Finding Spatial Clusters

Friedrich Gebhardt

GMD – German Research Center for Information Technology
Intitute for Applied Information Technology (FIT)
D-53754 Sankt Augustin
gebhardt@gmd.de

Abstract. The special challenge in analysing geographical data comes from the spatial distribution of the objects. We are interested here in finding out whether a given property is randomly distributed or concentrated somewhere. More exactly: consider a two-dimensional region subdivided into non-overlapping fields, e. g. a state divided into counties, and assume that some fields are marked for having a distinguishing property. Do the marked fields exhibit some spatial clustering?
Two tests feasible in data mining situations are proposed here, based on the number fields in clusters (defined by means of triplets, i. e. essentially three marked fields with a common boundary point) and on the number of edges of marked fields shared by another marked field. For regular settings such as honeycombs (sets of hexagons) some theoretical results are reported. In addition, simulations have been performed on honeycombs as well as on real subdivisions of a region and the tests have been applied to real data.

1 The problem

Geographical data can be treated with the usual procedures for data mining, too. However, they offer additional features in form of their (two-dimensional) geographical locations or derived properties such as neighbourhoods. Only rarely can the location be adequately handled by using the coordinates as an attribute – perhaps sometimes in a long-stretched country such as Norway.

There exist some (mostly elaborate) procedures for modelling the statistical dependence of neighbouring data items. Spatial autocorrelation models assume a more or less homogeneous correlation between neighbouring points or areas, perhaps with a directional component. Moran's I [Bailey and Gatrell, 1995, section 7.4.5] and other measures try to capture the overall spatial dependency.

Our goal here is to find spatial concentrations or clusters. We restrict ourselves to area data as opposed to point data. Point data could also be treated, if desired, by building an area around each point, e. g. by Dirichlet tesselation (Voronoi polygons), or equivalently by determining the neighbours by Delaunay triangulation.

Our problem now is the following. Given a twodimensional *region* divided into many non-overlapping *fields* with some of the fields being *marked,* is there a

significant clustering of the marked fields or are they randomly distributed over the whole area?

Some examples for marked fields are: election districts where a candidate of a certain party has been elected; plots in a region where a certain plant (or animal) has been spotted; counties with an unemployment rate above a limit. In practice the fields should be roughly comparable; thus a clustering of states on Earth, ranging in size from Russia to Vatican City, may be less meaningful.

The problem is of course not restricted to geography. Other potential examples include infected cells under a microscope, environmental damages in plots of a forest, crystals of various types on the polished surface of a mineral and conceivingly pattern recognition to distinguish real features from random deviation in colour or intensity.

One might just look at the corresponding map. However, this procedure has some severe deficiencies:

- Large fields catch the eye much more than tiny ones.
- One is easily inclined to detect expected (or easily explainable) structures and to ignore others.
- There is no hint to the statistical significance.
- Random distributions of marked fields exhibit generally some clustering; a seemingly random distribution may even be the result of an anti-clustering mechanism.
- Visual inspection cannot be automated for data mining.

Sometimes potential clusters are predefined such as industrial areas, and the question is whether the property under consideration is significantly related to these clusters, for instance whether the marked fields concentrate in these predefined clusters. This situation is *not* considered here; the problem can be handled with linear models treating the affiliation with the predefined area as an independent variable.

It seems that the questions of finding spatial clusters and statistics for deciding on random clustering have not been considered before in the literature.

Warning. In geographical data, a spatial autocorrelation is ubiquitous, see e. g. [Anselin, 1988, chapter 5] and [Bailey, 1995]. The underlying phenomenon just does not keep to the mostly artificial boundaries of the fields.

2 Connectivity regions

A possibility to define clusters of marked fields is to look for *connectivity regions:* maximal connected sets of marked fields.

