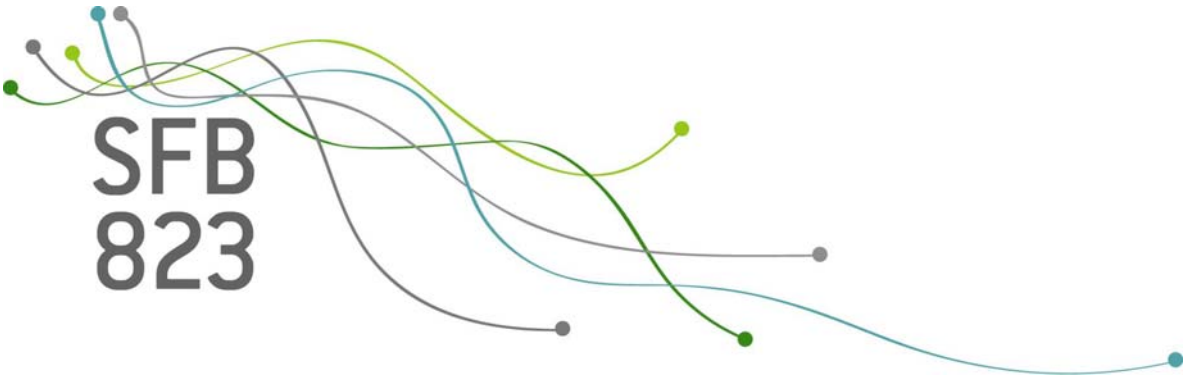
SFB 823

Testing for structural change in spatial regions at unknown positions

Matthias Arnold, Dominik Wied

Nr. 19/2012

Discussion Paper



Testing for structural change in spatial regions at unknown positions

by

Matthias Arnold

Fakultät Statistik, TU Dortmund
D-44221 Dortmund, Germany
arnold@statistik.tu-dortmund.de

and

Dominik Wied*

Fakultät Statistik, TU Dortmund
D-44221 Dortmund, Germany
wied@statistik.tu-dortmund.de

This version: May 10, 2012

Abstract

We propose a fluctuation-type procedure for detecting breaks in spatial regions. While such tests are common in the context of time series, it is not a priori clear how to apply them to spatial data as there is no natural order of the observations. We demonstrate how this order can be constructed from a spatial autoregressive model. Once such an order is derived, standard time series results apply and break points can be consistently identified.

Keywords: Change point; Fluctuation test; Spatial correlation; Spatial order

*Corresponding author. Phone: +49/231/755 5419, Fax: +49/231/755 5284.

1. Introduction

Testing for structural changes in data generating processes has been a standard topic in the statistical literature for a long time. This includes testing for parameter constancy in cross-sectional regression models (e.g. Chow, 1960) as well as in time series models (e.g. Brown et al., 1975). When applying such procedures it is necessary to group the data into possible regimes. This is mostly achieved by some natural order. In a time series regression, this is time. In a cross-sectional regression, this might be the size of some regressor, for instance when estimating the income effect on consumption, it is likewise natural to ask if the income effect is smaller for larger incomes.

However, there is no natural order of the data in spatial contexts. Anselin (1990) analyzes the effect of neglected spatial dependencies on Chow-like tests for structural stability. Here, the classification of the spatial units into different regimes is assumed to be known. López et al. (2010) or Mur et al. (2010) identify different regimes by performing Lagrange multiplier tests for different spatial classifications. Manly and Mackenzie (2000) and Manly and Mackenzie (2003) propose CUSUM methods for environmental monitoring.

The present paper proposes a method for the ordering of spatial data and suggests a fluctuation-type procedure for detecting structural changes in the model parameters. Given that we are able to transform our spatial data into a sequence $y_i, i = 1, \ldots, n$, we use CUSUM methods from the classical econometrics literature to detect structural changes and change points in the mean of this series. The "virtual" transformation into a ordered one-dimensional series guarantees spatially connected regions. Moreover, in contrast to Chow-like tests, we do not have to assume the position of potential change points to be known a priori.

The crucial point of course is to find a suitable transformation. Consequently, our second contribution is a method for concretely performing such a transformation using spatial autoregressive models. The basic idea is that the amount of spatial dependence in different directions is captured by the spatial correlation parameters.

The paper is organized as follows. Section 2 presents asymptotic results (null distribution,

consistency of the change point estimator, local power) for the change point test under the assumption that an order has already been found. Section 3 presents methods how to establish the order in the data. Section 4 provides simulation evidence for the procedures and Section 5 concludes. The proofs of our theorems are deferred to the appendix.

2. The Model and Main Results

This section presents asymptotic results for our procedure to detect spatial changes given that a spatial order is available.

