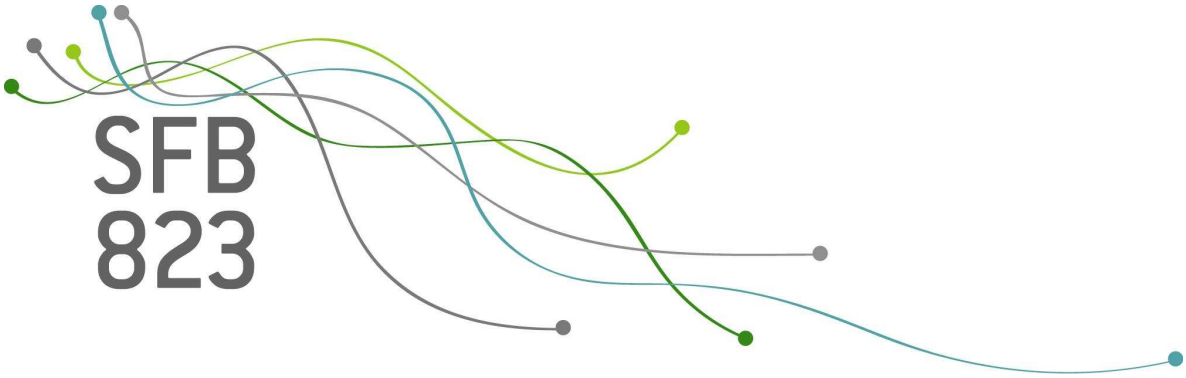
SFB 823

Discussion Paper

Identifying different areas of inhomogeneous mineral subsoil: spatial fluctuation approaches

Matthias Arnold, Nils Raabe,
Dominik Wied

Nr. 44/2012



Identifying different areas of inhomogeneous mineral subsoil: spatial fluctuation approaches

by

Matthias Arnold

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

Nils Raabe

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

and

Dominik Wied*

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

This version: October 22, 2012

Abstract

We use a recently proposed fluctuation-type procedure for detecting breaks in spatial regions to distinguish between hard and soft areas of inhomogeneous mineral subsoil like additives, air pockets and adhesion. For a proper application, some refinements of the procedure are necessary. Both simulation evidence of the refinement and the application on the subsoil yield favorable results.

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

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

1. Introduction

This paper refines a procedure from Arnold and Wied (2012) for detecting structural changes in spatial regions and applies it on inhomogeneous mineral subsoil in order to detect shifts from adhesion to additives or air pockets. Different levels of force signals allow to distinguish between hard and soft regions. The basic idea of the procedure is to transform the spatial data into a virtual time series by obtaining an ordering of spatial data which is mostly not natural in spatial contexts.

While e.g. López et al. (2010) or Mur et al. (2010) identify different regimes by performing Lagrange multiplier tests for different spatial classifications, Arnold and Wied (2012) propose an ordering approach based on spatial autoregressive modeling. The present paper extends this with a polygonal approach for detection of star-shaped regions in order to take the specific structure of inhomogeneous mineral subsoil into consideration. Whereas Euclidean or spatial approaches always detect regions which are point-symmetric around the assumed starting point, the polygonal approach turns out to be more robust against false starting point specification.

After obtaining an ordering, standard CUSUM methods from time series literature are applied on the transformed sequence. 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 paper is organized as follows. In Section 2, we present the practical problem and discuss the interest in distinguishing between hard and soft material areas. Section 3 presents the break detection method under the assumption that an ordering has already been found, Section 4 presents methods to find spatial orderings, Section 5 gives simulation evidence of the refined method and Section 6 presents the application on mineral subsoil. Finally, we give an outlook on possible further research in Section 7.

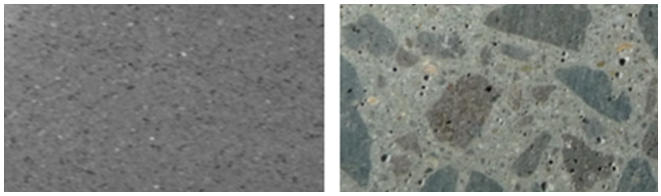


Figure 1: Two different materials. Left: homogeneous additive-free concrete. Right: inhomogeneous multi-phase material concrete.

