GI_Forum 2021, Issue 2
Page: 136 - 151
Full Paper
Corresponding Author:
julian.bruns@disy.net
DOI: 10.1553/giscience2021_02_s136

Multidimensional Exploratory Spatial Data Analysis

Oliver Hennhöfer[1], Julian Bruns[1], Peter Ullrich[1], Andreas Heiß[2], Galibjon Sharipov[2] and Dimitrios Paraforos[2]

[1]Disy Informationssysteme GmbH, Germany
[2]University Hohenheim, Germany

Abstract

The assessment of spatial autocorrelation is one of the primary tasks in geographical data analysis. Identifying and examining deviations from the expected autocorrelation is key to gaining a thorough understanding of the phenomenon under investigation. Traditional measures of geospatial sciences focus on the detection of spatial clusters or spatial heteroscedasticity, often in low-dimensional data. However, phenomena are often multidimensional and interdependent – both with and without their spatial dependency – and the toolbox of geospatial sciences is not yet well developed in this regard. The present study aims to contribute to this toolbox for scientists and practitioners. The proposed approach focuses on the detection of spatial discontinuity, considering heteroscedasticity by spatially contrasting residuals from a fitted spatial error model (SEM). This *contrast-enhancing* technique identifies locations whose attributes differ significantly from those of the surrounding features, and with that the technique indicate spatial breaks. The approach is evaluated using agro-ecological field data to identify anomalies and was originally motivated for application in the context of precision farming. Our results enhance understanding of the underlying spatial processes of agricultural fields. The findings contribute to advanced, multidimensional, exploratory, spatial data analysis and present an alternative approach to conventional methods.

Keywords:

heteroscedasticity, heterogeneity, spatial autocorrelation, spatial discontinuity, spatial error model, spatial outlier detection, spatial regression

1 Introduction

Exploratory data analysis is the first step in any data-driven analysis. It allows researchers to pose the question of why something is happening and provides the foundation for formulating hypotheses and for deriving confirmatory analysis (Tukey, 1977). Typical approaches to detect phenomena of interest include outlier or clustering detection methods. Within the field of spatial analysis, we extend our questions: not only why is this happening, but why is this

happening here? Therefore, we are interested in the detection of spatial anomalies. In the context of this work, we differentiate between outliers in general and what we define as anomalies. The difference is that an anomalous observation of interest belongs to the underlying population, whereas the term 'outlier' covers anomalies as well as erroneous observations.

We will focus solely on what we will define as contextual anomalies. These appear exclusively when observations are contextualized using certain observations of other variables – for example, when an unusually high deviation from the observed relationship between two or more observations in one location can be determined. In this study, the presence of potentially prevalent heteroscedasticity will be taken into consideration. A simple example from precision farming can illustrate this. While traditional hotspot analysis may indicate where soils are particularly fertile, contextual anomaly detection can explain where observations were expected based on (known) environmental conditions, and where the actual observations made may indicate an anomaly, according to a predefined model. These are rarely the focus of new methods in the field of typical spatial exploratory analysis, and few methods are in common use. However, in agro-ecological research, we often find that the parameter of interest is (spatially) dependent on several different external factors. For example, the anticipated crop yield may be dependent on weather, different qualities of the soil, and the spatial distribution of fertilizer. As each parameter influences the yield, we would expect their impact to be spatially homoscedastic. However, deviations from this can lead to interesting new questions and insights, often associated with the identification of factors that had been omitted. As the number of parameters adds up, it becomes increasingly difficult to understand the respective models and to identify anomalies – even for trained geographers and experts in this particular domain.

We aim to provide an intuitive and easy-to-understand method for the detection and visualization of these phenomena without extensive input by the analyst. We call our approach the Contrast-Enhancing Spatial Error Model (CESEM).

2 Related Work

The detection of spatial anomalies and unusual spatial patterns in multivariate ecological datasets is a research area that has remained largely untouched. Where powerful machine-learning approaches fail in practice due to low data availability, more traditional statistical methods for outlier detection are not particularly suitable either, since spatial data often violate important underlying statistical assumptions.

Spatial statistics tries to bridge this gap by offering a wide range of tools that are adapted to the characteristics of spatial data, and able to take spatial autocorrelation into account. Two of the most popular methods using spatial statistics for outlier detection – specifically for detecting statistically-significant clusters of higher or lower values of a certain variable in geographic space – are Local Moran's I (Anselin, 1995) and Getis-Ord G[*] (Getis & Ord, 1992) (Ord & Getis, 1995). Several modifications exist for both (univariate) methods, allowing for bivariate or even temporal data analysis, namely the Local Bivariate Moran's (Anselin et al., 2002) and the Local Differential Moran's I (Anselin et al., 2020). However, multivariate

data cannot be analysed appropriately this way, and demands for alternative approaches are still prevalent.

