geographical analysis

Check for updates

*Geographical Analysis* (2023) **55**, 621–650

# A Simulation Study to Explore Inference about Global Moran's *I* with Random Spatial Indexes

René Westerholt

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

*Inference procedures for spatial autocorrelation statistics assume that the underlying configurations of spatial units are fixed. However, sometimes this assumption can be disadvantageous, for example, when analyzing social media posts or moving objects. This article examines for the case of point geometries how a change from fixed to random spatial indexes affects inferences about global Moran's I, a popular spatial autocorrelation measure. Homogeneous and inhomogeneous Matérn and Thomas cluster processes are studied and for each of these processes, 10,000 random point patterns are simulated for investigating three aspects that are key in an inferential context: the null distributions of I when the underlying geometries are varied; the effect of the latter on critical values used to reject null hypotheses; and how the presence of point processes affects the statistical power of Moran's I. The results show that point processes affect all three characteristics. Inferences about spatial structure in relevant application contexts may therefore be different from conventional inferences when this additional source of randomness is taken into account.*

## Introduction

Spatial autocorrelation is a statistical property that operationalizes Tobler's First Law of Geography. The latter describes the idea that near things are more similar than distant things (Tobler 1970), and the concept of spatial autocorrelation, rooted in the quantitative revolution (Haining 2009), puts this commonly observed property of geographical data[1] on a formal, statistical footing: it refers to the property of spatially referenced random variables to be correlated with each other when they are spatially close. Besides the possibility of quantifying spatial dependencies, there are a variety of practical and scientific applications of spatial autocorrelation statistics. These include identifying spatial outliers, testing assumptions of spatial heterogeneity and stationarity, and assessing the role of space in stochastic processes (Getis 2007). The concept of spatial autocorrelation has also influenced and stimulated spatial empirical research outside geography. Spatial epidemiology (see Auchincloss et al. 2012; Kirby, Delmelle, and Eberth 2017; Eberth et al. 2021), ecology (see Legendre 1993; Diniz-Filho,

Correspondence: René Westerholt, Department of Spatial Planning, August-Schmidt-Straße 10, TU Dortmund University, 44227 Dortmund, Germany
e-mail: rene.westerholt@tu-dortmund.de

Submitted: December 8, 2021. Revised version accepted: September 14, 2022.

doi: 10.1111/gean.12349

Bini, and Hawkins 2003), the various quantitative social sciences (see Páez and Scott 2004; Townsley 2009), and other cognate fields make extensive use of the concept in a number of explicitly spatial research approaches. Yet, as with other statistical estimators, the mere calculation of spatial autocorrelation measures is not sufficient to assess the significance of observed spatial dependencies. Drawing inference about respective statistics is needed, but corresponding procedures are subject to assumptions and constraints.

One constraint in testing for spatial autocorrelation is that the spatial index set $S \subseteq \mathbb{R}^n$, on which spatially referenced random variables, $Y_i$, are mapped, is held fixed geometrically in inference procedures. An index set is a set that labels members of another set (see Munkres 2014, p. 36 f.) and borrows them structure in the sense of a relative arrangement. Common examples of index sets are the set of natural numbers for countable sets, or the timeline in the case of temporal considerations. In the spatial case, random variables are usually assigned to geometric entities such as points, lines, and polygons. In contrast to the conceptually simpler set of natural numbers, such geometric collections represent real-world geographical features and therefore offer qualities that should be taken into account in spatial analyses. Following the notation of Cressie (1993), the set $\left\{\left(Y_i, S_i\right) : Y_i \in \mathbb{R}, S_i \in S, i \in \mathbb{N}\right\}$ defines a general notion of spatial process from which different types of more specialized spatial processes can be derived. The latter depends on the nature of the spatial index set $S$. If the index set is considered nonrandom (i.e., fixed) and discrete, the resulting process is of the type of a lattice and concerns spatial structures such as census areas or administrative units. These types of units are usually not tied to the stochastic processes responsible for the attribute values $Y_i$. Census areas, for example, are designed to reflect demographic features, while geographical characteristics may be ignored. It is therefore reasonable to fix these types of spatial units in inference procedures for spatial estimators (e.g., in spatially randomizing attribute values in Monte Carlo procedures), as they provide an extrinsic spatial context but not an intrinsic property of the process under study. Similar arguments apply to the geostatistical school of thought, which deals with continuous spatial variation. Here, too, the set of coordinates possibly delimited by an observation area constitutes an a priori given structure. In contrast, the spatial point process school of thought is by definition concerned with random spatial index sets (see Cressie 1993). Analyzing attributes (called marks) in this context is usually conducted with specializations of the so-called mark correlation function (see Illian et al. 2008, p. 341 ff.). For example, Shimatani (2002) presents a point-pattern version of Moran's $I$ and Shimatani and Takahashi (2003) discuss some properties, respectively. Their main focus is on extending Moran's $I$ to continuous distance domains. However, there is no detailed methodological investigation of how specific point processes affect the properties of Moran's $I$. In fact, even in the domain of point processes, "[m]arked point patterns with quantitative marks are resampled by *random reallocation*, [ … ], [that is], the points are fixed but are allocated new marks" (Illian et al. 2008, p. 467, italic font in original).

The restriction to fixed spatial indexes is usually imposed for two reasons. First, fixing the spatial index simplifies inference procedures and allows the construction of approximate (e.g., Cliff and Ord 1972; Cliff and Ord 1981; Tiefelsdorf 2002) or exact null distributions (Tiefelsdorf and Boots 1995). Second, fewer assumptions need to be made regarding possibly complex geographical concepts. For example, using the classic inference framework for Moran's $I$ introduced by Cliff and Ord (1981)[2], one only has to decide whether the mapped random variables all come from the same normal distribution or whether the observed vector of variates should instead be fixed and only randomized across all sites without resampling. The spatial weights connecting the sites still need to be carefully chosen to avoid possible misspecification

issues, but no additional assumptions need to be made about how to vary the underlying sites in the inference procedure and in accordance with some presumed geographically meaningful mechanism. That is, the spatial weights are also considered fixed (Ripley 1981, p. 98). If instead the underlying spatial index would be considered random, this would also affect the spatial weights that formalize the potential for spatial interaction (active) or connectedness (passive) (Dray 2011) and represent the formal constructs through which properties of the spatial index enter the equations of spatial statistical methods. Various types of weights exist, including distance-based, hierarchical, and nearest-neighbor notions (Getis and Aldstadt 2004; Getis 2009), but those are typically derived from the underlying index set formed by the spatial units, and are thus functions of the latter. This relationship means that in the case of random index sets, the spatial weights become random variables and can no longer be treated as deterministic coefficients. This complicates drawing inferences about spatial autocorrelation and other statistics, as knowledge about the distributional properties of the associated spatial weights is required and existing inference frameworks can no longer be applied in the established way. However, the assumption of a fixed spatial index still ought to be dropped for the sake of meaningfulness of spatial estimators, especially in cases where the spatial units and their associated attributes are co-constitutive.

Spatial autocorrelation scenarios exist in which considering spatial indexes random would be helpful. Recent examples include the spatial analysis of human-generated information (especially so-called ambient geospatial information; Stefanidis, Crooks, and Radzikowski 2013) like that from social media and mobile-collected data such as nonstationary sensor measurements. Many of these data sets are collected through smartphones that involve the use of Assisted GPS (A-GPS) combining signals from built-in GPS receivers with those derived from the cellular network (Vallina-Rodriguez et al. 2013). A-GPS sometimes reduces the positional accuracy of retrieved coordinates, for example, when a location is acquired in an app such as Twitter only shortly after the GPS function has been activated. Average horizontal accuracies ranging between 6.5 and 13 m have been reported for different devices and in different urban settings (Zandbergen and Barbeau 2011; Garnett and Stewart 2015; Merry and Bettinger 2019). Randomness attached to locations can also originate from complex psychological and social processes. The strong embedding of social media in everyday practices has led to conflation of material and digital spaces (Wagner et al. 2021), which affects social norms and behaviors (Quesnot and Roche 2015; Kitchin, Lauriault, and Wilson 2017). For example, Saker (2017) found that many Foursquare users check in to places in strategic ways in order to, among other things, express their identities or to curate a particular digital alter ego. The geometries associated with social media data are hence the results of (often little understood) complex stochastic processes. To avoid these complexities, data from social media are often analyzed in aggregated form (e.g., van Zanten et al. 2016; Resch, Usländer, and Havas 2018; de Andrade et al. 2022). Aggregating, however, also means that a lot of potentially relevant microgeographic information is lost. At the same time, the spatial analysis of social media data at the level of individual posts poses challenges regarding scale, distributions of statistical measures, and in general the simultaneity of different phenomena being reflected (Westerholt, Resch, and Zipf 2015; Westerholt 2019). A third example where randomness in the spatial index plays a role is competition for land, resources, market areas, or other geographically distributed commodities. Both, Griffith and Arbia (2010) as well as Griffith (2006) show that corresponding situations can be represented with the help of random geometries, in these cases with Thiessen polygons. The randomness in polygon sizes, shapes, and locations (resulting from the randomness of the underlying point configurations) leads to

negative spatial autocorrelation, the characterization of which in turn allows conclusions to be drawn about spatial competition scenarios. In line with these results, and also using Thiessen polygons, Valcu and Kempenaers (2010) found that the degree of spatial autocorrelation in territory sizes decreases when the latter result from interactions between underlying neighbors. However, Valcu and Kempenaers (2010) only tested for positive spatial autocorrelation. In all demonstrated cases, the site randomness may be considered in the inference about spatial autocorrelation statistics, but it remains unclear how the additional variation contributed by random spatial indexes changes the distributional properties of spatial autocorrelation statistics like Moran's $I$.

This article falls within the scope of location variability. According to Jacquez (1999), the term *variability* should be distinguished from the more common term *uncertainty*. Both refer to random locations, but uncertainty can be reduced by more precise measurements, while variability refers to an inherent property of a system under study. Nevertheless, it is instructive to situate this article in the literature of the intersection of the topics of spatial autocorrelation and uncertainty, as these works share some commonalities with what is presented below. Originally dubbed "error," thinking in GIScience has shifted in recent years toward the notion of "uncertainty" (Goodchild 2010), and with it the acceptance that representing geographies is an inherently uncertain endeavor. In a series of early works, Geoffrey Jacquez explored the impact of uncertain locations on exploratory spatial data analysis (Jacquez 1996; Jacquez 1999; Jacquez and Jacquez 1999). These works address the effects of imprecise coordinates on spatial statistics, focusing on the Mantel test but also briefly on Moran's $I$. Building on so-called location models that describe corrupted point locations, these papers find that inferences drawn from established methods are inaccurate because they ignore an important additional source of variation. However, the focus is mainly on finding more appropriate reference distributions using Monte Carlo methods rather than understanding Morans $I$ in detail. Burra et al. (2002) investigate Moran's $I$ and the hotspot technique Getis-Ord $G_i^*$ in the context of geocoding errors. They conclude that global statistics are less prone to location uncertainty and that Moran's $I$ is more affected than $G_i^*$. Focusing on the same local statistics, Griffith, Chun, and Lee (2016) confirm the result that Moran's $I$ is more affected by location uncertainty. However, both studies are mainly concerned with differences in significant outcomes and revealed clusters, rather than more comprehensive aspects of the statistics studied. Jacquez and Rommel (2009), who also focus on geocoding inaccuracies, develop a scatterplot that combines estimates of error sensitivity and leverage, that is, the propensity of a point to spatial error propagation in spatial analyses. However, their work is based on the corruption of point coordinates with unknown distributions, which is different from the work presented in this article which deals with well-defined point processes. There are also uncertainty-based results for the related but somewhat differing field of spatial regression (e.g., Griffith et al. 2007; Lee, Chun, and Griffith 2018). The results of these studies show that spatial uncertainties there also have a large impact on parameter estimates. Also, there are approaches to quantify spatial variance, for example, the LOSH (Ord and Getis 2012; Xu, Mei, and Yan 2014) and LSD (Westerholt et al. 2018) estimators. These heterogeneity works are related but not directly comparable to the work presented below. In summarizing a series of papers on uncertainty presented at the 2017 AAG meeting in Boston, MA, including work on spatial autocorrelation, Griffith (2018, p. 1504) concludes that "[m]uch remains unknown about uncertainty in spatial as well as space-time data, with respect to both attribute and location error." The present article contributes to a better understanding of the latter

location-related aspect – not of uncertainty, but of variability – by examining the link between Morans *I* and well-defined point processes.

