for instance if the variable is derived from samples of different sizes (different population sizes in the districts). Here one can sometimes assume that the distributions belong to the same class (Poisson, normal, etc.) and that, under the null hypothesis, the pertinent parameter is known up to a common constant. For instance, the expectation of the normally distributed variables could be μ (unknown) and the variance σ^2/d_i with unknown σ and known d_i, the population size of district i, if the variables are district means.

Most statistics for testing the independence against spatial dependencies use a weight matrix (association matrix) W with elements w_{ij}; w_{ij} is a measure for the association or neighbourhood between districts i and j. Some examples are:

- w_{ij} = 1 if districts i and j have a common boundary (or, alternatively, at least a common boundary point), otherwise w_{ij}=0.
- w_{ij} is the proportion of the boundary of district i that is shared with district j. This weight matrix is unsymmetric.
- w_{ij} = 1 if the distance between districts i and j is less than a threshold; otherwise, w_{ij} = 0. The distance may be the distance between the capitals of the districts or between the geographical centers or between the interior points farthest from the boundary or between the areas (i.e., direct neighbours have distance 0).
- w_{ij} is a decreasing function of the distance (and usually equal to zero if the distance exceeds a threshold).
- w_{ij} = 1 if the center of district j is one of the nearest k to the center of district i; otherwise, w_{ij} = 0. This weight matrix is unsymmetric.
- w_{ij} reflects the reachability of district j from district i. Here two major cities that are geographically far apart may have a large value for w_{ij} because they are connected by train or airline. Such models are used for the spread of infectious diseases.

In the binary case based on distances, small districts tend to have many associated districts (districts with corresponding $w_{ij} > 0$), while large districts may have few associated districts or none at all. An example can be found in Unwin (1996). If, however, the common boundary is the criterion, small districts tend to have few associated districts (i.e. direct neighbours);

tiny local shifts in the geographical boundary can change the adjacency, for instance if four districts come nearly together in a point.

2.2 Counter hypotheses

The null hypothesis assumes independence of the spatial variables. In geographic data, this is nearly never true; neighbouring districts or points are almost always correlated. This fact hampers already the design or the evaluation of statistical tests. The situation becomes worse as soon as one tries to choose a counter hypothesis or a set of counter hypotheses. Usually one is not interested in just any deviation from independence (which is not given anyway) but in certain types of dependence, but how to specify them?

There are many possibilities to define counter hypotheses. One general class consists of distributions that are still independent but the distribution parameter(s) differ from one district to the other. For example, the variables for the districts may be normally distributed with a common variance but with means shifting from north to south (or in any other direction), the so-called *trend-surface analysis*. Such situations can be tested by linear models containing latitude and longitude as independent variables in the usual way.

Peculiar to spatial data is the assumption that the variables for near-by points or districts are correlated. Similar models are known from time series; however, these have only one dimension (the time) and a causal dependence in one direction. With spatial data, one has two (or even more) dimensions and causal dependencies in all directions. This implies more complicated models than with time series.

One particular often used model for spatial variables is Whittle's model (simultaneous autoregressive model), see e.g. Cliff and Ord (1981, Section 6.2.3), where the independent variables ε_i are hidden and only derived and mutually dependent variables $X_1, \ldots, X_n$ can be observed:

$$X_i = \rho \sum_{j \neq i} w_{ij} X_j + \varepsilon_i \qquad (1)$$

or in matrix notation

$$X = \rho\, W\, X + \varepsilon$$

where usually W is known while ρ and ε are to be estimated; the variables ε_i are assumed independent. The last equation can be rewritten as

$$X = (I - \rho\, W)^{-1}\, \varepsilon.$$

With $\rho = 0$ one gets the null hypothesis.

Another model is the moving-average model, Cliff and Ord (1981, Section 6.2.5),

$$X_i = \varepsilon_i + \rho \sum_{j \neq i} w_{ij} \varepsilon_j, \qquad X = (I + \rho W)\varepsilon. \qquad (2)$$

Both models assume an autocorrelation that is in principle the same (it is 'stationary') in all parts of the region. Any deviations in a subregion must be reflected beforehand in the choice of the weights w_{ij}. The models do not permit for instance different but unknown strengths of the correlations in different areas or a different mean in an area or global trends in the mean. Different strengths of the correlation between neighbours may be a peculiarity of the particular variable; it may also be an artefact, for instance if the districts are on the average smaller in some regions than in others.

Still other models and estimators for the parameters are treated in Cliff and Ord (1981, Section 6), in particular the conditional autoregressive model (Bartlett's model) in Section

6.2.4, see also Venema (1988, 1989, 1993) and Pernuš (1989). Besag (1974) suggests a binary Markov random field model called autologistic model. More complicated models consider a time series of spatial data. These will not be discussed here; see e.g. Sections 1.6 and 1.7.3 of Cliff and Ord (1981), Markov connected components fields in space and time by Møller (1998) and a Bayesian model in space and time by Knorr-Held and Besag (1998).

2.3 Form of test statistics

Let us denote the random variable measured for district i by x_i. Then many test statistics used in practice have the form

$$C\sum_{i,j} w_{ij} f_{ij} \qquad (3)$$

with a constant C and $f_{ij} = f(x_i, x_j)$. The matrix W used here need not necessarily be the same as in the counter hypotheses of Section 2.2 but to use the same one (as far as it is known) seems a good choice according to some asymptotic results for such test statistics as Moran's I and Geary's c (see below), Cliff and Ord (1981, Section 6.4.3).

Example. Assume w_{ij} is 1 for direct neighbours and 0 otherwise and x_i is binary, $f_{ij} = x_i \cdot x_j$. Then $\sum w_{ij} f_{ij}$ counts the pairs of neighbouring marked (or black) districts ($x_i = 1$); each common boundary of marked districts is counted twice. $\frac{1}{2}\sum w_{ij} f_{ij}$ is called the black-black count statistic.

A variety of weights W and functions f have been proposed in the literature; only some of them – the more important ones – are investigated in this survey. Some others are sketched in Marshall (1991). Recommendations for choosing the weight matrix W are given by Griffith (1995). According to that article, an incorrect choice inflates the standard error of the model and gives a wrong estimate for the autocorrelation, particularly for small sample sizes. Generally a small effective area outside which the weights are small or 0 seems to be better than one that is too large. In case of doubt it is advisable to use several weight matrices and to compare the results, for instance several thresholds if w_{ij} = 1 as long as the distance is below the threshold.

Often one or more of the following properties are assumed for the weight matrix W. Note that not all properties are compatible.

1. w_{ij} is binary, i.e. either 1 or 0.
2. $w_{ii} = 0$. It seems that this assumption is sometimes tacitly made in the literature. Sometimes it is made by using summations (as in formula (3) above) over $i \neq j$.
3. $\sum_j w_{ij} = 1$ (or some other common constant).
4. $w_{ij} = w_{ji}$. Note that in general W is not assumed to be symmetric.

Usually one has $w_{ij} \geq 0$, but this is not required.

2.4 Assumptions on the distribution

The distribution of the test statistics under the null hypothesis is mostly computed under one of the following assumptions:

N, normality. The values for the districts are derived from a common distribution function, for instance a normal distribution function with a common expectation and variance.

Under the null hypothesis, all districts are independent. An alternative distribution is: binomial with common probability *p* for outcome 1.

R, randomization. The set of values for the *n* districts is given; the assignment to the individual districts is random. There are *n!* permutations. In the case of binary variables this amounts to a fixed number of marked districts.

Rk, rank statistics. The statistic is based on the ranks of the *n* values rather than on the values themselves. This makes the statistic distribution independent.

A test under assumption R is a conditional test: it is performed under the condition that the numerical values of the outcomes for all districts are known, only the proper assignment to the districts is random. Conditional tests are always valid tests in the following sense: if α, the error probability of first kind, is the same under each condition (in our case: for each possible set of outcomes), then α is also the error probability of first kind for the unconditional test. However, the best conditional test need not be the best unconditional test (for a specified meaning of 'best').

The null hypothesis assumes usually a common variance for all districts. This condition is not fulfilled if the variables are quotients m_i/d_i with largely differing denominators (individuals at risk in the medical literature); the numerator (numbers of cases in the medical literature) counts the subpopulation with a distinct property. The diversity in the variances affects even the randomization assumption R, see Besag and Newell (1991, Section 2.1), as well as the proposed remedy to replace the quotient m_i/d_i by the probability of exceeding this quotient under the assumption of a Poisson distribution with an expectation proportional to d_i (i.e. expectation λd_i with $\lambda = \sum m_i / \sum d_i$). The reason is that districts with a large population tend to be more homogeneous (under the null hypothesis) so that small as well as large quotients tend to be concentrated in areas with small denominators, e.g. rural areas, with an increased chance of exhibiting there a cluster of either small or large values. On the other hand, small deviations from the null hypothesis are more likely to become apparent in districts with large denominators favouring there a spatial clustering.

The effect of different denominators d_i on some tests is checked in an example given by Waller and Turnbull (1993); sometimes the test results differ markedly.

To reduce the influence of differing variances, the quotients m_i/d_i may be replaced by their square roots or, in case of small numerators, by the Freeman-Tukey square root transformation $\sqrt{m_i / d_i} + \sqrt{(m_i + 1) / d_i}$; an example is explored in Cressie and Read (1989).

2.5 Monte Carlo tests

For designing an exact test of a hypothesis, it is necessary to know the distribution function of the test statistic under the null hypothesis or at least its α percentile. This is the case only in few situations, see e.g. Section 4.1, Moran's *I*; however, the computation is quite involved.

Often one knows only the first two moments of the distribution. Then its percentiles are approximated from those of the normal distribution with the same expectation and variance or, if higher moments are available, from those of a χ^2- or beta distribution; for an example based on a particular f_{ij} in (3), see Costanzo et al. (1983).

Another possibility is to find an approximation for the distribution function by Monte Carlo simulations, for instance by computing 10000 random values. This is in particular feasible if the distribution function does not depend on further parameters (or perhaps on a single parameter; then the computation has to be repeated for several parameter values and inter-

polated in between). To find reliable values for the 5% boundary, 1000 repetitions may suffice, but in data mining situations one needs the boundaries for error probabilities below 1% and thus more repetitions.

Alternatively on can apply a *Monte Carlo test*. From the null hypothesis, $N - 1$ samples are drawn and the real data set is taken as N-th sample. The N statistics computed from the samples are ordered and for a one-sided upper test at level α, the αN largest values are rejected. In particular, if the value computed from the real data is among them, the null hypothesis is rejected for the real data. Correspondingly one proceeds for one-sided tests at the lower end and for two-sided tests. This Monte Carlo test has exactly the error probability α (if αN is an integer), but it is a randomized test: it involves a random component (roughly speaking, in certain cases the test outcome is determined by a chance algorithm); repeating it (with other random samples in the chance algorithm) may lead to a different result. Thus it is not advisable to use $N = 20$ for $\alpha = 5\%$ but rather a much larger sample size.