2.1. Testing procedure and asymptotic null distribution For $i = 1, \ldots, n$, let $y_i \in \mathbb{R}^d$ be a random vector of observations. The observations are taken at locations $l_1, \ldots, l_n \in \mathbb{R}^k$ in a k-dimensional space which is equipped with a distance measure $d_{ij} := d(l_i, l_j)$. We maintain the following assumptions:

Assumption 1. *a)* $y_i = \mu_i + \epsilon_i$ *with* $\epsilon_i \sim (0, \Sigma) \forall\ i = 1, \ldots, n$, *where* $\mu_i \in \mathbb{R}^d$ *is a constant vector and* Σ *is a constant* $(d \times d)$*-matrix of full rank.*
b) For $i = 1, \ldots, n$, $l_i \in S$ *where* S *is a compact subset of* $\mathbb{R}^k$.
c) For $i, j = 1, \ldots, n,\ i \neq j$, ϵ_i *and* ϵ_j *are independent.*

Part (a) of Assumption 1 implies finite second moments of the observations, but does not impose any additional restrictions on the distribution of the y_i. We focus on changes in the expectation of the y_i, but in principle our procedure might also be extended to settings where the μ_i are explained by covariates. Note that S, the space of locations, does not depend on the sample size n. Consequently, the corresponding asymptotic concept can be described as in-space-asymptotics rather than S growing with the sample size.
We want to test the null hypothesis of constant expectations across space, i.e.

$$H_0 : \mu_i = \mu_j \ \forall\ i, j = 1, \ldots, n.$$

Let l_0 be a starting point which need not coincide with one of the l_i's. The locations

are ordered with respect to their distance to the starting point such that $l_{(i)}$ denotes the location with the i-th smallest distance to l_0. Thus for $l_{(i)}$ we have that

$$i = \sharp\{k \in \{1, \ldots, n\} : d_{k0} \leq d_{i0}\}.$$

The observation taken at location $l_{(i)}$ shall be denoted by $y_{\{i\}}$.

The idea for the test procedure is as follows: For $j = 1, \ldots, n$, we successively estimate μ from the j observations which are closest to the starting point:

$$\hat{\mu}_j := \frac{1}{j} \sum_{i=1}^{j} y_{\{i\}}.$$

Then, we compare these to the estimate $\hat{\mu}_n$ which is calculated from all observations. The null hypothesis is rejected if the sequence of suitably scaled differences,

$$\frac{j}{\sqrt{n}} \hat{\Sigma}^* \left(\hat{\mu}_j - \hat{\mu}_n \right), \tag{1}$$

fluctuates too much. The unbiased estimator for Σ,

$$\hat{\Sigma} = \frac{1}{n-1} \sum_{i=1}^{n} \left(y_i - \hat{\mu}_n \right) \left(y_i - \hat{\mu}_n \right)^T,$$

is calculated from all observations. Here, A^T stands for the transpose of a vector of matrix A and $\hat{\Sigma}^*$ is a matrix for which $\hat{\Sigma}^* \hat{\Sigma} \hat{\Sigma}^* = I_d$ is fulfilled, e.g. $\hat{\Sigma}^* = \hat{\Sigma}^{-\frac{1}{2}}$. The weighting factor $\frac{j}{\sqrt{n}}$ scales down differences where $\hat{\mu}_j$ is calculated only from few observations close to the starting point.

For $s \in [0, 1]$, (1) can be rewritten as

$$W_n(s) := \frac{[sn]}{\sqrt{n}} \hat{\Sigma}^* \left(\hat{\mu}_{[sn]} - \hat{\mu}_n \right) = \frac{[sn]}{\sqrt{n}} \hat{\Sigma}^* \left(\frac{1}{[sn]} \sum_{i=1}^{[sn]} y_{\{i\}} - \frac{1}{n} \sum_{i=1}^{n} y_{\{i\}} \right), \tag{2}$$

where $[.]$ denotes the floor function.

To derive the asymptotic null distribution of $W_n(s)$, we impose an additional assumption.

Assumption 2. *For* $i, j = 1, \ldots, n$, $i \neq j$, $P(d_{i0} = d_{j0}) = 0$.

Assumption 2 implies that there are no ties in the order of the locations with respect to their distance to the starting point. This guarantees that in the sequence $\hat{\mu}_j$, in each step exactly one observation is added to calculate $\hat{\mu}_j$ which simplifies deriving the asymptotic null distribution of (2). If the locations are random draws from a continuous distribution in $\mathbb{R}^k$, Assumption 2 is automatically fulfilled. It would be violated if the locations formed a regular grid and l_0 coincided with one of the grid points. In such a situation, the problem of ties in the d_{i0} can be circumvented by adding some noise to l_0 so that the new starting point is randomly drawn from a small neighborhood around the grid point. As long as this new starting point does not lie on the connection between two grid points, ties in the d_{i0} can be avoided by this small shift of l_0 and Assumption 2 is fulfilled.
The following proposition yields the asymptotic distribution of W_n.

Proposition 1. Under H_0 and assumptions 1 and 2, for $n \to \infty$,

$$\sup_{s\in[0,1]} ||W_n(s)|| \xrightarrow{d} \sup_{s\in[0,1]} ||B_d(s)||,$$

where $\xrightarrow{d}$ stands for convergence in distribution, $B_d(.)$ is a d-dimensional Brownian bridge and $||.||$ denotes Euclidian norm.