The present article investigates how the consideration of stochastic spatial indexes affects inferences drawn about global Moran's *I*. Two avenues are pursued. In a first step, experiments are conducted with synthetic data representing two types of homogeneous point processes: the Thomas and Matérn cluster processes. For each of these, 10,000 point patterns are simulated and populated with normal random variates, the latter drawn from spatial autoregressive models with varying degrees of spatial autocorrelation. In a second step, another two sets of 10,000 simulated point patterns each are generated, but this time based on a two-dimensional intensity map derived from tweets collected in Canary Wharf, London, UK. These latter simulations do allow the study of inhomogeneous point patterns corresponding to actual social media posts. Three aspects are investigated for all types of simulated point patterns: the respective null distributions of Moran's *I* in contrast to the conventional case; deviations between conventional and simulated critical values at different significance levels; and the statistical power of Moran's *I* under the assumption of stochastic spatial indexes. In general, the results obtained for the homogeneous case are more generalizable than those for the inhomogeneous case. The reason for this is that the latter depend on very specific real-world data, which is therefore both an advantage and a disadvantage. The empirical investigation put forward sheds light on the behavior of global Moran's *I* under the circumstances outlined, and thus provides a point of reference for assessments of the validity of identified spatial patterns in scenarios where the spatial index involves randomness.

## Moran's *I* and associated inference procedures

The method investigated in this article is global Moran's *I*, a frequently used measure of spatial autocorrelation. Moran's *I* extends the non-spatial Pearson coefficient with spatial weights that reflect pairwise interaction potentials between individual spatial units (Getis 2009; Dray 2011). The global version of Moran's *I* is given as (Getis 2010)

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{1}$$

where the $y_i$ denote $n$ observations with mean $\bar{y}$ that are spatially connected via weights $w_{ij}$. The expected value of the statistic is $\mathrm{E}[I] = -1/(n-1)$, but the variance depends on the hypothesis tested and the corresponding modeling of the null distribution (see Experimental setup Section). Hypothesis testing in a spatial autocorrelation context requires knowing (theoretically) or approximating (empirically) the probability of an observed pattern under spatial uncorrelatedness[3]. Various procedures for drawing inferences about Moran's *I* exist.

The classical inference framework used in conjunction with global Moran's *I* comprises a so-called normal and a randomization hypothesis (Cliff and Ord 1981). The normal hypothesis (referred to as hypothesis N hereafter) assumes that all variables have independently been drawn from the same normal distribution. With repeated sampling and allocation to sites, the full range of values drawn from the respective normal distribution could then hypothetically be realized at equal chance in the limit and at any given site. Alternatively, and relaxing distributional constraints, the randomization hypothesis (referred to as hypothesis R hereafter) holds the vector of observed values fixed and redistributes them on the spatial index. The latter type of hypothesis testing is useful when no values other than the ones observed could reasonably occur. Irrespective

of using assumption N or R, the focus of hypothesis testing is on testing for spatial associations in the attribute values given a fixed arrangement of spatial units. Under both hypotheses, Moran's *I* is asymptotically normal and respective closed-form, analytic expressions exist for evaluating the variance for the purpose of standardizing the *I* values. Further, for the case of hypothesis N, Griffith (2010) has studied the behavior of Moran's *I* under different distributional regimes and has found that convergence of *I* to normality is still acceptable as long as the analyzed random variables are drawn from approximately symmetric distributions. For nonsymmetrical cases that exhibit pronounced skewness and kurtosis, Tiefelsdorf (2002) provides a saddlepoint approximation. However, the distribution of *I* depends not only on the statistical properties of the random variables, but also on the spatial arrangement of the associated spatial units as it is manifested by the weights. The normality property may no longer hold if the analyzed data set is small or if very unfavorable spatial linkages are present. For the former, the exact distribution of Moran's *I* worked out by Tiefelsdorf and Boots (1995) can be used. Alternatively, Cliff and Ord (1972) have suggested the use of Beta approximations. In the latter cases of unfavorable spatial linkages, the shape of the null distribution of *I* is strongly influenced by the eigenvalue spectrum associated with the spatial weights matrix (de Jong, Sprenger, and van Veen 1984; Tiefelsdorf, Griffith, and Boots 1999). These eigenvalues reflect topology-induced variance, which can be interpreted as the potential of individual spatial units to interact with the rest of the map.

Inferences about Moran's *I* can be compromised by violations of the assumption of second-order stationarity. For example, if the variance of random variables differs greatly between different parts of a map, this compromises the assumptions for both hypotheses N and R. While the former assumes a uniform normal distribution for the resampling, the latter is based on redistribution across all sites and therefore expects similar conditions everywhere with respect to the first two moments at least in the null hypothesis of no autocorrelation. These assumptions are not met in case of heteroscedasticity. Furthermore, Moran's *I* uses global estimators for both the mean and the variance of the variables $y_i$. These are also no longer reliable if the variance and expected value are not uniform. A frequently studied case of unstable variance is the analysis of rate variables where the underlying populations may exhibit strong spatial heterogeneity. It has been shown that both Type-I and Type-II error inflation can result (Walter 1992a; Walter 1992b). Techniques and modified estimators have been proposed to cover a range of scenarios, including the rates mentioned but also those where uncertainty and nonstationarity come from other sources (Oden 1995; Waldhör 1996; Assuncao and Reis 1999; Jackson et al. 2010; Zhang and Lin 2016; Jung, Thill, and Issel 2019; Bucher et al. 2020). Sometimes nonstationarity-related effects are further reinforced, for example, when a data set reflects not only multiple but even spatially superimposed processes that are difficult to distinguish from each other. The resulting topology reflected in the spatial weights can then be complex and may lead to individual observations having an excessive influence on an analysis, which also affects the shape of the null distribution and respective inferences drawn (Westerholt, Resch, and Zipf 2015; Westerholt et al 2016a; Westerholt 2018). The latter is often the case with data from social media, for example, as these are neither collected according to any sampling scheme nor for the purpose of spatial analysis.

## Inference with random spatial indexes

Accounting for randomness not only attribute-wise but also in the geographic sites (i.e., in the underlying spatial index) changes the nature of null hypotheses when testing for spatial

autocorrelation. In all cases outlined in Moran's *I* and associated inference procedures Section, the hypotheses being tested in the case of Moran's *I* are about the extent to which spatial structures in attributes are possible random outcomes conditional on given spatial linkages between fixed geographical sites. In many cases, this restriction makes a lot of sense, such as in the analysis of census tracts or municipal districts the geometries of which are neither the result of random processes nor generated from the underlying observed spatial processes responsible for the attributes. In other cases, however, it may be considered a restriction to consider the spatial index as fixed given the genesis of certain data sets. For example, when analyzing social media data like tweets or Flickr photos whose locations depend to some extent on individual choices, or when considering mobile sensor measurements that are subject to GPS inaccuracies, it would be reasonable to include the randomness of locations in the hypothesis testing framework of Moran's *I*. Instead of testing only for associations between attribute values and a particular spatial index, including respective associated weights, the perspective changes to a joint assessment of the probabilities of spatial configurations as the outcome of more than one random process.

Varying the locations of a spatial index set in addition to randomizing the attribute confronts the analyst with a number of additional challenges. One challenge that arises is that meaningful mechanisms for relocating sites to new locations must be identified. On the one hand, such mechanisms should reflect meaningful realistic random properties of some spatial process under study. On the other hand, comprehensible and tractable ways must be found to reallocate the sites. If the attribute vector is to be held fixed too, as in the randomization hypothesis R, an additional constraint is that the number of points should be constant at each randomization. Otherwise, it would not be clear how to either generate additional attribute values or how to thin out additionally drawn ones. In contrast, if there are good reasons to believe that some more complex generation model is required, it would become difficult to keep the number of locations generated per randomization constant (as these are subject to randomness) and thus to preserve the observed attribute vector. The latter would require an additional commitment to also the distribution of attribute values, as is the case with hypothesis N. Another challenge is that the term random index here does not refer to mere uncertainty of coordinates as elaborated in some of the GPS location examples further above. Instead, the main focus of this article is on considering stochastic spatial processes, which represent real-world properties of geographic processes and thus go beyond technically induced uncertainty. The latter may be seen as a special case, but is not exhaustive in the sense of the discussion presented here.

Technically speaking, drawing inferences about Moran's *I* with random spatial indexes requires statistical knowledge about the spatial weights. The weights can no longer be considered as deterministic coefficients, but must be treated as an additional source of variation in inference procedures. Looking at the different types of spatial weights, most of them emerge as functions of characteristics of the underlying spatial units. The commonly used inverse distance weighting scheme, for example, is given as a function of the distance measures between members of the index set. The latter distance measures are easy to calculate and use in the case of fixed index sets, but are associated with randomness under the assumption of a random spatial index. Similarly, binary contiguity weighting schemes of first or higher order can be interpreted as functions of the nearest neighbor distribution under the assumption of a random spatial index set. Furthermore, the number of spatial units $n$ is also subject to randomness, which must be considered as well. These aspects, together with the fact that they are of different nature under different assumed underlying data generating processes, make it difficult to find an all-encompassing analytical solution to the problem at hand using the Pitman–Koopman Theorem. The latter would require

consideration of the specific combination of type of point process and chosen weighting scheme, as these together determine the distribution of spatial weights. Yet, these distributions are further complicated by merging, increasingly overlapping, growing clusters, as the latter introduces another random process. Moreover, the spatial weights used in the following are rather complex combinations of different schemes that are intended to mitigate certain possible confounding factors in the interest of the interpretability of the results obtained. For the reasons outlined, deriving analytical solutions to equations for the mean and variance of Morans *I* is beyond the scope of this article, but it is hoped that the discussion offered in this paragraph will be beneficial to respective future research.

It is instructive to shed more light on the problem at hand by comparing it with the randomization hypothesis presented in Moran's *I* and associated inference procedures Section. When the randomization hypothesis is applied, the assignment to existing fixed localized spatial units is considered random. Thus, the locations of random variables change and this happens randomly, but with the restriction to a finite set of possible locations. The spatial index itself including the number of spatial units and their pairwise links by means of spatial weights therefore remain unchanged and are not subject to randomness. In comparison, in the present case studied in this article, the restriction to a finite set of locations is abandoned. Instead, new locations are drawn for each sampling, following the logic of well-defined point processes. However, challenges arise in doing so. In point processes, by definition, the numbers of points in the samples drawn are not constant. Moreover, in the cluster processes considered in this article, overlaps of generated clusters occur frequently (see Point process models Section). The latter means that the spatial weights are subject to complex dynamics with potentially unknown distributions. As stated before, spatial weights are often functions of properties of the underlying spatial index such as pairwise distances, the probability of having a certain number of nearest neighbors on distance bands, and so on. However, if random point processes strongly complicate these dynamics, analytical solutions may not be available for every type of spatial weights. This complex case shall be considered in the following and approximated by means of simulations.

In this article, the influence of random spatial index sets on inferences about Moran's *I* is investigated. Since existing inference regimes assume fixed locations, no formal framework exists for the present case of interest. Therefore, approximations are presented below that rely on large numbers of simulations. The main interest of this study is to investigate the performance of established inference techniques when the phenomenon under investigation is actually also subject to randomness in terms of its locations. Three objectives are pursued: (i) the conventional normal approximations under the assumptions N and R are contrasted with the approximated distribution of *I* using random locations; (ii) deviations between the respective derived critical values for rejecting the null hypothesis of no significant spatial autocorrelation are investigated; and (iii) the statistical power of Moran's *I* is discussed in the outlined context. The following section outlines the methodology through which the subsequent experimental results are obtained.

## Methodology

The experiments reported in this article are based on two types of point process models. The following subsection introduces the Matérn and the Thomas cluster processes. Subsequently, the approach for the simulation of a large number of point patterns including their parameterization is outlined. In addition to homogeneous point processes, this article also studies the effects of inhomogeneous point pattern configurations on inference about Moran's *I*. For this reason, the

generation of corresponding simulations based on a real Twitter data set is presented in a further subsection. Finally, the analytical steps carried out are motivated and explained.

**Point process models**

Two different types of point process models are used to study the influence of random locations on inferences drawn about Moran's $I$: the Matérn and Thomas cluster processes. This choice is guided by their usefulness in analyzing a range of different phenomena, such as pine tree patterns in natural forests (Tanaka, Ogata, and Stoyan 2008), small cells as used in telecommunications (Wang and Zhu 2016), and the dispersal of tree seedlings (Fedriani, Wiegand, and Delibes 2010), among numerous other application scenarios. Both models are specializations of the so-called Neyman–Scott process, which describes random locations (daughter process) organized around cluster centers (parent process). Both the parent cluster centers and the daughter offspring points are outcomes of Poisson point processes and the points are thus generated independently of each other. The two processes considered in this article differ in terms of the placement of the locations of the offspring points.