The Monte Carlo test is widely used. Besag and Diggle (1977) illustrate its application in spatial data analysis with several examples and discuss its usefulness; see also Cliff and Ord (1981, Section 2.7).

3 Area statistics for binary variables

Just to look at a map where the districts are coloured black or white is not a reliable way to find peculiatities such as clusters of black districts; the eye is deceived for instance by different district sizes. The tests collected in this chapter detect primarily concentrations of districts with one colour (or one value of a binary variable). They may also detect extreme lack of clustering by extremely low values of the test statistic.

Generally one has to distinguish *global tests* for the whole region stating just a deviation from randomness and *local tests* testing a particular district (and its surrounding). All tests for binary data in this section are global; the tests for real-valued variables in Section 4 are in part global, in part local.

3.1 Black-black count test

Probably the best-known test for clustering in binary data is the black-black count test. The districts are either black or white. The test compares essentially the number of neighbourhoods of two black districts with the total number of neighbourhood relations. The black-black count belongs to the join counts admitting more than two classes of districts.

Moments for the distribution of the black-black count statistic are given in Cliff and Ord (1981, Section 2.2) and in many other textbooks. The value for a black district is 1, that for a white district 0. The matrix W is binary; for instance, $w_{ij} = 1$ if districts i and j are neighbours and $w_{ij} = 0$ otherwise; $w_{ii} = 0$. Then the general form of the statistic (3) is

$$BB = \frac{1}{2}\sum\nolimits_{i,j} w_{ij} x_i x_j \ .$$

We need the abbreviations

$$n_1 = \sum\nolimits_i X_i , \qquad S_0 = \sum\nolimits_{i,j} w_{ij} ,$$

$$S_1 = \frac{1}{2}\sum\nolimits_{i,j} \left(w_{ij} + w_{ji}\right)^2 , \qquad S_2 = \sum\nolimits_i \left(w_{i.} + w_{.i}\right)^2 \qquad (4)$$

with $w_{i.} = \sum_j w_{ij}$ and $w_{.j}$ correspondingly, as is customary in statistics. For black-black counts, $w_{ij} = w_{ji}$ and $w_{i.} = w_{.i}$, which is the number of neighbours of that district. If the w_{ij} take only the values 0 and 1 and $w_{ij} = w_{ji}$, then one gets $S_1 = 2S_0$.

Under Assumption N of Section 2.4 (of course with binomial distribution, parameter p) the first two moments under the null hypothesis are:

$$\mathbf{E}(BB) = \frac{1}{2} S_0 p^2,$$

$$\operatorname{var}(BB) = \frac{1}{4} p^2 (1-p)\left[S_1(1-p) + S_2 p\right] \qquad \text{under Assumption N.}$$

This test assumes that the proportion p is known in advance (or estimated independently of the sample to be analyzed).

Under Assumption R the total number of black districts, n_1, is given. The formulae for the moments are more complex due to dependence between the districts.

$$\mathbf{E}(BB) = \frac{1}{2} S_0 \frac{n_1(n_1-1)}{n(n-1)}, \tag{5}$$

$$\begin{aligned} 4 \cdot \operatorname{var}(BB) = {} & S_1\left[\frac{n_1(n_1-1)}{n(n-1)} - 2\frac{n_1(n_1-1)(n_1-2)}{n(n-1)(n-2)} + \frac{n_1(n_1-1)(n_1-2)(n_1-3)}{n(n-1)(n-2)(n-3)}\right] \\ & + S_2\left[\frac{n_1(n_1-1)(n_1-2)}{n(n-1)(n-2)} - \frac{n_1(n_1-1)(n_1-2)(n_1-3)}{n(n-1)(n-2)(n-3)}\right] \\ & + S_0^2 \frac{n_1(n_1-1)(n_1-2)(n_1-3)}{n(n-1)(n-2)(n-3)} \\ & - \left[S_0 \frac{n_1(n_1-1)}{n(n-1)}\right]^2 \qquad \text{under Assumption R.} \end{aligned} \tag{6}$$

This test is the correct one if p is unknown and would have to be estimated from the sample to be analyzed. It may also be used if p is known, see Section 2.4.

Assuming that an approximation by the normal distribution is good enough, expectation and variance of *BB* can be used in the usual way for testing whether the number of black-black counts is too large indicating clustering or too low indicating a trend to regularity (the black districts are too evenly spread). According to simulations with several areas (37 to 171 districts) in Gebhardt (1998b, Appendix A), the deviations from a normal distribution are in the order of the expected sample fluctuations for 1000 iterations; however, for 10000 iterations, some deviations become apparent: the upper tail is still thicker, the lower one thinner than for a standard normal variable. To be more precise, we give an example. The tail probability at $z = 2.5$ varies mostly between 0.8% and 1.6% depending on the percentage of marked districts (for the standard normal distribution, the tail probability is 0.62%); at $z = 3.0$ it varies between 0.2% and 0.5% (0.14% for the normal distribution). The largest deviations occur if about 10% of the districts are marked. A theorem stating sufficient conditions for asymptotic normality is stated in Cliff and Ord (1981, Section 2.4.2).

Taking as an example the counties with an alien rate > 0.11 in Figure 1, we find 24 black districts out of 171. There are 876 neighbour relations, i.e. $S_0 = 876$; furthermore, $S_2 = 22\,896$, $BB = 37$, $\mathbf{E}(BB) = 8.63$, $\operatorname{var}(BB) = 19.98$ under Assumption N. Thus, *BB* is more than

6 standard deviations off its expectation; the result is highly significant, as one would expect. Under Assumption R, **E**(*BB*) = 8.32 and var(*BB*) = 30 so that the result is even more significant.

3.2 Variant of the black-black count test

An excess of black-black counts may be due to a clustering of the marked (black) districts; it may also come from a concentration in districts with many neighbours (usually large districts). Similarly, a deficit of black-black counts may be caused by a concentration on districts with few neighbours: small districts or districts along the border of the whole region.

These are not the deviations from randomness that one is interested in if one is looking for spatial clusters. Therefore a variant of the black-black count is proposed in Gebhardt (2000, 1998b, 1997b).[1]

The expectation of *BB* in (5) is computed under the conditions that the total number of black districts is given (Assumption R of Section 2.4) and that the total number of neighbours of the black districts is given as well. We will call this *"Assumption R*"*. Let $T_0 = \sum_{i,j} w_{ij} x_j$ denote the total number of neighbours of the marked districts (x_j =1) and $T_2 = \sum_j \left(\sum_i w_{ij} x_j \right)^2$. Then with a slight approximation one obtains

$$\mathbf{E}(BB) = \frac{1}{2} \frac{T_0^2 - T_2}{2S_0 - T_0 / n_1} \qquad \text{under Assumption R}^*.$$

The variance under these two conditions is unknown but it should be smaller than that of *BB* under Assumption R in Section 3.1. This has been verified by numerous tests with different regions, see Gebhardt (1998b, Appendix B). In these cases it has been found that equation (6) overestimates the variance slightly (usually by 0 to 3%). So we are on the safe side (the real error of first kind is slightly smaller than expected).

From our example, Figure 1, we now find **E**(*BB*) = 10.0 under Assumption R*. This is larger than the corresponding value in Section 3.1 and reflects the fact that the black districts have an above-average number of neighbours (5.62; the average is 5.43). The clustering is still highly significant.

3.3 Black-white count test

Instead of looking for an excess of black-black counts one might search a deficit of black-white counts *BW*, i.e. neighbourhood relations between a black and a white district. In our general expression (3) we have to insert $f(x_i,x_j) = (x_i - x_j)^2$. Thus we define

$$BW = \frac{1}{2} \sum_{i,j} w_{ij} \left(x_i - x_j \right)^2 .$$

Defining *WW* analogously to *BB*, one has $BB + BW + WW = S_0$.

Under Assumption N of Section 2.4, the first two moments are according to Cliff and Ord (1981, Section 2.2), using $q = 1 - p$,

$$\mathbf{E}(BW) = S_0 pq, \qquad \text{var}(BW) = S_1 pq + \frac{1}{4} S_2 pq(1 - 4pq).$$

[1] Note that the statistic *b* in these publications counts each neighbourhood relation twice, i.e. $b = 2BB$.

Under Assumption R, using $n_2 = n - n_1$, one has according to Cliff and Ord (1981, Section 1.5)

$$\mathbf{E}(BW) = S_0 \frac{n_1 n_2}{n(n-1)},$$

$$var(BW) = \frac{1}{4} S_2 \frac{n_1 n_2}{n(n-1)} + \left(S_0^2 + S_1 - S_2\right) \frac{n_1(n_1-1) n_2(n_2-1)}{n(n-1)(n-2)(n-3)} - \mathbf{E}(BW)^2.$$

Similar formulae hold for more than two classes of districts, see Cliff and Ord (1981, Section 1.5). The complement, the number of neighbours belonging to the same class, is approximated by a χ^2 distribution in Ohno et al. (1979) and applied to five classes of cancer mortality in 1123 Japanese districts. The generalization of the *BW* count to more than two classes, counting then the neighbourhoods involving different classes, seems to be more accurate, while Ohno's approximation is easier to compute.

Under Assumption N, we find for the example in Figure 1 (Section 1.3) *BW* = 61, **E**(*BW*) = 105.7, var(*BW*) = 568.7, i.e. *BW* is too low by only 1.9 standard deviations. This insignificant result is due to the low portion of marked districts, 14%.

3.4 Triplet-based cluster test

A completely different approach to searching concentrations of marked districts is proposed in Gebhardt (1997b, 1999). Rather than counting neighbourhood relations between marked districts, the size of marked subregions is examined. The first idea that comes to mind is to consider the size of a connected subregion that consists of marked districts only and whose neighbours are all unmarked; such a subregion is called a *connectivity region*. It turns out, however, that this is a bad choice. Except if there are very few marked districts (perhaps less than 5 %), marked districts show a considerable degree of connectedness even under the null hypothesis of random distribution. A stronger requirement for clustering is needed.

For this purpose, marked triplets of districts are introduced. Essentially a triplet consists of three (marked) districts with common boundaries. Sometimes four or more districts meet at a point. For such situations a more involved definition is needed.

A *triplet* is a set of three (marked) districts with a point in common such that one of the districts shares with either of the other two a boundary ending in that common point. A *triplet-based cluster* or in short a *triplet cluster* is a maximal set of overlapping triplets. The test statistic *TC* used is the total number of districts belonging to triplet clusters.

Sometimes there are districts with only one neighbour; these could never belong to a cluster. Therefore the cluster definition has to be expanded: if the single neighbour of a marked district belongs to a cluster, this district is added also.