In a honeycomb of 169 hexagons, the fields have been marked with a given probability. A typical result is shown in figure 1. At the lower left, there is a connected set of 11 marked fields; two other continuity regions consist of eight and six fields. If such an aggregation would occur in reality, one would be inclined to "interpret" the underlying data somehow, but it is pure chance.

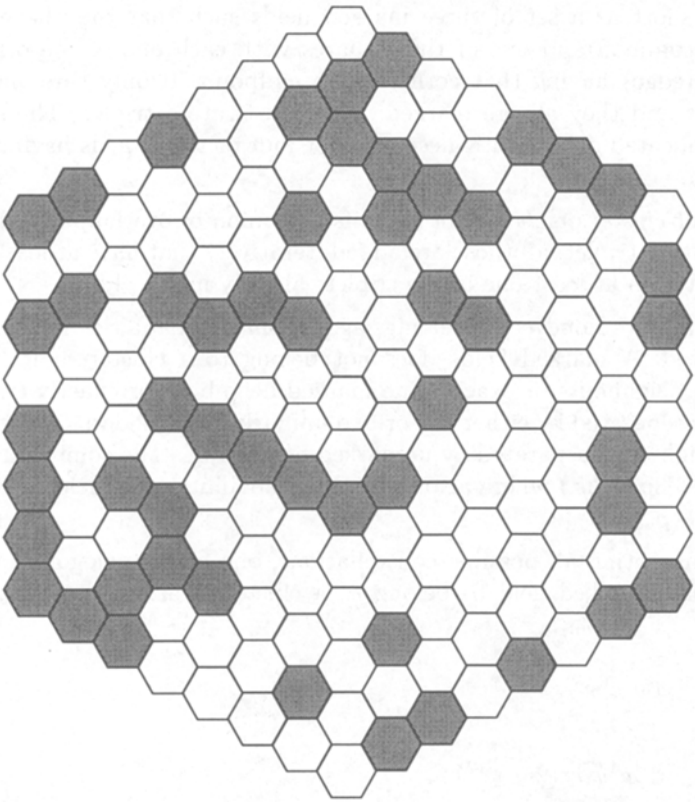

Fig. 1. A hexagon of 169 fields. Each field has been marked with probability 0.3 by a chance algorithm.

The behaviour of connectivity regions can be studied to some extent in the special situation of an infinite honeycomb. We assume that each hexagon has the same probability p of being marked and that all markings are independent. Then for instance the probability that all six neighbours of a given (marked) hexagon are unmarked is q^6 with $q = 1 - p$. With increasing effort, the probability of belonging to a connectivity set of 2, 3, 4 hexagons can be computed; the results are given in [Gebhardt, 1996] together with a graph showing these probabilities as functions of p. It turns out that for instance for $p = 0.2$ about 28% of all marked fields belong to a connectivity region of at least 5 fields.

Thus declaring all connectivity regions of size at least 3 (or 4, or even 5) to be regional clusters makes sense for rather small p only, perhaps for $p < 0.05$.

3 Triplet-clusters

We need a more restrictive definition of clusters. We define a *field-triplet* or *triplet* for short as a set of three marked fields such that they have a corner (node) in common and one of them shares with each one of the other two a boundary (edge) having that corner as an endpoint. If only three fields meet at a corner and they all are marked, they constitute a triplet. The somewhat more complicated definition is necessary for four or more fields having a corner in common.

A *triplet-cluster* or *cluster* for short is the union of overlapping triplets, i. e. starting with a triplet, all those are added iteratively that have at least one field in common with at least one of the triplets already in the cluster.

In the infinite honeycomb setting, again some probabilities for clusters can be computed. A marked fields does not belong to a cluster, if it has either no marked neighbours or exactly one marked neighbour or exactly two marked neighbours separated by either one or two unmarked neighbours or if it has three marked neighbours separated by unmarked neighbours. The combined probability is $q^3(1 + 3p - 2p^2)$ where again p is the probability for a field to be marked and $q = 1 - p$.