The offspring locations of Matérn cluster processes are uniformly and homogeneously distributed around cluster centers (Matérn 1960). This process type is suitable for modeling the assumption that there are local clusters of points, but the corresponding locations of the offspring are scattered locally at random, without any further specified spatial structuring. An example of this is the subjective and independent data collection via social media, for example through tweets. The urban topography causes more tweets to be posted in some places, such as shopping streets, nightlife areas, or major train stations, than elsewhere. However, since the tweets are posted independently of each other and often without any obvious centrally localized trigger, the result is a spatially random arrangement around an imagined, often not physically existing cluster center. Let $c_i$ be cluster centers of a Matérn cluster process $C_M$. Also, let $r = d(t, c_i)$ be the Euclidean distance of any point $t \in \mathbb{R}^2$ from a cluster center $c_i$ and $R$ be an upper distance limit beyond which the daughter process is no longer defined. Since there is no local spatial substructuring, the point intensity describing the expected number of daughter points on a disk with radius $r$ and center $c_i$ is constant for all $r \leq R$ and is given by (Lawson and Denison 2002, p. 69) (the inhomogeneous case is outlined in Simulation of inhomogeneous point patterns Section)

$$h_M(r|c_i) = \begin{cases} \mu, & \text{if } r \leq R, \\ 0, & \text{otherwise.} \end{cases}$$

The offspring locations of Thomas cluster processes are scattered around cluster centers with a Gaussian distance decay (Thomas 1949). In contrast to Matérn cluster processes, Thomas cluster processes thus model different assumptions regarding the genesis of geometric arrangements of a spatial index. Here, it is not assumed that the local behavior with regard to the realized locations is detached from local events beyond environmental geographical conditions. Instead, the Gaussian distance decay represents a centrally located origin triggering local clustering. Linking to the social media example given in the previous paragraph, one such use case would be the modeling of tweets about a car accident. The latter represents an incentive to post messages that is located in the cluster center. The further away potential Twitter users are from the incident, the less likely they are to post about it. Again, let $c_i$ be cluster centers, but this time of a Thomas process $C_T$. Using an analogous notation as above and with $\sigma$ as the standard deviation of the random offspring point displacement, the respective point density on the discs around the cluster

centers is given as (Lawson and Denison 2002, p. 69) (again, the inhomogeneous case is found in Simulation of inhomogeneous point patterns Section)

$$h_T(r|c_i) = \frac{\mu}{2\pi\sigma^2} e^{-r^2/2\sigma^2}.$$

The choice of the processes tested is not without limitations. One limitation is that both processes are Neyman–Scott processes and thus based on the assumption of independence of the parent as well as the offspring processes. The following results can therefore not be transferred to more complex process models representing proactive point interaction. At the same time, however, the assumption of independence is useful especially with regard to the analysis of user-generated data such as that from social media (to which much reference is made given the empirical data used in this article), to which users located near each other usually contribute independently. Furthermore, independence in the parent processes for cluster centers allows clusters to overlap, which, for example, occurs with social media data. This latter gives rise to a further limitation when testing the two processes in isolation, as they can also occur together under real-world conditions. However, thoroughly understanding the influence of first these individual processes on Moran's $I$ reference distributions requires careful isolation from confounders like complex mixtures, otherwise it would be difficult to draw meaningful conclusions about the influence of either of the two tested processes on spatial autocorrelation assessments. Nevertheless, the investigation of the complex interplay of different types of random point processes in connection with Moran's $I$ is interesting and important and therefore recommended as a useful follow-up research task. The next subsection will now present the simulation of actual homogeneous point patterns using the two models presented above.

### Simulation of homogeneous point patterns

The first investigation of Moran's $I$'s inference behavior is conducted with synthetic data reflecting homogeneous point processes. Two sets of estimated patterns are generated from the homogeneous Thomas and Matérn cluster processes outlined above. Each of these sets contains 10,000 random patterns that reflect a wide range of possible realizations of the two processes. For each pattern generated, the target average size is 400 points overall consisting of four clusters of 100 points each. The exact numbers of points per cluster are not controlled, as, per definition of the processes involved, these are random and will follow Poisson distributions. For both types of processes, the parent processes generating the cluster centers are set as homogeneous Poisson processes with intensity $\lambda = 6.673692e-07$. This intensity is obtained by dividing 4 (the average number of desired clusters) by the area of the rectangular window used. The radius $r$ chosen for the Matérn process is set to 400 m. To harmonize the scales of both processes, the standard deviation of the random point displacement for the Thomas process is set to $\sigma = 150$, resulting in comparable cluster sizes. In both cases, the mean values of the respective Poisson daughter processes are set to 100 points. The reasoning behind these parameterizations is to quantitatively resemble the set of tweets introduced below in Simulation of inhomogeneous point patterns Section. Fig. 1 illustrates two exemplary generated patterns.

### Simulation of inhomogeneous point patterns

The second investigation conducted reflects the inhomogeneous case and comprises a tweet data set extracted from a one-year Twitter corpus from London, UK. The data is available online (see Westerholt et al. 2016b) and has been used in previous studies (e.g., Steiger et al. 2015;

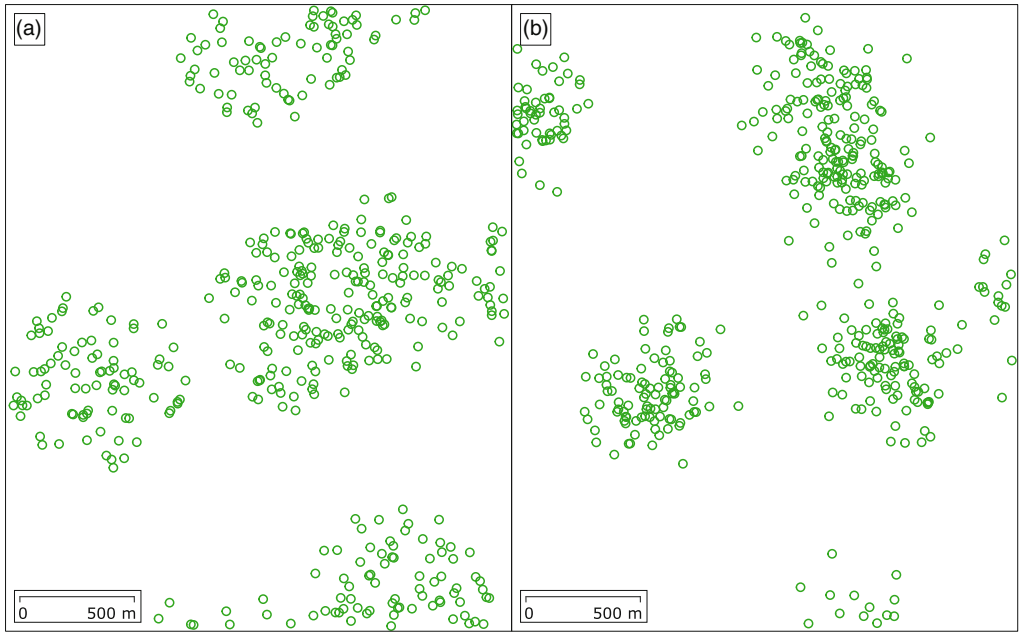


**Figure 1.** Illustration of two homogeneous point processes. (a) One of the simulated Matérn cluster processes. (b) One of the simulated Thomas cluster processes.

Steiger, Resch, and Zipf 2016; Westerholt 2021). The full preprocessing chain can be found in Steiger et al. (2015) and includes stop word removal, tokenization, and stemming. A classification of the tweet texts was performed using Latent Dirichlet Allocation (Blei, Ng, and Jordan 2003) in combination with a Gibbs sampling strategy to preidentify the most appropriate number of topics. This classification resulting in scores from zero to one and capturing work-related tweets, forms the basis for selecting the subset of data used. The data used in this article come from the Canary Wharf business district in London's docklands and it is mapped in Fig. 2a.

The reason for using Twitter data in this article is to investigate inhomogeneous point patterns and how these affect inferences about Moran's *I*. Spatial urban processes are subject to a commonly observed tendency to be nonuniform due to topographic characteristics, population distributions, and other underlying spatially relevant processes (Páez and Scott 2004). For this reason, inhomogeneous clusters are modeled to account for the resulting spatial heterogeneity. The additional experiments conducted in this way therefore complement the results obtained using the patterns presented in Simulation of homogeneous point patterns Section for the homogeneous cases. In order to operationalize the tweets, the data set is transformed into an intensity grid that reflects, at a fine-scale level, the probability that tweets are observed. This is achieved by performing kernel density estimation based on a target grid cell size of $10 \times 10$ m. The resulting grid is then converted into an ASCII grid, which has been loaded into the R statistical computing ecosystem for the analysis. With regard to the simulation of point patterns, inhomogeneous Matérn and Thomas processes are modeled. As with the homogeneous processes, the cluster centers again follow homogeneous Poisson processes but with intensity $\lambda = 0.00165$. This is to reflect the possibility that individual tweets can in principle occur anywhere in a city, and is not problematic as the focus is exclusively on an urban area. In contrast to the homogeneous cases,

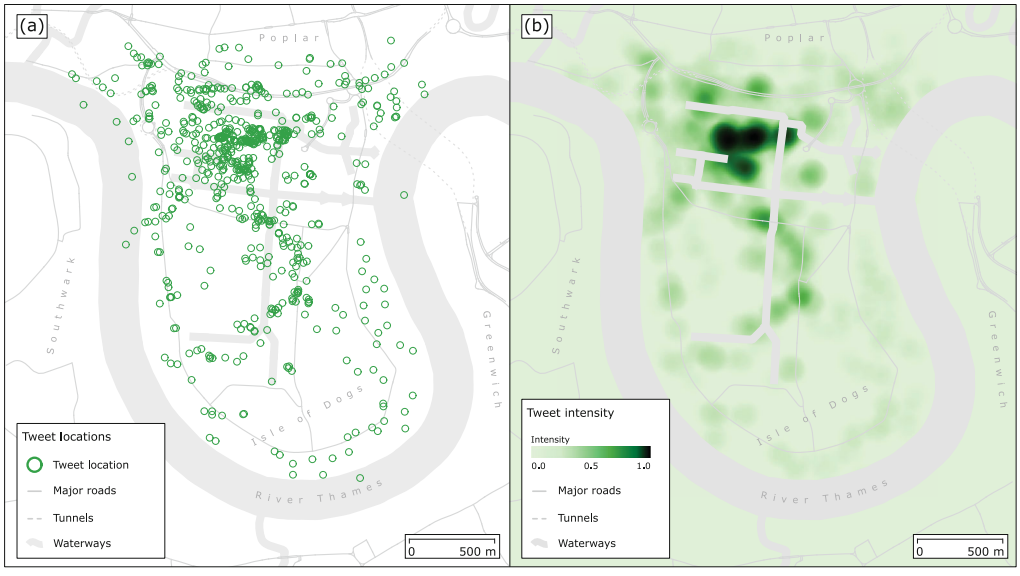


**Figure 2.** Tweet data set from Canary Wharf, London, UK. (a) Tweet locations. (b) Corresponding tweet intensity used for simulating inhomogeneous point patterns.

the densities of both daughter processes follow the intensity grid outlined. Thereby, the daughter process for the Matérn case is parameterized with a scale of 250 m, while the offspring points of the Thomas process are based on a Gaussian distance decay for a random point displacement of $\sigma = 100$. These parameters resemble the real-world tweets and ensure that 400 points are generated on average, making the results obtained more or less comparable to those obtained for the homogeneous patterns.

The tweets used to derive the intensity grid are tied to the urban topography and infrastructure of Canary Wharf. This means that the inhomogeneous patterns generated are not fully representative of all possible kinds of scenarios. However, the tweets used to generate the inhomogeneous intensity grid represent typical clustering behavior commonly found in social media data from large cities including other parts of the same data set from London. Clusters occurring in comparable areas elsewhere show similar geometric scales in clustering, although certain geometric structures are inherently local, such as those pertaining to the port landscape in Canary Wharf. The use of real-world data is therefore both an advantage (because it allows the features of interest to be tested in a real-world scenario) and a limitation (because it restricts the scope for generalizability). Follow-up studies could focus specifically on the inhomogeneous case and use a set of intensity maps that are independent of real observations and have different degrees of clustering and scale (or other parameters). These could then be combined with different types of inhomogeneous point processes to simulate various types of possible scenarios in a systematic way. However, this is beyond the scope of the present article (and would narrow it thematically) and is therefore deferred for future work.