To use *TC* one needs its distribution under the null hypothesis or at least its expectation and variance (assuming the significance limits of a corresponding normal distribution are sufficiently close). These depend on the region (on its topology, not just the number of districts) and on the probability of a district being marked or, under Assumption R of Section 2.4, on the number of marked districts n_1. A theoretical derivation is intractable; therefore one has to find the distribution by simulations.

Using different regions it turns out that expectation and variance of *TC* under Assumption R are approximated surprisingly well by fourth degree polynomials in n_1 with missing low-order terms:

$$\mathbf{E}(TC) \approx a_3 n_1^3 + a_4 n_1^4,$$
$$\mathrm{var}(TC) \approx b_2 n_1^2 + b_3 n_1^3 + b_4 n_1^4.$$

Here we assume $n_1 \le n/2$; otherwise one should look for clusters in the unmarked districts. To find the coefficients one needs simulations at three different values of n_1; using four or five values one gets in addition a confirmation for the fit of the polynomials.

The test statistic *TC* can also be used for detecting regularity, i.e. marked districts that are too evenly spread, if n_1 is so large that the lower confidence limit exceeds 0. In the examples used, this is roughly the case for $n_1 > n/4$.

The values for *TC* are relatively small integers. As a result, only few distinct error probabilities of first kind are available for a given region and number of marked districts. In addition, the approximation by the normal probability function is relatively poor. Therefore, the use of *TC* is not recommended for small *n* (smaller than about 50), and for rather small values of n_1 because then only rather strong clusterings can be detected. For regions with 50 to 100 districts, *TC* is applicable if moderate precision of the error probabilities suffices; this is usually the case in data mining situations. For larger regions, these shortages of *TC* become less restricting and should not exceed the imponderabilities inherent in this kind of tests anyway.

Using again the example of Section 1.3, Figure 1, we get two clusters with 19 districts out of the 24 black districts. The expected number of black districts in clusters is only 1.9 with a standard deviation of 2.2. This means that (applying a continuity correction) the actual value is by 7.6 standard deviations off the expected value and therefore again highly significant.

3.5 Comparison of the test statistics

All statistics mentioned above take essentially only integer values. When using an approximation by a normal distribution one should therefore apply a continuity correction. This is particularly important for *TC* because its values are markedly smaller than those for *BB* or *BW*.

The standard way to compare several test statistics is to examine their performance under the counter hypothesis or counter hypotheses – if these can be specified. If there is only one counter hypothesis, the maximum likelihood test is best in the sense of minimizing the error probability of second kind for a given error probability of first kind. Usually there are lots of possible counter hypotheses; then little more can be done than selecting some of them and examining the performance of the proposed tests with respect to these.

This procedure has been used for comparing the four tests introduced in the preceeding sections. For more details see Gebhardt (1998b, Appendix B). Several regions have been used, mainly a honeycomb of 91 hexagons (and for some simulations honeycombs of 37 to 169 hexagons), Bonn with 62 statistical districts and demographic data, 80 election districts in eastern Germany with election data, 94 départements in France with disease data, 171 counties in north-western Germany with demographic data and an artificial region (called 100-web) with 100 districts with either very few (mostly four) or very many neighbours (mostly twelve).

Three groups of counter hypotheses have been used. In the first group, *model A*, the marked districts have been selected randomly with changing probabilities: if a neighbour of a district has just been marked, the weight of that district is multiplied by a factor. The probabilities for selecting the next district are proportional to these weights.

The second group, *model B,* uses general autocorrelation models, i.e. the autocorrelation follows the same law in all parts of the region. The variable is continuous and transformed into binary data by selecting the n_1 districts with the largest values. The models (1) and (2) belong to this type.

Such a dichotomy is also used in the third group, *model C,* but the continuous variables are independent (standard Gaussian) and a fixed value, typically 1.0 to 1.5, is added to a randomly selected district, a smaller value to its neighbours, still a smaller value to their neighbours. This creates a moderate hump around the first district which however is obscured by the original random variables. For all three groups, one or two parameters determine the degree of deviation from the null hypothesis. The simulations have mostly been performed for about 10%, 20%, 30%, 40% and 50% of the districts marked.

As was to be expected, none of the test statistics turned out to be generally the best one with respect to power (the complement of the error of second kind for given error of first kind).

Comparing the power of all four tests at the upper tail, i.e. used as a test for above-average clustering, the main conclusions are the following. The variant of *BB*, that is *BB* under Assumption R* (Section 3.2), is called *BB** in this comparison.

- The statistic *BB* and its variant *BB** have about the same power; sometimes *BB* is better, sometimes *BB**. However, when *BB* is better, this is mainly due to finding constellations where the marked districts have above-average neighbours. The differences between *BB* and *BB** are larger for areas where the districts have quite different numbers of neighbours (100-web, Northwest).
- The statistic *BW* has about the same power as *BB** for 40% and 50% of the districts marked and becomes worse for smaller percentages.
- The triplet count statistic *TC* has almost always a markedly smaller power than *BB* and *BB**. It is comparable to *BW* for 20% marked, mostly better for 10% and worse for 40% and 50%.

At the lower tail, i.e. used as a test for below-average clustering, model C is not applicable. The main conclusions are the following.

- Again sometimes *BB* has larger power than its variant *BB**, sometimes vice versa, but the over-all advantage of *BB** now seems clearer: *BB* is somewhat better at model A with large areas, but markedly worse at model B with all areas.
- The statistic *BW* has about the same power as *BB** for 50% of the districts marked and becomes worse for smaller percentages.
- The statistic *TC* is comparable to *BW* at model A and to *BB* at model B for 30% or more marked; it is not applicable for small percentages (roughly, below 25%) due to the strong skewness of its distribution.

So which test statistic to take? That depends, obviously. The theoretically best founded tests are the standard black-black count and the black-white count; the former one is clearly better for small percentages of marked districts. Without additional information 35 to 40% may be the break-even. The variant *BB** may be considered if either one wants a single statistic for all proportions of marked districts (up to 50%) or if one knows that the counter hypothesis is model B or if one wants to compensate for concentrations of marked districts with either rather few or rather many neighbours. The triplet test has no justification on grounds of the test power but should be taken into account if the explainability of deviations from randomness is an issue: it is easier to interpret a compact cluster of marked districts detected by *TC* than long filaments or a general tendency for small clusters causing significance in *BB* or *BW*.

It has been stated in Section 6.4.4 of Cliff and Ord (1981) that *BW* is asymptotically (large regions size *n*) better than *BB* using the asymptotic relative efficiency as a measure for comparison. However, this is not generally true. The proof uses particular counter hypotheses (autocorrelation models) and, what is more important, it assumes that the variable in question is actually a continuous variable and the marked districts are those where the original variable exceeds a threshold and one is free to choose any threshold. In this particular situation, one should use a threshold that yields $n_1 \approx n/2$ marked districts and for this special constellation *BW* should be at least as good as, or better than, *BB* under Assumption R and considerably better under Assumption N.[2] Our simulations show that for small portions of the marked districts *BB* is clearly better than *BW*.

This general result is illustrated by our example: *BB* (both versions) and *TC* are higly significant, while *BW* is not due to the low portion (14%) of marked districts.

4 Area statistics for real-valued variables

Real-valued variables can be treated as binary variables by using a threshold, but this ignores part of the original information. It is not so obvious how a test statistic for real-valued variables should look like, and several alternatives have been proposed. A choice among them should consider their advantages and disadvantages, beside their statistical power for instance whether they just say yes or no or indicate which part of the area is suspicious. A drawback of some statistics is the sensitivity against deviations from the normal distribution.

As mentioned above, one has to distinguish *global tests* for the whole region stating just a deviation from randomness and *local tests* testing a particular district (and its surrounding); in the latter case, the use as a *general test* looks for clusters anywhere in the region while the use as a *focused test* examines one or more predefined districts, perhaps places with a putative environmental hazard.

Focused tests are not considered in this survey except that some general tests may also be used as focused tests. Some pertinent references are Bithell (1995), Waller and Lawson (1995), Tango (1995), Hills and Alexander (1989). A recent review is Lawson and Waller (1996), which treats, despite its title, not only point data but also area data.

4.1 Moran's I

The standard statistic for testing real-valued area data (*n* districts) on independence is Moran's *I*. It is mostly written as

$$I = \frac{n}{S_0 \sum_{i,j} z_i^2} \sum_{i,j} w_{ij} z_i z_j \qquad \text{with}$$

$$z_i = x_i - \bar{x}, \qquad S_0 = \sum_{i,j} w_{ij}, \qquad w_{ii} = 0.$$

It has been extended to regression residuals z_i and to matrices *W* without the restriction $w_{ii} = 0$, Anselin (1988, Section 8.1.1). It is of the general form (3) with $f_{ij} = z_i z_j$. Obviously, *I* takes on large values if there is a high correlation between neighbouring values of the spatial variable, i.e. if either the large values or the small ones (or both) are spatially clustered.

---

[2] In the cited book the asymptotic relative efficiency of *BB* is given as 0.307 at best. I cannot verify this figure; from the formulae given I compute values around 0.77.

Moran's I indicates a departure from independent observations but does not tell where this departure occurs nor even whether large or small values or both are affected. There may be no clustering at all; for instance, a geographical trend also leads to significant values of I.

The expectation of I is

$$\mathbf{E}(I) = -\frac{1}{n-1},$$

the variance, $\mathbf{E}(I^2) - \mathbf{E}(I)^2$, depends on the assumption on the distribution (Section 2.4):

$$\mathbf{E}(I^2) = \frac{n^2 S_1 - nS_2 + 3S_0^2}{\left(n^2-1\right)S_0^2} \qquad \text{under Assumption N,}$$

$$\mathbf{E}(I^2) = \frac{1}{(n-1)(n-2)(n-3)S_0^2}\left\{ n\left[\left(n^2-3n+3\right)S_1 - nS_2 + 3S_0^2\right] - \frac{n\sum z_i^4}{\left(\sum z_i^2\right)^2}\left[\left(n^2-n\right)S_1 - 2nS_2 + 6S_0^2\right]\right\}$$

under Assumption R

with S_0, S_1 and S_2 according to formula (4), see Sections 1.5.1 and 2.3 of Cliff and Ord (1981), where also the third moment under Assumption N is given.[3]

The second assumption takes the values z_i as fixed and therefore the variance depends on these values while under Assumption N the variance can be computed once and for all for each area.

