



Studying critical values for global Moran's I under inhomogeneous Poisson point processes

René Westerholt ¹

¹Department of Spatial Planning, TU Dortmund University, Dortmund, Germany

Correspondence: René Westerholt (rene.westerholt@tu-dortmund.de)

Abstract. Spatial autocorrelation is a fundamental statistical property of geographical data. A number of estimators have been introduced, with Moran's I being one of the most commonly used methods. The characterisation of spatial autocorrelation is useful for a number of applications, including finding clusters, testing model assumptions, investigating spatial outliers, and others. Most estimators of spatial autocorrelation are based on assessing the degree of correspondence between structures in an attribute and structures among spatial units, both of which are operationalised in matrix form. Associated inference procedures then rely on holding the spatial configuration fixed, but varying the attribute values over the geometries. Although fixing the geometries is useful in many scenarios, there are cases where it would be more appropriate to allow the geometries to vary as well, such as in the analysis of social media feeds or mobile sensor observations. In this short paper, the case is considered where geometries are the result of inhomogeneous spatial Poisson processes. Using diagonal and circular types of spatial structuring, it is investigated how random geometries affect critical values used to assess the significance of global Moran's I scores. It is shown that the critical values resulting from an established inference framework often underestimate the bounds that would result if geometric randomness were taken into account. This leads to type-I errors and thus potential false positive patterns.

Keywords. spatial autocorrelation, Moran's I , Poisson point process, spatial statistics

1 Introduction

Global spatial autocorrelation measures are summary statistics used to quantify spatial structure in attribute values. The explicit incorporation of spatial relations is achieved by using spatial weights that formalise pairwise associations between discrete units such as points, polygons, and lines. Moran's I and Geary's c (Cliff and Ord,

1981) are two very commonly used methods (with implementations available in mainstream software packages such as ESRI ArcMap, R-based spatial statistics libraries, and others), but there are a number of other statistics, including Γ (Hubert and Golledge, 1981), Ripley's K (Ripley, 1976), and the correlogram associated with a significance test Q (Oden, 1984). Typical use cases for spatial autocorrelation statistics include testing the geographical nature of data, testing for model misspecification in regression contexts, and investigating spatial outliers (Getis, 2010, 2008).

The present contribution is concerned with inferences about global Moran's I in the context of random geometric configurations. Drawing inference about Moran's I assumes that the underlying geometric units are fixed (Cressie, 1993). This is reflected in associated inference procedures based either on the assumption of normality (repeated sampling from a stationary normal distribution) or on randomisation of the observed attribute values over the geometric units present (see Westerholt, 2022b, for an overview of inference procedures including more specialised approaches). The assumption of a fixed spatial configuration simplifies inferences and often also makes sense from an empirical point of view. For example, administrative or census units are not the results of stochastic processes but predefined. However, there are application scenarios in which this assumption is not realistic. Examples include analyses of social media feeds (Steiger et al., 2016), mobile sensor data (Bucher et al., 2020), wildlife tracking observations (Demšar et al., 2015), and in scenarios of competition for resources or land (Griffith and Arbia, 2010; Griffith, 2006). Resorting to methods of point pattern analysis (which is concerned with random geometries) is not a way out since the covariance-based version of the mark correlation function is technically equivalent to Moran's I , with respective inference techniques also being based on holding geometries fixed (Illian et al., 2008; Shimatani, 2002).

Analytical solutions for inference about global Moran's I applied to random geometries would be complex due to the

need to include distributions of spatial weights. The latter depend on the underlying stochastic process generating the geometries, the types of spatial weights, and possible normalisations. Applying the Pitman-Koopman theorem – the relevant theorem for deriving the moments of the null distribution of Moran’s I (Griffith, 2010; Cliff and Ord, 1981) – is therefore not straightforward and can lead to different results depending on the nature of the weights.

