Essays on Distance Based (Non-Euclidean) Tests for Spatial Clustering in Inhomogeneous Populations

Adjusting for the Inhomogeneity through the Distance Used

ULLA ROMILD
Dissertation presented at Uppsala University to be publicly examined in Hörsal 2, Ekonomikum, UPPSALA, Wednesday, May 17, 2006 at 10:15 for the degree of Doctor of Philosophy. The examination will be conducted in English.

Abstract

This thesis consists of four papers dealing with distance based (non-Euclidean) tests for spatial clustering in inhomogeneous populations.

The density adjusted distance (DAD), which considers the underlying density, is defined in the first paper. The proposed distance can be used together with any of the old distance based methods developed for traditional homogeneous spatial patterns.

The test statistics in distance based tests can all be seen as a weighted sum of distance measures for distances between \( n \) cases with known co-ordinates. DAD based test statistics are developed and their performance is compared with the performance of previously suggested tests by simulation in the second paper. The tests are compared in different types of data set and for various kinds of clustering. It is shown that no test is the optimal choice for all alternative hypotheses and that the tests are unequally sensitive to the structure of the underlying data. Tests based on the DAD are often a good alternative.

Test statistics and graphical tools for the Empirical Distribution Function of DAD are developed and examined in the third paper. We show that the result of an EDF test combined with EDF plots provides more information about the possible nature of clustering in a sample than the result of a parametric test only.

Keywords: Clustering, Spatial Point Pattern, Inhomogeneous Density

Ulla Romild, Department of Information Science, Box 513, Uppsala University, SE-75120 Uppsala, Sweden

© Ulla Romild 2006

ISSN 1652-9030
ISBN 91-554-6548-X
urn:nbn:se:uu:diva-6829 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-6829)
To Alfons, David, Emma and Hillevi
with love

This thesis could have been finished years ago without you in my life, but I sincerely doubt that it would have made me happier.
Papers summarized in the dissertation

I  Romild, U., Kulldorff, M. A Density Adjusted Distance for Spatial Statistics

II Romild, U., A Class of Spatial Clustering Tests Based on a Density Adjusted Distance

III Testing for Clustering with Empirical Distribution Functions of Density Adjusted Distances in Inhomogeneous Populations

IV Can Attitudes and other Characteristics amongst Mountain Hikers be Explained by Place of Residence?
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>9</td>
</tr>
<tr>
<td>Spatial Statistics</td>
<td>9</td>
</tr>
<tr>
<td>Inhomogeneous distributions</td>
<td>9</td>
</tr>
<tr>
<td>Clustering</td>
<td>10</td>
</tr>
<tr>
<td>Distance Based Methods</td>
<td>11</td>
</tr>
<tr>
<td>Test statistics for spatial data</td>
<td>12</td>
</tr>
<tr>
<td>Power Comparison</td>
<td>12</td>
</tr>
<tr>
<td>Summary of the papers</td>
<td>14</td>
</tr>
<tr>
<td>I  A Density Adjusted Distance for Spatial Statistics</td>
<td>14</td>
</tr>
<tr>
<td>II A Class of Spatial Clustering Tests Based on a Density Adjusted Distance</td>
<td>15</td>
</tr>
<tr>
<td>III Testing for Clustering with Empirical Distribution Functions of Density Adjusted Distances in Inhomogeneous Populations</td>
<td>17</td>
</tr>
<tr>
<td>IV Can Attitudes and other Characteristics amongst Mountain Hikers be Explained by Place of Residence?</td>
<td>20</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>22</td>
</tr>
<tr>
<td>References</td>
<td>23</td>
</tr>
</tbody>
</table>
## Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSR</td>
<td>Complete Spatial Randomness</td>
</tr>
<tr>
<td>DAD</td>
<td>Density Adjusted Distance</td>
</tr>
<tr>
<td>EDF</td>
<td>Empirical Distribution Function</td>
</tr>
<tr>
<td>NN</td>
<td>Nearest Neighbour</td>
</tr>
<tr>
<td>SAD</td>
<td>Sum of Absolute Deviations</td>
</tr>
<tr>
<td>SD</td>
<td>Standard Deviation</td>
</tr>
<tr>
<td>SE</td>
<td>Standard Error</td>
</tr>
<tr>
<td>SSD</td>
<td>Sum of Squared Deviations</td>
</tr>
</tbody>
</table>
Introduction

Spatial Statistics

Spatial statistics and statistical analysis of spatial point patterns have been of interest for quite some time. One early reference is Student (1907), who compared the distribution of yeast cells or blood corpuscles over N units of area to the binomial distribution. Similarly, many studies have been conducted over homogenous areas and for quadrat counts, i.e. the number of events from a large number of smaller regions.