One common way to detect unusual data is regression analysis. This allows for the incorporation of several variables and, like most statistical methods, it can be adapted to a geographically weighted form that meets the requirements of spatial analysis and spatial data. One of the most popular spatial regression methods is Geographically Weighted Regression (GWR) (Brunsdon et al., 1996), which seeks to fit a local model to every feature in a dataset, taking into account locally varying relationships between observations in a study area, preferably for several dozen explanatory variables. This method is therefore able to handle spatially heteroscedastic relationships between the input variables. However, GWR comes with some drawbacks, as discussed by Bivand (2012) (Bivand, 2012) and Wheeler et al. (2005) (Wheeler et al., 2005). Some of these issues have been addressed by the Multiscale Geographically Weighted Regression (MGWR) (Fotheringham et al., 2017), whose developers allowed for each parameter to have a different spatial lag. MGWR learns the different lags dynamically through an iterative approach: after an initial distribution of spatial lags and then fixing all but one spatial lag, the optimal lag for this parameter is computed and then iterated for every parameter. After one overall iteration, a defined convergence criterion is checked for, and the process is repeated. This allows (M)GWR to be a powerful tool for the exploration of high-dimensional spatial datasets.

Other spatial regression approaches are commonly applied in econometrics and can model the data in a fashion similar to the non-spatial regression model (OLS) while considering the effects of spatial autocorrelation. These models extend the non-spatial linear regression model by three different spatial effects (Manski, 1993):

- Endogenous Effect: The behaviour of a spatial analysis unit in geographic space (regressand) depends on the behaviour of other spatial analysis units in proximity.[1]
- Exogenous Effect: The behaviour of a spatial analysis unit in geographic space (regressand) depends on the behaviour of the independent explanatory variables (regressors) of other spatial analysis units in proximity.
- Correlated Effect: The behaviours of a spatial analysis unit in geographic space are alike because they 'face similar institutional environments' (Manski, 1993, S. 533), but they do not directly influence each other by their own behaviour.

The theoretical model that incorporates every spatial effect is called the Manski-Model. It can be restricted to a range of other models that incorporate different combinations of spatial effects, applicable for the different phenomena under investigation:

$$Y = \rho WY + X\beta + WX\theta + u$$

$$u = \lambda Wu + \epsilon$$

where:

- WY describes the endogenous effect of spatially lagged variable y on Y,

[1] Cf. *peer pressure* in social contexts.

- ρ is the spatial autoregressive coefficient controlling the effect of the spatially lagged variable y,
- $X\beta$ describes the exogenous effects,
- WX describes the exogenous interaction effect, or the effect of the spatially lagged variable X on Y,
- θ controls the effect of the spatially lagged variable of all variables in X,
- u and ϵ are error terms of unobservables,
- Wu describes the correlated effect or the effect of the spatially lagged variable of the unobservables[2],
- λ controls the effect on Y of the spatially lagged variable of the residuals.

Many model selection procedures exist that help to determine which effects to exclude, as described in (Anselin et al., 1996), (LeSage et al., 2009), (Darmofal, 2015), (Floch et al., 2016) and (Elhorst, 2010), amongst others. Although those effects (and models) were established in more of a socio-economic context, they can be adapted to spatial data in other contexts.

In addition to the methods reviewed so far, which are comparatively well-known in spatial analytics, numerous algorithms exist that focus on spatial outlier detection. Some examples are described in the works of (Kou, Lu, & Chen, 2006), (Sun & Chawla, 2004), (Takeuchi & Yamanishi, 2006), (Liu, Ting, & Zhou, 2012) and (Chen, Lu, & Boedihardjo, 2010). Each suggests a different outlier detection algorithm, but each identifies outliers by the deviation from neighbouring points. While the outlier detection algorithm introduced in (Kou, Lu, & Chen, 2006) was applied to cleanse the yield data in our work, the approaches generally focus on spatial outlier detection as part of data preparation, rather than for actual spatial data exploration.

The findings in the literature suggest that the range of tools available for multivariate spatial data exploration is limited (the detection of potentially erroneous points aside). Furthermore, each method suitable for spatial anomaly detection is based on a different concept, making direct comparisons difficult. Thus, additional methods are needed for the identification of anomalous behaviour in spatial datasets.

[2] For example, in cases of spatial heteroscedasticity the residuals ϵ are autocorrelated.

3 Methodology

Data A spatial dataset containing one dependent variable and at least one independent variable. Spatial data from different observations should be interpolated in a common regular grid raster.
Result Statistical significance (z-scores) of the residuals from the fitted spatial error model.

define spillover-function;
    #weighting function for spatial neighbors
fit spatial error model;
    #direct neighborhood
calculate *Global Moran's I* and *SOH;*

while *Global Moran's I* ≲ *0.6* & *SOH* ≲ *0.9*:

    increase neighborhood;
    #i.e., the *spillover*
    calculate *G. Moran's I* and *SOH;*

end
calculate residuals;
calculate residual significances;

It should be noted that for a spill-over function that puts exponentially more weight on direct neighbors than on more distant neighbors, both indicators will converge sooner towards their global maximum, as illustrated in Figure 3 (right).

Figure 1: Brief outline of the CESEM algorithm.