2. Description of the problems in concrete machining

In general tools used in concrete machining operations are not adapted to the particular machining processes, whereas tool wear and production time are the main cost causing factors. A geometrical simulation model describing cutting forces and wear of both diamond and workpiece was proposed in the past (Raabe et al., 2011). This model takes the abrasive nature of the machined material into account by modeling the microparts of diamond and workpiece as delaunay tessellations of points randomly distributed within the workpiece and simulating the process iteratively. By fitting the model to a series of force signals measured during real experiments the general appropriateness of the model was shown.

An implicit assumption of these fittings is that the connected processes are stationary. However, after investigating real process data in the time domain it turns out that this assumption does not hold. Instead, the forces are obviously affected by material heterogeneity, which is not taken into account in the first stage model. To fill this gap, we now introduce an extension of the simulation model, where the material heterogeneity is modeled and simulated by Gaussian Random Fields.

However, by modeling the material heterogeneity by Random Fields the heterogeneity is implicitly assumed to be continuous. While this assumption is fulfilled for comparably homogeneous materials like additive-free concrete, the case is different for more complex composites like inhomogeneous multi-phase material concrete due to its contained mineral additives and air pockets (compare Figure 1). For this reason a procedure has to be developed to automatically detect shifts from adhesion to additives or air pockets.

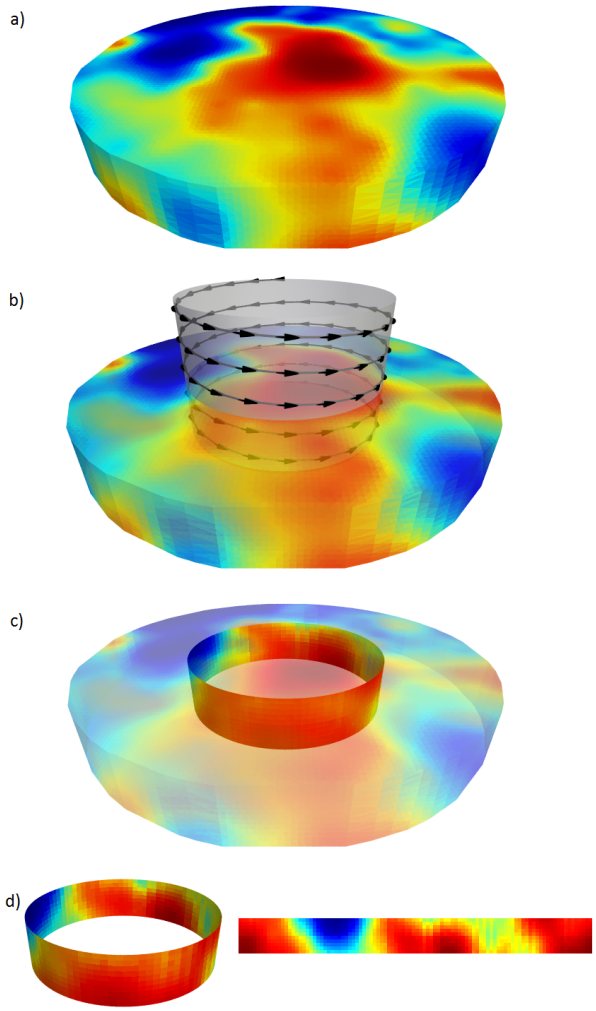


Figure 2: a) Simulated workpiece, b) Course of diamond during process, c) Cylindric bore hole, d) Cylinder cut free (left) and unrolled (right).

The heterogeneity-affected force signals measured during the machining processes are time series in nature. However, by considering their origination from a rotation around the center of the produced hole with fixed radius, rotational frequency and feed, these time series can be matched to the cylindric hole wall and by unfolding the wall to a two-dimensional image (compare sketch in Figure 2).

These 2D-images then can be used as a basis for the detection of spatial regions corresponding to additives or air pockets which are mean shifted in comparison to the concrete adhesion. The following sections describe the procedure we propose to solve this task.