## Experimental setup

All experiments conducted require the generation of attribute values and the specification of spatial weights. For each generated point pattern, a set of attribute vectors reflecting different

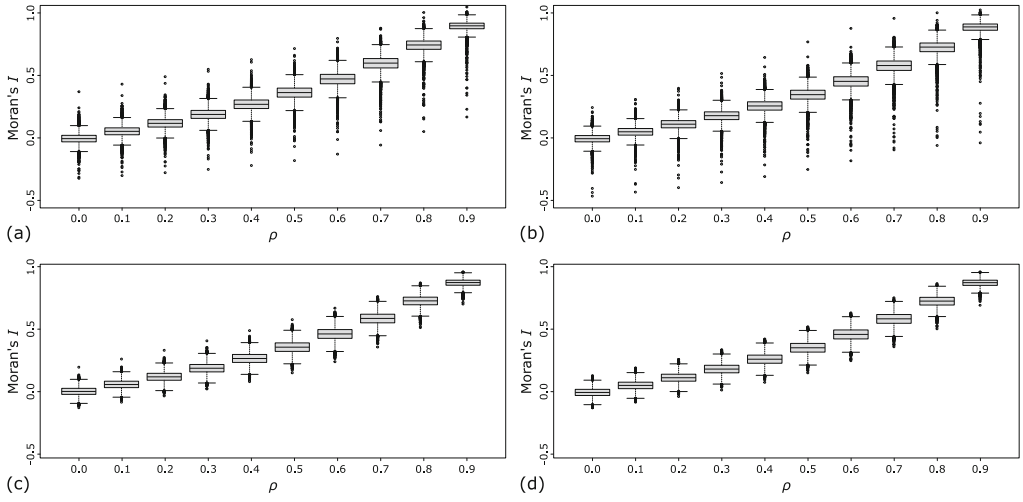


**Figure 3.** Boxplots of simulated Moran's *I* values for different spatial autoregressive parameters $\rho$. Homogeneous (a) Matérn and (b) Thomas processes. Inhomogeneous (c) Matérn and (d) Thomas processes.

levels of spatial autocorrelation is generated. These vectors are based on standard normal random variables, which are then left-multiplied by SAR generating operators associated with $\rho$ values in the range $[0, 0.9]$ (incremented in steps of 0.1). The seed of the pseudo random number generator used is set to 1 for each vector in order to exclude possible numerical variations as an additional factor and possible confounder. If instead the distributions of the attributes were also varied, it would be more difficult to discern the effects of geometric randomness from other sources of variation. The SAR generating operators turn the standard normal variates into spatially autocorrelated random variables according to simultaneous spatially autoregressive modeling. All SAR generating operators are given by $(\mathbb{I} - \rho W)^{-1}$ (Anselin 2001, p. 316), whereby $\mathbb{I}$ is the identity matrix[4] and $W$ is the spatial weights matrix as introduced in Moran's *I* and associated inference procedures Section. The different autocorrelation levels are needed for the assessment of statistical power outlined below, as this requires knowledge of the distribution of Moran's *I* in the alternative hypothesis of significant spatial associations. A summary of the Moran's *I* values generated is found in Fig. 3.

The spatial weights are determined in a two-step procedure: First, for each geometric point $x_i$, the 10 nearest neighbors are identified; second, these neighbors $x_j$ are assigned weights using an inverse distance weighting with a distance decay of $d(x_i, x_j)^{-2}$, where $d(\cdot, \cdot)$ again represents the Euclidean distance function. This approach ensures that all points have comparable analytical neighborhoods, which would not necessarily be the case with neighbors derived from distances alone. In addition, edge effects are mitigated, since points at cluster boundaries are also guaranteed to have 10 neighbors, most of which are from the same cluster. The latter means that the generated distance weights are not extreme even in boundary cases. To further rule out influences of possible topological imbalance, all weights have been row standardized using the W coding scheme (see Bavaud 2014).

One analytical step concerns the comparison of the null distributions of Moran's *I* for the conventional case of a fixed spatial index with the distribution obtained taking into account random spatial indexes. For this comparison, the null distributions of *I* under the assumption

of a fixed spatial index are constructed under both the N and R hypotheses and for all patterns generated. Given the convergence behavior of Moran's $I$, the respective asymptotic normal distributions $p_{0;k}(I)$ for each point pattern $k$ are constructed with the mean $\mathrm{E}[I] = -1/(n-1)$ (which is the same under both hypotheses) and using the variance terms (Cliff and Ord 1981, pp. 42 ff.)

$$\mathrm{Var}_{\mathrm{N}}[I] = \frac{1}{\left(n^2-1\right)S_0^2}\left(n^2S_1 - nS_2 + 3S_0^2\right) - \mathrm{E}[I]^2, \tag{2a}$$

$$\mathrm{Var}_{\mathrm{R}}[I] = \frac{n\left(S_1\left(n^2-3n+3\right) - nS_2 + 3S_0^2\right) - \frac{m_4}{m_2^2}\left(\left(n^2-n\right)S_1 - 2nS_2 + 6S_0^2\right)}{(n-1)(n-2)(n-3)S_0^2} - \mathrm{E}[I]^2, \tag{2b}$$

with $n$ being the number of spatial units contained in a pattern and

$$S_0 = \sum_{i,j}^{n} w_{ij}, \quad S_1 = \frac{1}{2}\sum_{i,j}^{n}\left(w_{ij}+w_{ji}\right)^2, \quad S_2 = \sum_{i}^{n}\left(\sum_{j}^{n} w_{ij} + \sum_{j}^{n} w_{ji}\right)^2,$$

$$m_2 = \sum_{i}^{n}\left(y_i - \bar{y}\right)^2 \quad \text{and} \quad m_4 = \sum_{i}^{n}\left(y_i - \bar{y}\right)^4.$$

Additionally, the null distribution $q_0(I)$ for the case of random spatial index sets is estimated from the simulations. Estimation is necessary because closed-form analytical expressions are unknown for this case. The attributes generated with $\rho = 0$ are used, as these represent the case of no spatial autocorrelation. For both point processes and for each simulated pattern, the resulting null distributions are compared pairwise. This is done using the Kullback–Leibler (KL) divergence (Kullback and Leibler 1951) as a measure of the amount of information lost when the observed null distribution $q_0(I)$ assuming randomness in the index is approximated by the conventional null distributions $p_{0;k}(I)$ for each individual fixed pattern $k$. The KL divergence associated with each simulated pattern $k$ is given as

$$D_{\mathrm{KL}}(q_0, p_{0;k}) = \int_{-\infty}^{\infty} q_0(u) \cdot \log_2\left(\frac{q_0(u)}{p_{0;k}(u)}\right) du. \tag{3}$$

Estimating the KL divergence of continuous distributions is more involved than for discrete variables. In this article, we operationalize the solution proposed by Pérez-Cruz (2008). This solution requires only that the samples used are independent and identically distributed. Both presumptions are satisfied for the samples drawn in the null hypothesis. Technically, the implemented estimation uses the empirical cumulative distribution functions of the variables involved, instead of first estimating their densities as it occurs in many other solutions and which introduces further uncertainties.

Utilizing all null distributions introduced, the critical values for rejecting the null hypothesis of no significant spatial autocorrelation are then determined in a second analytical step. The critical values are calculated in each case for the commonly used significance levels $\alpha \in \{0.1, 0.05, 0.01\}$. On this basis, the respective distributions of the deviations between the critical values for the fixed and the random index cases are estimated from all simulated samples. These distributions provide insight into how conservative or overly loose the commonly used critical values are when the data at hand originate (knowingly or unknowingly) from a process that is subject to randomness in the corresponding spatial index. The risk of possibly spurious

assessments of significant spatial autocorrelation can be discerned this way for Thomas and Matérn cluster processes.

The third analytical step carried out addresses the statistical power of the conventional global Moran's $I$ estimators in the presence of random spatial indexes. Statistical power means the probability of rejecting the null hypothesis when the alternative hypothesis is true (Cohen 1992). Investigating power requires knowledge about the distribution of an estimator in the alternative hypothesis. In the present case, the respective alternative hypothesis distributions of Moran's $I$ under different levels of spatial autocorrelation are estimated from the spatially autocorrelated random variables whose generation is described above in the first paragraph of this section. For each $\rho$ value, a vector of Moran's $I$ estimates is calculated over all simulated patterns and for both types of processes. The actual calculation of statistical power is then performed by first estimating the empirical cumulative distribution function $Q_1$ from these $I$ values and for each $\rho$ greater than zero. In a second step, the powers are determined by calculating $1 - Q_1\left(C_\alpha; \rho\right)$, where $C_\alpha$ is the critical value for the significance level $\alpha$ that would conventionally be used to evaluate the theoretically constructed distributions. What is tested here, therefore, is the statistical power of the usual global Moran's $I$ when inadequately applied to a data set originating from a process that is subject to randomness in the spatial index.

## Results and discussion

The following subsections start out discussing the comparisons between the null distributions obtained under the assumptions of fixed and random spatial indexes. Subsequently, the related topic of critical values is addressed, which is of great importance for the application of Moran's $I$. In the last subsection, the statistical power of Moran's $I$ is discussed when a fixed spatial index is spuriously assumed even though the actual process under investigation is tied to a random index set.

### Null distributions

The first investigation concerns possible differences between the null distributions of $I$ with and without random spatial indexes. All histograms shown in green in Fig. 4 depict the KL divergences between the theoretical null distributions under hypothesis $N^5$ and the respective empirically determined null distribution estimated from the simulated point patterns with randomness in the spatial index. Shown in gray for comparison are histograms of the divergences based on the same theoretical distributions but compared with the empirical distributions obtained for each simulation step using only one respective arrangement of spatial units through repeated redrawing (hypothesis N) or randomization (hypothesis R) of the attributes. The gray histograms thus represent the performance of the estimator for $I$ as is conventionally used.

The normal approximations of the null distributions of Moran's $I$ are affected by the presence of stochastic spatial index sets. The green and the gray histograms in Fig. 4a and b reveal that the theoretically and empirically determined distributions can differ considerably in case of homogeneous point processes. A large number of KL divergences accumulate near the respective mean values. However, in both green histograms there is a strong positive skewness of about $g \approx 6$. One consequence of this is that in the simulations the left-hand side of the symmetrical gray reference histograms remains unreached. At the same time, extreme values occur that lead to a heavy right tail. For inhomogeneous point patterns, the behavior is slightly different. The

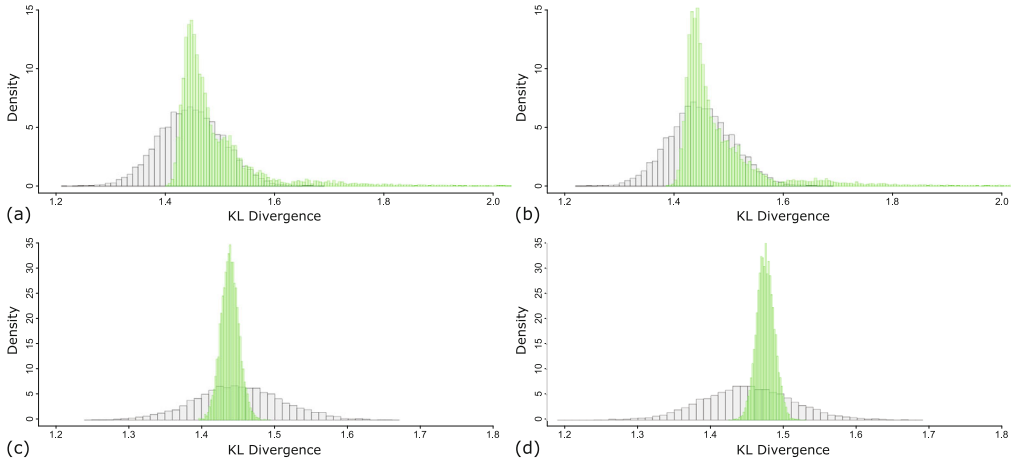


**Figure 4.** Histograms of Kullback–Leibler divergences. Homogeneous (a) Matérn and (b) Thomas processes. Inhomogeneous (c) Matérn and (d) Thomas processes. Green histograms depict divergences between empirical null distributions with random spatial indexes and corresponding normal approximations (hypothesis N); gray histograms depict divergences between conventional empirical null distributions without assuming random spatial indexes and corresponding normal approximations (hypothesis N).

gray reference histograms behave similarly to the homogeneous cases, but the green histograms appear more concentrated and symmetrical, as can be seen in Fig. 4c and d. The concentration around the means suggests that the use of the normal approximations does not lead to strong deviations from corresponding empirical distributions in case of inhomogeneous Matérn and Thomas processes. However, the narrow ranges around the mean values also imply that very small divergences are not achieved either. Another notable difference to the homogeneous cases is that the null approximation performs less optimally under the Thomas process than under Matérn-based patterns. While the estimator under hypothesis N performs slightly above average for Matérn processes, the mean value of the green histogram in the case of Thomas processes is on the right side of the symmetrical gray reference histogram. Overall, the KL divergences explained suggest that the conventional estimators for $I$ often perform better in the case of inhomogeneous point patterns than in the homogeneous cases.