From the middle of last century, the alternative approach of distance-based methods has been developed. In these the basic sampling unit is a point, and information is recorded in the form of distances to neighbouring events or points.

Application areas for statistical analysis of spatial point patterns are typically found within biology and ecology, but also in, e.g., archaeology, astronomy, geography, epidemiology, economics and social sciences. With applications in social sciences, one common problem is non-homogeneous or inhomogeneous underlying distributions, which must be taken into account.

Inhomogeneous distributions

A spatial distribution as such can either be uniform, clustered or random. Points in a uniform distribution are evenly distributed over the area in question, whereas clustered patterns are characterised by ag-
Aggregation of points. A random distribution can be described as a combination of uniform and aggregated distributions (Ward et al, 2000).

Diggle (1983) defines Complete Spatial Randomness (CSR) as a pattern in which the number of events in any planar region $A$ with area $|A|$ follows a Poisson distribution with mean $\lambda|A|$. When the distribution is inhomogeneous, it can be described as a Poisson model with varying $\lambda_i$, which varies according to the underlying population. Another alternative when the data consists of binary counts is the Bernoulli model. Instead of estimating the exact distribution, Monte Carlo simulation can be applied to assess possible clustering in such a population.

Aggregation of events in space and/or in time can appear because of aggregation in the population as such or because of clustering: it is important to tell the difference.

Clustering

Clustering can be defined as a non-random distribution of events over space, such that events of a particular kind are aggregated in space and/or in time. A cluster is a collection of events in an unusual aggregation, whereas clustering is the underlying process which gives the overall propensity of events to cluster together (Besag and Newell, 1991). Clustering can thus appear without unusual aggregates if, e.g., the clustering is in the form of pair-wise clusters.

Accordingly, tests for clustering can be of varying types. Besag and Newell (1991) differentiate between tests for overall clustering and tests to detect clusters. The former type is aimed to detect general deviations from absence of clustering and only answers the question whether the phenomenon of clustering occurs in data. Examples of this are found in Moran (1948), Whittemore et al (1987), Cuzick and Edwards (1990) and Diggle and Chetwynd (1991). Tests to detect clusters often utilize multiple tests and can typically determine which collection of cases that represent the most significant cluster (see, e.g., Openshaw et al, 1988, Besag and Newell, 1991 or Kulldorff and Nagarwalla, 1995).
Another distinction pointed out by Besag and Newell (1991) is between general tests and focused tests. General tests test for clustering anywhere in the study area; focused tests, on the other hand, test for cluster/clustering around predefined foci of suspected increased risk. Most procedures proposed in the literature are aimed for either cause; however, only a few procedures are applicable to two types of clustering in an integrated manner. Tango (1995) suggested a class of tests applicable to the detection of disease clustering, either ‘focused’ or ‘general’ clustering.

The tests suggested here are aimed to detect general overall clustering, but the theory can easily be modified to cover focused tests as well.

Distance Based Methods

Population data are often available in aggregated form for different sub-regions, or cells, of an area of interest. Distance Based Methods require more specific information. In fact, we need information to identify the exact position of every element for optimal results. If we have to rely on data placing individuals somewhere in a sub-region, the methods become less reliable.

One way of overcoming the problem with population data in unwanted form is to sample controls to represent the population. If we are interested in a particular group, here called cases, we can use a random sample of controls from the same population to evaluate the distribution of the population.

The choice of distance is also crucial. Euclidean distances are a usual choice, but they have several disadvantages, especially in inhomogeneous populations. We also need to consider which distances we should take into account. We can measure nearest neighbour distances, inter-event distances or some other choice of distances.
Test statistics for spatial data