A recent paper by Westerholt (2022a) has investigated empirically the impact of random geometries on inferences about global Moran’s I . The focus of that paper is on homogeneous and inhomogeneous Matérn and Thomas cluster processes, both special cases of the Neyman-Scott process modelling offspring points around unobserved Poisson parent points (Yau and Loh, 2012). In that paper, more emphasis is on the homogeneous case as inhomogeneity is modelled after tweets and thus based on a rather specific intensity surface. The results show that neglecting the contribution of the geometries to the variability of Moran’s I has implications for the application of critical values used to decide about statistical significance and for the statistical power of the estimator. The current paper complements these results by shedding light on the inhomogeneous case using Poisson processes and thus no offspring points. This type of modelling can be useful for investigating spatial urban structures without assuming specific punctiform events as causes for clustering. Using 20,000 simulated inhomogeneous point patterns with different types of trends in the point intensity functions, it is shown in what way the inhomogeneous geometric randomness has an impact on critical values. Simulated point patterns are used to eliminate potential confounding factors that might otherwise be caused by specific geographical contexts. The latter would make it difficult to interpret the results obtained, to attribute them to an underlying mechanism, and would limit the generality of the results. The results obtained are of both practical and theoretical value. They are useful in empirical contexts informing how to interpret Moran’s I in respective application scenarios and contribute to a better understanding of the interplay between methodology and random geometry.

2 Methods

The following subsection introduces the methodology as well as the availability of the data and the software used.

2.1 Moran’s I

The Moran’s I statistic estimates spatial autocorrelation and is given as (Getis, 2010, p. 264)

$$I = \frac{n}{\sum_{i=1}^n \sum_{j=1}^n w_{ij}} \frac{\sum_{i=1}^n \sum_{j=1}^n w_{ij} (x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad (1)$$

with $x_i, x_j \in X \subseteq \mathbb{R}$ as n attribute values with mean \bar{x} and $w_{ij} \in \mathbb{R}_{\geq 0}$ denoting spatial weights. The spatial weights are determined based on 10 nearest neighbours and using inverse distance weighting in the form d_{ij}^{-2} , where d_{ij} denotes the Euclidean distance between observations i and j . Positive and negative I values indicate positive (clustering of similar values) and negative spatial autocorrelation (adjacent contrasting values), respectively. There are several inference mechanisms associated with I , including special procedures for small sample situations (Tiefelsdorf and Boots, 1995; Cliff and Ord, 1972) and to account for skewness (Tiefelsdorf, 2002). However, the two most important inferential frameworks (based on large samples) are those built on either the assumption of normality or randomisation of observed values, as put forward by Cliff and Ord (1981, p. 46 ff.). The results presented in Section 3 are based on the normalisation assumption. It was shown in the article by Westerholt (2022a) that the differences between the two types of assumptions are negligible with respect to the objectives of this work, so the second variant is omitted in the following due to space constraints.

2.2 Sampling from Inhomogeneous Spatial Poisson Processes

The *homogeneous* Poisson process is the reference process that reflects complete spatial randomness and to which geometric patterns are usually compared in point pattern analysis. This type of process is determined by an intensity parameter λ that reflects the average number of points per unit area (Illian et al., 2008, p. 66). In the present work, two spatially varying intensity parameters are used to simulate *inhomogeneous* spatial Poisson processes. These parameters are based on the geometric coordinates u and v and are given as

$$\lambda_1 = 175 \cdot |u + v|, \quad (2a)$$

$$\lambda_2 = 50 / \sqrt{(u - 0.5)^2 + (v - 0.5)^2}. \quad (2b)$$

Equation 2a generates point patterns with an increasing southwest-to-northeast trend. Equation 2b, on the other hand, yields point patterns with a radial trend starting from the midpoint (0.5, 0.5) of 1×1 unit windows in descending form. The multiplicative factors 175 and 50 control the average numbers of points and keep them close to 175 to ensure comparability. For both types of point processes, illustrated in Figure 1, 10,000 random samples were drawn.