The point processes used lead to variable point pattern sizes. It is therefore interesting to consider the above findings in the light of the sample sizes in the respective simulations. Fig. 5 illustrates this comparison for the case of Matérn processes (Thomas processes behave very similarly). It can be seen that only for the homogeneous case and in the presence of stochastic geometries is there a clear connection between KL divergence and $n$. As the pattern size increases, the KL divergence decreases, since the desired clusters are fully formed up to about the mean value. In the higher $n$ ranges, the KL divergence then increases again, which probably has to do with partly overlapping clusters and increasing point density (see Statistical power Section for further comments on point density). In the inhomogeneous case (Fig. 5b), the same trend can be seen in principle, although it is much less pronounced. The results shown in Fig. 5c and d are instead based on simulations under hypothesis N for each simulated pattern, performed in the conventional way, that is, on the repeated drawing of normal random variates, neglecting the

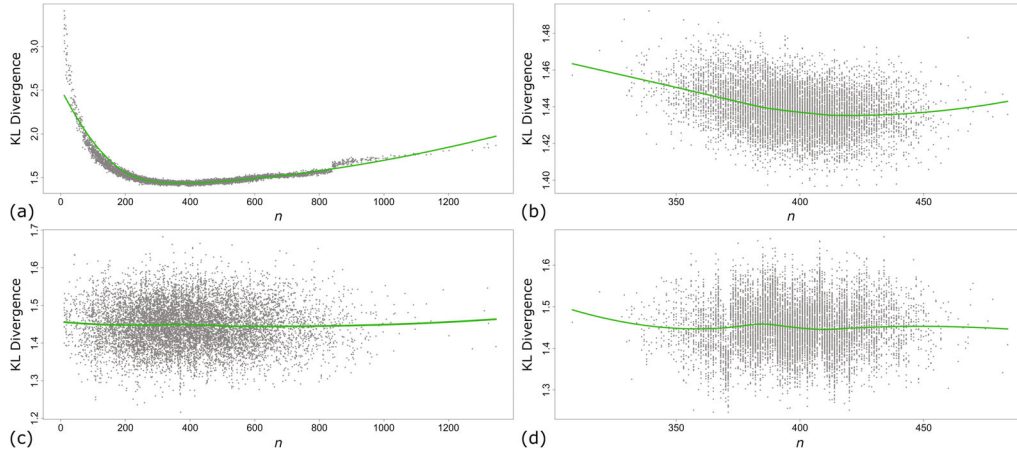


**Figure 5.** Kullback–Leibler divergences plotted against point pattern sizes $n$. (a) Homogeneous and (b) inhomogeneous Matérn processes with random spatial indexes. (c) Homogeneous and (d) inhomogeneous Matérn processes without assuming random spatial indexes. All divergences shown are calculated against corresponding normal approximations (hypothesis N). The green trend lines depict LOESS regressions.

existence of all other simulated patterns. We can see that in these cases without the presence of random sites, there is no discernable correlation with $n$. Overall, the picture is similar under both the conditions of homogeneous and inhomogeneous point processes, but the estimators of $I$ perform different in terms of tail behaviors.

Possible explanations for the divergences in the performances of the $I$ estimators tested are explored in more detail. Knowledge about the distribution of KL divergence is limited, but Belov and Armstrong (2011) provide some interesting results that bear relevance to the insights outlined above. Among other configurations, they report that KL divergence (i) is $\chi^2$-distributed when the mean differences between the studied distributions are normal and when the variability in the variance terms of the (in the present case) null approximations is 0. Furthermore, Belov and Armstrong (2011) prove that (ii) the KL divergence is non-centrally $F$-distributed if the variability of the addressed variance follows a noncentral $\chi^2$ distribution. There are indications that case (i) may apply to the gray histograms in Fig. 4 describing the performance of the conventional null approximations with fixed geometries. These histograms appear to resemble the normal distribution[6]. However, the variability of the empirically estimated variances among all drawings is indeed close to 0, and the central limit theorem suggests that the mean differences are normal since the analyzed attributes are all drawn from normal distributions (the latter follows from the Normal Sum Theorem, see Lemons (2003, p. 34 f.)). It is therefore not implausible to assume that the values follow a $\chi^2$ distribution with the mean being far away enough from zero to resemble normality. In contrast, visual inspection of the green histograms reflecting the performance of the estimators under random point processes shows a close resemblance to the $F$-distribution, that is, to case (ii). Moreover, the dispersion of the variance of the empirical estimator is larger, since different point patterns were used for estimation in each comparison. The difference between the green and gray histograms could therefore be due to the greater variability caused by the variation in the points in each drawing, which means that the mean deviations may not be normal and the variances may not be near 0. Deductive research beyond

the scope of this simulation study is needed to confirm these conjectures, but the present study supplies indicative evidence.

The results presented above are of importance for empirical statistical research on spatial structures. In many cases, the conventional null approximations for Moran's $I$ come close to the observed null distributions even when the underlying geometries are outcomes of point processes. However, the observed deviations between the approximation and the empirical distribution are seldom better than the average performances that would be observed in fixed geometry cases when comparing those respective theoretical and empirical null distributions. In both homogeneous cases, values above 1.6 are observed in about 10% of KL divergences when point processes are involved, while this number drops to below 0.1% in the fixed geometry cases. At the same time, the minimum values obtained with random geometries are close to 1.4, while 14.5% (Thomas) and 22.2% (Matérn) of all cases perform better with the conventional variants. Overall, there is thus often an increased risk of drawing erroneous conclusions about spatial autocorrelation when the null approximations for $I$ are used in cases where the analyzed data come from point processes. The differences in the tails of the distributions discussed above motivate investigating the impact of stochastic spatial indexes on critical values for assessing the significance of Moran's $I$. This is done in the following subsection.

### Critical values

Critical values delineate the bounds of certain $(1-\alpha)$ percentiles of the distribution of a test statistic. They thus limit the range of values above (or below) which observed values are considered significant. It is therefore important to understand how reliable the critical values obtained from the normal approximations for Moran's $I$ are in the light of empirical evidence with random index sets. Two types of critical values are considered and are reflected in Fig. 6: critical values $C_{q;\alpha}$ based on sample percentile estimates from $I$ values obtained in the simulations and for $\rho = 0$, hence reflecting empirical evidence taking into account geometric randomness; and critical values $C_{p;\alpha}$, which are based on theoretical percentiles of normal distributions with $\mu = \mathrm{E}[I]$ and $\sigma^2$ being either equation (2a) or (2b). Most of the point patterns considered are large enough to assure that $\mathrm{E}[I]$ is close to zero (more than 96% of all simulated point patterns have $n > 100$). Therefore, it is not necessary to subtract the mean and we can consider the values obtained directly. However, the positions of the critical values are case dependent and therefore difficult to interpret. Fig. 6 therefore shows the differences between the respective $C_{p;\alpha}$ and $C_{q;\alpha}$ expressed as multiples of the conventionally used critical values $C_{p;\alpha}$. This enhances the interpretability of the critical value deviations calculated due to a more relative character. Based on these considerations, the determined critical values show interesting behaviors with relevance for statistical inference about $I$.

The critical values resulting from the null approximations of $I$ and the empirical distribution with geometric randomness in the index differ markedly. Considering the case of homogeneous point patterns shown in Fig. 6a and b, a general negative trend in the differences of the critical values over sample size is noticed. This trend is observed independently of the significance levels $\alpha$. All observed differences start from a common origin area for small sample sizes and show a strong negative slope, which changes quite rapidly to a more uniform rate of change from medium sized point patterns near $n = 400$. This observation is stronger for the critical values associated with $\alpha = 0.01$. The distance between this type of critical values and other less stringent values attached with higher $\alpha$ increases with the sample size, indicating a stronger downward trend, especially since the other two types of critical values studied for $\alpha = 0.05$ and

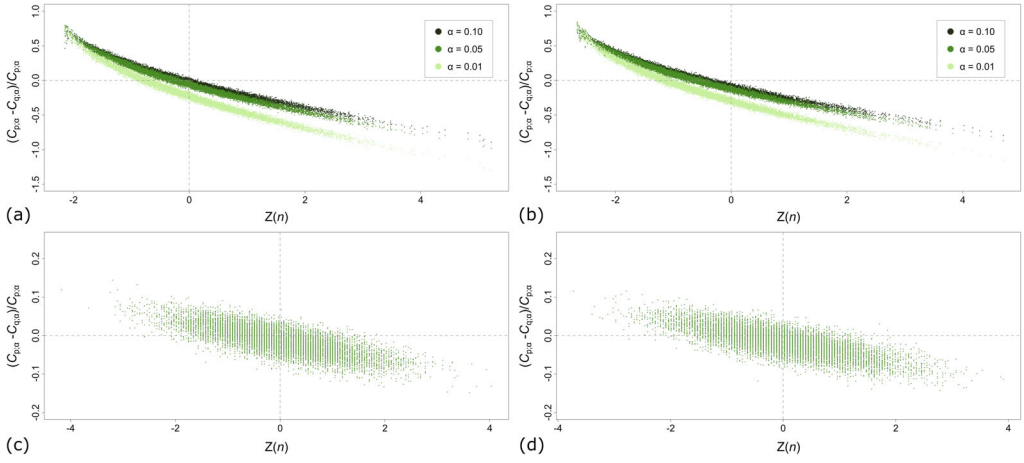


**Figure 6.** Differences between critical values $C$ for different significance levels $\alpha$ determined using conventional null approximations $p$ (hypothesis N) and based on empirical null distributions $q$ accounting for stochastic geometries. All differences are given relative to the critical values of the null approximations $p$. Homogeneous point patterns: (a) Matérn and (b) Thomas processes. Inhomogeneous point patterns: (c) Matérn and (d) Thomas processes. The data for the inhomogeneous cases are only shown for $\alpha = 0.05$. The point pattern sizes $n$ are standardized to allow for better comparability within the rows of the figure. Please note that the $z$-scores are not comparable one-by-one between homogeneous and inhomogeneous point patterns. The mean sizes ($z = 0$) are identical for all point patterns (at $n = 400$). The step sizes on the $x$-axes are equivalent to 181 (Matérn) and 187 (Thomas) for the homogeneous processes and to 21 (Matérn and Thomas) for the inhomogeneous processes. Please refer to the $x$-axes of other figures like Fig. 5 for the absolute point pattern sizes.

$\alpha = 0.1$ behave more similarly to each other. A closer look at the two sub-figures reveals that the outlined trends continue in a very similar way even for the few very large outlier point patterns. All the observations described are comparable for both Thomas and Matérn processes. The picture that emerges for the inhomogeneous cases, however, seems to be different. The Fig. 6c and d shows only the cases of $\alpha = 0.05$ for both types of processes. The reason that the other two cases are not shown is that the points are so close together that it would be impossible to discern any patterns if all the points were shown. In both cases we see a negative slope as in the homogeneous case. However, if we look at the value ranges, we see that the differences between the compared critical values are not as large as for homogeneous point patterns. These observations are interesting, but require further interpretation.

The critical values applied using the traditional inference framework for Moran's $I$ in many cases both overestimate and underestimate the effective bounds of the relevant percentiles for significance. Negative values in Fig. 6 represent Type-I errors, since in these cases the critical values assessed using the empirical distribution from the point patterns suggest a stricter limit for rejecting the null hypothesis. Analogously, positive values represent a higher risk of Type-II errors, that is, being too conservative and possibly missing spatial effects. The positive values observed in Fig. 6a and b are associated with point patterns of up to 500 points. Especially in combination with the commonly used significance level of $\alpha = 0.05$, this is a very relevant

result for a number of application scenarios, as many point patterns will be smaller than $n = 500$ in practice. The risk of rejecting the null hypothesis too conservatively is therefore a serious concern when analyzing point patterns without accounting for the additional variation contributed by geometric randomness in the inference procedure. If the point patterns are large, the risk of detecting a large number of possibly uninteresting spatial effects is also considerable. In both cases, overestimation and underestimation of critical values, the figures suggest that the extent to which the critical values are misspecified is in many cases up to almost 100% of the value suggested by the conventional null approximations. While these extreme results only occur for very small ($n < 50$) or very large patterns ($n > 900$), the still noteworthy discrepancy of about 50% of the conventional critical value is not uncommon according to Fig. 6. The values obtained for inhomogeneous point patterns are more difficult to interpret since these may be influenced by slightly different characteristics including sample sizes. More patterns there are near the mean size of $n = 400$, which is an effect of the intensity map used that is difficult to control. It is tempting to think that the results obtained for these cases may correspond to the zoomed region near the intersections of the $x$ and $y$-axes in Fig. 6a and b. However, closer inspection (not visualized) reveals that this is not the case and that the values given for $\alpha = 0.05$ are higher than for homogeneous point patterns in the same value range and that the order (i.e., from top to bottom) for the different $\alpha$ values is not the same. Next, we should take a look at the statistical power of Moran's $I$ in the context of random geometries, since the locations of the discussed critical values directly influence the degree of separability between the distributions in the null and the alternative hypothesis, which is relevant for concluding significance about Moran's $I$.