Our approach of multivariate spatial anomaly detection is based primarily on the application of spatial regression models and the assessment of the models' residuals. In short, residuals that deviate significantly from all other residuals will be identified as anomalous.
For the detection of multivariate spatial anomalies by means of spatial regression, we propose a contrast-enhancing technique based on the Spatial Error Model (SEM) (Anselin L. , 1988). The ordinary SEM extends the linear model[3] by an error term $u = \lambda W u + \varepsilon$. This allows the residuals to spill over spatially, with $y = X\beta + u$. Usually, the error term in the SEM adjusts for local deviations from the non-spatial linear model (OLS) by incorporating the error term as an omitted or unobserved, but spatially autocorrelated, variable.

[3] The SEM can in turn also be seen as a restriction of the *Manski-Model*.

The technique proposed here provides for an artificial increase4 of the number of neighbours affected for each observation, and with that for the expansion of the spill-over in the error term. In practice, this leads to overlapping spill-overs of errors from different points that are in proximity to each other. For similar points, this spill-over of errors does not greatly impact the predictions made by the model, which can adjust for slight deviations from the predictions. However, when points differ greatly, numerically speaking, from their surroundings, the model cannot adapt. This results in even higher model deviations for these points, since their predicted value further increases or decreases. Model deviations that are detected indicate spatial discontinuity and will subsequently be identified as anomalous, provided that the deviation of the residuals is statistically significant. Due to the locality of the error spill-overs, the method is not overly prone to heteroscedasticity if the change happens gradually (continuously) and not suddenly (discontinuously).

For purely predictive purposes (for which the SEM is usually applied), considering direct neighbourhood would most often lead to the highest model precision of the SEM, since the model can adapt readily to the most minute variation within a given area. By extending the spill-over, there is a trade-off: the SEM becomes practically impaired in exchange for more spatially autocorrelated residuals, without a complete generalization back to the non-spatial linear model (OLS).

In practice, increasing the area defined as a neighbourhood can be realized by two parameters: (i) modelling the spatial influence of the spill-over in the error term; (ii) the extent (distance) of the spill-over itself. As part of this work, an inverse distance function is defined that increases the impact of error spill-overs on immediate neighbours. The greater the impact of the error on more distant neighbours, the larger the spatial anomalies obtained may be. In the extreme case of every residual from every point spilling over to the rest of the points, the SEM could potentially become comparable to the non-spatial linear model.

A more detailed example of the effects of the CESEM is presented in Section 3.2.

3.1 Parametrization

For the parametrization of the CESEM, two measures were calculated for two different powers of the inverse distance function by gradually increasing neighbourhood radiuses: (i) the Global Moran's I (Moran, 1950) (Cliff & Ord, 1972) for the quantification of the global spatial autocorrelation of the residuals across the entire study area; (ii) the Stability of Hotspots (SOH) (Bruns & Simko, 2017) for the quantification of the magnitude of change between (significant) residuals obtained from the iteratively computed SEM.

[4] *Artificial* in the sense that the predictive abilities of the fitted SEM would potentially be more accurate with a sparser spatial weights matrix (i.e., if a smaller neighbourhood was defined).

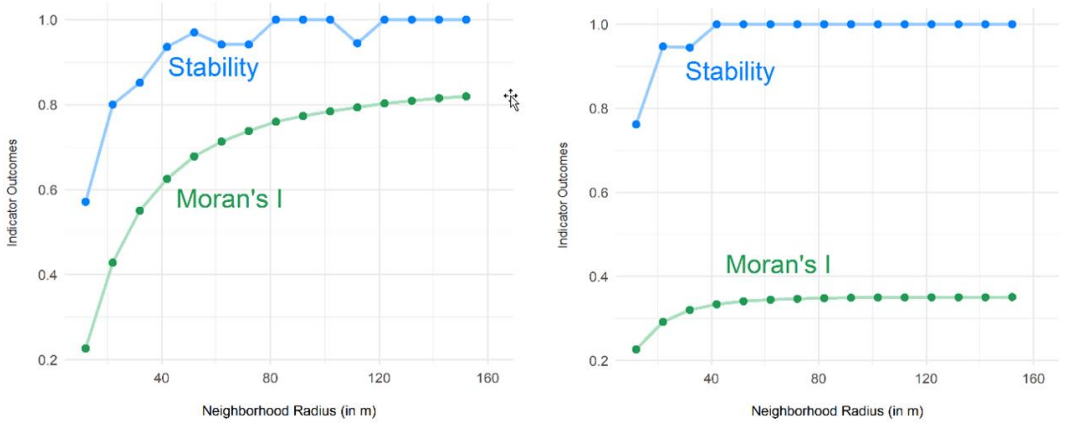

Figure 2: Stability of Hotspots and the Global Moran's I of significant residuals ($\alpha = 0.5$) of a SEM calculated for an increasing neighbourhood size of ½ (left) and an inverse distance power of 4 (right). The effect of the neighbourhood radius for higher inverse distance powers is much smaller, since the impact of the error term peters out faster, with no meaningful impact whatsoever on more distant neighbours.

The calculation of the Global Moran's I gives information about the spatial autocorrelation of the residuals of the SEM. The smaller the defined neighbourhood, the less spatially autocorrelated the residuals will be. While precisely this outcome is usually intended when the calculation is applied in order to obtain the underlying SEM (for the sake of an accurate spatial prediction), it may be less useful for the detection of spatial anomalies, since the SEM adapts to the most minute spatial variations for these smaller neighbourhoods. On the other hand, the larger the neighbourhood, the more similar the model will be to the non-spatial linear model, since the errors will cancel each other out until the predictions resemble the regression mean.

