Exploring and characterising irregular spatial clusters using eigenvector filtering

René Westerholt*[1]

[1]School of Spatial Planning, TU Dortmund University, Dortmund, Germany

January 15, 2021

Summary

The statistical study of spatial clusters is an important part of the exploratory data analysis toolbox. Spatial autocorrelation and hotspot statistics are now routinely used to better understand the arrangement of variables on maps. Spatial clusters, however, are often not internally homogeneous, but may exhibit interesting spatial heterogeneities. In this contribution, approaches to explore irregular spatial clusters are applied to a recent mapped index of food deserts. Both approaches are based on hotspot and heteroscedasticity measures, but one of the methods additionally uses eigenvector filtering. The results show that the latter contributes to the disclosure and understanding of spatial cluster irregularities.

KEYWORDS: Spatial filtering; spatial analysis; spatial autocorrelation; cluster analysis

1. Introduction

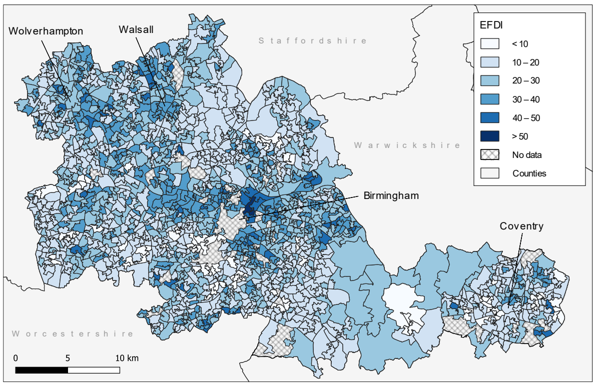
Spatial analysis has entered the mainstream of data analytics. Popular methods include measures of spatial autocorrelation (Anselin, 1995), hotspot statistics (Ord and Getis, 1995), spatial heteroscedasticity measures (Ord and Getis, 2012; Westerholt et al., 2018), analysing flows (Liu et al, 2015; Tao et al, 2019), and user-generated data (Westerholt et al, 2015; Westerholt et al, 2016). Such methods allow to detect spatial repulsion (negative) and clustering (positive spatial autocorrelation) in high (hotspots) or low attribute values (coldspots). Common diagnostics of disclosed spatial structures include the Moran scatterplot (Anselin, 1996), local spatial statistics (Anselin, 1995), and evaluations of the defensibility of spatial weights (Getis and Aldstadt, 2002). Possible spatial irregularities within clusters, however, are often not investigated in detail. This abstract proposes exploratory diagnostics for irregular clusters. The application of spatial amplifier filtering (Westerholt, 2021) is compared with the joint use of hotspot statistics and LOSH (Ord and Getis, 2012), as proposed by Aldstadt et al. (2012) for analysing internal spatial cluster heterogeneities.

2. Materials and Methods

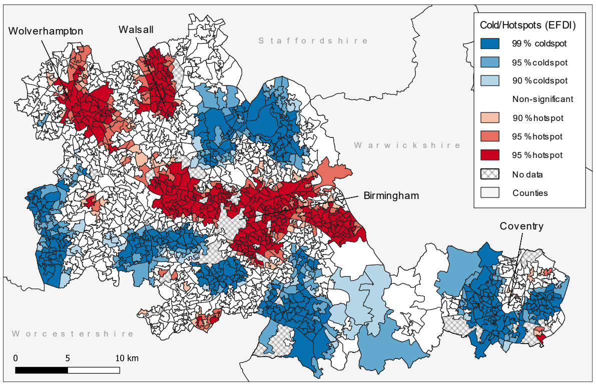
The dataset used is the e-food desert index (EFDI), which considers traditional dimensions of food deserts together with the provision of home delivery services (Newing and Videira, 2020; Videira et al., 2020). The EFDI is linked to Lower Super Output Areas (LSOAs) and is a numerical composite combining twelve indicators covering four domains. The region considered is the West Midlands Metropolitan County, from which a total of 1613 LSOAs are examined. One approach applied in the remainder uses the hotspot method G_i^*—a measure of local spatial attribute value concentration—and LOSH—a measure of spatial heteroskedasticity based on local spatially weighted variance estimations—both of which are applied to raw attribute data as proposed by Aldstadt et al. (2012). The alternative approach is also based on G_i^*, but applies LOSH to spatially pre-filtered attributes. Positive spatial autocorrelation and spatial randomness are removed from the data via an initial regression step that uses the eigenvalues of the spatial weights matrix associated with positively