3. A method for detecting structural breaks in spatial regions

This section shortly describes the change point detection procedure from Arnold and Wied (2012) under the assumption that a spatial order is available. In Section 4, we discuss methods to find this order.

For $i = 1, \ldots, n$, let $y_i \in \mathbb{R}$ be force signals observed at locations $l_1, \ldots, l_n \in \mathbb{R}^2$ in a 2-dimensional space which is equipped with a distance measure $d_{ij} := d(l_i, l_j)$ (the choice of the distance measure is discussed in the following subsection). Assuming that $y_i = \mu_i + \epsilon_i$ with $\epsilon_i \sim (0, \sigma^2) \forall\ i = 1, \ldots, n$, where μ_i and σ^2 are scalar constants, the basic idea of the procedure (which is described in more detail in Arnold and Wied, 2012) is to localize changes in the expectations μ_i.

To this end, 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 $y_{\{i\}}$.

We make the assumption that the expectations are constant in a surrounding area of the starting point l_0, but different for locations with larger distances from l_0, so that

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

for some $d_* \in \mathbb{R}$ and $\mu_1 \neq \mu_2$. Localizing the change in expectations is then equivalent to estimating d_*. However, as there are only n observations available, it is not possible to consistently estimate this parameter. Instead, our goal is the estimation of $s_* = \lim_{n\to\infty} \frac{n_1}{n}$ (the limit is assumed to exist), where n_1 is the number of observations which

are taken at locations with $d_{i0} \leq d_*$. d_* is uniquely related to s_* so that separation of S into the two subareas with different expectations can be achieved by consistent estimation of s_*.

The main tool for the estimation of s_* is the function

$$W_n(s) := \frac{[sn]}{\sqrt{n}\hat{\sigma}} \left(\hat{\mu}_{[sn]} - \hat{\mu}_n \right). \tag{1}$$

Here,

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

is the estimator for μ from the j observations which are closest to the starting point and

$$\hat{\sigma} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (y_i - \hat{\mu}_n)^2}$$

estimates the standard deviation σ.

A natural estimator for s_* is then provided by the point where $\left| \frac{W_n(s)}{\sqrt{n}} \right|$ is largest, that means

$$\hat{s} = \underset{s \in [0,1]}{\operatorname{argmax}} \left| \frac{W_n(s)}{\sqrt{n}} \right|. \tag{2}$$

Under additional assumptions, this estimator is consistent for s_*.

4. Finding spatial orderings

Section 3 describes the break detection procedure under the assumption that an ordering of the data points has already been found. This section presents several ways how the locations can be ordered. While the two first approaches (Euclidean distances, spatial autoregressive modeling) have already been discussed in detail in Arnold and Wied (2012), the third subsection proposes a refined polygonal approach for detection of star-shaped regions in order to take the specific structure of inhomogeneous mineral subsoil into consideration.

4.1. Euclidean distances An obvious way is provided by standard distance measures like Euclidean distance:

$$||l_i - l_0|| = \left(\sum_{j=1}^{2} \left| [(l_i - l_0)]_j \right|^2 \right)^{\frac{1}{2}}. \tag{3}$$

4.2. Spatial autoregressive modeling A second way to obtain such an ordering relies on spatial autoregressive modeling. A spatial autoregressive model with different kinds of spatial dependencies is fit to the observations, and the shape of regions is then determined by the amount of the different spatial dependencies in the data.
The spatial autoregressive model for the force signals is

$$y = \rho_1 W_1 y + \rho_2 W_2 y + \rho_3 W_3 y + \rho_4 W_4 y + \varepsilon, \tag{4}$$

where y is the n-vector of observed force signals, W_w, $w = 1, 2, 3, 4$, are $(n \times n)$-dimensional spatial weighting matrices, ε are n-dimensional vectors 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. The so called spatial lags $\rho_w W_w y$ capture dependencies in four different directions: horizontal, vertical both diagonals. 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.
The unknown parameters ρ_w can be estimated by generalized method of moments (GMM). Since

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

GMM-estimates for the ρ_w are given by

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

These estimates provide a plug-in estimate for $\mathrm{Cor}(y)$. In the final step, the locations are ordered with respect to their estimated correlation to $l_{(1)}$.