The coefficient I is sometimes called spatial autocorrelation; however, it is no correlation coefficient. Depending on the weights and the assignment of the z-values to the districts, the maximal value of I is mostly less than, but occasionally larger than, 1. In order to make I more similar to a correlation coefficient, it is sometimes divided by the expectation of its maximal value under Assumption R, Bailey and Gatrell (1995, Section 7.4.5), but still its range is not exactly [-1, 1]. The *exact* limits of I, if needed, can be computed from the eigenvalues of a matrix involving W, see Tiefelsdorf and Boots (1995) and the correction Tiefelsdorf and Boots (1996).

Moran's I is a quotient of two quadratic forms. Its exact distribution under the null hypothesis and normally distributed variables is known, see the cited articles by Tiefelsdorf and Boots; the probability $P(I>I_0)$ for any I_0 can be written as a one-dimensional real integral of a function of the eigenvalues of an n-dimensional matrix involving the matrix W and, in the general case, the regression matrix X (a column of 1's if the z_i are simply the deviations from the mean).

For moderate deviations from normality, Cliff and Ord (1981) state that the distribution of I is quite robust. This is confirmed by Walter (1992a) for the case of incidence rates in districts with substantially different population sizes. However, Waldhör (1996) finds in a simulation large deviations of the mean, standard deviation and significance levels from the nominal ones (often by a factor > 2) if the variances in the districts are quite different (this is for instance the case if the variable is an incidence rate and the base population differs widely). He

[3] The formula for var(I) pertaining to assumption N given by Bailey and Gatrell (1995, Section 7.5.3) is wrong.

computes the variance of *I* under these conditions and shows in the simulations that the use of this variance improves the significance levels considerably.

According to simulations by the author the distribution function of *I* is quite robust against deviations from normality, see Gebhardt (1998b, Appendix A). Simulations include several regions and symmetric as well as unsymmetric distributions, among them Student's t_3 (a heavy-tailed distribution with variance but no higher moments), uniform distribution, one-sided normal distribution. Therefore *I* can be used whenever an approximate test suffices, e.g. in data mining.

Many authors assume that the distribution of *I* can be approximated by a normal distribution already for $n > 20$. This has also been checked by Gebhardt (1998b, Appendix A). The approximate significance boundaries for the standardized value of *I*, i.e. $\left(I - \mathbf{E}(I)\right) / \mathrm{var}^{1/2} I$, derived from 10000 iterations (in part 20000 iterations) for seven different distribution functions and seven different regions (37 to 327 districts) are given in the following table; most of the simulation results for these distributions and regions deviate from the boundary by no more than the accuracy given in the last line. They show almost no dependency on the number of districts.

significance level (%)	0.5	1	2	5	95	98	99	99.5
significance boundary	-2.37	-2.18	-1.93	-1.57	1.70	2.17	2.48	2.77
accuracy	0.21	0.19	0.13	0.07	0.08	0.12	0.15	0.18

According to these simulations the nominal error probability (from the normal approximation) may be wrong by a factor up to 2 even for $n > 100$. The variances of *I* under both assumptions (N and R) differ little except for rather extreme value sets, in particular if one or very few values are much larger or smaller than all others. In these situations, only time-consuming randomization experiments with the given data values can approximate the significance boundaries.

The distribution of Moran's *I* assumes equal distribution for all districts. This is often not true, in particular if the data are incidence rates and the population at risk in the districts varies widely. The statistics I_{pop} and I^*_{pop} take this into account, Oden (1995), Oden et al. (1996). Conceptually each person is a 'district' and $w_{ij} = 2$ if both persons reside in the same geographical district, = 1 for neighbouring geographical districts and = 0 otherwise. This leads to a statistic in the form of Moran's *I* for the geographical districts but with main diagonal elements ≠ 0. Therefore the approximation converges only very slowly to a normal distribution and it is not appropriate in most cases; an approximation by a χ^2 distribution (using third moments to determine the degrees of freedom) should be used. A closely related statistic is Tango's 'general' test T_G, Tango (1995); for a discussion on both, see Tango (1998).

The example of Section 1.3, using now the real-valued variable "share of aliens among the workers" itself rather than its bisection, yields $I = 15.8$, which is highly significant. Looking at Figure 1, the clusters of high values seem to be the reason, but other causes could contribute, too: there is a general trend (not shown in Figure 1) from low values in the north to large values in the south, and most low values are clustered between Aurich (AUR) and Braunschweig (BS); another cluster with not quite so low values is found around Trier (TR) in the south-west. This is of course no surprise: the low values concentrate in rural areas.

4.2 Local version of Moran's I

Any statistic of the form (3) can be rewritten as

$$\Gamma = C_1 \sum \Gamma_i \qquad \text{with} \qquad \Gamma_i = C_2 \sum_j w_{ij} f_{if} \tag{7}$$

with convenient factors C_1 and C_2. The local indicators of spatial association (LISA) Γ_i are the local contributions to the global statistic Γ. They are also used as local test statistics to find small regions (essentially a district and ist neighbourhood) that deviate from the general pattern. For more details, see later in Section 4.9.

Using Moran's I in (7), the local statistics I_i are defined in Anselin (1995) as

$$I_i = \frac{z_i}{m_2} \sum_j w_{ij} z_j \qquad \text{with}$$

$$z_i = x_i - \bar{x}, \qquad m_2 = \frac{1}{n} \sum_i z_i^2, \qquad \mathbf{E}(I_i) = -\frac{\sum_j w_{ij}}{n-1}$$

and a variance under hypothesis R (Section 2.4) given in Anselin (1995) for general w_{ij}. In our special case (binary symmetric weights, $w_{ii} = 0$) this reduces, using $b_i = \sum_j w_{ij}$ and $m = n \sum z_i^4 \,/ \left(\sum z_i^2 \right)^2$, to

$$\operatorname{var}(I_i) = b_i \left[\frac{n-m}{n-1} + (b_i - 1) \frac{2(2m-n)}{(n-1)(n-2)} - \frac{b_i}{(n-1)^2} \right].$$

Instead of the randomization hypothesis R, a conditional randomization hypothesis may be used for the local statistics: the value x_i for district i is held fixed while the other values are permuted over all other districts. These permutations need, in a simulation study, only be performed to assign the values for the neighbours of district i (the j with $w_{ij} \neq 0$). In this case, the factor z_i / m_2 is irrelevant and one has to consider only the permutations of $\sum_j w_{ij} z_j$ which happen to be equivalent to those of the corresponding Getis-Ord statistics (Section 4.4 below), i.e. the conditional permutations yield the same local statistics.

According to the example and some simulations given in Anselin (1995), the distribution of the local statistics is far from that of a normal distribution so that an approximation by the latter one is not possible. In addition, the distribution of the I_i becomes more and more skewed if the global autocorrelation becomes larger. An approximation by a χ^2 distribution with proper third moment will not help, however, since the problem does not lie in the skewness but rather in the extremely large fourth moment.

The extremely bad approximation by a normal distribution is confirmed by a series of simulations in Gebhardt (1998b, Appendix A) using various regions with 37 to 327 districts. The expectation and variance (under the null hypothesis) are quite stable even under distributions for the spatial variable that are far from normal (the variance is somewhat larger than 1 for small regions, e.g. about 1.14 for a honeycomb with 37 hexagons and 1.08 for Bonn with 62 statistical districts). The third moment of the standardized statistic $\left(I_i - \mathbf{E}(I_i)\right) \operatorname{var}^{1/2} I_i$ is somewhere near –0.5 for the small regions and between –0.2 and 0.2 for the larger ones under normal distribution but very sensitive to the distribution (e.g. it varies for 327 German counties between –0.4 and 0.7 for five distributions that have been investigated). The fourth moment of the standardized I_i varies even more: between 7 and 13 for normal distributions, smaller for uniform distribution and seemingly arbitrarily large for other distributions (for

comparison: the fourth moment of a normal distribution is 3). The confidence limits vary accordingly; while the 1%-boundary in the simulations was mostly between 2.9 and 3.2, the 0.1%-boundary fluctuated between 4 and 8. Since small nominal error probabilities must be used due to the large number of tests (one per district), large boundary values (beyond 5) should be chosen and even then the test is very unreliable.

Strictly speaking, these significance boundaries do not pertain to *the* distribution of the standardized local I_i since the distribution depends on i, at least on b_i. However, simulations show here almost no variability: for a region with 327 districts, the percentage points have been determined separately for districts with 1, 2, …, 8, 9 to 11 neighbours (in each group at least 10000 function values) and the percentage points show no trend whatsoever.

A similar statistic (using different standardization) has been studied in some simulations by Munasinghe and Morris (1996).

The example data of Section 1.3 yield significant values for several counties near Frankfurt. The highest value is 16.0 for Frankfurt itself. Outside this area, the highest value is 7.1 for Solingen (SG) near Köln; due to the very long-tailed distribution of the local statistic, this may not yet be significant. The areas with low values of aliens yield no significant local statistics.

4.3 Geary's c

Geary's c seems to be used as a competitor to Moran's I. It is also of the general form (3) with $f_{ij} = (x_i - x_j)^2$:

$$c = \frac{n-1}{2S_0 \sum z_i^2} \sum_{i,j} w_{ij} \left(x_i - x_j\right)^2, \qquad w_{ii} = 0.$$

Geary's c emphasises differences between neighbours comparable to variograms, while Moran's I resembles a correlation.

The expectation is $\mathbf{E}(c) = 1$; the variance is given in Sections 1.5.1 and 2.3 of Cliff and Ord (1981) and elsewhere. The black-white count (Section 3.3) is a special case: the x_i have only two distinct values.

This statistic takes on large values if the variability in the neighbourhoods is large; thus roughly large values of I correspond to small values of c and vice versa.

4.4 Getis-Ord statistics G_i and G_i^*

Getis and Ord (1992) proposed two statistics for finding local concentrations, G_i and G_i^* for $i \le n$, the number of districts. They differ in that all summations related to G_i are to be taken for $j \ne i$ only. The definitions and the moments are

$$G_i \text{ resp. } G_i^* = \frac{\sum_j w_{ij} x_j}{\sum_j x_j},$$

$$w_{ij} = 0 \text{ resp. } 1, \qquad w_{ij} = w_{ji,} \qquad x_j > 0,$$

$$N = \begin{cases} n-1 & \text{for } G_i, \\ n & \text{for } G_i^*, \end{cases}$$

$$W_i = \sum_j w_{ij}, \qquad X_i = \sum_j x_j \quad \text{(independent of } i \text{ for } G_i^*\text{)},$$

$$Y_i = \frac{\sum_j x_j^2}{N} - \left(\frac{X_i}{N}\right)^2,$$

$$\mathbf{E}(G_i^{(*)}) = W_i / N, \qquad \mathrm{var}(G_i^*) = \frac{W_i(N - W_i)Y_i}{(N-1)X_i^2}.$$

The variances are computed under Assumption R of Section 2.4.[4]

The statistics are scale-invariant but not location-invariant. They are intended for use only for variables possessing a natural origin. Large values occur if the neighbours of district *i* (excluding district *i* itself in the case of G_i, including it in the case of G_i^*) have large values. It seems more natural to include district *i* in the statistic, i.e. to use G_i^* rather than G_i. Getis himself (1994) proposes only G_i^*.[5]

The distribution of G_i and G_i^* may be far from normal; this is suggested in Anselin (1995) due to their similarity to the local version of Moran's *I*. The use of G_i seems a bit awkward if some districts have no neighbours, for instance if $w_{ij} = 1$ for districts within a given distance (between their centers).