**Statistical power**

The statistical power of Moran's $I$ is affected by the presence of underlying point processes. As described in Experimental setup Section, the power of $I$ is not examined in this article in terms of the alternative hypothesis that would normally complement the null hypothesis in the context of conventional inference. Instead, alternative distributions simulated on the basis of point processes and representing several effect sizes $\rho$ are used. In this way, it is possible to see to what extent the conventional test statistic for $I$ is able to detect spatial effects when the normal approximation is mistakenly expected to be true in the null hypothesis, but the data actually come from spatial point processes. Fig. 7 provides four series of boxplots showing the distributional characteristics of all calculated power estimates both over the values of the spatial autoregressive parameters $\rho$ and for three significance levels $\alpha$. In Fig. 7a and b, it can be seen for the homogeneous cases that regardless of the type of point process, the power estimates calculated for each simulated point pattern are mostly acceptable after $\rho$ exceeds the 0.4 mark. Below $\rho = 0.4$, there are quite a few outliers and the whiskers also extend considerably downwards. For example, looking at $\rho = 0.3$ in Fig. 7a, about 50% of all estimates of power are arranged in a wide interval that extends downwards to 0 for both $\alpha = 0.01$ and $\alpha = 0.05$. These are relevant results, as Moran's $I$ is often used to detect spatial autocorrelation of the order of $\rho = 0.3$, for instance in regression scenarios. It is thus important that the estimator offers acceptable power, which is only true to a limited extent given the results obtained. Inhomogeneous processes generally show similar trends, but not the same outlier behavior. The latter could be due to the smaller range of sample sizes in these cases, as already mentioned above.

Looking more closely at the boxplots in Fig. 7 including the underlying power estimates, additional features of inference about Moran's $I$ in the presence of point processes become apparent. One such aspect is the finding that the average statistical power with Thomas processes

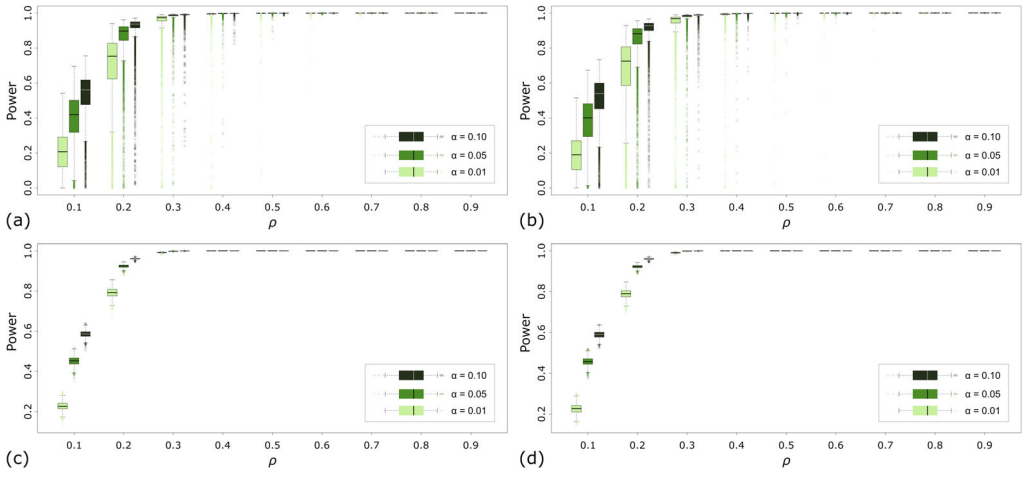


**Figure 7.** Boxplots of statistical powers for different spatial autoregressive parameters $\rho$. Homogeneous (a) Matérn and (b) Thomas processes. Inhomogeneous (c) Matérn and (d) Thomas processes.

is lower across all autocorrelation levels $\rho$ for the homogeneous point patterns studied. These differences are not very large (the largest difference is found for $\rho = 0.1$ and $\alpha = 0.05$), but are consistent. Considering the geometric properties of point dispersal expressed in both processes, one likely reason could be that the Matérn processes allow for a larger geometric variability. The latter are not locally constrained by an additional geometric point dispersal mechanism in the daughter processes. This is in contrast to the Thomas processes, which are locally governed by Gaussian distance decay and are based on the same parameters in each draw. This difference suggests that, especially for alternative hypotheses associated with low $\rho$ values, the distributions of $I$ may not be so different from the null hypothesis restricted to a particular Thomas outcome derived from the same geometric principle as is the case for Matérn processes. This difference between the two process types is not observed in the inhomogeneous point patterns. A possible reason for this could be the faster filling up with points in rather restricted spaces imposed by the intensity grid. Another feature of statistical power in the given context is the consideration of the dispersion behavior of the estimates. Using $\alpha = 0.1$, the variance is greatest when $\rho = 0.1$. The statistical power in this case is in the middle range, but there are many outliers, most of which range downwards. On the other hand, if we consider the stricter inference criteria at $\alpha = 0.01$, the highest uncertainty is associated with the powers estimated for $\rho = 0.2$, which corresponds to a deviation from the mean variance of 2.4 times the standard deviation of all the variances determined. For weaker spatial structuring with $\alpha = 0.1$, the statistical power is so low that not many downward outliers are possible. For $\rho = 0.3$, the variance is still strongly above average, but then moves slowly toward the mean. These results are again of importance for applications, since quite high uncertainties are attached to practically relevant ranges of autocorrelation values.

Statistical power is closely related to sample size. Fig. 8 shows a series of power estimates plotted against the sizes $n$ of the corresponding point patterns. Only plots for Matérn processes are shown, as the Thomas-based point patterns show similar behavior as a function of sample size. Fig. 8a–d show the relationships between statistical power and $n$ for $\rho = 0.1$, 0.2, 0.3, and 0.4. Beyond $\rho = 0.4$, the plots would not be very informative, as can also be inferred from the

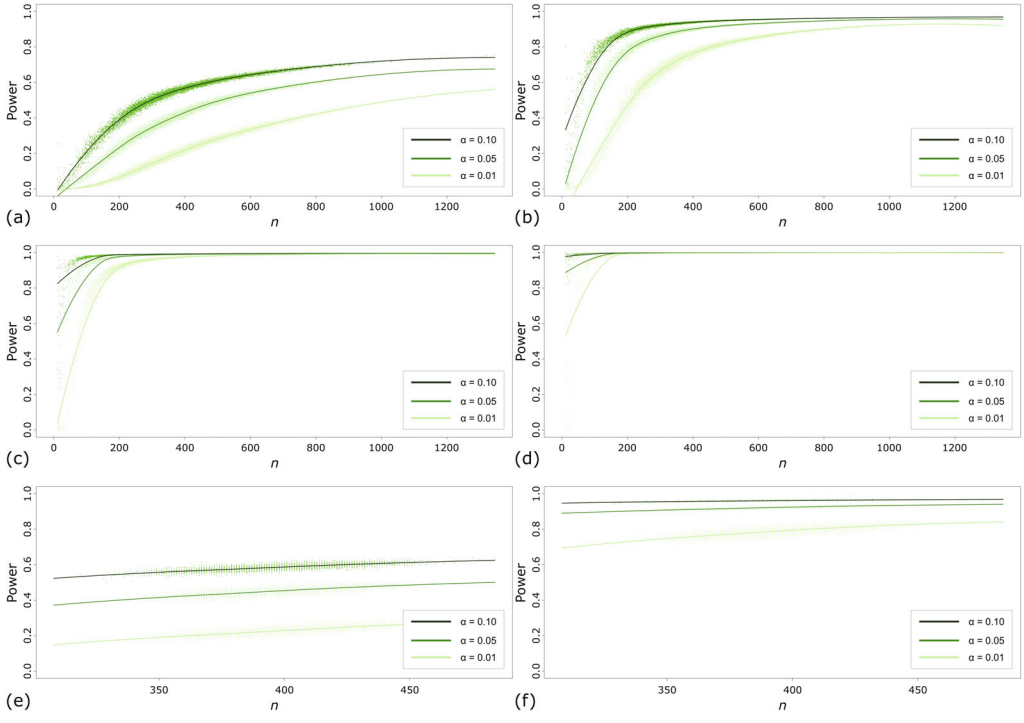


**Figure 8.** Statistical power of Moran's $I$ under different alternative hypotheses and plotted against point pattern sizes $n$. Homogeneous Matérn cluster processes with (a) $\rho = 0.1$, (b) $\rho = 0.2$, (c) $\rho = 0.3$, and (d) $\rho = 0.4$; and for inhomogeneous Matérn processes with (e) $\rho = 0.1$, and (f) $\rho = 0.2$. The points in the background represent statistical power for each simulated point pattern and three significance levels (separated in three tones of green). The green trend lines represent LOESS regressions.

boxplots in Fig. 7. Looking at Fig. 8a for $\rho = 0.1$, it is noticeable that the trend line for the strictest significance level $\alpha = 0.01$ deviates from the other trend lines shown. While the trends for $\alpha = 0.05$ and 0.1 initially rise rapidly before flattening out to a progressively lower slope rate, the trend for $\alpha = 0.01$ takes on a sigmoidal shape. As a result, the curve remains close to 0 almost until $n = 200$. Thus, it can be said with some confidence that the detection of weak spatial patterns, assuming a low error rate, is very difficult when the data generating process induces geometric randomness. In all cases, the trend lines tend toward plateaus of statistical power. These plateaus are quite high, reaching maximum values of 0.57 ($\alpha = 0.01$), 0.69 ($\alpha = 0.05$), and 0.76 ($\alpha = 0.1$), but these values are only reached when $n$ is very large. The inflection points after which the curves flatten out are in the range between 200 and 400 points. As the degree of spatial autocorrelation increases (Fig. 8b–d), these inflection points shift to the left, that is, toward smaller point pattern sizes. But even in these cases they are still between $n = 100$ and 200, which, as has been pointed out several times, can sometimes be high hurdles for lots of applications.

To better understand the outlined relationship between sample size and power, it is necessary to relate it to the experimental setup of this study. As described in Simulation of homogeneous point patterns Section, the simulated point patterns in this article are generated using a finite window of fixed size. This arrangement results in the growth of the point pattern realizations

following an infill pattern resulting in increasing point density as $n$ grows with this effect being stronger for the inhomogeneous patterns due to the additional spatial constraints added by the intensity map. This type of strategy is representative of certain real-world processes including the arrangement of trees in forests, scenarios related to crowding, and also the analysis of tweets in different urban environments. Instead of the infill strategy, however, an increasing-domain strategy would also have been possible, in which the density of the points is controlled and only the mere number of points increases by expanding the boundary of the analysis window (cf. Cressie 1993, p. 350 ff.). This is important for the results obtained in that the inflection points reported in the previous paragraph, which are quite distinct, may be characterized by a complex interplay of degree of spatial autocorrelation, point pattern size $n$, and point density. This should be investigated in more detail in follow-up studies, but should be taken into account when interpreting the results described in this subsection and beyond. Another matter is the interpretation of the results for the inhomogeneous cases shown in Fig. 8e and f for $\rho = 0.1$ and 0.2. Given the limited range of point pattern sizes obtained from the intensity grid, it is only possible to describe the middle range of $n$. Zooming into the same range of $n$ for the homogeneous cases[7] shows that the inhomogeneous cases behave very similarly to the homogeneous point processes. However, it remains unclear whether and how the tails of the statistical power would behave with respect to $n$, which is again left to future research at this point.

## Conclusions and future research

This article investigates the performance of the established and widely applied inference framework of Cliff and Ord (1981) for global Moran's $I$ in cases where the analyzed data are drawn from point processes, that is, whose spatial index is subject to geometric randomness. Two types of Poisson cluster processes, namely Matérn and Thomas processes, were investigated for both homogeneous and inhomogeneous scenarios. For each of these cases, 10,000 random point patterns were generated and assigned normally distributed random values. The patterns thus generated were then used to analyze three important characteristics using Monte Carlo simulation. One analyzed property is the deviation of parametric normal approximations of the null distribution of $I$ from the empirically generated null distributions taking into account point process induced randomness. Furthermore, the differences between conventional (with fixed spatial index) and empirically estimated critical values (with stochastic spatial index) as used for rejecting the null hypothesis of no spatial autocorrelation were determined. The third investigation concerns the statistical power of Moran's $I$ when the actual distribution in the alternative hypothesis is expected to contain geometric randomness due to point processes. From the results, it can be concluded that point processes affect the performance of the estimator for $I$. The empirical null distributions with stochastic spatial indexes deviate more from the theoretical null approximation than those obtained under conventional conditions. The critical values assessed indicate that the traditional inference framework may be too rigid or too liberal, depending on the sample size. Also, the statistical power is often quite low (for small sample sizes and weak spatial processes), but can reach acceptable values as the sample size increases. A number of conclusions can be drawn from the study conducted.