By enumerating all possible constellations, one finds analogously the probabilities for a marked field to belong to a cluster of size n as follows; we set $r = 1 + p$.

$$
\begin{aligned}
n = 3: &\quad 6p^2q^6r^3 \\
n = 4: &\quad 12p^3q^6r^2 \\
n = 5: &\quad 15p^4q^7r^2(2 + qr^2) \\
n = 6: &\quad 12p^5q^7r^2(3 + 9q - q^2 - q^3)
\end{aligned}
$$

The probabilities of not belonging to a cluster are considerably higher than those for not belonging to a connectivity region. For $p = 0.20$, the probability of a marked field to belong to a cluster of at least five fields now is 4.4%. This holds for infinite honeycombs; for finite ones, the clusters are somewhat smaller due to boundary effects. For more details including simulation results with finite honeycombs and similar theoretical evaluations as well as simulations for a chessboard see again [Gebhardt, 1996]. In figure 1 there are just three tripletts, one at the lower left and two near the top. The latter two form *two* clusters because they do not overlap.

Let us compare connectivity regions and clusters. For the former ones in an infinite cluster, it is more tedious than for clusters to compute formulae corresponding to those given above; it has been done for regions of size one to four only. Therefore we plot in figure 2 the probability for a marked field to belong to a connectivity region or a cluster of size ≥ 5, given probability p for a field being marked.

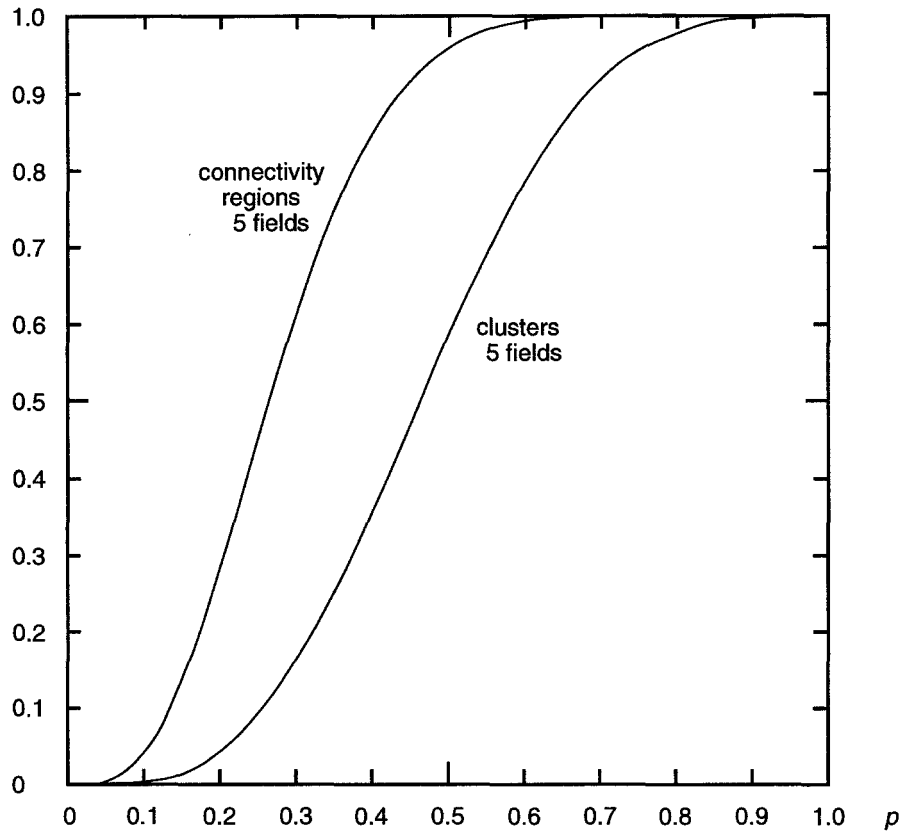

Fig. 2. Comparison between connectivity regions and clusters: probability of a marked field (probability p for being marked) to belong to a connectivity region or a cluster of size at least 5.