2.3 Analysis of critical values

The present study compares critical values for determining the statistical significance of Moran’s I . These critical values indicate certain percentiles of the null distribution of Moran’s I under spatial randomness. Values above the critical value (or below, in the case of a two-sided test) are considered significant and denote spatial configurations

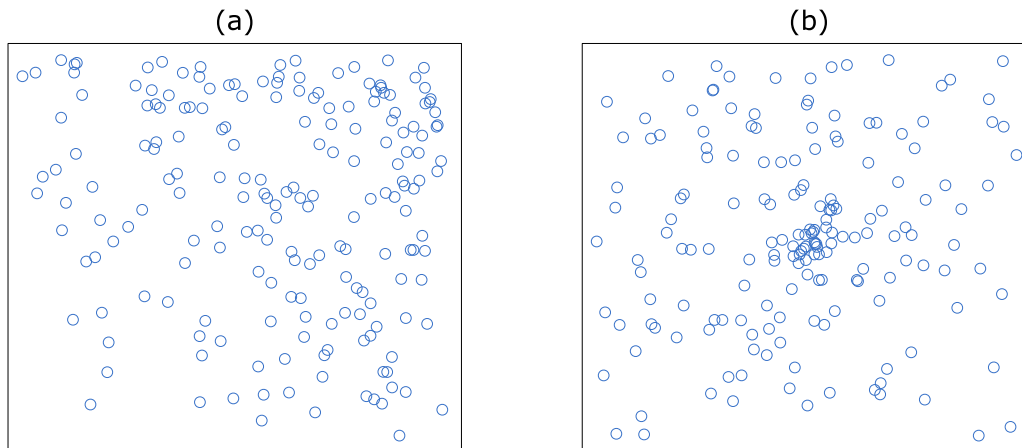


Figure 1. Illustration of realisations of inhomogeneous spatial Poisson point patterns, where the intensity λ leads to either (a) a diagonal trend or (b) a circular pattern.

that are unlikely to occur under random conditions. Critical values are therefore of high practical relevance as they form the basis for judgements about potentially interesting or uninteresting spatial structures in data. Attribute vectors X_k assigned to each of the $1 \leq k \leq 20,000$ simulated point patterns are drawn from a standard normal distribution, whereby each sample is drawn with the same seed in order to eliminate distributional fluctuations and to ensure that the observed differences are due to geometric randomness. In this way, each simulated sample is assigned a vector of standard normal variates from the same distribution and the same pseudo-random number generation process, thereby controlling potential mere technical confounders.

The critical values that would be used in the conventional way (without taking geometric randomness into account) are determined for each individual simulated pattern and assuming asymptotic normality of Moran's I in the null hypothesis (see Section 2.1). The 90th, 95th, and 99th percentiles of respective normal distributions fitted with the means of I (which is $\mu = -1/(n-1)$) and the respective variances of I for each individual pattern (the equation of the variance would be too bulky to be reproduced here, see Griffith (2010)) are calculated. The counterparts for the cases of geometric randomness are obtained empirically from vectors of Moran's I estimates for the respective 10,000 patterns of both point processes. Due to space constraints, only positive spatial autocorrelation is considered, as this is of greater practical importance (since it represents accumulation in space and thus clustering of similar values). The critical values are then analysed using boxplots that allow a visual comparison of the distributional characteristics assuming common significance values $\alpha \in \{0.10, 0.05, 0.01\}$. Respective lines in the boxplots allow to look into the possible deviations between the two types of inference. This visual analysis is supported numerically by calculating skewness and kurtosis.

2.4 Data and Software Availability

The data used in this work consists of 20,000 simulated point patterns that are generated using the `spatstat` R package. The code to perform the analysis and to create the raw versions of the figures (which were manually post-processed using professional vector graphics editing software) is based on the R libraries `sf`, `spdep`, `spatialreg`, `foreach`, and `doParallel`. All code is provided through Zenodo: <https://doi.org/10.5281/zenodo.7824967>.