When using these statistics one should be aware that one is performing *n* tests (which, though, are not independent) with obvious effects on the total error probability. So it is not really surprising that Getis and Ord (1992) find five significant counties out of 100 at the 0.05 level in their first example (sudden infant death syndrome in North Carolina).

If a G_i^* (or a G_i) exceeds the confidence limit, a local concentration for this variable has been found. Note that the potential 'clusters' are predefined by the rows of *W* (and unions thereof).

A software system to analyze spatial data using G_i^* is REGARD, Unwin (1996).

4.5 Rank statistics

Instead of the original variables, one can use the ranks in the general formula (3). Of course, the ranks are not independent, even under the null hypothesis of independent original variables. Nevertheless, mean and variance for this rank statistic can be computed at least for some functions f_{ij}.

This has been done by Walter (1994) for $f_{ij} = |r_i - r_j|$ in our notation for symmetric and binary weights. This statistic goes back to Kemp et al. (1985). For

$$D = \frac{1}{S_0}\sum_{i,j} w_{ij}\left|r_i - r_j\right|$$

one gets

$$\mathbf{E}(D) = \frac{n+1}{3}, \qquad \mathrm{var}(D) \approx \frac{n(n-1)-2}{18S_0}.$$

Note that *small* values of *D* indicate a positive correlation of the ranks and thus a clustering of the original variables. The equation for $\mathbf{E}(D)$ holds for any weights; Walter (1994) gives also an approximation for var(*D*) for arbitrary weights w_{ij}. In addition he shows by means of three

---

[4] The original article contains a serious error. In Table 1, the expression for Y_{i2}^* should read $\sum_j x_j^2 / n - (Y_{i1}^*)^2$.

[5] In this article, the definition of G^* contains a misprint: the summation in the numerator is to be from 1 to *n*.

examples that the distribution can be approximated quite well by a normal distribution. This is confirmed by two examples in Möhner (1991) (219 districts in east Germany, 56 districts in Scotland) with relative errors below 8% and 10%, respectively; his formula for var(D) is in error, however.

This rank test is a global test. A significantly low value of D may have quite different reasons (just as large values for Moran's I), among them a clustering of high (or low) values of the original variable or a concentration of medium values in part of the region while the values are randomly distributed in the rest or a global trend with otherwise random distribution so that high values tend to occur at one end of the region, low values at the opposite. In the case of clustering, there is no apparent indication whether small or large values are clustered and how far the cluster extends.

The rank statistic D seems to be widely used in medical applications such as health atlases although its properties are largely unknown, Walter (1992a). According to this investigation, the actual tail probability of D can differ substantially from the nominal one if the values for the districts have widely differing variances. On the other hand, even a moderately large number of ties, in particular zero counts, has no severe effects.

4.6 Comparison of I, c, and D

According to Cliff and Ord (1981, Section 6.4.3), Geary's c is asymptotically somewhat worse than Moran's I in the following sense: under a specific counter hypothesis (Whittle's model (1)) and for large regions, the asymptotic relative efficiency is ARE$(c, I) \leq 1$ while for the maximum likelihood statistic λ one finds ARE$(I, \lambda) = 1$, i.e. Moran's I is asymptotically fully efficient. The difference between I and c may be small, and in fact it vanishes for certain regular regions (e.g. honeycombs and chess boards). The asymptotic results also do not imply that for finite regions or other counter hypothses I is better than c although in some examples this seems to be the case.

Walter (1992b) finds in some simulations that the power of I is somewhat better than that of c. In addition, c is much stronger affected by varying variances in the districts (for instance incidence rates based on different population sizes) than I, Walter (1992a).

Thus if there is no strong other criterion one should prefer I to c.

A series of simulations in Walter (1992a) suggests that D has generally somewhat less power than Moran's I or Geary's c, which can be attributed to the loss of information (ranks rather than the numerical values in the districts). In addition, its power is severely lower in certain configurations such as very small hot spots (a few neighbouring districts with high values) or long filaments of districts with high values (in this case counties in Ontario with high values along the Great Lakes).

4.7 Triplet clusters

Sometimes one is interested in the question whether a spatial variable deviates in a certain region substantially from the rest of the area. If the region is given in advance, for instance for county data a Bundesland (state) or an industrial area, this can be handled as a linear model, in the presence of autocorrelations using the procedures mentioned at the end of Section 4.9.

Obviously this is not possible for unknown regions. There are just too many connected subsets of a given area, many thousands of them, that could qualify as a test region. To alleviate this problem, one can combine two ideas: to reduce the number of potential test areas by requiring some degree of compactness and to take account of the number of such test areas.

This has been tried in Gebhardt (1998c). The procedure restricts the test areas to the sets of overlapping triplets (as in Section 3.4) and involves a heuristic for the rough number of such test areas in a region. Conspicuous regions are searched in a kind of beam search: starting with triplets, one finds the most outstanding clusters of size k by adding a triplet to the most outstanding clusters of size $k - 1$ and $k - 2$ (the new triplet has 2 or 1 district, resp., in common with the old cluster). The error probability pertaining to a cluster is estimated under the randomization hypothesis R using an approximation to the number of permutations of the observed values that would yield a higher average for that cluster. Simulations have been performed with several regions of 62 to 171 districts, some of them with rather extreme connectivity properties, and with several distribution functions for the district data, again some of them with rather extreme properties. The actual error probability of first kind for a given k stays in most cases below the nominal one; the combined error probability (all k between 3 and 8 to 10) mostly exceeds the nominal one by a factor 1.5 to 3, which should be tolerable in data mining situations.

A comparison of the power with Moran's I, again by simulations, depends heavily on the counter hypothesis chosen. In general autocorrelation models, I has, not surprizingly, the higher power; in a model adding a moderate hill to otherwise random data sometimes I and sometimes the triplet cluster test has higher power, but of course I does not indicate the reason for its significance (let alone show a cluster) while the triplet test identifies one or more (in general, overlapping) clusters.

An earlier proposal, Gebhardt (1997a, 1998a), is not recommended since it is too sensitive against deviations from the normal distribution.

The example data of Section 1.3 yield an extremely significant cluster (error probability well below 10^{-7}) around Frankfurt: F, OF, DA*, GG, and OF*. Successively larger clusters are also highly significant, e.g. F, OF, DA*, GG, HG, MTK, OF*, and RÜD still with an error probability below 10^{-6}. Considering low values of the share of aliens, the cluster EMD, OL, WHV, WST, AUR, CLP, EL, FRI, LER, OL*, and WIT in the north-western corner has an error probability of 0.0002. Note that these are the error probabilities for the combined test randomness vs. *any* triplet-based cluster while the local statistics test *one* district (and its neighbourhood) at a time. Thus these clusters stand out much more clearly than with the local version of Moran's I.

4.8 Other cluster tests

There exist various other proposals for cluster tests. I want to mention here one by Kullforff and Nagarwalla (1995). Under the null hypothesis, the probability to be a case is the same for all individuals; under the alternative, it is increased by an unknown factor in one region out of a family of regions (e.g. circles of arbitrary radius around any one of a number of grid points). The test statistic is the *maximum likelihood statistic*. Its distribution is unknown; therefore the error probability is found by a Monte Carlo testing procedure, see Section 2.6.

A test for *clusters in rare diseases* is proposed by Besag and Newell (1991). Let us assume that there is a large number n of districts with populations at risk t_i and number of cases y_i for a rare event (e.g. disease) such that the number of cases in a district is Poisson distributed under the null hypothesis of independence. The test statistic counts the number of nearest districts to any case (including the district of that case) such that the total number of cases exceeds a limit k. The distribution of this statistic depends on the populations at risk in these districts, i.e. it differs from one district to the next. If for a case the number of districts to exceed the given limit is too *small*, this case indicates the center of a cluster. The example

presented has 16183 districts and 496 cases. Choosing different limits k may sometimes yield substantially different results, see the example by Waller and Turnbull (1993). Obviously, the tests for neighbouring districts are correlated. Using a moderate significance level for each single test (say, 5%) implies a sizeable expected number of cases with positive outcome, 25 in the example. Thus the test can only be preliminary, pointing at regions with a potential clustering.

There exist other tests for clusters in rare diseases. A test by Stone (1988) uses essentially the maximum incidence rate in a sequence of growing surroundings of a 'hot spot', e.g. a contamination source. A test by Waller et al. (1992) tests uniform incidence rates (null hypothesis) against incidence rates decreasing with the distance from the hot spot; it is locally uniformly most powerful.

Given two variables for the districts of an area, one is interested in any possible association between them. The usual tests are not applicable. Due to spatial correlation of each variable separately, the ordinary estimator s for the standard deviation σ is not unbiased; a seeming association may be an artefact produced by the autocorrelations of both variables.

A test of correlation between two variables with autocorrelation therefore needs a modification of the standard t-test; σ must be estimated differently. The somewhat clumsy formulae are given in Cliff and Ord (1981, Section 7.2). Similarly, the standard procedure is invalid for spatial regression. The variances must be computed from more complicated models as in Cliff and Ord (1981, Section 7.3.2) and in Bailey and Gatrell (1995, Section 7.5.4).

For a test of the null hypothesis 'no correlation between two spatial processes measured at the same points or districts' see Clifford et al. (1989).

4.9 Global and local tests

There are two types of statistical tests for real-valued spatial data, global and local ones.

A global test checks the whole area at once. If it is significant, one does not know in general where in the area the deviation from randomness occurs. In fact, there need not be a particular region producing the significance; the reason can be a strong correlation between neighbours throughout the area or a trend from one edge to the opposite one.

A local test checks whether a particular district and its neighbours as they are specified by the proper row of the weight matrix W deviate from randomness; which property exactly deviates is specified by the function f_{ij} in (3). So one can either speak of suspicious districts or of suspicious clusters comprising the district and its neighbours. The results are more informative than those of a global test pointing to the relevant region within the area. However, one needs n tests, which is not only more work but, more seriously, diminishes the worth of the tests: either one must use a tiny error probability of first kind for each district or one gets a huge over-all error probability and therefore lots of chance results.