4 Testing for clusters

4.1 Number of fields in clusters

We will present two tests on clustering. The first one is based on the number of marked fields in clusters.

The test procedure is as followsw. A spatial distribution is regarded as non-random if the number of (marked) fields in clusters, M_C, differs from the expected value m_C by more than k standard deviations s_C, where k may be set to 2 or 2.5 in single tests and should be set to 3 or larger in extended series of tests such as in data mining. Written as formula: M_C is considered significant if $|M_C - m_C| > ks_C$. This is a standard procedure in statistics; the problem lies in determining the proper expectations and standard deviations (or their squares, the variances).

m_C and s_C depend on p, the probability of a field being marked, and of course on the size and structure of the arrangement of fields. It seems hopeless to compute expectation and standard deviation of this statistic; so the method of choice is simulation. This may not make sense for single tests, but for a series of tests on the same region (with different sets of marked fields) this procedure is feasible. In particular, it is useful in a data-mining context.

Simulations with different regions have shown that expectation and standard deviation can be approximated with sufficient precision by polynomials in p of fourth order with missing low-order terms, i. e.

$$m_C \approx c_3 p^3 + c_4 p^4,$$

$$s_C \approx d_2 p^2 + d_3 p^3 + d_4 p^4,$$

where the coefficients depend on the topology of the region and $p \leq 0.5$.

Thus, given a new region with F fields, one has to choose four or five values for the number M of marked fields (the largest one near $F/2$), generate perhaps 1000 random arrangements of M fields, compute M_C for each arrangement and the mean and sample standard deviation for each M and finally the parameters c_i and d_i by linear regression. For finding the triplets one has to know not only which fields are neighbours but also which neighbours of a field are adjacent.

When using the test one should apply a continuity correction since the number of fields in clusters is of course always an integer.

4.2 Neighbouring marked fields

An alternative test is based on the number of pairs of marked fields that have an edge in common. We leave out the details here, see [Gebhardt, 1996], and just give the results. We need the quantity

$$T := \frac{M-1}{F-1}\,(2G - M\,\bar{B})$$

where F is the number of fields, M is the number of marked fields, G is the total number of neighbours of the marked fields and $\bar{B}$ is the average number of neighbours of a field in the whole region. Let

$$b_i := \begin{cases} \text{number of marked neighbours} & \text{if field } i \text{ is marked} \\ 0 & \text{else} \end{cases}$$

$$b := \sum_i b_i.$$

Since the neighbourhood of marked fields is a reflexive property, b is always even. It turns out that given the number of marked fields and the total number of neighbours of the marked fields, $b/2$ is approximately Poisson distributed with parameter $T/2$ and thus, in a further approximation if T is not too small, b is normally distributed with expectation T and standard deviation $\sqrt{2\,T}$. Again a

continuity correction should be applied to b, but here using 1.0 rather than 0.5 because b is always even (i.e., the regular continuity correction to $b/2$).

T and b can be computed if all the neighbourhoods are known; one does not need any simulations as in section 4.1.

Several simulations have been performed. The resulting standard deviations for b were not larger than given above, but often smaller; in many cases they were close to $\sqrt{T}$, i. e. without the factor $\sqrt{2}$. However, it was not obvious under which circumstances our formula overestimates the standard deviations.

While this second test is simpler to compute, it does not directly use our restricted notion of clusters. However, there is an affinity because three fields in a triplet contribute more to b than three connected marked fields not forming a triplet.

Both tests, in this as well as in the preceding section, decide only whether there is a non-random clustering but they do not identify the clusters. For this purpose the largest clusters based on triplets should be used.

If p is not too small, any real (non-random) cluster will have attached by chance some marked fields or strings of marked fields. Therefore one should be restrictive in the extension of clusters considered non-random; i. e. marked fields neighbouring a cluster should be assigned in retrospect, for purposes of interpretation, to that cluster only very cautiously.