* rene.westerholt@tu-dortmund.de

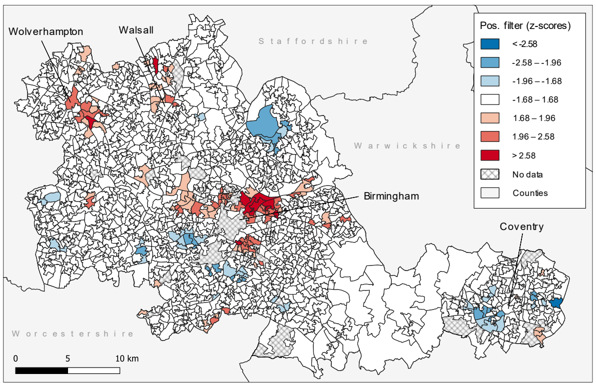
autocorrelated and random structures. What remains are residuals in which locally occurring negative spatial autocorrelation is strongly emphasised (Westerholt, 2021). This remaining variation associated with negative spatial autocorrelation is of a systematic, fluctuating nature revealing heterogeneities associated with measurement error, unfavourable spatial units, endogenous effects like competition, or exogenous effects such as differently structured covariates (Griffith, 2006). The latter effects are interesting as they may reveal additional insights into internal cluster heterogeneities beyond mere random fluctuations. In addition, the filtering process is reversed to also map positive spatial autocorrelation and see where the strongest clusters are located. Second-order, binary spatial weights based on queen's contiguity are used in all spatial analyses outlined, which are implemented using the R-based *spdep* package (Bivand et al., 2019).



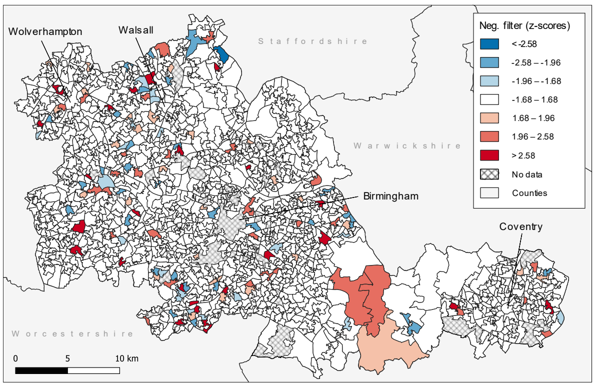
(a) The raw EFDI indicators.



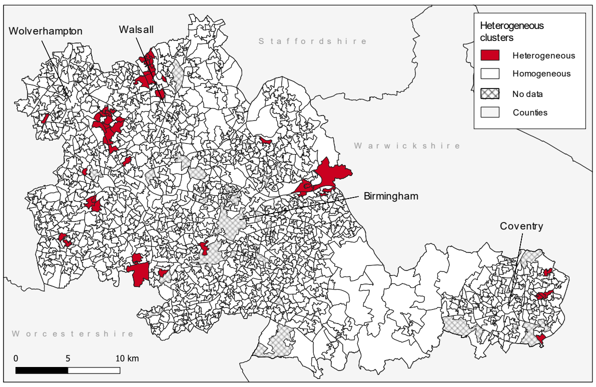
(b) G_i^* cold and hotspots of the EFDI indicators.



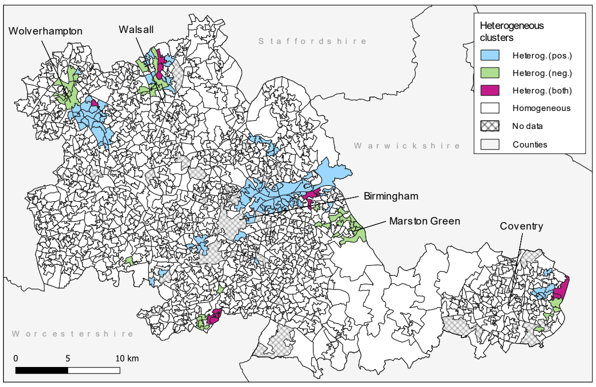
(c) Filtered positive spatial autocorrelation.



(d) Filtered negative spatial autocorrelation.



(e) Heterogeneous clusters (G_i^* and LOSH).



(f) Heterogeneous clusters (pre-filtered).

Figure 1 Maps of the raw, filtered, and analysed EFDI indicators.

3. Results and Discussion

Figures 1a and 1b visualise the mapped EFDI index and the respective cold and hotspots. The most prominent hotspots are located around Birmingham and along an axis connecting West Bromwich with Wolverhampton and Walsall. Coldspots are located around Coventry and in the more rural areas, which are generally less socio-economically deprived than the city cores hence showing better healthy food accessibility. We shall take a closer look at the results of both the positive and the negative filter in conjunction with LOSH and G_i^* and at the outcomes of the unfiltered approach.

The characterisation of irregular clusters with LOSH and G_i^* applied to raw data (Aldstadt et al., 2012) often identifies boundaries instead of internal cluster fluctuations (Figure 1e). Most of the areas identified in this way are located either between stronger and less pronounced parts of the elongated central axis, or at the edges of cold/hotspots. These results are not entirely unexpected given the nature of the methodology used. LOSH identifies areas where the strongest residuals above locally estimated means accumulate, which, in an urban context with spatial regimes, naturally occurs at cluster boundaries. However, this means that this approach is prone to overlooking intra-cluster irregularities, unless the latter stand out strongly. Moreover, it is not possible to distinguish between intrinsic variations in the intensity of the phenomenon under study and possible interesting anomalies such as local outliers or extrinsic confounders. Therefore, this approach should be used to identify interesting cluster boundaries rather than to characterise internal cluster irregularities.