2.2. Detecting change points As far as deviations from the null hypothesis are concerned, our main goal is to detect changes in the μ_i where the expectations are constant in a surrounding area of the starting point l_0, but different for locations with larger distances from l_0, e.g. alternatives of the form

$$H_1 : \ \mu_i = \begin{cases} \mu_1, & d_{i0} \leq d_* \\ \mu_2, & d_{i0} > d_*, \end{cases}$$

for some $d_* \in \mathbb{R}$ and $\mu_1 \neq \mu_2$. Our main focus is to localize the change in expectations, i.e. we want to estimate d_*. Let n_1 be the number of observations which are taken at locations with $d_{i0} \leq d_*$ and $n_2 = n - n_1$ be the corresponding number of observations at locations with $d_{i0} > d_*$.

Assumption 3.

$$\lim_{n\to\infty} \frac{n_1}{n} = s_*.$$

Assumption 3 uniquely relates d_* to s_* so that separation of S into the two subareas with different expectations can be achieved by consistent estimation of s_*.
A natural estimator for s_* is provided by the point where $||\frac{W_n(s)}{\sqrt{n}}||$ is largest:

$$\hat{s} = \underset{s\in[0,1]}{\operatorname{argmax}} ||\frac{W_n(s)}{\sqrt{n}}||. \tag{3}$$

Theorem 1. *Under assumptions 1, 2 and 3, for $n \to \infty$,*

$$\hat{s} \xrightarrow{P} s_*.$$

Theorem 1 guarantees that the subarea around l_0, where the μ_i are different from the expectations in the rest of S, can consistently be identified from the data.

2.3. Local alternatives In this section, we derive the asymptotic distribution of our test statistic under a sequence of local alternatives. This allows for a fine analysis of how the test behaves under the alternative hypothesis. To this end we basically keep the setting from the former sections with the only difference that the random variables y_i now form a triangular array $y_1^n, y_2^n, \ldots, y_n^n$, thus depend on n, and modify Assumption 1.*a* to

Assumption 4. $y_i^n = \mu_i + \frac{1}{\sqrt{n}} g\left(\frac{i}{n}\right) + \epsilon_i$ *with* $\epsilon_i \sim (0, \Sigma) \forall\ i = 1, \ldots, n$, *where* Σ *is a constant* $(d \times d)$*-matrix of full rank.*

In the following, we oppress the additional index for simplicity.

The function g is a bounded function which can be approximated by step functions and which is not identically 0 such that the function

$$\left(\int_0^s g(u)du - s\int_0^1 g(u)du\right)$$

is different from 0 for at least one $z \in [0,1]$. One particular example might be $g(s) = \mathbf{1}_{\{s \geq \frac{1}{2}\}}$ corresponding to a level shift in the middle of the sample. The rate $\frac{1}{\sqrt{n}}$ with which the local alternatives converge to the null hypothesis serves for deriving the asymptotic result

Theorem 2. *Under the sequence of local alternatives, assumptions 1.b, 1.c, 2 and 4, for* $n \to \infty$,

$$\sup_{s\in[0,1]} ||W_n(s)|| \xrightarrow{d} \sup_{s\in[0,1]} \left|\left|B_d(s) + \Sigma^{-1/2}\left(\int_0^s g(u)du - s\int_0^1 g(u)du\right)\right|\right|.$$

The supremum is now taken over the stochastic process $B_d(s)$ plus a deterministic function which depends on the variance-covariance matrix of y_i and the function g.
As a byproduct of the local power analysis we can easily derive consistency of our test. To this end, we rewrite the function g to $g(s) = Mh(s)$ for a bounded function h and a factor M. The function h represents the structural form of the alternative, whereas M captures its amplitude.

Corrolary 1. Let $P_{H_1}(M)$ be the rejection probability for given M under the alternative. Let $\delta > 0$. Then there is a M_0 such that

$$P_{H_1}(M) > 1 - \delta$$

for all $M > M_0$.

This means that local rejection probabilities become arbitrarily large as structural changes are increasing.

3. Spatial distances

Section 2 shows that standard results about breakpoint detection in the time series literature carry over to the spatial context. A central ingredient in our analysis is the order of locations with respect to their distance to the starting point. In the time series context, the observations are taken at different points in time so that the time line provides a natural order. In our spatial context however, observations are taken at points in $\mathbb{R}^k$ so that there is no unequivocal order of the observations. This section presents several ways how the locations can be ordered.

An obvious way is provided by standard distance measures like Euclidean distance $||l_i - l_0||$ or, more generally, a weighted p-norm of the difference between l_i and l_0,

$$\left(\sum_{j=1}^{k} \left| [A(l_i - l_0)]_j \right|^p \right)^{\frac{1}{p}}, \tag{4}$$

with a weighting matrix $A \sim (k \times k)$ and a positive scalar p. Many different distance measures are covered by (4). Still, in practical applications the question arises which one of these measures should be used. Hence, in the following subsection we present an alternative approach to obtain an order of the locations.