3 Results

The critical values as derived from the established inference framework often underestimate the critical values that do account for geometric randomness. Figure 2 shows boxplots summarising the conventional critical values for all 20,000 estimated patterns and for the three significance levels $\alpha \in \{0.10, 0.05, 0.01\}$. Sub-figure 2a shows the results for the patterns with diagonal intensity trend, while sub-figure 2b presents boxplots for the circular pattern. Both types of point patterns show similar results but differ in some details. A general observation is that a large number of the conventionally determined values (the boxplots) underestimate the critical values obtained under geometric randomness (dotted lines in the figure). This is most pronounced for $\alpha = 0.01$ and thus at a strict significance level. For both types of point patterns, the boxes containing the mean 50% of estimated conventional critical values of all simulated patterns are completely below the empirically estimated critical value that takes into account geometric randomness. Underestimation here means that in many situations we would be confronted with statistical type-I error inflation (i.e. the null hypothesis of spatial randomness is falsely rejected too often) because conventional values below the one accounting for the additional randomness indicate a too permissive decision criterion. The presence

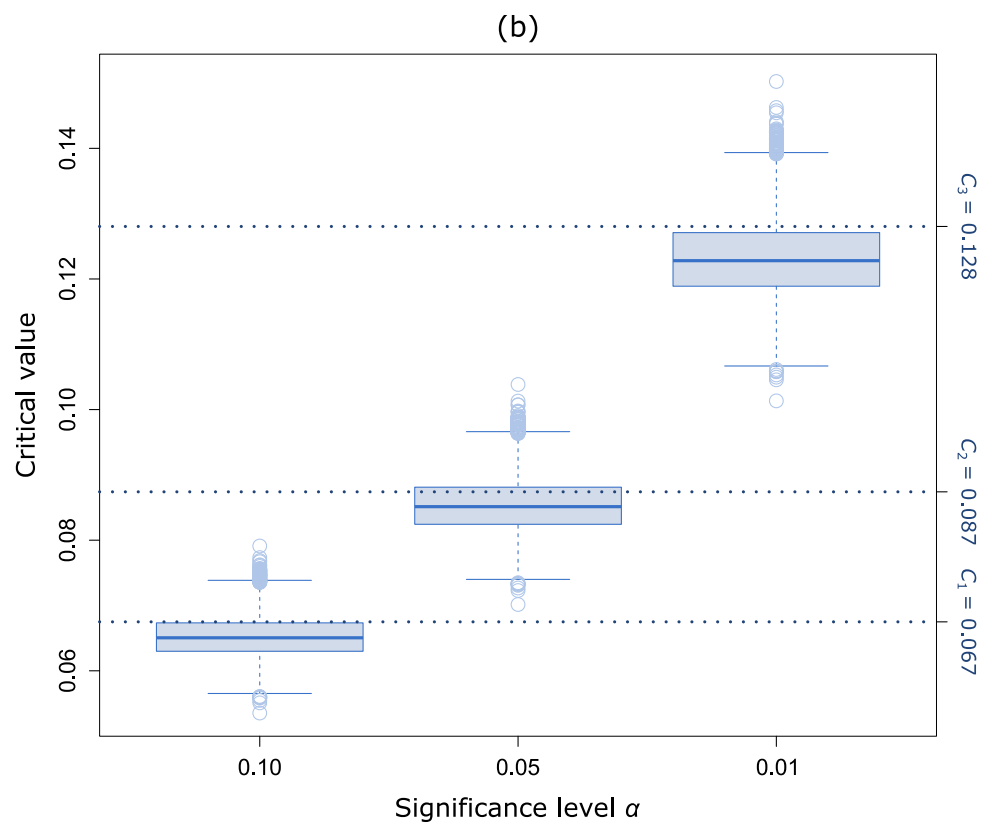
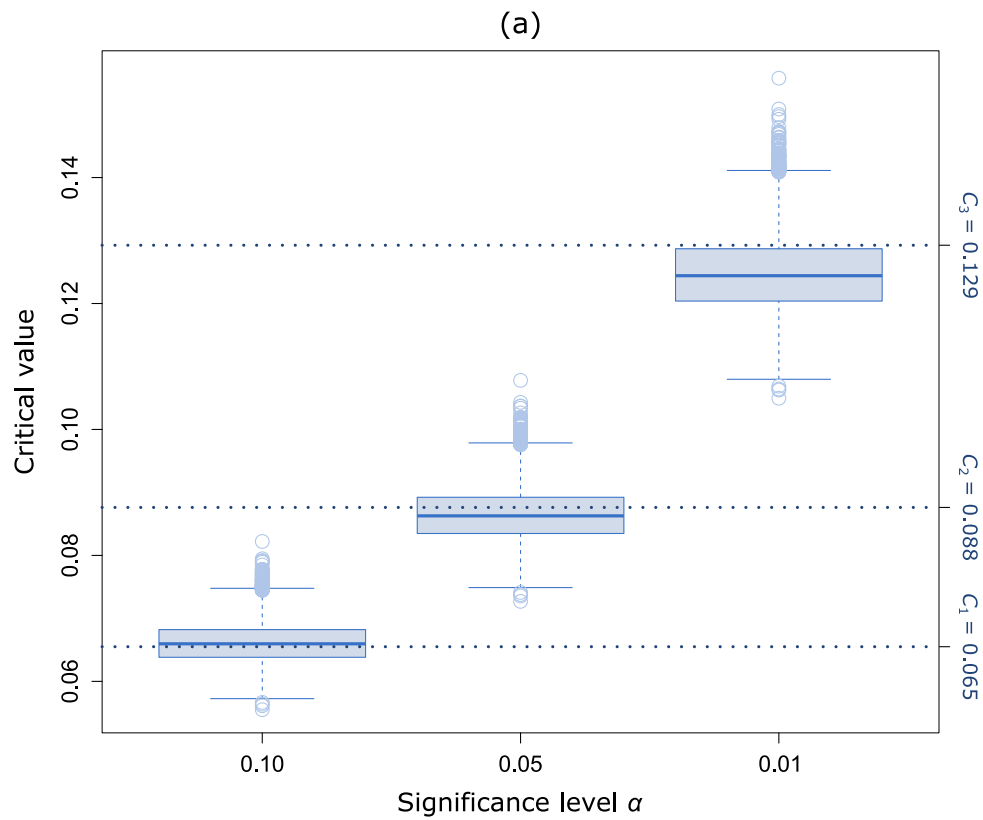