One conclusion that can be drawn from the present simulation study is that the geometric constraints added by point processes have an impact on inferences about global Moran's $I$. Geometric constraints is thereby not to be understood in absolute terms. Compared to the traditional assumptions for estimators of spatial autocorrelation, both point processes studied

allow for more rather than less variation in the spatial weights. What is meant here is that the Thomas process is more geometrically constraining with respect to small distances than its Matérn process counterpart. This difference seems to lead to consistent differences in terms of global Moran's $I$. This conclusion is supported by a number of results: The statistical power is consistently lower under the influence of Thomas processes than for their Matérn counterparts. Analogously, the deviations of the empirical null distributions from the null approximation are also larger for geometric configurations derived from Thomas processes. The main difference between Thomas and Matérn processes is the geometric point dispersal mechanism in the daughter processes. Matérn processes do not geometrically constrain the local offspring points, apart from an upper distance limit. In contrast, the Thomas process adds a geometric constraint that translates into a Gaussian distance decay. The local random placement associated with the Matérn process thus allows for a wider range of possible point configurations, resulting in a wider span of spatial weights than the more constrained bell-shaped clusters that form under the Thomas process. Translating the notion of a stochastic spatial index to the idea of nonfixed, stochastic spatial weights, and considering the close relationship between these with the shape and possible range of the distribution of Moran's $I$ (de Jong, Sprenger, and van Veen 1984), the results obtained support the conclusion that it is probably the geometric random mechanism that determines the differences between the two types of processes studied. Influences on the estimation of spatial autocorrelation by variation of characteristics of underlying geometries have also been found in previous studies. For example, Griffith and Arbia (2010) and Griffith (2006) found a tendency toward negative spatial autocorrelation in the case of geographic competition for resources, whereas negative spatial autocorrelation is otherwise rare and often indicative of measurement error or other technical problems. The results obtained here extend these findings to the cases of the two point processes considered. The differences obtained are not very large, as the simulations are designed to obtain similar cluster characteristics in terms of size and scale. However, future studies may investigate the differences while varying these characteristics as well, in order to determine their role in the inference about Moran's $I$.

A second conclusion concerns the statistical power of Moran's $I$. In a study of different types of spatial structures, Bivand, Müller, and Reder (2009) explored the power properties of global and local Moran's $I$, including hypothesis N, which is also considered in the present article. The types of spatial structures examined in that earlier study differ, but considering them all together allows comparison with the results obtained here. Global Moran's $I$ is studied only for small spatial configurations. For a regular $5 \times 5$ lattice, Bivand, Müller, and Reder (2009) find that the power increases rapidly with the spatial autoregressive coefficients $\rho$. In contrast, for an irregular structure consisting of eight points, it is observed that the statistical power of $I$ remains low. The cases presented in this article are larger in size but have the same irregular character as the latter case studied by Bivand, Müller, and Reder (2009). Considering only the subset of very small simulated point patterns with $n \leq 20$, one also finds that the statistical power for small values of $\rho$ is low, even lower than the values given in the article by Bivand, Müller, and Reder (2009). The difference, however, is that the statistical power increases to high values close to 1 even for small point patterns when $\rho$ is large. What is different between the present study and the one referred to is that here the distributions generated under the alternative hypotheses allow for a greater degree of variability due to the point processes. The range of possible spatial weights increases and so does the range of $I$. Both the regular grid (which could occur as a result of the Matérn process, albeit with very low probability) and the irregular lattice studied by Bivand, Müller, and Reder (2009) are possible special cases of the simulations studied here, analogous to perhaps

picking two extreme cases out of all the simulations and fixing on their geometries. In a recent study, Luo, Griffith, and Wu (2019) examined the statistical power of Morans *I* not only in the context of small samples, but also when applied to large data sets such as tweets, which are often referred to in this article (the authors use a remotely sensed image). While their results for small samples and using different distributional and spatial connectivity regimes confirm previous findings, the authors show that the statistical power for very large data sets approaches one. In the study presented here, these previous results are complemented by a stochastic geometry perspective. The results obtained here thus extend the findings of the previous study by showing that the statistical power of *I* for weak spatial processes seems generally low for small to medium sized samples, even if the range of considered possible random geometries is extended.

The results obtained and the conclusions drawn are subject to a number of limitations that give rise to possible follow-up investigations. Matérn and Thomas processes are only two possible types of cluster processes. There are other, more complex processes such as Gibbs processes that model proactive point interaction (Illian et al. 2008, p. 137 ff.). It would be interesting to see how the results obtained in this article would differ systematically under different types of point processes. This should also include possible combinations of them, as well as "dilutions" of "pure" model results, for example by using thinning techniques or random infusions of points. Both can occur in practical applications, as point process models often do not occur in their pristine form. A systematic understanding of these scenarios would be helpful in assessing the quality of the corresponding spatial autocorrelation evaluations. Furthermore, the Matérn and Thomas point processes tested in this article are relatively similar in design in terms of scale, number of points, and the attribute values assigned to them. This choice was made for reasons of comparability, but it also means that some of the effects reported in this article might be more pronounced if the parameters mentioned are varied. In follow-up studies, the characteristics mentioned should be examined individually and the respective interactions worked out.

Finally, the vigilant reader will have noticed that local Moran's *I* (Anselin 1995) has been omitted from this article for reasons of space. Indeed, the local counterpart of *I* implies a number of additional inference systems attached to various kinds of null hypotheses (for an overview, see Sauer et al. 2021). Certainly, the investigation of these null hypotheses would also be relevant and interesting, but it should be carried out in a separate investigation in order to give it due prominence. Similar arguments can be made for the case of inhomogeneous point processes. The results obtained here show that the case of inhomogeneous point patterns deserves a separate investigation in order to give full attention to the specific properties of the associated process types and their interaction with Moran's *I*. Overall, there is scope for much follow-up research, and this article is hopefully informative both for empirical interpretations of spatial autocorrelation and for methodological follow-up work.

## Acknowledgements

My gratitude goes to Florian Klopfer (TU Dortmund University), Natalie Rothwell (University of Warwick), and Andra Sonea (University of Oxford) for carefully proofreading a draft of this article before submission. I am further grateful to Franz-Benjamin Mocnik (University of Twente) as well as Liudmila Slivinskaya, Víctor Cobs-Muñoz, Ibrahim Mubiru, and Mathias Schaefer (all TU Dortmund University) for their constructive comments on the figures. Further, I would like to thank the anonymous reviewers for their helpful comments, which have further

strengthened the material presented in this article. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. Open Access funding has been enabled and organized by Projekt DEAL.

## Note

1 See Sui (2004), Miller (2004), and Tobler (2004) for a debate on the degree of universality and the scientific-theoretical status of Tobler's First Law of Geography.
2 This monograph continues a previous version of their book: Cliff and Ord (1973). See Griffith (2009) and the corresponding special issue for an overview of the influence that Andrew Cliff's and Keith Ord's books have had (and continue to have) on the field of spatial analysis.
3 I intentionally avoid the stronger notion of "independence." The latter has wider implications and cannot always be guaranteed.
4 The notation $\mathbb{I}$ is used for the identity matrix instead of the usual notation $I$ to avoid confusion with the symbol used for Moran's $I$.
5 Both hypotheses N and R have been tested but only the results for hypothesis N are reported in the figures. The reason is that the results are very similar for both assumptions.
6 Normality was investigated with Shapiro–Wilk tests. For both types of point processes and for hypotheses N and R, the respective 10,000 KL divergences were each randomly divided 1,000 times into subsets of 250 to avoid the Shapiro–Wilk test being too sensitive due to large $n$. The normality hypothesis could not be rejected in well over 90% of all cases.
7 These zoomed-in sections are not graphically illustrated in this article.

## References

Anselin, L. (1995). "Local Indicators of Spatial Association – LISA." *Geographical Analysis* 27, 93–115.
Anselin, L. (2001). "Spatial Econometrics." In *A Companion to Theoretical Econometrics*, 310–30, edited by B. H. Baltagi. Hoboken, NJ: Wiley.
Assuncao, R. M., and E. A. Reis. (1999). "A New Proposal to Adjust Moran's $I$ for Population Density." *Statistics in Medicine* 18, 2147–62.
Auchincloss, A. H., S. Y. Gebreab, C. Mair, and A. V. Diez Roux. (2012). "A Review of Spatial Methods in Epidemiology, 2000–2010." *Annual Review of Public Health* 33, 107–22.
Bavaud, F. (2014). "Spatial Weights: Constructing Weight-Compatible Exchange Matrices from Proximity Matrices." In *Proceedings of the 8th International Conference on Geographic Information Science (GIScience 2014)*, 81–96, edited by M. Duckham, E. Pebesma, K. Stewart, and A. U. Frank. New York, NY: Springer.
Belov, D. I., and R. D. Armstrong. (2011). "Distributions of the Kullback–Leibler Divergence with Applications." *British Journal of Mathematical and Statistical Psychology* 64, 291–309.
Bivand, R., W. G. Müller, and M. Reder. (2009). "Power Calculations for Global and Local Moran's $I$." *Computational Statistics & Data Analysis* 53, 2859–72.
Blei, D. M., A. Y. Ng, and M. I. Jordan. (2003). "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3, 993–1022.
Bucher, D., H. Martin, D. Jonietz, M. Raubal, and R. Westerholt. (2020). "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*. Wadern, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik.
Burra, T., M. Jerrett, R. T. Burnett, and M. Anderson. (2002). "Conceptual and Practical Issues in the Detection of Local Disease Clusters: A Study of Mortality in Hamilton, Ontario." *The Canadian Geographer/Le Géographe Canadien* 46, 160–71.
Cliff, A. D., and J. K. Ord. (1972). "Testing for Spatial Autocorrelation among Regression Residuals." *Geographical Analysis* 4, 267–84.
Cliff, A. D., and J. K. Ord. (1973). *Spatial Autocorrelation*. London, UK: Pion.
Cliff, A. D., and J. K. Ord. (1981). *Spatial Processes: Models & Applications*. London, UK: Pion.

Cohen, J. (1992). "Statistical Power Analysis." *Current Directions in Psychological Science* 1, 98–101.
Cressie, N. (1993). *Statistics for Spatial Data*. Hoboken, NJ: Wiley.
de Andrade, S. C., J. P. de Albuquerque, C. Restrepo-Estrada, R. Westerholt, C. A. M. Rodriguez, E. M. Mendiondo, and A. C. B. Delbem. (2022). "The Effect of Intra-Urban Mobility Flows on the Spatial Heterogeneity of Social Media Activity: Investigating the Response to Rainfall Events." *International Journal of Geographical Information Science* 36, 1140–65.
de Jong, P., C. Sprenger, and F. van Veen. (1984). "On Extreme Values of Moran's *I* and Geary's *c*." *Geographical Analysis* 16, 17–24.
Diniz-Filho, J. A. F., L. M. Bini, and B. A. Hawkins. (2003). "Spatial Autocorrelation and Red Herrings in Geographical Ecology." *Global Ecology and Biogeography* 12, 53–64.
Dray, S. (2011). "A New Perspective about Moran's Coefficient: Spatial Autocorrelation as a Linear Regression Problem." *Geographical Analysis* 43, 127–41.
Eberth, J. M., M. R. Kramer, E. Delmelle, and R. S. Kirby. (2021). "What is the Place for Space in Epidemiology?" *Annals of Epidemiology* 64, 41–6.
Fedriani, J. M., T. Wiegand, and M. Delibes. (2010). "Spatial Pattern of Adult Trees and the Mammal-Generated Seed Rain in the Iberian Pear." *Ecography* 33, 545–55.
Garnett, R., and R. Stewart. (2015). "Comparison of GPS Units and Mobile Apple GPS Capabilities in an Urban Landscape." *Cartography and Geographic Information Science* 42, 1–8.
Getis, A. (2007). "Reflections on Spatial Autocorrelation." *Regional Science and Urban Economics* 37, 491–6.
Getis, A. (2009). "Spatial Weights Matrices." *Geographical Analysis* 41, 404–10.
Getis, A. (2010). "Spatial Autocorrelation." In *Handbook of Applied Spatial Analysis*, 255–78, edited by M. M. Fischer and A. Getis. Heidelberg, Germany: Springer.
Getis, A., and J. Aldstadt. (2004). "Constructing the Spatial Weights Matrix Using a Local Statistic." *Geographical Analysis* 36, 90–104.
Goodchild, M. F. (2010). "Twenty Years of Progress: GIScience in 2010." *Journal of Spatial Information Science* 1, 3–20.
Griffith, D. A. (2006). "Hidden Negative Spatial Autocorrelation." *Journal of Geographical Systems* 8, 335–55.
Griffith, D. A. (2009). "Celebrating 40 Years of Scientific Impacts by Cliff and Ord." *Geographical Analysis* 41, 343–6.
Griffith, D. A. (2010). "The Moran Coefficient for Non-Normal Data." *Journal of Statistical Planning and Inference* 140, 2980–90.
Griffith, D. A. (2018). "Uncertainty and Context in Geography and GIScience: Reflections on Spatial Autocorrelation, Spatial Sampling, and Health Data." *Annals of the American Association of Geographers* 108, 1499–505.
Griffith, D. A., and G. Arbia. (2010). "Detecting Negative Spatial Autocorrelation in Georeferenced Random Variables." *International Journal of Geographical Information Science* 24, 417–37.
Griffith, D. A., Y. Chun, and M. Lee. (2016). "Locational Error Impacts on Local Spatial Autocorrelation Indices: A Syracuse Soil Sample Pb-Level Data Case Study." In *Proceedings of Spatial Accuracy 2016*, 136–43, edited by J.-S. Bailly, D. A. Griffith, and D. Josselin. Montpellier, France: HAL Open Science.
Griffith, D. A., M. Millones, M. Vincent, D. L. Johnson, and A. Hunt. (2007). "Impacts of Positional Error on Spatial Regression Analysis: A Case Study of Address Locations in Syracuse, New York." *Transactions in GIS* 11, 655–79.
Haining, R. P. (2009). "Spatial Autocorrelation and the Quantitative Revolution." *Geographical Analysis* 41, 364–74.
Illian, J., A. Penttinen, H. Stoyan, and D. Stoyan. (2008). *Statistical Analysis and Modelling of Spatial Point Patterns*. Chichester, UK: Wiley.
Jackson, M. C., L. Huang, Q. Xie, and R. C. Tiwari. (2010). "A Modified Version of Moran's *I*." *International Journal of Health Geographics* 9, 33.
Jacquez, G. M. (1996). "Disease Cluster Statistics for Imprecise Space-Time Locations." *Statistics in Medicine* 15, 873–85.