3.1. Spatial autoregressive modeling The spatial autoregressive approach for obtaining an order of the locations does not require prior knowledge of the shape of regions, but uses the data to order the locations. To this end, a spatial autoregressive model with different kinds of spatial dependencies is fit to the observations. Each of the spatial dependencies represents dependence in one direction: vertical, horizontal or diagonal. The shape of regions is then determined by the amount of the different spatial dependencies in the data: For example, large horizontal dependence produces regions with large horizontal extent. Different combinations of the spatial dependencies correspond to regions of different shapes so that this flexible approach covers many different possible shapes instead of assuming one particular shape a priori.

We first illustrate this procedure for the case of one-dimensional observations ($d = 1$) and two-dimensional locations ($k = 2$) on a regular grid. Thereafter, some generalizations are discussed. The model is a spatial autoregressive model with

$$y = \rho_1 W_1 y + \rho_2 W_2 y + \rho_3 W_3 y + \rho_4 W_4 y + \varepsilon, \tag{5}$$

where y is the n-vector of observations, W_j, $j = 1, 2, 3, 4$, are $(n \times n)$-dimensional spatial weighting matrices, ε is an n-dimensional vector of innovations with $\mathrm{E}(\varepsilon) = 0$ and $\mathrm{Cov}(\varepsilon) = \sigma_\varepsilon^2 I_n$ and the scalar parameters ρ_1, ρ_2, ρ_3 and ρ_4 have to be estimated from the data. Each row and column of the spatial weighting matrices represents one of the locations. In each row of W_j the nonzero elements correspond to those locations which are direct neighbors in the respective direction. For example, in W_1, the element in row m and column p is nonzero if location l_p is a direct horizontal neighbor of l_m as illustrated in the upper left part of Figure 1.

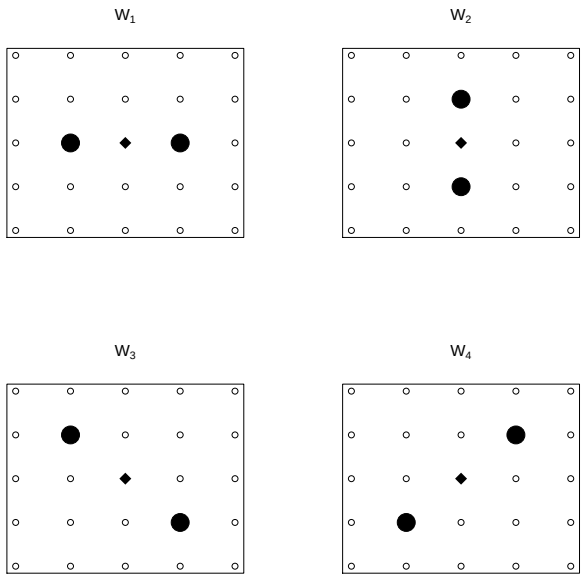


Figure 1: Spatial dependencies in four different directions

Accordingly, the elements $W_{2,mp}$ are nonzero if l_p is a vertical neighbor of l_m and the

nonzero elements of W_3 and W_4 correspond to neighbors in one of the diagonal directions, respectively. Finally, the matrices W_j are row-standardized such that the row sums are equal to one. This way the so called spatial lags $\rho_j W_j y$ capture dependencies in four different directions. The idea to get an order of the locations is motivated by the fact that the amount of spatial dependence in the four directions is determined by the ρ_j. A large value for ρ_1 e.g. corresponds to strong horizontal dependence and will produce regions with large horizontal extent. The formal implementation arranges locations in terms of correlations to the starting point. Note that (5) leads to

$$\mathrm{Cov}(y) = \sigma_\varepsilon^2 (I_n - \rho_1 W_1 - \rho_2 W_2 - \rho_3 W_3 - \rho_4 W_4)^{-1} (I_n - \rho_1 W_1^T - \rho_2 W_2^T - \rho_3 W_3^T - \rho_4 W_4^T)^{-1}. \tag{6}$$

This matrix is well-defined if

$$(\rho_1, \rho_2, \rho_3, \rho_4) \in U := \{(\rho_1, \rho_2, \rho_3, \rho_4) || \rho_1| + |\rho_2| + |\rho_3| + |\rho_4| < 1\}$$

so that we use the parameter restriction given by U in the following.
The unknown parameters ρ_j can be estimated by generalized method of moments (GMM). Since

$$\mathrm{E}(\varepsilon^T W_j \varepsilon) = \mathrm{tr}(\sigma_\varepsilon^2 W_j) = 0,$$