The chief intention of the model parametrization here is to generalize the SEM by a spill-over expansion until the residuals start to cluster spatially. The main idea of the CESEM is that the residuals that start to cluster first are the ones that represent the most severe model deviations, which are particularly difficult to predict for the SEM due to their variable characteristics.5 The parametrization process seeks to determine for which parameter the SEM falls apart first, and which residuals first start to cluster spatially.

The SOH, on the other hand, gives information about the stability of the resulting hot- and coldspots. The measure compares the clusters for the smaller neighbourhood with those of the next-largest neighbourhood and quantifies the similarity between them (comparable to the computation of a difference map). For higher values of the SOH, the clusters of significant residuals remain stable for the next-largest neighbourhood. We aim for a convergence of the SOH values, which would indicate that the gradual extension of the spill-over has stopped. To date, we have not been able to find a fully standardized and automated approach for this

[5] Here, defined by an [assumed] linear relationship between the input variables.

problem. However, plots like the ones in Figure 4 can serve as a reference for how to parametrize the CESEM.

In this example, the choice of an inverse distance power is driven mainly by the spaciousness of potential anomalies. Lower inverse distance powers, which are able to compensate for smaller irregularities, will result in more extensive trend-deviating (regional) clusters as the error term is propagated to more distant spatial neighbours. In turn, for higher inverse distance powers, the current spatial analysis unit is compared to spatially closer neighbours, resulting in the detection of spatially less extensive (local) clusters.

The spill-over itself can be modelled by practically any function and is required for the computation of the CESEM. The primary effect of the distance band for neighbourhood definition is that it increases computational performance, since for an inverse distance function the area of meaningful impacts by the error term is spatially restricted anyway.

Based on this initial approach, the CESEM was parametrized with a neighbourhood radius of 40 metres and an inverse distance power of $\frac{1}{2}$. From the corresponding plot based on this parametrization (see Figure 2a), it can be observed that at about 40 metres (1) the SOH starts to peak, and (2) the residuals are about to cluster spatially.

3.2 Contrast-Enhancing Effects

The following example illustrates the contrast-enhancing effects that can be achieved by extending the spill-over of an error term in a SEM.

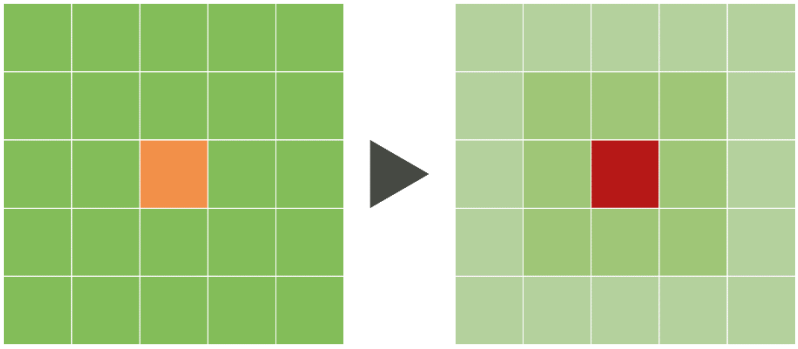

Figure 3: Visualization of the impacts of an extended (contrast-enhancing) spill-over effect of an error term from a SEM.

The visualization of residuals for an exemplary non-spatial linear regression (OLS) – as depicted in Figure 3 (left) – between a dependent variable and one or more explanatory variables shows an overestimated spot (orange; negative residual) in the centre of an exemplary geographic space, surrounded by underestimated spots (green; positive residual). The application of CESEM depicted in Figure 3 (right) now incorporates a spatially dependent error term spilling over to adjacent sections in geographic space. As a result, the overestimated spot in the centre becomes more significant, since the positive error terms spilling over from

the surrounding underestimated spots mean that this central spot deviates even more strongly from the predicted value of the SEM.

For the very same reason, underestimated spots now appear to be less significant, due to the compensation by overlapping (positive) errors adjusting for their collective underestimation. Since the spill-over has a greater effect on immediate neighbours (because of the impact modelled by an inverse distance function), the spots adjacent to the centre are in turn also affected by the negative error term of the centre counteracting (to some extent) the adjustment. As a result, the location of the highest spatial discontinuity (i.e. the centre and adjacent spatial analysis units) becomes more significant compared to the entirety of the residuals in the hypothetical area of study, which appear to be spatially continuous due to their similarity. The second consequence is that the central spot is penalized additionally since it is located between spatial units that appear to have strongly differing characteristics. The application of the CESEM, therefore, primarily penalizes spatial discontinuity based on a global and linear relationship between regressand and regressor(s).

Due to the focus exclusively on model residuals for anomaly detection, the SEM is predestined to be the foundation for CESEM, as it tries to adapt the model to the data by means of spatially autocorrelated residuals only. Another advantage is that the interpretations of model coefficients for the SEM and for the non-spatial linear model (OLS) correspond to each other. This stands in contrast to other spatial regression models: in other models, the interpretation of coefficients (and hence model comprehensibility) can become non-trivial because of the incorporation of spill-over effects for the dependent and/or independent variables.