5 A geographical arrangement: six Bundesländer

Real arrangements of fields show some peculiarities that need special treatment. To find out what might happen and to compare honeycombs (169 fields) as used in simulations with a real situation, six Bundesländer with 171 counties (Kreise and kreisfreie Städte) have been selected: Bremen, Niedersachsen (Lower Saxony), Hessen (Hesse), Nordrhein-Westfalen (North Rhine-Westphalia), Rheinland-Pfalz (Rhineland-Palatinate) and Saarland.

In the stylized map, the counties are named by their motor vehicle codes. Sometimes a city (kreisfreie Stadt) and the adjacent county (Landkreis) have the same code; then an asterisk has been appended for the Landkreis (e. g. HH for Stadt Hannover, HH* for Landkreis Hannover). Occasionally, a county has a small exclave. This has been neglected.

Some counties (always Stadtkreise) have only one neighbour; examples are Hannover and Aachen (AC). According to our definition, they could never belong to a cluster. Other counties have exactly two neighbours; they could only belong to a cluster, if both neighbours are also marked. Examples are Bonn (BN) almost entirely surrounded by Siegburg (SU), Dannenberg (DAN) and Wolfsburg (WOB). Therefore we extend the definition of a cluster slightly: a field with only one neighbour is added to a cluster if the neighbour belongs to the cluster and the field itself is marked. For each field with two neighbours only, one of them is predefined as major neighbour. Such a field is added to a cluster if the major neighbour belongs to the cluster and the field itself is marked.