GMM-estimates for the ρ_j are given by

$$(\hat\rho_1,\ \hat\rho_2,\ \hat\rho_3,\ \hat\rho_4)^T = \underset{(\rho_1,\rho_2,\rho_3,\rho_4)\in U}{\mathrm{argmin}} \ \sum_{j=1}^4 \ \left[y^T (I_n - \rho_1 W_1 - \rho_2 W_2 - \rho_3 W_3 - \rho_4 W_4)^T \right.$$
$$\left. W_j (I_n - \rho_1 W_1 - \rho_2 W_2 - \rho_3 W_3 - \rho_4 W_4) y \right]^2.$$

Note that the variance parameter σ_ε^2 need not be known to calculate the $\hat\rho_j$. These estimates are plugged into (6) to obtain an estimate for $\mathrm{Cov}(y)$ (apart from the scalar factor σ_ε^2), which can then be standardized to an estimate for $\mathrm{Cor}(y)$. In the final step, the locations are ordered with respect to their estimated correlation to $l_{(1)}$ such that the

observation with the highest correlation to $l_{(1)}$ attains $l_{(2)}$ and the observation with the smallest correlation is located at $l_{(n)}$.

3.2. Some generalizations To complement this section we shortly discuss generalizations to (i) more than one-dimensional observations, (ii) more than two-dimensional locations and (iii) situations where the locations do not form a regular grid.

In the case of d-dimensional observations ($d > 1$), the spatial autoregressive model (5) can be formulated separately for each of the d components, respectively. The GMM-estimates for the ρ_j are then given by

$$(\hat{\rho}_1,\ \hat{\rho}_2,\ \hat{\rho}_3,\ \hat{\rho}_4)^T = \underset{(\rho_1,\rho_2,\rho_3,\rho_4)\in U}{\operatorname{argmin}} \quad \sum_{r=1}^{d}\sum_{j=1}^{4} \quad \left[y_r^T(I_n - \rho_1 W_1 - \rho_2 W_2 - \rho_3 W_3 - \rho_4 W_4)^T \right.$$
$$\left. W_j(I_n - \rho_1 W_1 - \rho_2 W_2 - \rho_3 W_3 - \rho_4 W_4) y_r\right]^2,$$

where $y_r \sim (n \times 1)$ contains the respective r^{th} components of the y_i. Note that the spatial dependence parameters as well as the spatial weighting matrices remain constant over r in order to provide one unique order of the locations.

A generalization to more than two-dimensional locations ($k > 2$) is equally straightforward - it only leads to more than four directions of spatial dependence. In the case of $k = 3$, if the locations formed a regular grid, there would be 13 different directions leading to 13 different weighting matrices and dependence parameters, respectively. Although the number of different directions grows rapidly with k, the flexible approach remains manageable for larger k. Note that k will rarely exceed 2 in practical applications.

If the locations do not form a regular grid, specification of the weighting matrices W_j gets slightly more involved. There might be no "perfect" direct neighbors like in figure 1 so that one has to be more generous to determine neighbors. The following way seems promising. Suppose we are looking for e.g. a horizontal neighbor to the starting point l_0, but there is no location which exactly lies horizontal to l_0. We suggest to consider locations as horizontal neighbors to l_0, if they are on the one hand close to l_0 (measured

e.g. by Euclidean distance) and are on the other hand at least approximately horizontal to l_0. The latter can be measured by the angle between the difference $l_i - l_0$ and the required direction. This angle should be close to zero or 180 degrees. A formal implementation could determine all locations with angles in $(-5, 5)$ or $(175, 185)$ degrees and select from these the neighbor as the location with smallest Euclidean distance to l_0. To further refine this approach, one could allow for different weights in the spatial weighting matrices W_j such that neighbors get higher weights if the angle is especially close to 0 and Euclidean distance is small. That way one could give more weight to perfect neighbors than to approximate neighbors.

4. Simulation Study

4.1. Setting We simulate the effectiveness of the spatial fluctuation test to detect subareas with different expectations. The locations form a regular $n_x \times n_y$ grid on $[0, 1]^2$ and the starting point is $l_0 = (0.5 + \pi/(n_x^2 n_y), 0.5 - e/(n_x n_y^2))$ which guarantees a unique order of the locations. The locations are ordered with respect to one of the following distance measures:

(I) Euclidean distances $d_{i0} = (l_{i,1} - l_{0,1})^2 + (l_{i,2} - l_{0,2})^2$

(II) elliptical distances $d_{i0} = (l_{i,1} - l_{0,1})^2 + 5(l_{i,2} - l_{0,2})^2 - 4(l_{i,1} - l_{0,1})(l_{i,2} - l_{0,2})$

(III) weighted city block distances $d_{i0} = 4|l_{i,1} - l_{0,1}| + |l_{i,2} - l_{0,2}|$

(IV) spatial distances $d_{i0} = 1 - Cor(y_i, y_{\{1\}})$ with $\rho_1 = 0.2$, $\rho_2 = 0.2$, $\rho_3 = 0.2$, $\rho_4 = 0$.