Jacquez, G. M. (1999). "Spatial Statistics When Locations are Uncertain." *Geographic Information Sciences* 5, 77–87.

Jacquez, G. M., and J. A. Jacquez. (1999). "Disease Clustering for Uncertain Locations." In *Disease Mapping and Risk Assessment for Public Health*, 151–68, edited by A. B. Lawson, A. Biggeri, D. Böhning, E. Lesaffre, J.-F. Viel, and R. Bertollini. Chichester, UK: Wiley.

Jacquez, G. M., and R. Rommel. (2009). "Local Indicators of Geocoding Accuracy (LIGA): Theory and Application." *International Journal of Health Geographics* 8, 60.

Jung, P. H., J.-C. Thill, and M. Issel. (2019). "Spatial Autocorrelation Statistics of Areal Prevalence Rates Under High Uncertainty in Denominator Data." *Geographical Analysis* 51, 354–80.

Kirby, R. S., E. Delmelle, and J. M. Eberth. (2017). "Advances in Spatial Epidemiology and Geographic Information Systems." *Annals of Epidemiology* 27, 1–9.

Kitchin, R., T. P. Lauriault, and M. W. Wilson. (2017). *Understanding Spatial Media*. London, UK: SAGE.

Kullback, S., and R. A. Leibler. (1951). "On Information and Sufficiency." *The Annals of Mathematical Statistics* 22, 79–86.

Lawson, A. B., and D. G. Denison. (2002). *Spatial Cluster Modelling*. Boca Raton, FL: CRC Press.

Lee, M., Y. Chun, and D. A. Griffith. (2018). "Error Propagation in Spatial Modeling of Public Health Data: A Simulation Approach Using Pediatric Blood Lead Level Data for Syracuse, New York." *Environmental Geochemistry and Health* 40, 667–81.

Legendre, P. (1993). "Spatial Autocorrelation: Trouble or New Paradigm?" *Ecology* 74, 1659–73.

Lemons, D. S. (2003). *An Introduction to Stochastic Processes in Physics*. Baltimore, MD: The Johns Hopkins University Press.

Luo, Q., D. A. Griffith, and H. Wu. (2019). "Spatial Autocorrelation for Massive Spatial Data: Verification of Efficiency and Statistical Power Asymptotics." *Journal of Geographical Systems* 21, 237–69.

Matérn, B. (1960). *Spatial Variation: Stochastic Models and Their Application to Some Problems in Forest Survey and Other Sampling Investigations*. Stockholm, Sweden: Meddelanden från Statens Skogsförsöksanstalt.

Merry, K., and P. Bettinger. (2019). "Smartphone GPS Accuracy Study in an Urban Environment." *PLoS One* 14, e0219890.

Miller, H. J. (2004). "Tobler's First Law and Spatial Analysis." *Annals of the Association of American Geographers* 94, 284–9.

Munkres, J. R. (2014). *Topology*, Vol 2. Harlow, UK: Pearson.

Oden, N. (1995). "Adjusting Moran's *I* for Population Density." *Statistics in Medicine* 14, 17–26.

Ord, J. K., and A. Getis. (2012). "Local Spatial Heteroscedasticity (LOSH)." *The Annals of Regional Science* 48, 529–39.

Páez, A., and D. M. Scott. (2004). "Spatial Statistics for Urban Analysis: A Review of Techniques with Examples." *GeoJournal* 61, 53–67.

Pérez-Cruz, F. (2008). "Kullback-Leibler Divergence Estimation of Continuous Distributions." In *Proceedings of the 2008 IEEE International Symposium on Information Theory*, 1666–70. Toronto, Canada: IEEE.

Quesnot, T., and S. Roche. (2015). "Platial or Locational Data? Toward the Characterization of Social Location Sharing." In *Proceedings of the 48th Hawaii International Conference on System Sciences*, 1973–82. Kauai, HI: IEEE.

Resch, B., F. Usländer, and C. Havas. (2018). "Combining Machine-Learning Topic Models and Spatiotemporal Analysis of Social Media Data for Disaster Footprint and Damage Assessment." *Cartography and Geographic Information Science* 45, 362–76.

Ripley, B. D. (1981). *Spatial Statistics*. Hoboken, NJ: Wiley & Sons.

Saker, M. (2017). "Foursquare and Identity: Checking-in and Presenting the Self Through Location." *New Media & Society* 19, 934–49.

Sauer, J., T. Oshan, S. Rey, and L. J. Wolf. (2021). "The Importance of Null Hypotheses: Understanding Differences in Local Moran's $I_i$ Under Heteroskedasticity." *Geographical Analysis*.

Shimatani, K. (2002). "Point Processes for Fine-Scale Spatial Genetics and Molecular Ecology." *Biometrical Journal* 44, 325–52.

Shimatani, K., and M. Takahashi. (2003). "On Methods of Spatial Analysis for Genotyped Individuals." *Heredity* 91, 173–80.

Stefanidis, A., A. Crooks, and J. Radzikowski. (2013). "Harvesting Ambient Geospatial Information from Social Media Feeds." *GeoJournal* 78, 319–38.

Steiger, E., B. Resch, and A. Zipf. (2016). "Exploration of Spatiotemporal and Semantic Clusters of Twitter Data Using Unsupervised Neural Networks." *International Journal of Geographical Information Science* 30, 1694–716.

Steiger, E., R. Westerholt, B. Resch, and A. Zipf. (2015). "Twitter as an Indicator for Whereabouts of People? Correlating Twitter with UK Census Data." *Computers, Environment and Urban Systems* 54, 255–65.

Sui, D. Z. (2004). "Tobler's First Law of Geography: A Big Idea for a Small World?" *Annals of the Association of American Geographers* 94, 269–77.

Tanaka, U., Y. Ogata, and D. Stoyan. (2008). "Parameter Estimation and Model Selection for Neyman-Scott Point Processes." *Biometrical Journal: Journal of Mathematical Methods in Biosciences* 50, 43–57.

Thomas, M. (1949). "A Generalization of Poisson's Binomial Limit for Use in Ecology." *Biometrika* 36, 18–25.

Tiefelsdorf, M. (2002). "The Saddlepoint Approximation of Moran's *I*'s and Local Moran's *Ii*'s Reference Distributions and Their Numerical Evaluation." *Geographical Analysis* 34, 187–206.

Tiefelsdorf, M., and B. Boots. (1995). "The Exact Distribution of Moran's *I*." *Environment and Planning A: Economy and Space* 27, 985–99.

Tiefelsdorf, M., D. A. Griffith, and B. Boots. (1999). "A Variance-Stabilizing Coding Scheme for Spatial Link Matrices." *Environment and Planning A: Economy and Space* 31, 165–80.

Tobler, W. R. (1970). "A Computer Movie Simulating Urban Growth in the Detroit Region." *Economic Geography* 46, 234–40.

Tobler, W. R. (2004). "On the First Law of Geography: A Reply." *Annals of the Association of American Geographers* 94, 304–10.

Townsley, M. (2009). "Spatial Autocorrelation and Impacts on Criminology." *Geographical Analysis* 41, 452–61.

Valcu, M., and B. Kempenaers. (2010). "Is Spatial Autocorrelation an Intrinsic Property of Territory Size?" *Oecologia* 162, 609–15.

Vallina-Rodriguez, N., J. Crowcroft, A. Finamore, Y. Grunenberger, and K. Papagiannaki. (2013). "When Assistance Becomes Dependence: Characterizing the Costs and Inefficiencies of A-GPS." *Mobile Computing and Communications Review* 17, 3–14.

van Zanten, B. T., D. B. Van Berkel, R. K. Meentemeyer, J. W. Smith, K. F. Tieskens, and P. H. Verburg. (2016). "Continental-Scale Quantification of Landscape Values Using Social Media Data." *Proceedings of the National Academy of Sciences* 113, 12974–9.

Wagner, C., M. Strohmaier, A. Olteanu, E. Kıcıman, N. Contractor, and T. Eliassi-Rad. (2021). "Measuring Algorithmically Infused Societies." *Nature* 595, 197–204.

Waldhör, T. (1996). "The Spatial Autocorrelation Coefficient Moran's *I* Under Heteroscedasticity." *Statistics in Medicine* 15, 887–92.

Walter, S. (1992a). "The Analysis of Regional Patterns in Health Data: I. Distributional Considerations." *American Journal of Epidemiology* 136, 730–41.

Walter, S. (1992b). "The Analysis of Regional Patterns in Health Data: II. The Power to Detect Environmental Effects." *American Journal of Epidemiology* 136, 742–59.

Wang, Y., and Q. Zhu. (2016). "Modeling and Analysis of Small Cells Based on Clustered Stochastic Geometry." *IEEE Communications Letters* 21, 576–9.

Westerholt, R. (2018). "The Impact of the Spatial Superimposition of Point Based Statistical Configurations on Assessing Spatial Autocorrelation." In *Proceedings of the AGILE'2018 International Conference on Geographic Information Science*, 63, edited by A. Mansourian, P. Pilesjö, L. Harrie, and R. van Lammeren. Lund, Sweden: Association of Geographic Information Laboratories in Europe.

Westerholt, R. (2019). "Methodological Aspects of the Spatial Analysis of Geosocial Media Feeds: From Locations Towards Places." *gis.Science: Die Zeitschrift für Geoinformatik* 31, 65–76.

Westerholt, R. (2021). "Emphasising Spatial Structure in Geosocial Media Data Using Spatial Amplifier Filtering." *Environment and Planning B: Urban Analytics and City Science* 48, 2842–61.

Westerholt, R., B. Resch, and A. Zipf. (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, 868–87.

Westerholt, R., B. Resch, F.-B. Mocnik, and D. Hoffmeister. (2018). "A Statistical Test on the Local Effects of Spatially Structured Variance." *International Journal of Geographical Information Science* 32, 571–600.

Westerholt, R., E. Steiger, B. Resch, and A. Zipf. (2016a). "Abundant Topological Outliers in Social Media Data and their Effect on Spatial Analysis." *PLoS One* 11, e0162360.

Westerholt, R., E. Steiger, B. Resch, and A. Zipf (2016b). Twitter Sample. https://doi.org/10.1371/journal.pone.0162360.s001.

Xu, M., C.-L. Mei, and N. Yan. (2014). "A Note on the Null Distribution of the Local Spatial Heteroscedasticity (LOSH) Statistic." *The Annals of Regional Science* 52, 697–710.

Zandbergen, P. A., and S. J. Barbeau. (2011). "Positional Accuracy of Assisted GPS Data from High-Sensitivity GPS-Enabled Mobile Phones." *The Journal of Navigation* 64, 381–99.

Zhang, T., and G. Lin. (2016). "On Moran's *I* Coefficient Under Heterogeneity." *Computational Statistics & Data Analysis* 95, 83–94.