However, for the sensitivity of the application of the CESEM, the theoretical requirements for the application of an ordinary SEM must be met and assessed by the relevant model selection processes, as stated in Section 2.

4 Evaluation

4.1 Dataset

The technique will be demonstrated for cleansed[6] and interpolated[7] yield data and apparent soil electrical conductivity measurements (ECa), both sets of data collected in an experimental field (Lammwirt[8]) at the Ihinger Hof research farm of the University of Hohenheim.

[6] The unprocessed yield data were cleansed using the *Averaged Difference Algorithm* as proposed in (Kou, Lu, & Chen, 2006).

[7] The processed yield data were interpolated using *Ordinary Kriging* (spherical variogram function considering the 50 nearest observations) and then converted onto a hexagonal grid with a resolution of 10 m.

[8] Location: N48°44'50'', E8°54'33''.

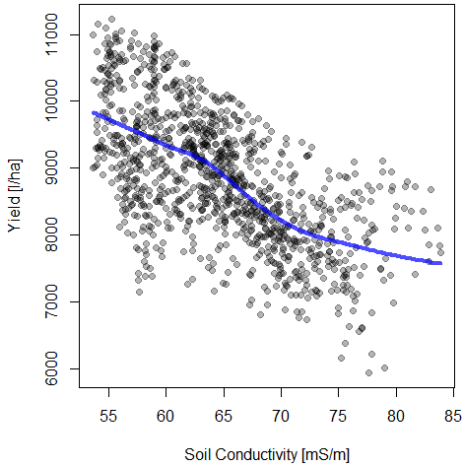

Figure 4:
Relationship between crop yield and electrical conductivity of soil. A locally weighted scatterplot to which smoothing (LOWESS) was added for the identification of potential structural breaks in the data.

The two variables studied exhibit a fairly strong negative correlation to each other (compare Figure 4 and Table 1), which will be examined in order to identify any anomalous field sections. In general, it can be stated that higher values for soil electrical conductivity typically correspond to soil characteristics that lead to an impaired crop yield (Kitchen, Sudduth, & Drummond, 2003). Examples of such soil characteristics may be waterlogging due to high clay content, or overly saline soil. Both high clay content and saline soil exhibit higher electrical conductivity.

Table 1: Correlation coefficients between soil conductivity and yield obtained for the non-spatial Pearson correlation coefficient (r_p), the Spearman's rank correlation coefficient (r_s), and for the two spatial correlation measures Moran's I and Lee's L (Lee, 2001). The global bivariate Moran's I is identical to the mean of the total of the local bivariate for Moran's I for an area.

Year	r_p	(r_s)	p	Moran's I global, bivariate	p	Lee's L	p
2007	-0.571	(-0.537)	<0.0001	-0.537	<0.0001	-0.530	<0.0001
2008	-0.504	(-0.515)	<0.0001	-0.455	<0.0001	-0.440	<0.0001
2010	-0.583	(-0.577)	<0.0001	-0.548	<0.0001	-0.541	<0.0001
2012	-0.493	(-0.505)	<0.0001	-0.483	<0.0001	-0.487	<0.0001
average	-0.630	(-0.635)	<0.0001	-0.595	<0.0001	-0.588	<0.0001

The data were interpolated upfront and converted onto a regular grid for data homogenization as an efficient way to define neighbourhoods. Regular hexagonal grids are particularly suitable for this purpose, since the neighbourhood definition becomes unambiguous.

4.2 Approach

In general, evaluating the results of the CESEM by using results from other approaches is difficult since no comparable approach is readily available. We compared our results to those for Local Bivariate Moran's I (LBMI) as the parametric and computational restrictions of this very method led to the conception of the approach of this paper. However, the CESEM is not intended to enhance the results obtained for the LBMI; rather it aims to provide an approach that complements the findings and tries to overcome some of the main (computational) limitations of the LBMI's detective abilities.

The LBMI represents a modification of the (univariate) Local Moran's I and allows for the incorporation of a dependent and an independent variable; it is based on the correlation of observations from one data layer and the spatially lagged observations of another data layer for identical locations.

$$I_{B,i} = x_i \sum_{j=1}^{n} w_{i,j} y_j$$

where:

- x_i is an original observation and y_j the corresponding spatially lagged observation of another variable
- $w_{i,j}$ are the weights assigned to the neighbour (row-standardized; $\sum w_{i,j} = 1$)

The cluster designations obtained using this calculation are comparable to those of the Local Moran's I. As already stated, the computational limitation of the LBMI lies in its great dependence on the (normalized) means of the respective variables of both layers, which results in a limited detective ability (see Figure 5). Consequently, while the method can detect clusters based on the correlation between an original observation and a spatially lagged observation of another variable, it is unable to differentiate between strong or moderate occurrences within such trends. The statistical significance is determined by a random permutation test for which merely pseudo z-scores can be obtained, which of course comes with certain drawbacks.