Tests for spatial randomness are either based on counts in non-overlapping equal size areas, or on the Euclidean distances between cases. When adjusting for an underlying inhomogeneity, current tests are of either four types:

i) **Area based**: Test statistics based on counts in non-overlapping areas, with unequal expected number of cases

ii) **Distance based (Euclidean), adjusts for inhomogeneity in the test statistic distribution**: The same Euclidean distance-based test statistics as in the homogeneous situation are used, however, adjusted for the inhomogeneity in the distribution of the test statistic (e.g., Whittemore et al, 1987)

iii) **Distance based (Euclidean), adjusts for the inhomogeneity in the construction of the test statistic**: A new type of test statistic using Euclidean distances not simply between cases, but between cases as well as controls in some suitable combination (e.g., Cuzick and Edwards, 1991; Diggle and Chetwynd, 1991) are created.

iv) **Distance based (Non-Euclidean), adjusts for the inhomogeneity through a transformation of the space**: The inhomogeneous background density is transform into a differently shaped region such that the underlying process is now homogenous. After performing the transformation, any of the distance-based tests constructed for the homogeneous case can be used.

In this thesis a fifth class is proposed and studied: **Distance based (Non-Euclidean), adjusting for the inhomogeneity through the distance used.**

Power Comparison

Different tests, adjusting for an un-even background, can perform better or worse depending on the actual type of clustering process as well as the background population distribution.
To investigate these differences it is important to evaluate the power for different tests under different circumstances when the null hypothesis is indeed false (Wartenberg, 1990). So far, this type of evaluation is not a general procedure for suggested tests for spatial randomness (Kulldorff et al, 2003).
I A Density Adjusted Distance for Spatial Statistics

We define a density adjusted distance where the underlying density can be either discrete or continuous.

Let $S_{a,b}$ be the set of those $x$ such that the Euclidean distance between $a$ and $x$ as well as between $b$ and $x$ is smaller than the Euclidean distance between $a$ and $b$. That is

$$S_{a,b} = \{(x_1, x_2): d_{a,x}^{\text{eucl}} > d_{a,b}^{\text{eucl}} \quad \text{and} \quad d_{b,x}^{\text{eucl}} > d_{b,a}^{\text{eucl}} \}$$  \hfill (1)

We define $u_{a,b}^*(x) = 1$ if $x \in S_{a,b}$ and $u_{a,b}^*(x) = 0$ otherwise. The density adjusted distance is then

$$d^*(a,b|\mu(x)) = d_{a,b}^* = \sqrt{\int u_{a,b}^*(x) d\mu(x)}$$  \hfill (2)

Figure 1 illustrates how the distance between $a$ and $b$ is computed when the underlying density is of the Bernoulli type. There are four points or elements within the common neighbourhood of $a$ and $b$ where $u^*(x) = 1$. The Density Adjusted Distance (DAD) between $a$ and $b$ is thus computed: $d_{a,b}^* = \sqrt{\int u_{a,b}^*(x) d\mu(x)} = 2$. 

Summary of the papers
Figure 1: The Density Adjusted Distance

For applications concerning point patterns, typically with elements divided into cases and controls, the inhomogeneous population density is generally discrete. We show that the suggested distance is optimal according to a robustness criterion and fulfils several conditions when the distance function involved is a 0/1-function.

The suggested distance is then used for testing purposes in a British data set consisting of 62 cases of childhood leukaemia and lymphoma and 141 randomly selected controls.

II A Class of Spatial Clustering Tests Based on a Density Adjusted Distance

Two types of test statistics based on the Density Adjusted Distance are developed and compared in a Monte Carlo study with tests previously suggested by Whittemore et al (1987), Cuzick and Edwards (1990) and Diggle and Chetwynd (1991). All tests compared are aimed to detect general clustering rather than the location of specific clusters.

The events of interest, individuals with a particular characteristic for which the clustering test is undertaken, are called cases. The number of cases is \( n \), and the co-ordinates of every case are known. \( m \) observations called non-cases or controls, also with known co-ordinates, are used to evaluate and estimate the distribution in the population in
combination with the cases. The data are evaluated under the hypothesis that the cases come from the same (inhomogeneous) population as the controls.

The proposed DAD tests are either the mean DAD to the \( k \) nearest cases or the mean DAD to the \( k^{th} \) nearest case.

All tests, except the tests according to Cuzick and Edwards, were performed as Monte Carlo tests where the 5% critical values were based on 10000 replications each. The program code was written in FORTRAN 90.

Three data sets with deterministic locations of the “individuals” were used. They are all in the unit square, each representing a different type of population.

1) Random distribution with uniform density.
2) A high-density area surrounded by an area with lower density, representing a city surrounded by countryside. The density in the high-density area is 24 times the density in the low-density area.
3) A chess-like pattern with density differing between 0 to twice the density in set 1.