This approach can further be generalized 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 (compare Arnold and Wied, 2012 for details).

The spatial approach is more flexible than Euclidean distances in the sense that it can also detect non-circular regions. However, both approaches assume the inner region to be point-symmetric around the starting point which will rarely be known in applications. The next subsection suggests a way how to circumvent this drawback.

4.3. Two-dimensional extension for detection of star-shaped regions The originally proposed method for the detection of spatial structural changes has some crucial assumptions. Basically, the area is assumed to be point symmetric due to one specific metric around the starting point, which has to be known. As these assumptions are not realistic in the case of inhomogeneous subsoil, we now propose a two-dimensional extension for a wider class of area shape which furthermore is robust for starting point shift.

For this extension, first the assumption of point symmetric areas is weakened to star shaped areas, i.e. each consecutive area A for all possible rays starting from center $(l_{1;s}, l_{2;s})$ intersects the area borders exactly once. The idea of our extension is to equally distribute rays around the starting point and to determine the intersection point of border and ray in each rays direction. By connecting the crossing points to a polygon the true area border can then be approximated to any precision. If the assumption for the area shape is relaxed to convex shapes - a realistic assumption for concrete additives and air pockets - the shape can be approximated even when the starting point is not the center, as long as it is an inner point of the area.

Practically this approximation is obtained as follows. Consider the data set $y_1, ..., y_n$ with two-dimensional spatial coordinates $(l_{1;i}, l_{2;i}), ..., (l_{1;n}, l_{1;n})$. Set the starting point to

$(l_{1;s}, l_{2;s})$, the number of polygon vertices to P and $\gamma_P = 2\pi/P$. Next the data set is subdivided into P subsets where observation i is assigned to subset j, if

$$(j-1)\cdot\gamma_P < \gamma_i \leq j\cdot\gamma_P \text{ with } \gamma_i = asin\left(\frac{l_{1;i}-l_{1;s}}{\sqrt{(l_{1;i}-l_{1;s})^2+(l_{2;i}-l_{2;s})^2}}\right).$$

Now, the method described in the previous section is applied to each of the P subsets using Euclidean distance and the same starting point $(l_{1;s}, l_{2;s})$ for all subsets. By this, for each subset i a radius d_i of the inner area is obtained. The approximating polygon is then determined by connecting the P points

$$(p_{1;j}, p_{2;j}) = (d_i\cdot sin[(i-0.5)\cdot\gamma_P], d_i\cdot cos[(i-0.5)\cdot\gamma_P]), \qquad i = 1, ..., P.$$

The data set finally is classified by assigning each inner point of the polygon to the inner area.

In a modified version of the extension, a finer polygon is obtained by not only subdividing the data set once. Instead, the reference angle of the subdivision is subsequently varied and, for each reference angle, the data set is subdivided and the method is applied to each temporary subset. By this, formally, the polygon is defined by P_m points, $(m_{1;i}, m_{2;i}), i = 1, ..., P_m$, which are obtained by setting

$$(m_{1;j}, m_{2;j}) = d_{m;j}\cdot sin[(j-0.5)\cdot\gamma_{P_m}], d_{m;j}\cdot cos[(j-0.5)\cdot\gamma_{P_m}]), \qquad j = 1, ..., P_m,$$

where $\gamma_{P_m} = 2\pi/P_m$ and $d_{m;j}$ is the radius that results from applying the original procedure using Euclidean distance and the j-th subset, consisting of the observations i for which

$$\gamma_{P_m}\cdot j - \gamma_P/2 < \gamma_i \leq \gamma_{P_m}\cdot j + \gamma_P/2$$

holds.