A way out could be to use a global test for finding out whether there is an irregularity at all and if so to find the conspicuous region by the most outstanding local statistic. However, Moran's I as the most prominent global spatial statistic may be insignificant while the largest local Moran I is strongly significant or vice versa; this can be demonstrated by using random values for the districts (standardized gaussian) and adding either a global trend or a fixed constant for a small or large region. While both Moran's I and the largest local statistic are highly correlated with that constant, there is little correlation between both if the constant is held fixed (unpublished simulations by the author).

The distinction between global and local tests does not hold for the triplet-based tests, Sections 3.4 and 4.7. These are global tests (one test per data set), but at the same time they indicate the suspicious region. This is an advantage in data mining where one is interested in *interpretable* peculiarities of the data.

4.10 Descriptive analysis

Classification of spatial data requires that not only the data in a class should be similar but in addition the groups should be contiguous. This problem is treated in Johnston (1976).

Spatial hierarchies given in advance are utilized for data mining, for instance for finding characteristic rules, in the database mining system prototype GeoMiner, see Han et al. (1997), Koperski and Han (1995).

Descriptive spatial statistics include *autocorrelograms*, the correlation of a variable between areas (or points) as a function of the distance, in particular points that can be reached in 1, 2, … steps. The correlograms are probably more useful for regular grids than for irregular tesselations. Similarly, *variograms* show the variance of the difference of two values as a function of the distance. Autocorrelograms and variograms can also be constructed to show the dependency on distance and direction.

Another means of descriptive statistics is the *Moran scatterplot*, a scatter diagram of z_i vs. $\sum_j w_{ij} z_j$, see Anselin (1995). Points lying astray from the others could be measurement errors or outliers.

5 Some remarks to point data and spatially continuous data

This chapter gives some hints to the treatment of point data and continuous data as far as the methods are related to those for area data.

5.1 Point data

Sometimes point data (in particular samples from a continuous variable) can be treated as area data. Districts are constructed by assigning each point of the region to its nearest sample point (Dirichlet tesselation, also called Voronoi or Thiessen polygons).

Point data as we introduced them are in fact a collection of different data types requiring different methods for analysis.

Locational data consist purely of the points where certain events occurred. This is also referred to as event data or a point process. If several types of events are involved, it is called a *marked point process*. Methods for analysing point processes are explained in several textbooks, e.g. Diggle (1983), Upton and Fingleton (1985), Ripley (1988). The latter one assumes good knowledge in stochastic processes and Bayesian analysis. It shows that under various circumstances a simple adoption of time series results leads to wrong conclusions; two- or more-dimensional processes behave quite different from one-dimensional ones, in particular they tend to be much less robust (e.g. against misspecifications and the border effect). Simple descriptive tools are quadrat counts, the number of points in regularly or irregularly spaced quadrats, see Cliff and Ord (1981, Section 4.1) and Upton and Fingleton (1985, Sections 1.1 to 1.3).

A different problem is to find clusters in a multi-dimensional set of points. This can be done by various means of cluster analysis, Ester et al. (1998). Another approach is based on a minimum density of points for forming a point cluster, see Ester et al. (1996), Sander et al. (1998).

In the case of marked point data (two types: black and white), an often used test seems to be that of Cuzick and Edwards (1990). For each point, the *k* nearest neighbours are determined; the test counts how many black points are among these nearest neighbours of black points. Here *k* is a small integer, possibly even 1. Obviously this test has some similarity to the black-black cout test in Section 3.1 using points rather than districts and an unsymmetric matrix *W* (if point *i* is among the *k* nearest neighbours of point *j*, the reciprocal need not be true).

Attribute data consist of attributes attached to points. The attributes can often be analyzed with the methods described for area data. These methods assume neighbourhood (or distance) relations but it is not essential whether this is in reality a contiguity between districts or a conveniently defined neighbourhood between points. One way to define neighbourhood of points is the Dirichlet tesselation mentioned above.

Interaction data are data associated with two (or more) points such as travel time or exchange of goods. Chapter 9 of Bailey and Gatrell (1995) is devoted to this data type.

If the points are actually samples of a continuous function, the problem of interpolation arises, see the next section. Interpolation methods may also be applicable to point data, for instance to estimate the point density underlying the observed points (events) of a point process.

5.2 Continuous data

Sometimes continuous data are known only for selected points. Then methods for point or area data may have to be used. Another task is interpolation, that is to find estimates for the variable at points where it is not measured. Several smoothing techniques exist such as spatial moving averages or, for regular grids, median polish. A more sophisticated method is kernel smoothing where the smoothed value is essentially a weighted average over the values at all other points, the weights depending on the distance, see various sections in Bailey and Gatrell (1995).

Kriging is a class of estimation methods named after the South African mining geologist D. G. Krige who developed an early version of it. The idea is as follows. Consider the spatial process $y(s) = f(s; \beta) + u(s)$ with unknown parameter β and a zero-mean process $u(s)$ with known or estimated covariance $C(s, s')$ for points s and s'. The values $y(s)$ are measured for some points $s_1, \ldots, s_N$; from these values the estimate $\hat{\beta}$ is derived. Then one can do better in estimating $y(s)$ than using $f(s;\hat{\beta})$: $u(s)$ can also be estimated as a linear function of $u(s_i)$ where the coefficients turn out to be a linear function of the correlation matrix C between the points s_i and of the correlation vector c between point s and the points s_i. The method is widely referred to but rarely described in the textbooks on geographical data analysis; chapter 5 of Bailey and Gatrell (1995) is an exception. See also Oliver and Webster (1990), Isaaks and Srivastava (1989).

Uncritical use of kriging may lead to debatable results. In the French disease surveillance system Sentinelles, see Toubiana and Flahault (1998), available in WWW under address `http://www.b3e.jussieu.fr:80/sentiweb/en/sommaire.html`, area data (incidences of diseases per 100 000 inhabitants) are displayed alternatively by département or as smoothed contours created by adapting kriging to area data. Sometimes both displays look entirely different, e.g. for measles in the third quarter of 1996.

6 Discussion and conclusions

6.1 Related work

There exist several *textbooks* on geographical data analysis. Some of them will be briefly reviewed here.

The classical textbook, cited over and over, is by Cliff and Ord (1981), based on an even older book by the same authors, Cliff and Ord (1973). It is out of print. The major theme is the treatment of spatial data as a process with autocorrelation. The book is mainly on area data; one of the nine chapters covers point data. If a spatial correlation is established, there are two basic types of models to explain it, either interaction between neighbouring areas or dependency on other spatial variables, as well as a combination of both. These models are treated in the second half of the book. The work is mathematically oriented including proofs to all the propositions.

The textbook Bailey and Gatrell (1995) has the main parts introduction, analysis of point patterns, analysis of spatially continuous data, analysis of area data, analysis of spatial interaction data (meaning interactions between locations; examples are traffic or exchange of goods). Proofs are in general omitted. Attached to the book is a disk with the program INFO-MAP and some small data sets.

Econometric models in space or in space and time are the subject of Anselin (1988). Some of the problems treated are estimation and hypothesis testing with maximum likelihood methods, multiple regression with spatially dependent error terms, testing for spatial heterogeneity and space and time models. The procedures are in general too advanced for routine use as in data mining.

'Spatial Analysis and GIS' by Fotheringham and Rogerson (1994) is not a textbook but a collection of articles. The problem areas concerning the analysis of spatial data and the use of geographic information systems are treated rather informally.

Often cited is also Cressie (1993). The emphasis is on spatial or space-time processes, i.e. on the joint distribution of random variables (mostly Gaussian) under various covariance structures. The book cites almost 1400 references.

The pair of books Upton and Fingleton (1985), Upton and Fingleton (1989) is comparatively easy to read. Emphasis is on applying adequate techniques; thus there are many examples, mostly from biology and geography or a combination of both. The rationale of the methods is developed but the mathematical derivations are mostly omitted and many practical suggestions are added such as pitfalls in applying a method thoughtlessly.

Two other books on spatial data analysis are Haining (1990) that addresses itself primarily to social scientists and Isaaks and Srivastava (1989), which concentrates on continuous data (measured at selected points).

However, there exist many more books on the subject.

Spatial analyses have been performed already for a long time in *medicine*, such as the spread of infectuous diseases. There exist numerous publications, mostly, however, with standard statistical methods only; regional inspection is usually done by eye. If statistics like Moran's I or Gearie's c have been applied, then often uncritically. Many investigations concern cancer as well as some rare diseases (e.g. leukaemia with emphasis on the area around nuclear plants). The Journal *Statistics in Medicine* has several issues devoted to statistics and computing in disease clustering: Volume 15 no. 7–9, 1996 (conference at Vancouver, July 1994);

volume 14 no. 21-22, 1995; volume 12 no. 19–20, 1993 (workshop at Port Jefferson, New York, July 1992).

Incidence data on various diseases are collected in health atlases, see the survey of 49 atlases by Walter and Birnie (1991).

6.2 Some caveats

It is easy to come to wrong conclusions in data mining in general, and spatial analysis has some additional pitfalls.

There are many dangers in applying statistical procedures to conveniently available data (as opposed to data derived after appropriate experimental design), spatial or not, including choice of wrong models, overlooking latent variables, using (perhaps unknowingly) truncated data, performing inherently too many tests (so that many of them are bound to be formally 'significant'). I recommend to read Glymour et al. (1997) for dangers in data mining in general.

In geography, there are in general no natural objects for statistical analysis such as patients or crop fields or production units; the boundaries of geographical districts are more or less arbitrary, for instance historically grown, and not created for the problem at hand. Choosing different boundaries will yield different results. This situation has been coined the *modifiable areal unit problem* and has been widely discussed in the literature, see Openshaw and Taylor (1981), Fotheringham and Rogerson (1993). The underlying variables usually vary slowly and when they change somewhere abruptly this just does not occur at the artificial boundaries.

Different tesselations of a region will lead to different results. Even using the smallest possible districts may not help: larger-area effects will be hidden by the random fluctuations of the small districts. If on the other hand a region is divided into too few districts, the interesting phenomena disappear as inconspicuous deviations of a single data item (or even distributed on several neighbouring items).

There is some discussion in the literature whether the search for clusters is justified at all. Clusters occur either due to spatial autocorrelation; then the location of a cluster is random. Or the reason is that an influential variable has been forgotten in the model that should explain the variable under study. A critique of cluster tests, primarily for small (i.e., relatively homogeneous) regions with point data, with a review of this discussion is given by Elliott et al. (1995). Another possible cause of apparent clustering lies in different quality of the data, for instance if the area around a putative source of pollution has been screened more intensively for persons with a particular disease; this is called the post-hoc effect. Still another cause for clustering (or, possibly, for not finding the cluster one is looking for) is called *socioeconomic confounding*, the effect of a different socioecenomic structure of the population in part of the region. Thus the reason for cancer around a factory could be that workers are living there smoking more than other people and not pollution from the factory.