Figure 2. Summaries of the distributions of conventional critical values for Moran's I compared against critical values obtained from simulated point patterns. Critical values are given for point patterns with (a) a diagonal and (b) a circular pattern in the intensity function.

of inhomogeneous Poisson point processes thus seems to lead to a reduction in the decision quality of Moran's I as a test statistic.

The underestimation of the critical values is more consistent for point patterns with circular point intensities. For the diagonal case, it is noticeable that with decreasing α the boxes of the boxplots move increasingly below the critical value estimates that account for point geometry variation. While the dotted line for the latter threshold cuts through the middle of the box for $\alpha = 0.10$, it gradually moves upwards to eventually no longer hit the box. This behaviour is not observed for the point patterns with circular intensities, for which the conventional critical values already cause type-I errors even at weak significance levels. However, this behaviour seems to be stable and does not depend on the rigour of the test. Recall that for both types of samples the attribute values were drawn from the same normal distributions with the same seeds. The only difference between them is the spatial structuring of the intensity function, which affects the spatial weights. One possible interpretation of the two observations described above is therefore that the diagonal patterns presumably exert a greater influence on the tails of the distribution of Moran's I than the circular patterns. The more sensitive the statistical test is in terms of α , the more emphasis is placed on the tails of the distribution, and hence we see the trend in Figure 2a. The fact that both series of patterns examined come from the same point process, just with differently structured intensities, shows how sensitive spatial statistics are to geometric configurations, especially when these are attached with randomness.