For each of the four settings, observations are drawn from $N(5, 1)$ for $d_{i,0} > d_*$ and $N(10, 1)$ for $d_{i,0} \leq d_*$, with $d_* = 0.15$ in (I), $d_* = 0.2$ in (II), $d_* = 1$ in (III) and $d_* = 0.999$ in (IV). The different values for d_* ensure reasonable amounts for the respective inner and outer areas. In each setting, we perform the spatial fluctuation test with four different assumptions for the distance measures: the three distances according to (I), (II) and (III) and fourthly the alternative approach described in the last section where the order of locations is estimated from the data. This way we can assess accuracy for true distances as well as for cases where wrong distances are used. In the fourth approach, we do not

apply the parameter constellation of (IV) but separately estimate the ρ_j from the data in each replication. Since we are not interested in rejection probabilities in the first place, we do not specify a fixed significance level α. Instead, out major interest is to assess whether the procedure is able to classify locations into subareas. For each estimation approach, the realizations of (3) provide an estimate for d_* and thus partition the locations into two areas: an outer area (corresponding to locations with $d_{i0} > \widehat{d_*}$) and an inner area of locations which are closer to the starting point.

We choose $n_x = n_y = 15$ and 1000 replications for each setting. For a given location and each of the four approaches, the percentage of replications where the location is regarded as belonging to the subarea around l_0 describes how accurate the partitioning into subareas works. Optimal values of these percentages would be equal to one for locations with true distances $d_{i,0} \leq d_*$ and equal to zero for locations with $d_{i,0} > d_*$.

4.2. Results Table 1 shows the resulting misclassification rates, i.e., the absolute deviations of the simulated percentages from the optimal values, averaged over all locations.

		estimation approach			
	true distance	(I)	(II)	(III)	spatial
(I)	Euclidean	0.004	0.280	0.180	0.043
(II)	elliptical	0.257	0.004	0.396	0.199
(III)	weighted city block	0.187	0.421	0.009	0.133
(IV)	spatial	0.131	0.408	0.248	0.050

Table 1: Misclassification rates

Each row of Table 1 stands for one kind of true distances (I) to (IV) which was used to generate the data. The corresponding simulated misclassification rates for the four different estimation approaches are given in the columns. If the true distances are used, misclassification rates are very small for true distances (I) to (III) (less than one percent). For true spatial distances (IV), the misclassification rate is higher. This is due to the fact that here, instead of simply employing the true distances, the order of the locations is estimated from the data in each replication.

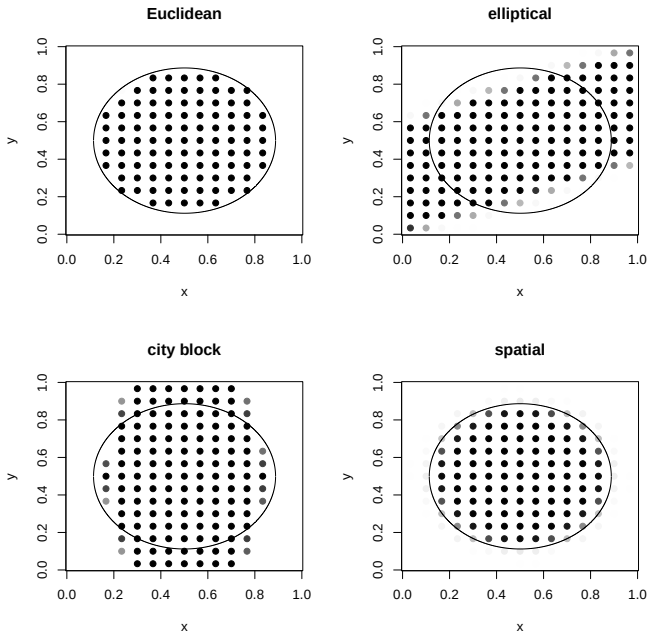


Figure 2: Identification of subareas, true distances Euclidean

If estimation relies on the wrong kind of distances, misclassification rates are much higher. This is caused by the fact that the assumed kind of distances determines the shape of regions which can be identified. For example, when Euclidean distances are used, the regions will always be circles so that elliptical areas cannot be identified. The highest misclassification rate results for true weighted city block distances when estimation relies on elliptical distances. The last column of Table 1 gives the misclassification rates for the spatial approach. Here, the order of locations does not follow a given kind of distances but is deduced from the data as explained in the last section. This leads to medium misclassification rates which are higher than in the optimal case (when the true distances are used), but always lower than in the other cases where wrong distances are assumed. We conjecture that the spatial approach provides some flexibility which of course is inferior to cases where true distances are used, but superior to cases where wrong distances are assumed to be true ones. Comparing the numbers in the last column to each other suggests that it is easier for the spatial approach to detect circles than ellipses or diamonds.
Figures 2, 3, 4 and 5 illustrate the performance of the different approaches.