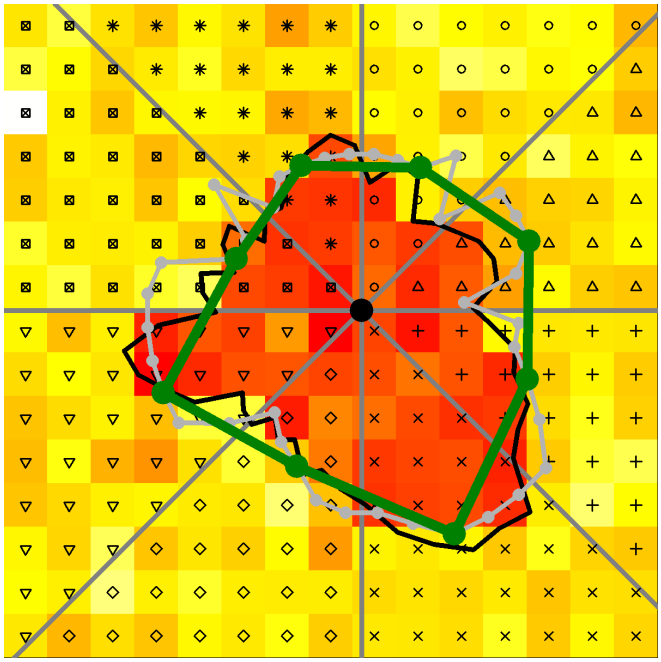


Figure 3: Visualization of the method extension. The symbols denote the eight subsets, the dark gray lines are the subset borders and the starting point is depicted in black. The color scale from yellow (low) to red (high) reflects the values of y_i. The black line shows the true area border, the approximating polygons are in green for the unmodified and in light gray for the modified extension.

Figure 3 visualizes the extension described here.

5. Simulation Study

The introduced method is analyzed on the basis of an extensive simulation study. For this, the following two factors are varied.

- The true shape of the inner area on two levels. The first level is circle-shaped, the second level is star-shaped, where the radius set is a realization of a Gaussian Random Field with exponential covariance function with parameters mean 0, variance 1, nugget 0 and scale 5.
- The (mean) radius of the areas on levels 1 and 1.5.

For all 4 factor combinations 20 data sets are generated by using a normal distribution for y_i with mean 10 and variance 1 for the inner and mean 5 and variance 1 for the outer

area. The coordinates $(l_{1;i}, l_{2;i})$ are given by a regular quadratic 30×30-grid with values between -3 and 3.

For each of these 80 data sets areas are estimated by the original procedure using Euclidean distance and spatial autoregressive modeling, the unmodified and the modified polygonal extension. For the polygonal extensions the parameter P is varied between 8 and 16, where the parameter P_m for the modified extension are fixed to 64. For each data set the starting point is estimated by taking the point for which the sum of y-values in a local neighborhood of size 0.5 is maximal.

By this in total 480 (2 true shapes, 2 shape radii, 4 estimation types, 2 resolutions for two of the estimation types, 20 repetitions) estimations are obtained. The misclassification rates of these estimations are analyzed in the following.

The data is analyzed by a logistic regression model with the misclassification of each single point y_i being the regressand. The factors of simulation and estimation are included as fixed effects, the distance between true center and estimated starting point as covariate and the specific data sets as random effect. Where possible, two-fold interactions are included. The results of this logistic regression model are summarized in Table 1.

The table shows that the majority of the coefficients is statistically significant on the level 5%. However, as two-way interactions are involved, signs and p-values cannot be interpreted directly. Therefore, specific contrasts are now investigated more closely.

Table 2 shows the odds ratio of misclassification rate between the four estimation methods under variation of each of the interacting factors and covariates. Figure 4 visualizes these contrasts by interaction plots.

All contrasts in the table except of the one between *Polygonal mod.* and *Polygonal* for true shape *Circle* are statistically significant on the level 5%.

As can be seen from the interaction plots, for nearly all factor/covariate combinations, the order from worst to best estimation type w.r.t. misclassification is *Spatial*, *Euclid*, *Polygonal*, *Polygonal mod.* This order is violated only once, namely for small distances between the estimated starting point and the true center. There, the unmodified Polyg-