The search for clusters seems generally acceptable in exploratory studies where any hypotheses found should afterwards be verified or rejected with data from a different region or possibly the same region at a different time.

Fotheringham and Rogerson (1993) discuss problems in spatial analysis, in particular in connection with the use of geographical information systems (GIS). The topics are: the modifiable areal unit problem; boundary problems; spatial interpolation; spatial sampling procedures; spatial autocorrelation; goodness-of-fit in spatial modelling; context-dependent

results and nonstationarity; aggregate versus disaggregate models. Anyone attempting to use statistical methods with spatial data should be aware of these problems.

A particular warning regards the estimation of the variance of a spatial variable. Usually the values at neighbouring areas or points are positively correlated. As a result, the conventional s^2 is not an unbiased estimator for the unknown σ^2; it is too small. Therefore the usual tests (assuming independent observations) yield too many formally significant results, for instance the test for comparing two means. Again the use of the smallest possible districts (in order to get a large sample size) is no remedy: the correlation between neighbours usually becomes even larger.

6.3 Conclusions

The topic of this report is finding suspicious clusters for spatial data. Briefly, the situation that has been analyzed is as follows.

We examine a region divided into n districts with a neighbourhood structure expressed in general as a weight matrix $W = \{w_{ij}\}$. For each district, the value of a variable of interest is given, either a binary or a real-valued variable. The null hypothesis, i.e. the uninteresting case, is the independence of the data for all districts (or the near-independence, since spatial data are always at least somewhat spatially correlated). We want to check if the given variable exhibits deviations from this null hypothesis; more exactly, we are looking not just for any deviations but for concentrations of high (or low) values in one or more subregions: spatial clusters.

Here we mean clusters that are not specified in advance but are derived from the data. Otherwise the conventional procedures for nominal or hierarchical variables are applicable.

As we have seen, there exist several global and local tests; which ones should be used? Certainly, that depends – but on what?

As long as one has no clues for using other preferences, I recommend for binary data either the *BB* test, Section 3.1, or the triplet cluster test, Section 3.4. The triplet test has theoretical disadvantages (mostly smaller power, fewer distinct significance levels), but the advantage of yielding results that are easier to interpret. If about half of the districts are marked, the *BW* test is somewhat better than the *BB* test.

For real-valued data, I recommend Moran's *I* to check whether there is a spatial dependence; this statistic is rather insensitive against deviations from the normal distribution of the underlying variables. However, a departure from independency need not mean clustering. The local version of Moran's *I* is very distribution sensitive and hardly to be recommended. Therefore I propose to combine Moran's *I* with the triplet cluster statistic of Section 4.7: it is more reliable than the local tests and finds clusters with more versatile shapes, not just a district with all its neighbours.

In data mining the variables for the districts are often a specialization of a much larger data set, for instance the proportion of persons with a special characteristic (voters of a party, owners of an appliance, customers of a business type etc.) within an age range, occupational group and income class. The tests are performed for many combinations of these parameters so that the total error probability becomes entirely vague. In general, many suspicious results are essentially the same due to largely overlapping subsets (either overlapping intervals of the variable used for selecting the subset or dependency between two such variables) so that the question arises which one of the suspicious results is most important and which ones should be suppressed, Gebhardt (1991). In our context, one has in addition overlapping regions (not

just different significance measures) as results; so which one to select? This problem has not been tackled so far.

Bibliography

Anselin, Luc: *Spatial Econometrics: Methods and Models.* Kluwer, Dordrecht, 284 pages, 1988, ISBN 90-247-3735-4.

Anselin, Luc: Local indicators of spatial association – LISA. *Geographical Analysis*, 27, pages 93–115, 1995.

Bailey, Trevor C.; Gatrell, Anthony C.: *Interactive Spatial Data Analysis.* Longman Scientific & Technical, Harlow, Essex, 413 pages, 1995, ISBN 0-582-24493-5.

Besag, Julian: Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B,* 36, pages 192–225, 1974.

Besag, Julian; Diggle, Peter J.: Simple Monte Carlo tests for spatial pattern. *Applied Statistics*, 26, pages 327–333, 1977.

Besag, Julian; Newell, James: The detection of clusters in rare diseases. *Journal of the Royal Statistical Society, Series A,* 154, pages 143–155, 1991.

Bithell, J. F.: The choice of test for detecting raised disease risk near a point source. *Statistics in Medicine*, 14, pages 2309–2322, 1995.

Cliff, Andrew D.; Ord, J. Keith: *Spatial Autocorrelation.* Pion, London, 178 pages, 1973, ISBN 0-85086-036-9.

Cliff, Andrew D.; Ord, J. Keith: *Spatial Processes: Models and Applications.* Pion, London, 266 pages, 1981, ISBN 0-85086-081-4.

Clifford, Peter; Richardson, Sylvia; Hémon, Denis: Assessing the significance of the correlation between two spatial processes. *Biometrics*, 45, pages 123–134, 1989.

Costanzo, C. Michael; Hubert, Lawrence J.; Golledge, Reginald G.: A higher moment for spatial statistics. *Geographical Analysis*,15, pages 347–351, 1983.

Cressie, Noel; Read, Timothy R. C.: Spatial data analysis of regional counts. *Biometrical Journal*, 31, pages 699–719, 1989.

Cressie, Noel A. C.: *Statistics for Spatial Data.* Wiley, New York, 900 pages, 1993, ISBN 0-471-00255-0.

Cuzick, J.; Edwards, R.: Spatial clustering for inhomogeneous populations. *Journal of the Royal Statistical Society (B)*, 52, pages 73–104, 1990.

Diggle, Peter J.: *Statistical analysis of spatial point patterns.* Academic Press (Mathematics in Biology), London, 148 pp., 1983. ISBN 0-12-215850-4.

Elliott, Paul; Martuzzi, Marco; Shaddick, Gavin: Spatial statistical methods in environmental epidemiology: a critique. *Statistical Methods in Medical Research*, 4, pages 137–159, 1995.

Ester, Martin; Kriegel, Hans-Peter; Sander, Jörg; Xu, Xiaowei: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Simoudis, Evangelos; Han, Jiawei; Fayyad, Usama, editors, *Proceedings Second International Conference on Knowledge Discovery and Data Mining*, pages 226–231. AAAI, Menlo Park, 1996, ISBN 1-57735-004-9.

Ester, Martin; Kriegel, Hans-Peter; Sander, Jörg; Xu, Xiaowei: Clustering for mining in large spatial databases. *KI – Künstliche Intelligenz,*12, No. 1, pages 18–24, 1998.

Fotheringham, A. Stewart; Rogerson, Peter A.: GIS and spatial analytical problems. *International Journal of Geographical Information Systems,*7, pages 3–19, 1993.

Fotheringham, A. Stewart; Rogerson, Peter A.: *Spatial Analysis and GIS.* Taylor & Francis, London, 281 pages, 1994, ISBN 0-7484-0104-0.

Gebhardt, Friedrich: Choosing among competing generalizations. *Knowledge Acquisition,*3, pages 361–380, 1991.

Gebhardt, Friedrich: *Clusters in spatial area data.* Arbeitspapiere der GMD, number 1068. GMD, Sankt Augustin, 21 pages, 1997a.

Gebhardt, Friedrich: Finding spatial clusters. In: Komorowski, Jan; Zytkow, Jan, editors, *Principles of Data Mining and Knowledge Discovery: First European Symposiun, PKDD '97, Trondheim, June 1997,* pages 277–287. Springer, Berlin, 1997b, ISBN 3-540-63223-9.

Gebhardt, Friedrich: Identifying clusters in spatial area data. In: Gierl, Lothar; Bull, Mathias et al., editors, *GEOMED'97: proceedings of the International Workshop on Geomedical Systems,* pages 260–270. Teubner, Stuttgart, 1998a, ISBN 3-8154-2311-2.

Gebhardt, Friedrich: *Survey on cluster tests for spatial area data.* GMD Report, number 7. GMD, Sankt Augustin, 52 pages, 1998b.

Gebhardt, Friedrich: *Spatial cluster test based on triplets of districts.* GMD Report, number 20. GMD, Sankt Augustin, 20 pages, 1998c.

Gebhardt, Friedrich: Cluster test for geographical areas with binary data. *Computational Statistics and Data Analysis,* 31, pages 39–58, 1999.

Gebhardt, Friedrich: Survey on cluster tests for spatial area data. *Computers & Geosciences*, in print, 2000.

Getis, Arthur; Ord, J. Keith: The analysis of spatial association by use of distance statistics. *Geographical Analysis,*24, pages 189–206, 1992.

Getis, Arthur: Spatial dependence and heterogeneity and proximal databases. In: Fotheringham, Stewart; Rogerson, Peter, editors, *Spatial Analysis and GIS,* chapter 6, pages 105–120. Taylor & Francis, London, 1994, ISBN 0-7484-0104-0.

Glymour, Clark; Madigan, David; Pregibon, Daryl; Smyth, Padraic: Statistical themes and lessons for data mining. *Data Mining and Knowledge Discovery*, 1, pages 11–28, 1997.

Griffith, Daniel A.: Some guidelines for specifying the geographic weights matrix contained in spatial statistical models. In: Arlinghaus, S. L., editor, *Practical Handbook of Spatial Statistics,* pages 65–82. CRC Press, Boca Raton, 1995.

Haining, Robert P.: *Spatial Data Analysis in the Social and Environmental Sciences.* Cambridge University Press, Cambridge, 431 pages, 1990, ISBN 0-521-38416-8.

Han, Jiawei; Koperski, Krzysztof; Stefanovic, Nebojsa: GeoMiner: a system prototype for spatial data mining. *SIGMOD Record,*26, No. 2, pages 553–556, 1997.

Hills, M.; Alexander, F.: Statistical methods used in assessing the risk of disease near a source of possible environmental pollution: a review. *Journal of the Royal Statistical Society (A),* 152, pages 353–363, 1989.

Isaaks, E. H.; Srivastava, R. M.: *An Introduction to Applied Geostatistics.* Oxford University Press, Oxford, 1989.

Johnston, R. J.: *Classification in Geography*. Geo Abstracts, Norwich, 43 pages, 1976, ISBN 0-902246-54-2.

Kemp, I.; Boyle, P.; Smans, M.; Muir, C.: *Cancer Atlas in Scotland 1975 – 1980*. IARC Scientific Publications, 72. IARC, Lyon, 1985.

Knorr-Held, Leonhard; Besag, Julian: Modelling risk from a disease in time and space. *Statistics in Medicine*, 17, pages 2045–2060, 1998.

Koperski, Krzysztof; Han, Jiawei: Discovery of spatial association rules in geographic information databases. In: Egenhofer, Max J.; Herring, John R., editors, *Advances in Spatial Databases: 4th International Symposium, SSD'95, Portland*, pages 47–66. Springer, Berlin, 1995, ISBN 3-540-60159-7.