The somewhat more extreme behaviour at the critical values caused by the random diagonal patterns is confirmed by additional distributional characteristics. Looking at Figure 2, it is noticeable in both cases that there are more outliers in the right tail than in the left. So in all cases there is a stronger propensity to underestimate the empirical critical values, but with a tendency towards more outliers at the other end of the spectrum. This observation suggests a right skewness in the distributions of critical values. The estimation of skewness and kurtosis for all six distributions summarised in Figure 2 shows that the conventional critical values obtained for the diagonally structured patterns are more right-skewed (the skewness lies in the interval $(0.3409, 0.3436)$ as opposed to $(0.2728, 0.2765)$) and are more leptokurtic (the kurtosis lies in the interval $(3.2405, 3.2531)$ as opposed to $(3.0687, 3.0751)$) than their counterparts for the radial patterns. On average, therefore, conventional critical values are more reliable for random Poisson processes with a diagonal intensity trend, at least if one accepts less stringent testing. However, this comes at the expense of a slightly higher chance of observing extreme patterns that would effectively lead to an overly stringent assessment. The latter leads to type-II errors (i.e. the null hypothesis of spatial randomness is falsely accepted too often) and thus to potentially overlooking interesting spatial structures in the

data. Thus, while the mean behaviours described in the previous paragraphs (i.e. the boxes) lead to unwanted patterns being detected, there is also a certain probability that a very rigid outcome of a spatial Poisson process may have been observed, leading to spatial effects in attributes being missed. Randomness in the points thus causes both type-I and type-II errors, the former being more common but the latter not unlikely.

4 Conclusions

This short paper investigated critical values associated with using global Moran's I as a test statistic in conjunction with geometries that are the outcome of inhomogeneous spatial Poisson processes. A total of 20,000 patterns were simulated, half with a diagonal trend in the intensity and the other half with a circular point intensity. All patterns were simulated in 1×1 unit windows, and each of them includes an average of 175 points, with attributes from normal distributions. The analysis consisted of two parts: a characterisation of the distributions of the critical values for I resulting from the established inference framework, and a comparison of these values with the critical values obtained empirically from the simulated patterns, thereby accounting for their inherent geometric randomness.

The results obtained in the present study indicate a reverse picture of the results presented by Westerholt (2022a). Considering only similarly sized point patterns, the earlier results suggest that type-II errors are of more concern and some type-I errors occur in outlier cases. In the present work, the analysis yields the opposite interpretation with more type-I errors and fewer type-II errors. The main difference between the two papers is the nature of the point processes studied. Westerholt (2022a) has studied two types of cluster processes based on additional mechanisms for point dispersal around parent points (Matérn and Thomas cluster processes). These are useful, for example, for modelling punctiform events that serve as spatial attractors. In contrast, this paper deals with spatially structured Poisson processes, but without additional dispersal mechanisms and hence only featuring spatial structure in what would be called the (unobserved) parent point process in Westerholt (2022a). The type of modelling considered here is suitable when there is (often externally conditioned) spatial structuring, for example, in the case of urban data structured by the general urban fabric but without additional locatable sources of clustering on top. The additional clustering process in Westerholt (2022a) seems to result for smaller point patterns in the clustering being overly pronounced and missing much of the surrounding point matrix that would increasingly occur for larger patterns. Consequently, relevant patterns are missed. In the present case, the opposite appears to be the case, as the data presumably often resembles unstructured (and thus more or less homogeneous) Poisson processes, which

means that structures are more quickly flagged significant. This paper is only a short contribution and further work is needed to gain a more comprehensive picture of the intersection of spatial autocorrelation statistics and random geometry.