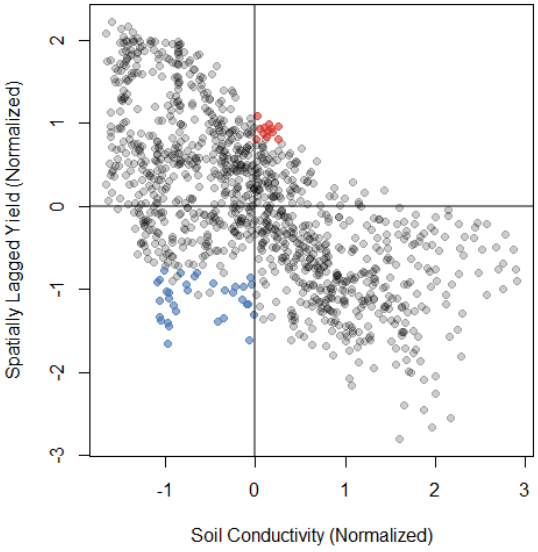

Figure 5:
Scatterplot showing the normalized input data and the coloured (pseudo-)significant data points. The strong break at the normalized means becomes evident.

Since the approaches are not directly comparable with each other, no difference maps or statistical comparisons were generated. Instead, the results will be compared visually, supplemented by expert knowledge of the locality itself (if available).

4.3 Results

What follows aims to demonstrate the operating principles of the CESEM, illustrated by a direct comparison of selected findings for both the CESEM and the LBMI.

One field section that could be identified a priori as a potential anomaly, according to domain experts, is the outer south-eastern corner of the field, the access point for any agricultural machinery used to till, fertilize, and eventually harvest the field. This area is characterized by overall lower plant productivity due to soil compaction from frequently being driven over, and mechanically-induced stress to the plants. The CESEM was able to identify a coldspot at this very location: the model prediction was significantly lower than the expected value, i.e. even after the automatic adjustment by the model to a lower value (spatial heteroscedasticity) to take into account the localized effects of the farm machinery. The LBMI does not indicate any anomalies for the same region.

For this example, the differences between the methods become evident. The CESEM identifies a suspiciously strong deviation from the modelled trend, despite local adjustments introduced by the error spill-over. The LBMI, on the other hand, is able to identify field sections that oppose the modelled trend directly. However, it is computationally incapable of identifying suspicious deviations within such trends. In this case, both the yield and the soil conductivity are below average (see Figure 5), but this does not constitute a violation of the (modelled) underlying trend. This example demonstrates how the CESEM can be applied to supplement and complement the findings of the LBMI.

The next example demonstrates the limitations of the CESEM which should be considered at the outset, before any interpretation of its findings. In this case, the notable accumulation of coldspots in the centre of the research field are examined more closely.

In the visualization of the average crop yields (Figure 6, upper right), an area of lower productivity can be identified. The LBMI identifies two cold spots for this region. Again, this can be explained by a positive correlation between crop yield and soil conductivity, a correlation which goes against the general trend. The very same areas can be identified when the CESEM is applied, although the geographical extent appears to be smaller. Additionally, the CESEM yields a third coldspot, one which protrudes into an area of increased conductivity and thus actually fits the model. Here, the effects of the spill-over become apparent. This third coldspot obtained by the CESEM does not conflict with the underlying model, but solely because it is surrounded by field sections of above-average productivity, especially towards the west. As a result, the expected yield value is much higher than the actual yield. Therefore, it should be kept in mind that the CESEM is not only a reverse correlation between variables but also a spatial discontinuity, even though the observations fit