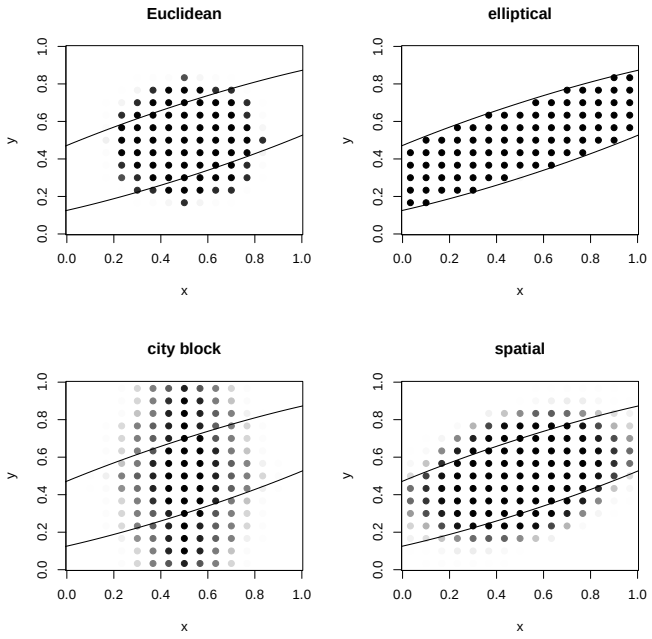


Figure 3: Identification of subareas, true distances elliptical

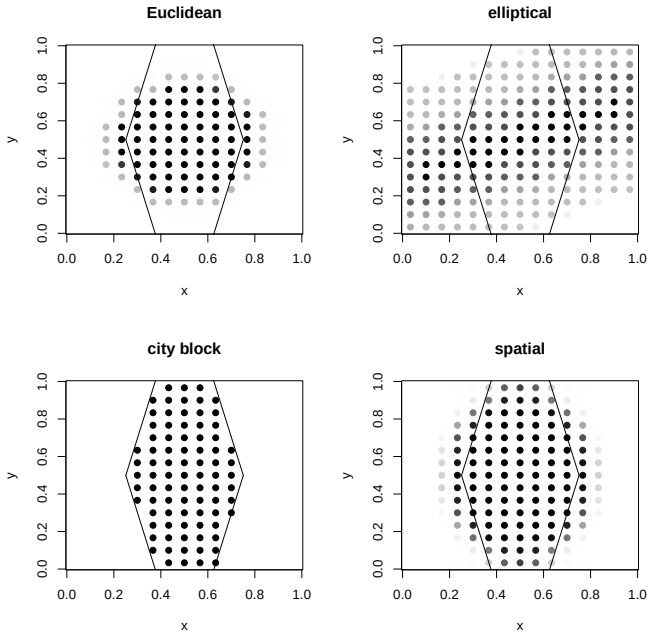


Figure 4: Identification of subareas, true distances weighted city block

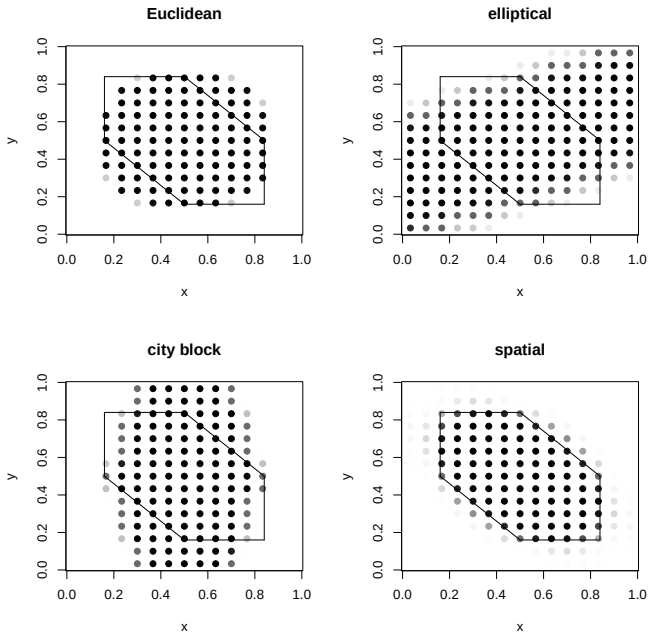


Figure 5: Identification of subareas, true distances spatial

For each of the locations, Figure 2 displays the percentage of the 1000 repetitions in which the location is classified into the inner subarea if the true distances are Euclidean. These percentages are represented by different shades of gray, where dark shades of gray correspond to high percentages. Locations which have been allocated to the inner region in all of the repetitions appear black, locations which have never been allocated to the inner region are not visible (white). The true border between the two subareas is displayed as a solid line which is given by a circle for Euclidean distances. The upper left panel of the Figure shows how well the partitioning works if the true Euclidean distances are used. On the contrary, the results are not satisfactory if elliptical or city block distances are used because this way the circles cannot be identified. The lower right panel of the Figure shows that the spatial approach works reasonably well.