Regressor	Coefficient	p-Value
Intercept	-2.9067	< 0.0001
Estimation type *Spatial*	0.8418	< 0.0001
Estimation type *Polygonal*	-0.1214	0.1600
Estimation type *Polygonal mod.*	0.3966	< 0.0001
True shape *Star*	0.3351	< 0.0001
Shape radius 1.5	0.7458	< 0.0001
$P = 16$	-0.011	< 0.0001
Distance to true center	0.3525	< 0.0001
Est. type *Spatial* / distance to true center	-0.1716	< 0.0001
Est. type *Polygonal* / distance to true center	0.2987	< 0.0001
Est. type *Polygonal mod.* / distance to true center	-0.0246	0.4238
Est. type *Spatial* / True shape *Star*	-0.0319	< 0.0001
Est. type *Polygonal* / True shape *Star*	0.0412	0.2552
Est. type *Polygonal mod.* / True shape *Star*	-0.139	0.00018
Est. type *Spatial* / Shape radius 1.5	-1.2304	< 0.0001
Est. type *Polygonal* / Shape radius 1.5	-1.4607	< 0.0001
Est. type *Polygonal mod.* / Shape radius 1.5	-0.3374	< 0.0001

Table 1: Results of logistic misclassification regression model.

Interacting Factor/Covariate	Level	*Spatial* $\rightarrow$ *Euclid*	*Euclid* $\rightarrow$ *Polygonal*	*Polygonal* $\rightarrow$ *Polygonal mod.*
Distance to true center	0	-0.4724	-1.3106	0.1075
	1	-0.3009	-1.0119	-0.2158
True shape	*Circle*	-0.3627	-1.1051	0.0207
	Star	-0.3307	-1.0638	-0.1596
Radius	1	-0.3307	-1.0638	-0.1596
	1.5	-0.1620	-1.679	-0.2747

Table 2: Misclassification rate odds ratio contrasts between estimation types.

onal extension outperforms the modified one slightly. This implies that the modification increases robustness w.r.t. starting point deviations.

Most of the misclassification rate shifts between levels have the same sign for all estimation types. The only exception here is that for higher radii misclassification increases for the original version, where it decreases in the case of the two extension variants. So, for vaster inner areas the polygonal extension is even more preferable as compared to the original version than for smaller areas.

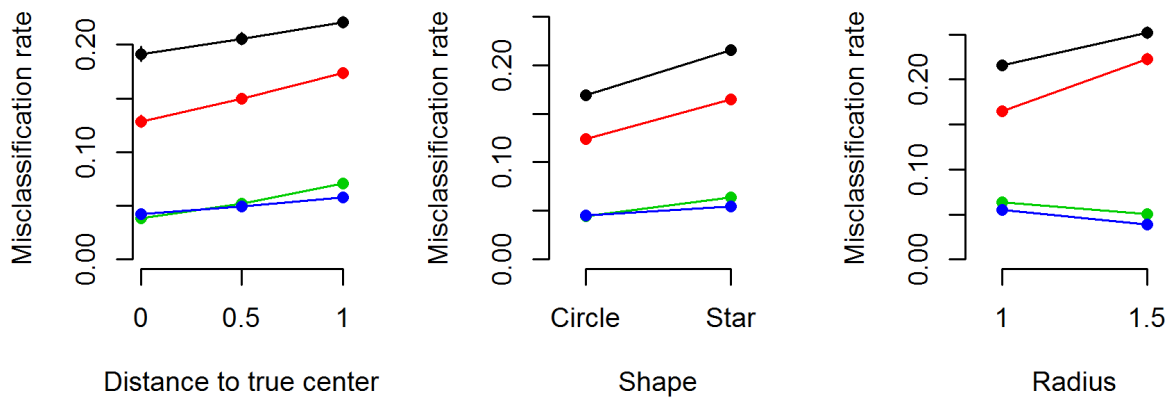


Figure 4: Interaction plots of logistic misclassification regression model. Black: *Spatial*, Red: *Euclid*, Green: *Polygonal*, Blue: *Polygonal mod.* Factors/Covariates which are not varied were fixed to 0.8261 (empirical mean) for distance to true center, *Star* for shape and 1 for radius.

6. Application of the procedure on inhomogeneous mineral subsoil

The procedure described in the previous sections is now applied to data measured during real grinding processes. As the true areas in the real data are not known, it therefore is a mere unsupervised method. To give an impression of how well the method works in the real application, we therefore apply it to a superposed image of a sample taken from a process of grinding into additive free basalt and a high area simulated as in the previous section. Figure 5 shows the superposed image.