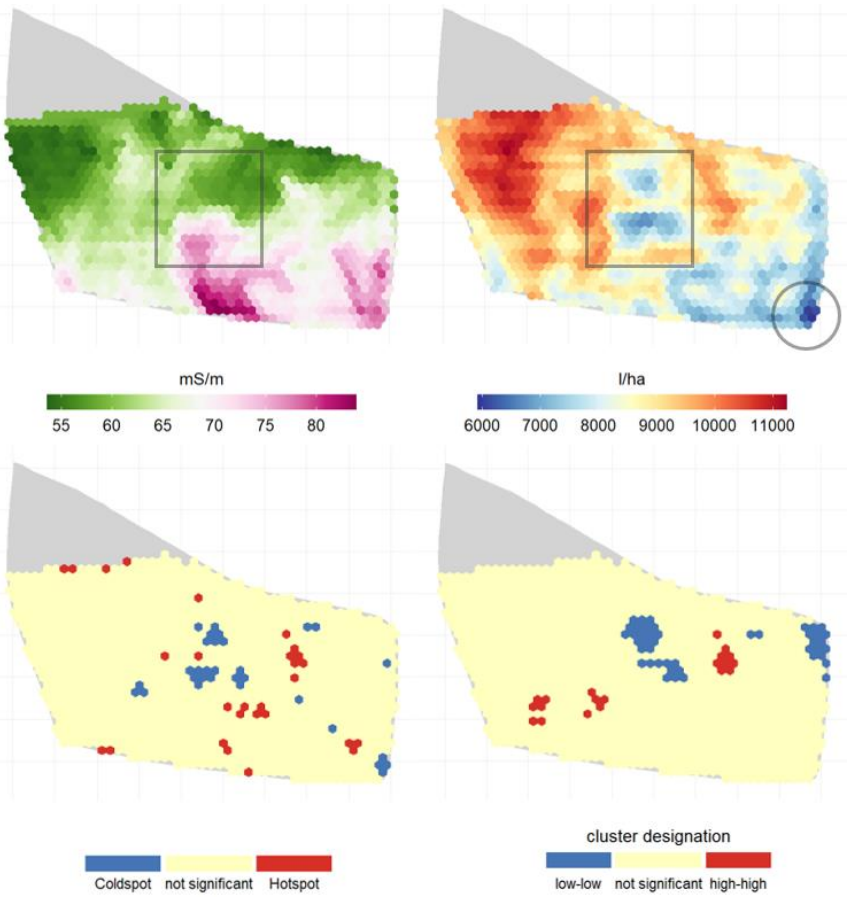

Figure 6: Input data consisting of the soil conductivity (upper left) and crop yield (upper right) and the respective results for the CESEM (lower left) and the LBMI (lower right).

The next example demonstrates the limitations of the CESEM which should be considered at the outset, before any interpretation of its findings. In this case, the notable accumulation of coldspots in the centre of the research field are examined more closely.

In the visualization of the average crop yields (Figure 6, upper right), an area of lower productivity can be identified. The LBMI identifies two cold spots for this region. Again, this can be explained by a positive correlation between crop yield and soil conductivity, a correlation which goes against the general trend. The very same areas can be identified when the CESEM is applied, although the geographical extent appears to be smaller. Additionally, the CESEM yields a third coldspot, one which protrudes into an area of increased conductivity and thus actually fits the model. Here, the effects of the spill-over become apparent. This third coldspot obtained by the CESEM does not conflict with the underlying model, but solely because it is surrounded by field sections of above-average productivity, especially towards the west. As a result, the expected yield value is much higher than the actual yield. Therefore, it should be kept in mind that the CESEM is not only a reverse correlation between variables but also a spatial discontinuity, even though the observations fit the underlying model.

In practice, these findings could now be used to identify other omitted or unknown ecological variables that might explain plant behaviour, and the CESEM could be applied in conjunction with other methods such as the LBMI, (M)GWR etc.

5 Discussion and Conclusion

Here, we have presented the CESEM as a new technique for the application of Spatial Error Models to identify multidimensional spatial hot- and coldspots, not only under the consideration of spatial autocorrelation but also spatial heteroscedasticity. The model compares neighbouring areas by applying the model correction of a SEM for one area to that of an adjacent area, in order to detect any sudden spatial breaks within a defined (linear) model and its inherent trend. Due to the spatially limited range of the comparisons that the technique allows, the technique is not overly prone to heteroscedasticity unless there are sudden changes (discontinuity) in the observed variances.

Our method is based on the combination of classical hotspot analysis and spatial regression analysis. By focusing more on the spatial discontinuity between the interaction of different dimensions and increasing the comparable neighbourhood, the CESEM contrasts the spatial residuals and resulting hot- and coldspots. It performs well in comparison to other approaches. We evaluated our method with real-world agricultural field data and demonstrated that the method provides advantages in the exploration of existing spatial phenomena. Discussion with agricultural experts allowed the results to be verified and explained. The approach improves visual detection of anomalies and therefore also improves time-efficiency for their analysis. An implementation of the method is already available as a webservice within the project iFAROS.

However, there are several limitations which need to be considered. First, the evaluation of methods for exploratory data analysis is quite difficult (see e.g. (Ben-David & Ackermann, 2008). While (Bruns & Simko, 2017) provide an approach using SOH, it was evaluated only for one-dimensional spatial data. Second, while the results were discussed with experts for the

specific use-case, a broader evaluation in other contexts, using datasets from different scientific fields, would enable a more general evaluation of the method and its applicability. Finally, we compared the approach solely to the Local Bivariate Moran´s I. During our literature research, we did not find any comparable approaches for the challenge we are aiming to meet, namely improving precision agriculture. An in-depth comparison with other spatial regression approaches, and analysis of variants of kriging or multi-stage hotspot approaches could be of interest. However, this is beyond the scope of the present study. Our focus is on providing a simple, intuitive approach for computation and visualization by researchers and practitioners, who often do not have the means to carry out the alternative approaches discussed here.

In the future, we aim to remedy the limitations to which we have pointed. A more in-depth evaluation using more datasets and different approaches would be highly interesting to the authors, as would testing the CESEM using more variables, which were reduced in number for the present study in favour of comparability to the Local Bivariate Moran's I. Classic ecological topics such as water or air pollution, or more human-centred ones such as mobility or irregularities in complex supply chains, are interesting fields that could generate numerous use-cases for future studies. These are areas that require urgent investigation, but also specialist knowledge to understand and evaluate the results. In addition to evaluating the method's reliability on real-world data, the use of synthetic datasets, for an even more reliable assessment is also planned.

Acknowledgements

We would like to thank Erik Haas for his valuable inputs and interesting discussions during the development of the CESEM method.