Figures 3, 4 and 5 illustrate the results for the cases where the true distances are elliptical and weighted city block, respectively. Again, the partitioning of locations into subareas is nearly perfect if the respective true distances are used. When the estimation approach makes use of wrong distances, the method fails to identify the true subareas since it

forces the inner subarea to be of a given shape. The lower right panel of the Figures shows the results for the spatial approach which does not require a priori knowledge of the distances. It seems to be some sort of compromise since it provides more flexibility in terms of subarea shapes although the results do not get close to the optimal results when true distances are used. These results conform to the last column of Table 1. It is interesting to see that there seems to be some tendency of the spatial approach to detect bulgy areas. Consequently, it performs better when the true area is a circle whereas it has more difficulties to detect narrow structures like ellipses or diamonds. Still, it seems to be superior to approaches where wrong distances are used.

5. Summary and Discussion

This paper provides statistical procedures how to detect changes in spatial structures. One key point is the transformation of spatial data, measured at different points in space, into a one-dimensional "line" on which classical econometric procedures can be applied. A promising approach to provide a reasonable transformation is the use of spatial autoregressive models with the objective to order the data points with respect to their correlation determined by the spatial autocorrelation parameters.

Another key point is the choice of the starting point l_0. The theory presented in this paper bases on the assumption that this point is known and indeed, there is a natural choice in some applications, e.g. if one is concerned with weather forecasts. If the statistician considers a certain city, he might be interested in the size of the region around this city in which the temperature can be considered as constant. In this case, the particular city would be the natural starting point.

If there is no natural candidate, it is possible to choose the point in which the estimated centered Mahalanobis distance $(y_i - \bar{\mu}_n)^T \hat{\Sigma}^{-1} (y_i - \bar{\mu}_n)$ is maximal. Of course, such a choice will then affect the nominal size of the fluctuation test, but not the resulting spatial regions.

For ease of exposition, we focused on spatial changes in the mean in this paper. Basically,

one might also be interested in changes of other parameters, e.g. in the degree of spatial dependence. We leave it as a task for further research to address these issues in detail.

Acknowledgements:

Financial support by Deutsche Forschungsgemeinschaft (SFB 823, project A1) is gratefully acknowledged. We are grateful to Walter Krämer for useful criticism and comments.

References

Anselin, L. (1990): "Spatial Dependence and Spatial Structural Instability in Applied Regression Analysis," *Journal of Regional Science*, 30(2), 185–207.

Bai, J. and P. Perron (1998): "Estimating and testing linear models with multiple structural changes," *Econometrica*, 66, 47–78.

Brown, R., J. Durbin, and J. Evans (1975): "Techniques for Testing the Constancy of Regression Relationships over Time," *Journal of the Royal Statistical Society B*, 37, 149–163.

Chow, G. (1960): "Tests of Equality Between Sets of Coefficients in Two Linear Regressions," *Econometrica*, 28(3), 591–605.

López, F., J. Mur, and A. Angulo (2010): "Local Estimation of Spatial Autocorrelation Processes," in *Progress in Spatial Analysis - Methods and Applications*, ed. by A. Páez, R. Buliung, J. L. Gallo, and S. Dall'erba, Springer, Heidelberg, Dordrecht, London and New York.

Manly, B. and D. Mackenzie (2000): "A cumulative sum type of method for environmental monitoring," *Environmetrics*, 11(2), 151–166.

——— (2003): "CUSUM environmental monitoring in time and space," *Environmental and Ecological Statistics*, 10, 231–247.

Mur, J., F. López, and A. Angulo (2010): "Instability in Spatial Error Models: an Application to the Hypothesis of Convergence in the European Case," *Journal of Geographical Systems*, 12, 259–280.

Ploberger, W., W. Krämer, and K. Kontrus (1989): "A new test for structural stability in the linear regression model," *Journal of Econometrics*, 40, 307–318.

6. Appendix section

Proof of Proposition 1

With Assumption 1 and 2 and an application of the functional central limit theorem and Slutzky's theorem, $W_n(\cdot)$ converges weakly against $B_d(\cdot)$. An application of the continuous mapping theorem then yields the proposition. ■

Proof of Theorem 1

This follows from the fact that, with Assumptions 1–3, $\frac{W_n(s)}{\sqrt{n}}$ uniformly converges against the function

$$W(s) = \begin{cases} s \cdot (1 - s^*) \cdot (\mu_1 - \mu_2) & s \leq s^* \\ -s \cdot s^* \cdot (\mu_1 - \mu_2) + s^* \cdot (\mu_1 - \mu_2) & s > s^* \end{cases}$$

whose absolute maximum is equal to s^*, compare the argument in Bai and Perron (1998), p. 77. ■

Proof of Theorem 2

The proof follows the arguments of the proof of Proposition 1 with the additional argument

$$\sup_{s \in [0,1]} \left| \frac{1}{n} \sum_{i=1}^{[sn]} g\left(\frac{i}{n}\right) \right| \rightarrow_{n \rightarrow \infty} \int_0^s g(u) du$$

from e.g. Ploberger et al. (1989). ■

Proof of Corollary 1

This follows from the fact that the test statistic becomes arbitrarily large for sufficient large M. ■