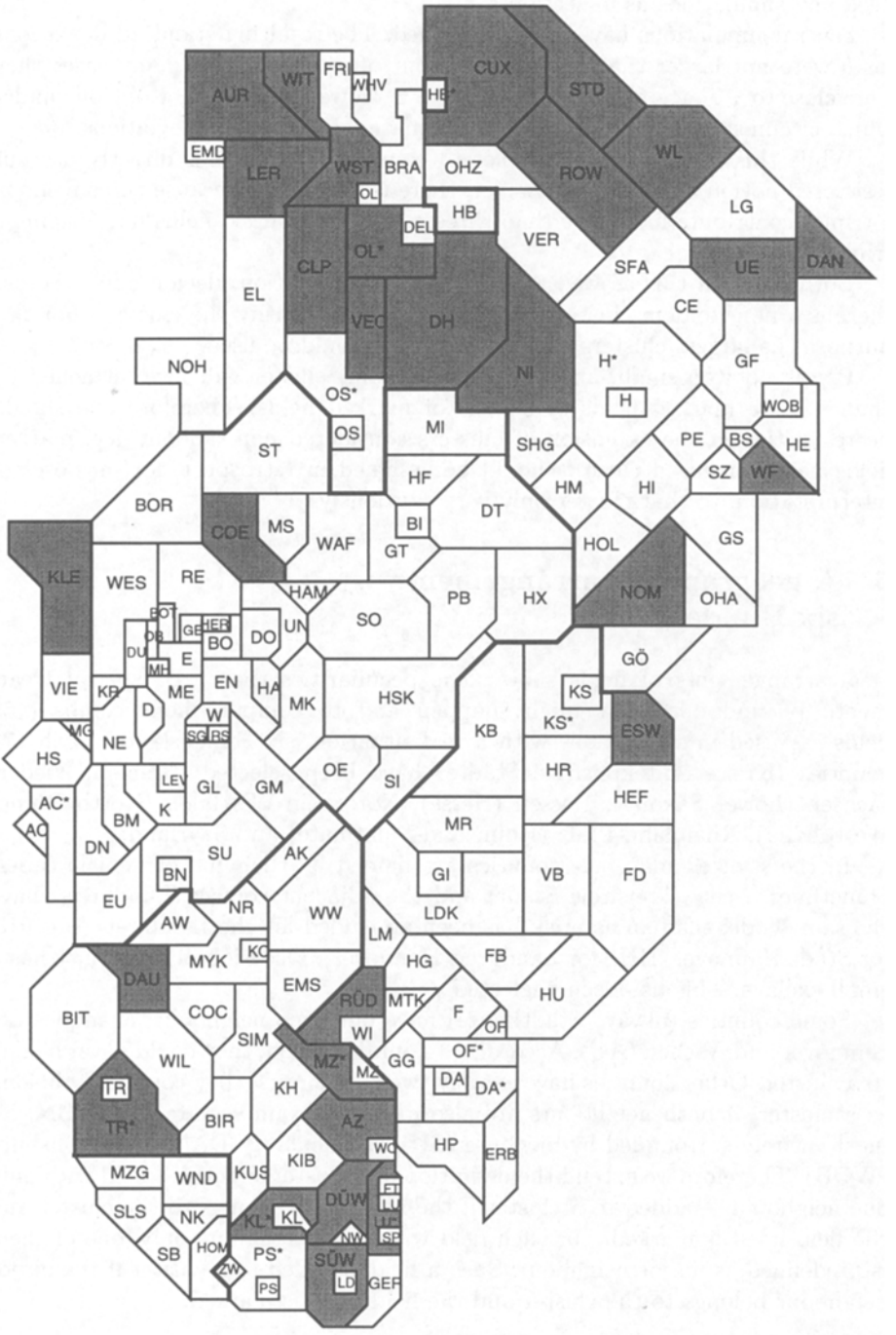

Fig. 3. Portion of agriculture (including forestry and fishery) among the workers $\geq 2\%$.

There are four counties with only one neighbour, 11 counties with two neighbours, and the maximum is 11 neighbours for Mettmann (ME). The number of edges (counted twice) is 876; thus each county has an average of $\bar{B} = 5.12$ neighbours. This is less than for the 169 hexagons (5.47): the region is less compact.

According to simulations, the average number of counties in clusters is about 10% lower than in the case of 169 hexagons, the average number of clusters even about 15%. A reason for this is supposedly the smaller average number of neighbours; perhaps the uneven distribution of neighbours (1 to 11 rather than 3 to 6) plays also a role. Similarly the number of small clusters is markedly lower while there is no deviation for large clusters.

The data used for some tests include population, area, number of persons working and counted for social security (sozialversicherungspflichtig Beschäftigte, henceforth called "workers" for short) and several subgroups of the latter one: men, blue-collar workers, aliens and a division into agriculture and fishery, production, commerce and traffic, others (in particular services trade). The workers are counted at the place of work (not of living). All data refer to 1993. The variables used are derived from those mentioned, e. g. population density or percentage of aliens in all workers. The data are taken from [Statistisches Bundesamt, 1994] and [Statistisches Bundesamt, 1995].

An example is shown in figure 3. There are 29 marked counties, among them 12 in two clusters (expected: 3.2 with standard deviation 2.9). The marked counties have 163 neighbours, among them 54 marked ones (expected: 29.2 with standard deviation 7.6). So both statistics have a z-value near 3 and are significant. There are two clusters in the north (the shaded counties in the south contain no triplet). Since up to about nine counties in clusters are insignificant here, only the larger cluster (AUR to DH; NI does not belong to it) should be regarded as conspicuous.

Incidentally, counties with low quotients are not clustered according to the tests of sections 4.1 and 4.2. In fact, if agriculture is $\leq 0.5\%$, there are 39 marked counties but no triplet at all. The test on neighbouring marked fields yields a sample statistic below the expected value.

6 Conclusions

We wanted to find out whether a set of marked fields within an assembly of fields shows a significant clustering. Connectivity regions do not provide a meaningful starting point except perhaps for a small share p of marked fields (a few percent at most), because for larger shares the marked fields will form connectivity regions of considerable size even in random assignments.