Kulldorff, Martin; Nagarwalla, Neville: Spatial disease clusters: detection and inference. *Statistics in Medicine*, 14, pages 799–810, 1995.

Lawson, Andrew B.; Waller, Lance A.: A review of point pattern methods for spatial modelling of events around sources of pollution. *Environmetrics*, 7, pages 471–487, 1996.

Marshall, Roger J.: A review of methods for the statistical analysis of spatial patterns of disease. *Journal of the Royal Statistical Society (A)*, 154, part 3, pages 421–441, 1991.

Möhner, M.: A global rank test for geographical clusters of disease. *Biometrical Journal*,33, pages 317–323, 1991.

Møller, Jesper: Markov connected component fields. *Advances in Applied Probability*, 30, pages 1–35, 1998.

Munasinghe, Rajika L.; Morris, Robert D.: Localization of disease clusters using regional measures of spatial autocorrelation. *Statistics in Medicine*, 15, pages 893–905, 1996.

Oden, Neal: Adjusting Moran's *I* for population density. *Statistics in Medicine*, 14, pages 17–26, 1995.

Oden, Neal; Jacquez, Geoffrey; Grimson, Roger: Realistic power simulations compare point- and area-based disease cluster tests. *Statistics in Medicine*, 15, pages 783–806, 1996.

Ohno, Yoshiyuki; Aoki, Kunio; Aoki, Nobuo: A test of significance for geographic clusters of disease. *International Journal of Epidemiology*,8, pages 273–281, 1979.

Oliver, M. A.; Webster, R.: Kriging: a method of interpolation for geographical information systems. *International Journal of Geographical Information Systems*,4, pages 313–332, 1990.

Openshaw, Stan; Taylor, P. J.: The modifiable areal unit problem. In: Wrigley, N.; Bennett, R. J., editors, *Quantitative Geography, a British View*, pages 60–69. Routledge and Kegan, London, 1981.

Pernuš, Franjo: Spatial distribution of fiber types in skeletal muscle: test for a random distribution. *Muscle & Nerve*, 12, page 696, 1989.

Ripley, B. D.: *Statistical Inference for Spatial Processes*. Cambridge University Press, Cambridge, 1988.

Sander, Jörg; Ester, Martin; Kriegel, Hans-Peter; Xu, Xiaowei: Densitiy-based clustering in spatial databases: the algorithm GDBSCAN and its applications. *Data Mining and Knowledge Discovery*, 2, pages 169–194, 1998.

Statistisches Bundesamt: *Bevölkerung und Erwerbstätigkeit, Fachserie 1, Reihe 4.2.1: Struktur der Arbeitnehmer 1993*. Metzler-Poeschel, Stuttgart, 72 pp., 1994.

Statistisches Bundesamt: *Statistisches Jahrbuch für die Bundesrepublik Deutschland*. Metzler-Poeschel, Stuttgart, 771 pp., 1995. ISBN 3-8246-0476-0.

Stone, R. A.: Investigations of excess environmental risks around putative sources: statistical problems and a proposed test. *Statistics in Medicine*, 7, pages 649–660, 1988.

Tango, Toshiro: A class of tests for detecting 'general' and 'focused' clustering of rare diseases. *Statistics in Medicine*, 14, pages 2323–2334, 1995.

Tango, Toshiro: Adjusting Moran's *I* for population density: letter to the editor and reply by N. Oden. *Statistics in Medicine*, 17, pages 1055–1062, 1998.

Tiefelsdorf, M.; Boots, B.: The exact distribution of Moran's *I*. *Environment and Planning A*,27, pages 985–999, 1995.

Tiefelsdorf, M.; Boots, B.: Letter to the editor: The exact distribution of Moran's *I*. *Environment and Planning A*,28, pages 1900, 1996.

Toubiana, Laurent; Flahault, Antoine: Monitoring the participation of Sentinel general practitioner with the health care workstation SITIE. In: Gierl, Lothar; Bull, Mathias et al., editors, *GEOMED'97: proceedings of the International Workshop on Geomedical Systems*, pages 194–203. Teubner, Stuttgart, 1998, ISBN 3-8154-2311-2.

Unwin, Anthony: Exploratory spatial analysis and local statistics. *Computational Statistics*,11, pages 387–400, 1996.

Upton, Graham J. G.; Fingleton, Bernard: *Spatial Data Analysis by Example. Vol. I: Point Pattern and Quantitative Data*. Wiley, Chichester, 410 pages, 1985.

Upton, Graham J. G.; Fingleton, Bernard: *Spatial Data Analysis by Example. Vol. II: Categorical and Directional Data*. Wiley, Chichester, 1989.

Venema, Henk W.: Spatial distribution of fiber types in skeletal muscle: test for a random distribution. *Muscle & Nerve*, 11, pages 301–311, 1988.

Venema, Henk W.: Spatial distribution of fiber types in skeletal muscle: test for a random distribution – a reply. *Muscle & Nerve*, 12, pages 697–698, 1989.

Venema, Henk W.: Modeling fiber type grouping by a binary Markov random field. *Muscle & Nerve*, 15, pages 725–732, 1992.

Venema, Henk W.: Estimation of the parameters of a binary Markov random field on a graph with application to fiber type distributions in a muscle cross-section. *IMA Journal of Mathematics Applied in Medicine and Biology*, 10, pages 115–133, 1993.

Waldhör, Thomas: The spatial autocorrelation coefficient Moran's *I* under heteroscedasticity. *Statistics in Medicine*, 15, pages 887–892, 1996.

Waller, Lance A.; Lawson, Andrew B.: The power of focused tests to detect disease clustering. *Statistics in Medicine*, 14, pages 2291–2308, 1995.

Waller, Lance A.; Turnbull, Bruce W.; Clark, L. C.; Nasca, P.: Chronic disease surveillance and testing of clustering of disease and exposure: application to leukaemia incidence and TCE-contaminated dumpsites in upstate New York. *Environmetrics*, 3, pages 281–300, 1992.

Waller, Lance A.; Turnbull, Bruce W.: The effects of scale on tests for disease clustering. *Statistics in Medicine*, 12, pages 1869–1884, 1993.

Walter, S. D.: The analysis of regional patterns in health data. I. Distributional considerations. *American Journal of Epidemiology*, 136, pages 730–741, 1992a.

Walter, S. D.: The analysis of regional patterns in health data. II. The power to detect environmental effects. *American Journal of Epidemiology*, 136, pages 742–759, 1992b.

Walter, S. D.: A simple test for spatial pattern in regional health data. *Statistics in Medicine*, 13, pages 1037–1044, 1994.

Walter, S. D.; Birnie, S. E.: Mapping mortality and morbidity patterns: an international comparison. *International Journal of Epidemiology*, 20, pages 678–689, 1991.

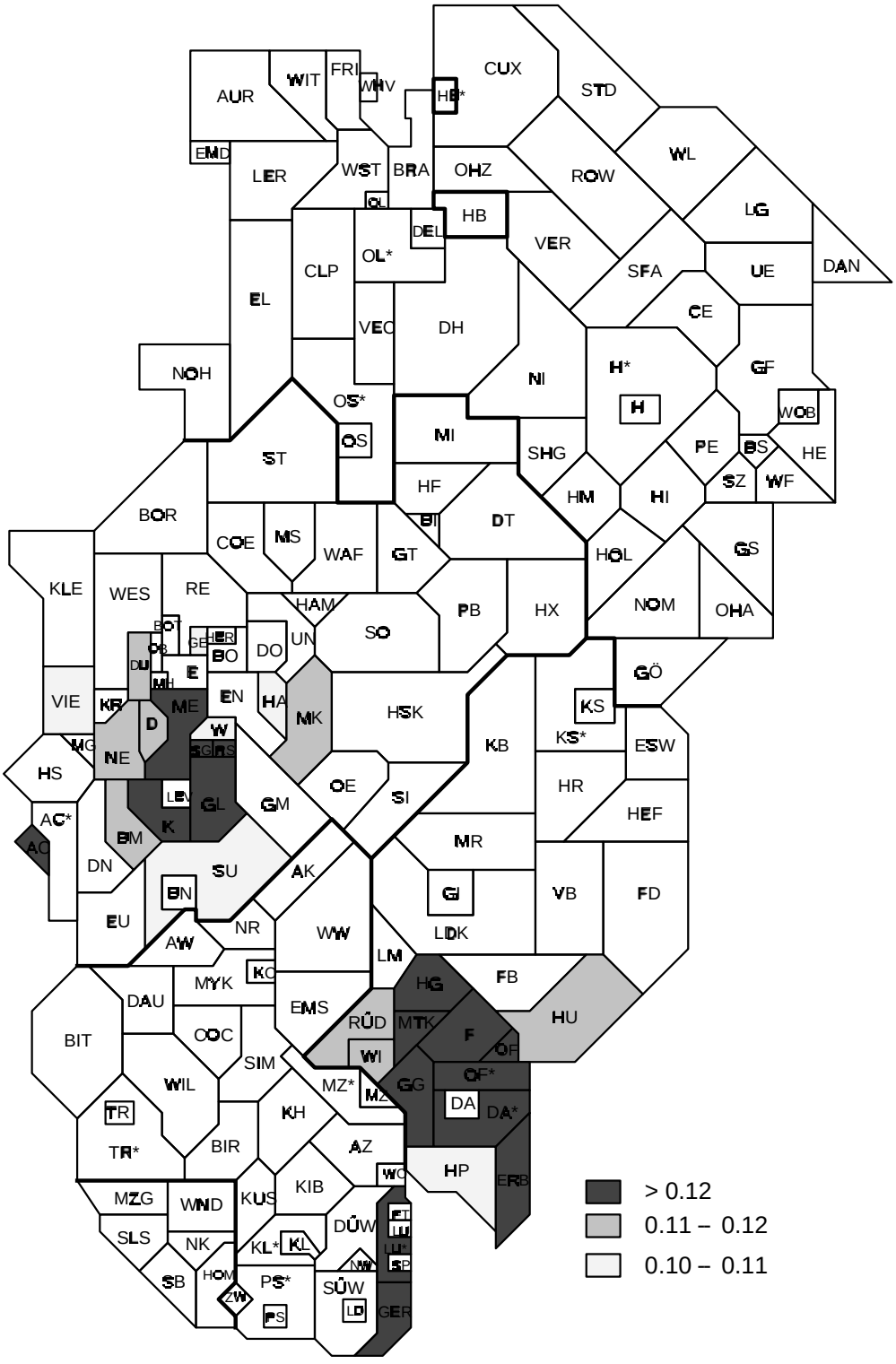

Figure 1. Counties with a high share of aliens among workers in 171 counties in north-west Germany. For explanation see Section 1.3.