References

- Bucher, D., Martin, H., Jonietz, D., Raubal, M., and Westerholt, R.: Estimation of Moran's I in the context of uncertain mobile sensor measurements, in: Proceedings of the 11th International Conference on Geographic Information Science (GIScience 2021) – Part I, edited by Janowicz, K. and Verstegen, J. A., Poznań, Poland, 2020.
- Cliff, A. and Ord, K.: Testing for spatial autocorrelation among regression residuals, *Geographical Analysis*, 4, 267–284, 1972.
- Cliff, A. D. and Ord, J. K.: *Spatial Processes: Models & Applications*, Pion, London, UK, 1981.
- Cressie, N.: *Statistics for Spatial Data*, Wiley, New York, NY, USA, 1993.
- Demšar, U., Buchin, K., Cagnacci, F., Safi, K., Speckmann, B., Van de Weghe, N., Weiskopf, D., and Weibel, R.: Analysis and visualisation of movement: an interdisciplinary review, *Movement Ecology*, 3, 5, 2015.
- Getis, A.: A history of the concept of spatial autocorrelation: a geographer's perspective, *Geographical Analysis*, 40, 297–309, 2008.
- Getis, A.: Spatial autocorrelation, in: *Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications*, edited by Fischer, M. M. and Getis, A., pp. 255–278, Springer, Berlin/Heidelberg, Germany, 2010.
- Griffith, D. A.: Hidden negative spatial autocorrelation, *Journal of Geographical Systems*, 8, 335–355, 2006.
- Griffith, D. A.: The Moran coefficient for non-normal data, *Journal of Statistical Planning and Inference*, 140, 2980–2990, 2010.
- Griffith, D. A. and Arbia, G.: Detecting negative spatial autocorrelation in georeferenced random variables, *International Journal of Geographical Information Science*, 24, 417–437, 2010.
- Hubert, L. J. and Golledge, R. G.: A heuristic method for the comparison of related structures, *Journal of Mathematical Psychology*, 23, 214–226, 1981.
- Illian, J., Penttinen, A., Stoyan, H., and Stoyan, D.: *Statistical Analysis and Modelling of Spatial Point Patterns*, Wiley, Chichester, UK, 2008.
- Oden, N. L.: Assessing the significance of a spatial correlogram, *Geographical Analysis*, 16, 1–16, 1984.
- Ripley, B. D.: The second-order analysis of stationary point processes, *Journal of Applied Probability*, 13, 255–266, 1976.
- Shimatani, K.: Point processes for fine-scale spatial genetics and molecular ecology, *Biometrical Journal*, 44, 325–352, 2002.
- Steiger, E., Westerholt, R., and Zipf, A.: Research on social media feeds – a GIScience perspective, in: *European Handbook of Crowdsourced Geographic Information*, edited by Capineri, C., Haklay, M., Huang, H., Antoniou, V., Kettunen, J., Ostermann, F., and Purves, R., pp. 237–254, Ubiquity Press, London, UK, 2016.
- Tiefelsdorf, M.: The saddlepoint approximation of Moran's I 's and local Moran's I_i 's reference distributions and their numerical evaluation, *Geographical Analysis*, 34, 187–206, 2002.
- Tiefelsdorf, M. and Boots, B.: The exact distribution of Moran's I , *Environment and Planning A*, 27, 985–999, 1995.
- Westerholt, R.: A simulation study to explore inference about global Moran's I with random spatial indexes, *Geographical Analysis*, 2022a.
- Westerholt, R.: Exploratory statistical analysis of spatial structures in urban datasets, in: *Metropolitan Research: Methods and Approaches*, edited by Gurr, J. and Parr, R. and Hardt, D., pp. 37–61, transcript, Bielefeld, Germany, 2022b.
- Yau, C. Y. and Loh, J. M.: A generalization of the Neyman-Scott process, *Statistica Sinica*, 22, 1717–1736, 2012.