We defined clusters by means of triplets, essentially three marked fields with a corner in common. The total number of fields in triplets is a useful test statistic. The distribution of this statistic under the null hypothesis (random distribution) has to be computed by simulations.

The test works for a wide range of p. It will rarely find a clustering for very small p, and for $p > 0.5$ one should inspect the complement (the unmarked fields).

A perhaps less evident test statistic is the number of edges between marked fields. Approximations to the expected value and its standard deviation can easily be computed from the characteristics of the assembly of fields (the number of fields with 1, 2, ... neighbours).

Several simulations with two assemblies, a honeycomb structure of 169 hexagons and 171 counties of six Bundesländer, support and augment the theoretical considerations. In particular they suggest that the approximations for the expected values just mentioned are surprisingly good while the standard deviation is often even smaller than anticipated. The tests have also been used for regions not reported here (62 statistical districts of the city of Bonn; 80 election districts in the eastern part of Germany; smaller honeycombs). It seems that for the tests to be meaningful there should be no less than about 50 fields.

The procedures are expected to work also for several hundred or even some thousands of fields; however, it is conceivable that for large regions (or, equivalently, finer subdivision of a given region) one needs another notion of cluster. It should, on the one hand, start with larger assemlies than triplets, but it should, on the other hand, be somewhat tolerant against scattered fields *not* belonging to the cluster. For a proposal see [Gebhardt, 1996, section 3.3].

In the majority of tests both test statistics behave similarly, but occasionally the results differ. The (perhaps subjective) impression is that the cluster test is somewhat closer to what one would expect.

The tests seem to be the first ones for binary data (marked vs. unmarked fields). Often, a numerical value is attached to each field such as the proportion of workers in agriculture and forestry; then the tests can be used by marking all fields whose values exceed a boundary as we have done above. A true cluster should show up at different boundary values though with different cluster sizes. A small cluster which is significant only for boundary values in a small interval should not be taken seriously: when the boundary is moved until the first small cluster builds, it is often significant formally, but then one has selected the test (i. e. the boundary) according to the data, an example for how to lie with statistics.

But certainly there should be tests using numerical data directly rather than only the coarsened property of exceeding a boundary. Several tests exist, mostly based on the overall correlation of the fields with the fields in the neighbourhood (not necessarily directly adjoining fields). An example is Moran's I [Bailey and Gatrell, 1995, section 7.4.5]. The statistics G_i and G_i^* by Getis and Ord [Getis and Ord, 1992] are designed to find fields whose neighbours are significantly above (or below) average. They are restricted to inherently positive variables excluding thus residuals from regression, among others. For a test statistic based on the mean values in clusters see [Gebhardt, 1997].

References

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

[Bailey and Gatrell, 1995] Trevor C. Bailey and Anthony C. Gatrell. *Interactive Spatial Data Analysis.* Longman Scientific & Technical, 1995, 413 pages, ISBN 0-582-24493-5.

[Bailey, 1995] Trevor C. Bailey. A review of statistical spatial analysis in geographical information systems. In Stewart Fotheringham and Peter Rogerson, editors, *Spatial Analysis and GIS*, chapter 2, pages 13–44. Taylor & Francis, London, 1995, ISBN 0-7484-0104-0.

[Gebhardt, 1996] Friedrich Gebhardt. Clusters in geographical distributions. Arbeitspapiere der GMD 1035, GMD, Sankt Augustin, 1996, 33 pages.

[Gebhardt, 1997] Friedrich Gebhardt. Clusters in spatial area data. Arbeitspapiere der GMD 1068, GMD, Sankt Augustin, 1997. In preparation.

[Getis and Ord, 1992] Arthur Getis and J. K. Ord. The analysis of spatial association by use of distance statistics. *Geographical Analysis*, 24:189–206, 1992.

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

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