This work has been partially funded by the German Federal Ministry of Food and Agriculture (Bundesministerium für Ernährung und Landwirtschaft, BMEL) through the ICT-AGRI ERA NET project iFAROS (Decision Support for Optimized Site-Specific Fertilization based on Multi-Source Data and Standardized Tools, grant # 2817ERA11H), https://www.ifaros-ictagri.com/.

References

Anselin, L. (1988). Spatial Econometrics: Methods and Model. Dordrecht: Kluwer.

Anselin, L. (1995, April). Local Indicators of Spatial Association. Geographical Analysis 27 (2), pp. 93-115.

Anselin, L. (2020, 10 10). geodacenter. Retrieved from Local Autocorrelation (2): https://geodacenter.github.io/workbook/6b_local_adv/lab6b.html#differential-local-moran

Anselin, L., Bera, A., Florax, R., & Yoon, M. (1996). Simple Diagnostic Tests for Spatial Dependence. Regional Science and Urban Economics, pp. 77-104.

Anselin, L., Syabri, I., & Smirnov, O. (2002). Visualizing Multivariate Spatial Correlation with Dynamically Linked Windows.

Ben-David, S., & Ackermann, M. (2008). Measures of Clustering Quality: A Working Set of Axioms for Clustering. In D. Koller, S. Dale, B. Yoshua, & B. Léon, Advances in Neural Information Processing System 21 (pp. 121-128). Vancouver, British Columbia, Canada: Curran Assicaites, Inc.

Bivand, R. (2012, February 16th). r-sig-geo. Retrieved from a question about gwr.morantest pvalue: http://r-sig-geo.2731867.n2.nabble.com/A-question-about-gwr-morantest-pvalue-td7292670.html

Bruns, J., & Simko, V. (2017). Stable Hotspot Analysis for Intra-Urban Heat Islands. GI_Forum Journal (pp. 79-92). Salzburg: Austrian Academy of Sciences Press.

Brunsdon, C., Fotheringham, A., & Charlton, M. (1996, October). Geographically Weighted Regression: A Method for exploring Spatial Nonstationarity. Geographical Analysis 28 (4), pp. 281-298.

Chen, F., Lu, C.-T., & Boedihardjo, A. (2010). GLS-SOD: A Generalized Local Statistical Approach for Spatial Outlier Detection. Proceedings of the 16th ACM SIGKDD Internationcal Conference on Knowledge Discovery and Data Mining (p. 1069). New York: ACM.

Cliff, A., & Ord, K. (1972). Testing for Spatial Autocorrelation among Regression Residuals. Geographical Analysis, pp. 267-284.

Darmofal, D. (2015). Analytical Methods for Spatial Research. New York: Cambridge University Press.

Elhorst, J. (2010). Applied Spatial Econometrics: Raising the Bar. Spatial Economic Analysis, pp. 9-28.

Floch, J.-M., & Le Saout, R. (2016). Econométrie Spatiale: Une Introdution Practique. Paris, France: Institut National de la Statistique et des Études Économiques.

Fotheringham, A., Yang, W., & Kang, W. (2017). Multiscale Geographically Weighted Regression. Annals of the American Association of Geographers 107 (6), pp. 1247-1265.

Getis, A., & Ord, J. (1992, July). The Analysis of Spatial Assocation by use of Distance Statistics. Geographical Analysis 24 (3), pp. 189-206.

Kitchen, N., Sudduth, K., & Drummond, S. (2003, 5). Soil Electrical Conductivity and Topography Related to Yield for Three. Agronomy Journal, pp. 483-494.

Kou, Y., Lu, C.-T., & Chen, D. (2006). Spatial Weighted Outlier Detection. Philadelphia: Society for Industrial and Applied Mathematics.

Lee, S.-I. (2001). Developing a Bivariate Spatial Association Measure: An Integration of Pearson's R and Moran's I. Journal of Geographical Systems, pp. 369-385.

LeSage, J., & Pace, R. (2009). Introduction of Spatial Econometrics. Statistics, Textbooks and Monographs.

Liu, F., Ting, K., & Zhou, Z.-H. (2012). Isolation-Based Anomaly Detection. ACM Transactions on Knowledge Discovery from Data, pp. 1-39.

Manski, C. (1993). Identification of Endogenous Social Effects: The Reflection Problem. The Review of Economic Studies, p. 531.

Moran, P. (1950). Notes on Continous Stochastic Phenomena. Biometrika.

Ord, J., & Getis, A. (1995). Local Spatial Autocorrelation Statistics: Distributional Issues and an Application. Geographical Analysis, 286-306.

Sun, P., & Chawla, S. (2004). On Local Spatial Outliers. Processdings / Foruth IEEE International Conference on Data Mining (pp. 209-216). Los Alamitos, California: IEEE Computer Society.

Takeuchi, J., & Yamanishi, K. (2006). A Unifying Framework for detecting Outliers and Change Points from Time Series. IEEE Transaction on Knowledge and Data Engineering (pp. 482-492). IEEE Computer Society.

Tukey, J. W. (1977). Exploratory data analysis. Pearson.

Wheeler, D., & Tiefelsdorf, M. (2005). Multicollinearity and Correlation among Local Regression Coefficients in Geographically Weighted Regression. Journal of Geographical Systems, 7(2), pp. 161-187.