Each data set consists of 200 observations, where the individuals will be randomised in different ways to be either a case or a control. The starting point for clusters and the principle for the clustering process were varied in three ways, which gave nine basic types of simulations.

Results from the Monte Carlo study show that no test is the overall solution in the mission to detect clustering. The comparisons between the tests were based on power, equality of power for a given test between different data sets and sensitivity to parameter specification.

The Cuzick and Edwards tests and the DAD \( k \) NN tests are the best alternatives when clustering is characterised by several clusters. These tests are also invariant to differences in the underlying density of the population.

The Whittemore test and the test suggested by Diggle and Chetwynd both depend on Euclidean distances between observations which the test suggested by Cuzick and Edwards and the DAD tests do not. It is
obvious that the latter is preferred in inhomogeneous populations. When the distribution of cases rather is caused by varying risk than actual clustering, the Whittemore test can sometimes be a suitable alternative, at least in more homogeneous data sets.

We show that the DAD measure provides a useful tool in the field of spatial statistics. The in-homogeneity is adjusted for through the distance used. Once this is done, more traditional distance-based spatial methods can be used for either testing purposes or for modelling.

The suggested test statistics for clustering performed well for most types of clustering in the Monte Carlo Study and considerably better than methods based on Euclidean distances.

III Testing for Clustering with Empirical Distribution Functions of Density Adjusted Distances in Inhomogeneous Populations

In this paper the Empirical Distribution Function (EDF) of DAD among clustered cases is described and tested for different types of clustering and underlying populations. The analysis is based on a Monte Carlo study with almost the same design as in II, enabling comparisons between the two studies.

The basic idea is to compute the EDF for an observed sample and compare it with the distribution defined in the null hypothesis. The EDF, denoted by $F_1(d)$, is in this context the cumulative relative distribution of the distances computed. For every distance $d$, this function represents the observed proportion of distances which are at most $d$:

$$F_1(d) = c^{-1} \# \left( d_{ij} \leq d \right)$$  \hspace{1cm} (3)

where $c$ is equal to the number of observed distances and $d_{ij}$ is the $j$th distance observed for the $i$th case.

The EDFs are either computed for the $k$ nearest cases or to the $k$:th nearest case.
The EDF of distances can be used for both graphical analysis and more formal testing of clustering. The null hypothesis to be tested here is that the distribution of the cases follows the same inhomogeneous process with spatially varying density as the population.

To compare the hypothetical \( F(d) \) and the observed \( F_1(d) \) distributions graphically we can either perform a plot of \( F_1(d) \) as ordinate against \( F(d) \) as abscissa. If the observed sample of cases is compatible with the population as a whole, the plot should be roughly linear. Critical limits can be evaluated simultaneously when \( F(d) \) is estimated by Monte Carlo simulations. If the observed distribution exceeds these borders, it is a signal that there are severe discrepancies between the sample of cases and the population.

An alternative graphical comparison of the distribution of distances in the observed sample and the expected distribution is to plot the distribution functions in the same graph against the values of \( d \). Critical values showing the borders for the \( \alpha \) percent most extreme values under the null hypothesis complete the graph. The perspectives are somewhat different in the two types of graphs, but reached conclusions should be the same.

An EDF statistic is a statistic measuring the difference between the observed \( F_1(d) \) and \( F(d) \) according to the null hypothesis.

The most well known statistic among the Supremum statistics is the Kolmogorov-Smirnoff statistic, which seemed to be a natural choice when testing the EDF:

\[
D = \sup_d |F_1(d) - F(d)| \quad (4)
\]

We also used two kinds of Quadratic Statistics:

\[
SSD = \sum_{i=1}^{r} (F_1(d_i) - F(d_i))^2 \quad (5)
\]

\[
SAD = \sum_{i=1}^{r} |F_1(d_i) - F(d_i)| \quad (6)
\]
All tests were carried out as Monte Carlo Tests. The basic idea is to define a statistic $U$ for which $u_i$ is observed in the sample in question. Corresponding values, $u_i$, are generated under the null hypothesis by random sampling from a set of relevant data. The rank of the observed test statistic $u_i$ among the set of values \{ $u_i : i = 1, \ldots, s$ \} determines the exact significance level of the test since each of the $s$ rankings for $u_i$ is equally likely under the null hypothesis. One of the advantages with a Monte Carlo test is that it is not necessary to specify the exact distribution under the null hypothesis. This is indeed very suitable here when the null hypothesis is that the cases follow an inhomogeneous Poisson process with varying parameters that are not specified explicitly and also because the observations are not independent.