Using the two types of filters (Westerholt, 2021) in combination with LOSH and G_i^* reveals additional structures beyond those described in the previous paragraph. Controlling for negative autocorrelation and randomness through filtering, we see that areas in the north-western parts of Birmingham and in Wolverhampton appear particularly spatially clustered when outliers and local anomalies are removed (Figure 1c). In contrast, using the filter that enhances negative spatial autocorrelation identifies areas where local irregularities exist beyond smooth neighbourhood effects (Figure 1d). To better understand what this means, consider the area of Marston Green north of Birmingham International Airport, which is highlighted in the combined map in Figure 1f. Closer inspection reveals that this area shows small-scale diverse characteristics regarding the Index of Multiple Deprivation and also with respect to the Internet User Classification in terms of propensity to order food online (both not shown in the map). The approach presented may thus allow to disclose spatially varying interactions of constitutive subcomponents of the EFDI. The negative filtering approach in combination with G_i^* and LOSH therefore seems promising to better understand the spatial details of (possibly unknown) subcomponents and how these play into observed data and indicators. It may thus contribute to a better understanding of how and why different revealed clusters differentiate.

4. Conclusions

This abstract briefly and exploratorily compares two approaches to characterise spatial clusters. One approach is based on a hotspot statistic in combination with the LOSH method to quantify spatial heteroscedasticity. The other approach additionally uses two variants of Moran eigenvector filtering to distinguish different types of spatial variation in cluster characterisation. The results show that the first of the two approaches is better suited to identify potentially interesting cluster boundaries where contrasts feature particularly strong. The second approach is of special interest in terms of the negative spatial autocorrelation amplifier, which allows to reveal clusters that may be affected by local anomalies or, like in the present case of the EFDI index, by possible subcomponents with a spatial structuring different from the finally observed data. Future research should explore better explanations of the negative amplifier results, which are more difficult to read and interpret than positively autocorrelated (i.e. clustered) results due to their negatively spatially autocorrelated nature.

References

Aldstadt J, Widener M and Crago N (2012) Detecting irregular clusters in big spatial data. In: Xiao N, Kwan MP, Goodchild MF and Shekhar S (eds) *Proceedings of the 7th International Conference on Geographic Information Science* (GIScience 2012).

Anselin, L (1995). Local indicators of spatial association—LISA. *Geographical Analysis*, 27(2), 93–115.

Anselin, L (1996). The Moran scatterplot as an ESDA tool to assess local instability in spatial association. In: Fischer, M. (ed.) *Spatial Analytical*. London: Routledge, 111–125.

Bivand R, Altman M, Anselin L, Assuncao R, Berke O et al. (2019). Package "spdep". https://cran.r-project.org/web/packages/spdep/spdep.pdf. Last accessed 23 December 2020.

Getis, A, Aldstadt, J (2002). Constructing the Spatial Weights Matrix Using a Local Statistic. *Geographical Analysis* 34 (2), 130–140.

Griffith, D A (2006). Hidden negative spatial autocorrelation. *Journal of Geographical Systems, 8*(4), 335–355.

Liu, Y, Tong, D, and Liu, X (2015). Measuring spatial autocorrelation of vectors. *Geographical Analysis*, 47(3), 300–319.

Newing, A and Videira, F (2020). EFDI User Guide. https://data.cdrc.ac.uk/dataset/e-food-desert-index/resource/efdi-user-guide. Last accessed 21 December 2020.

Ord, J K, and Getis, A (1995). Local spatial autocorrelation statistics: distributional issues and an application. *Geographical Analysis*, 27(4), 286–306.

Ord, J K, and Getis, A (2012). Local spatial heteroscedasticity (LOSH). *The Annals of Regional Science*, 48(2), 529–539.

Tao, R, and Thill, J C (2019). FlowAMOEBA: Identifying regions of anomalous spatial interactions. *Geographical Analysis*, 51(1), 111–130.

Videira, F, Newing, A, Hood, N and Lewis, J (2020). 'Sorry we do not deliver to your area': Geographical inequalities in online groceries provision. Manuscript submitted for publication.

Westerholt, R (2021). Emphasising spatial structure in geosocial media data using spatial amplifier filtering. *Environment and Planning B: Urban Analytics and City Science*, vol. and issue pending, online first.

Westerholt, R, Resch, B, and Zipf, A (2015). A local scale-sensitive indicator of spatial autocorrelation for assessing high-and low-value clusters in multiscale datasets. *International Journal of Geographical Information Science*, 29(5), 868–887.

Westerholt, R, Steiger, E, Resch, B, and Zipf, A (2016). Abundant topological outliers in social media data and their effect on spatial analysis. *PLOS ONE*, 11(9), e0162360.

Westerholt, R, Resch, B, Mocnik, F-B, and Hoffmeister, D (2018). A statistical test on the local effects of spatially structured variance. *International Journal of Geographical Information Science*, 32(3), 571–600.

Biographies

René Westerholt is Juniorprofessor at TU Dortmund University, Germany, and his research interests include geographic data science, spatial analysis, and representations of space and place in GIScience.