Next our procedure is applied to this data using the modified extension with parameters $P = 6$ and $P_m = 64$. Figure 6 shows a comparison of the true area and the estimated polygon. Replications of the procedures to different samples and series show similar results like Figure 6. It turns out that true areas are detected even with low signal to noise ratios, when identification by eye is difficult (compare 5).

Furthermore, the procedure turns out to be robust against the high spatial correlation caused by the material heterogeneity. This heterogeneity is subject of another part of the

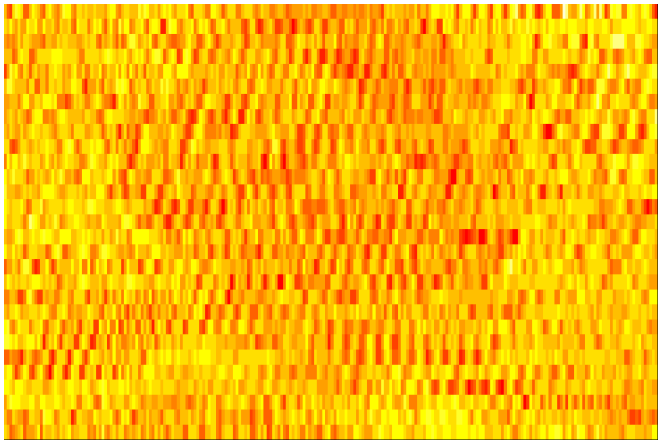


Figure 5: Superposed image of grinding data and simulated additive.

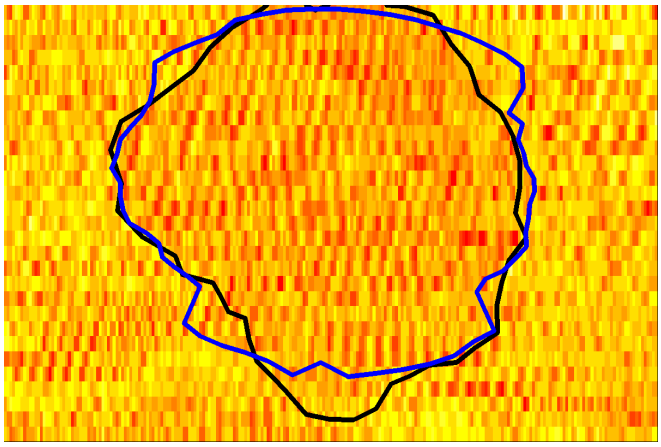


Figure 6: True area border (black) and polygon estimated by modified procedure extension (blue).

project the grinding data is taken from. One of the next steps in the project is to lead the identification of additives presented in this work and the modeling of heterogeneity together and to implement the results to a simulation model of the grinding process. Due to the robustness of the area identification it is straightforward to realize the combination of the two subjects by first identifying additives and then fitting material heterogeneity. As up to now for our procedure only one active consecutive area is assumed, the major task for the implementation will be the extension to multiple areas. A promising approach to solve this task is the pre-identification of additive centers by dense-based clustering methods.

7. Summary and Discussion

In this paper, we use a recently proposed fluctuation-type procedure for detecting breaks in spatial regions to distinguish between different areas of inhomogeneous mineral subsoil. In these regions which correspond to additives or air pockets, force signals are mean shifted in comparison to the concrete adhesion. For a proper application, some refinements of the procedure are necessary. New polygonal approaches are more favorable than approaches based on Euclidean or spatial distances. The superiority of the polygonal approaches is presumably due to robustness against false starting point specifications. While both simulation evidence of the refinement and the application on the subsoil yield favorable results, there are still some issues to be left for further research, e.g. it would be interesting to consider a setting with more than one break.

Acknowledgements:

Financial support by Deutsche Forschungsgemeinschaft (SFB 823, projects A1 and B4) is gratefully acknowledged.

References

Arnold, M. and D. Wied (2012): "Testing for structural change in spatial regions at unknown positions," *Discussion Paper 19/2012, SFB 823*.

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.

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.

Raabe, N., C. Rautert, M. Ferreira, and C. Weihs (2011): "Geometrical Process Modeling of Concrete Machining Based on Delaunay Tessellations," *Proceedings of The World Congress on Engineering and Computer Science*, 2, 991–996.