The performance of the EDF tests was also compared with the performance of the parametric tests evaluated in paper II.

The results from the EDF tests all together imply that when very closely connected clusters (either pair-wise clustering or larger formations) are suspected, the D test is favourable, and $k$ should be set to either 1 or 2.

With larger clusters, where the cases in a specific cluster are mixed with controls, a test based on the SSD statistic is the best choice. The value of $k$ should be set equal to the approximate number of cases in each cluster, if possible.

Finally, when we are dealing with varying incidence rather than actual clustering, the SSD test or the SAD test is the best.

The graphical analysis can be used as a diagnostic tool in itself, but also as a help to choose the proper type of test in a specific situation.

The EDF tests appeared to have a somewhat lower power compared with that of a parametric test under some circumstances. However, given that the EDF tests are performed simultaneously with graphical analysis of the EDF, they provide a more complete tool than a parametric test only. The test result of an EDF test combined with EDF plots provides more information about the possible nature of clustering in a sample than the test result of a parametric test only.
IV Can Attitudes and other Characteristics amongst Mountain Hikers be Explained by Place of Residence?

The last paper is an application of the methods. For this study a sample of 841 Swedish citizens sampled from a census among visitors in the Southern Mountain Region of Jämtland, Sweden during the summer of 1999 was used.

One matter of interest is the place of residence of the visitors. Visitors from remote parts of the country or even from abroad have different demands on maintenance of the mountain region as compared with the local visitors.

The visitors can also be characterised according to their attitudes towards management of wilderness areas as purists, neutralists and urbanists (Stankey and Lime, 1973). The different segments differ in attitudes towards maintenance, other visitors, traces of human actions and regulations. A purist is a visitor in the wilderness with high demands on perceived untouched wilderness. He/she wants to be alone in the wilderness and can react negatively to even slight disturbances from other people or traces of other human beings. The opposite is an urbanist with high tolerance towards meeting others along the trail. The urbanist not only tolerates different types of maintenance but also expects and asks for different kinds of services. Between those two extreme groups of purists and urbanists there is the great mass of neutralists with more moderate demands upon both untouched wilderness and the touch of human hand. But there are also other differences between the segments.

The methods based on DAD where used to test different aspects in search of clustering behaviour or other deviations from pure randomness amongst purists or urbanists or other subgroups of the sample concerning their place of residence.

The purists appeared to live as they think: they tend to be less clustered than other visitors in the mountain regions. This was indicated by the EDF plots but not confirmed by the tests undertaken.
Even though no obvious results concerning the purists and the urbanists appeared, some other interesting results came out of the study. We could, e.g., see that purists as a group are younger than other categories but the youngest strata (up to 29 years old) appeared to be significantly clustered, whereas the purists tend to be repulsive, which is somewhat contradictory. One explanation could be that the purists rather are rare among the eldest than very common among the youngest. Another finding is that first time visitors proved to be more clustered than others.

The results show that spatial analysis of data can provide useful insights that are not viable when other types of analysis are used.
Acknowledgements

I wish to express my sincere gratitude to all those who have encouraged and supported me and in many ways contributed to this work. In particular, I would like to thank:

**Martin Kulldorff**, my co-author in the first paper, who led me into the field of spatial statistics and generously shared his idea of a density adjusted distance with me.

**Anders Ågren**, my supervisor for the main part of the thesis, who never seemed to loose faith in me throughout the years and always supported me with time and valuable advice whenever needed.

**Tuomas Vuori** for providing the data set used in paper IV.

**Viveca Asproth** at Mid-Sweden University, who encouraged me and helped me to get the extra financial support that was needed when I felt I would never be able to finish this thesis.

**Anders Frank**, the man who took the DAD to the movies by animating parts of the thesis.
References


**Student**, 1907. On the error of counting with a haemacytometer. *Biometrika, 5*, 351-360.


Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Social Sciences 13

Editor: The Dean of the Faculty of Social Sciences

A doctoral dissertation from the Faculty of Social Sciences, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Social Sciences. (Prior to January, 2005, the series was published under the title “Comprehensive Summaries of Uppsala Dissertations from the Faculty of Social Sciences”.)