Exploratory Statistical Analysis of Spatial Structures in Urban Datasets

René Westerholt

1. Introduction

Metropolitan phenomena are often characterised by spatial structuring and arrangement. The term 'urban', understood as a "fabric in which […] sociocultural and political-economic relations […] are enmeshed" (Brenner/Schmid 2014, 751) and where collective action "comes from a plurality of sources" (Castells 1985, 3), implies a complex assemblage of social and physical processes that co-occur in nearby locations. This simultaneity of processes not found in the same way in rural areas makes urban phenomena interesting objects of study. Simultaneity and nearness in space and time are hence reflected in different forms throughout the entire volume at hand. Urban epidemiological incidents, as discussed in the chapter by Moebus, gain traction when spatial conditions facilitate contagion. Urban narratives embedded in the confluence of different ideas and conceptualisations, as discussed in the chapters by Sattler and Parr, have a spatial structure, either explicit (through the use of place names) or implicit (through spatial language). Methodological workflows in metropolitan research often reflect the spatial nature of urban phenomena through the presentation of maps, the explicit consideration of spatiality, and through the attribution of analytical results to different places. The elaborated literal 'throwntogetherness' (Massey 2005, 140) of metropolitan areas, combined with contextual geographical variation, almost inevitably leads to identifiable spatial structures, both observed in situ and reflected in data.

Understanding systematic spatial structuring is the core of statistical spatial analysis. In the context of geographic information systems (GISs), the term 'spatial analysis' is often used as a catch-all term for any kind of methods that involve space. In statistical analysis, however, the term has a stricter and narrower meaning and refers to the identification and characterisation of (possibly non-random) spatial structures (Fischer/Getis 2010). The core idea of this subfield in the nexus of geography, (analytical) cartography, statistics, economics (through regional science and econometrics), and (more recently) computer science (Brachman 2020; Singleton/Arribas-Bel 2021) is the application of a statistical epistemology to the analysis of spatial arrangements. The latter implies that

spatial analysis falls within the realm of the sciences, as it attempts in a nomothetic way to derive law-like statements from observed recurring regularities. The focus of spatial analysis is on the 'in-between', that is, on the interconnectedness of spatial units, which can stand for places, cities, regions, or any other kind of spatial entity (Fischer 2005). The perspective taken is a holistic one. The whole arrangement, manifested in map patterns, is considered together and at once rather than looking at individual places in isolation. The fundamental task at the heart of spatial analysis, thus, is to explore systematic geographical arrangements.

This chapter provides an accessible introduction to statistical spatial analysis aimed at interdisciplinary metropolitan researchers. It combines an applied approach with a rigorous, sufficiently technical consideration. This way, the reader is empowered in terms of establishing a solid basis for further engagement with spatial-statistical analysis. The following Section 2 is devoted to application examples from metropolitan research motivating the methodology introduced in the subsequent sections. Section 3 then discusses a number of concepts and presumptions on which spatial analysis is based. These are important for understanding the principles and contexts of application in which spatial analysis can and may be used. In a second step, in Section 4, Moran's I, a widely used measure of spatial autocorrelation, is explained in detail. This method has been researched for decades and serves here as a prototypical example for the nature and application of methods in spatial analysis. Section 5 introduces both additional established and novel techniques from the spatial statistical toolkit that are useful in different, selected application contexts. These include measures and tests for spatial configurations in numerical, categorical, vector, and multivariate variables.

2. Urban Applications

The methods introduced in this chapter are widely being applied in the context of metropolitan research. Three popular areas of application that have recently gained interest are exemplified below.

2.1 Census Analyses

Censuses have long been a rich source of information for metropolitan research. Among other things, they enable the study of population dynamics, socioeconomic characteristics, and demographic features. Spatial statistical methods play a central role in many census analyses. Using spatial regression models and autocorrelation methods, Manley et al. (2006) identify different scales and associated processes in British census variables. Spatial statistics here enable the distinction of finely graded sub-regions. In the same vein, census studies have been conducted with spatial statistics on topics such as employment (Martín-Román et al. 2020; Fingleton et al. 2020), housing (Barreca et al. 2018; Lin et al. 2014), and socioeconomic disadvantage (Andrews et al. 2020; Cebrecos et al. 2018), to name but a few. A young but active area of research is geodemography, the explicitly spatial study of demographic characteristics. Largely accelerated by increased data accessibility, studies have been conducted on the geography of surnames (van Dijk/

Longley 2020a; 2020b; Kandt/Longley 2018), small area characterisations (Yazgi Walsh et al. 2021; Singleton et al. 2020), and links between geodemographics and other domains (Kim et al. 2021; Liu/Cheng 2020). Although much spatial statistical work has been done in the field of census analysis, methodological concerns continue to be raised. For example, using the American Community Survey, Jung et al. (2019a) criticise the inappropriate use of spatial statistics, particularly in the context of rates, an issue discussed in Section 5.1. Yet, explicitly spatial techniques are increasingly being used but are still not as widespread as they probably should be, given the spatial nature of census data.

2.2 Urban Infrastructures

The analysis of urban infrastructures and their use is also a research area that makes frequent use of spatial statistics. In particular, mobility and transport infrastructure research make extensive use of spatial autocorrelation and regression, which are used to assess the spatial configuration of deployed infrastructures including their possible impacts on economic, social, and other characteristics (Wang et al. 2020; Potoglou et al. 2019; X. Gao et al. 2019). In addition, the utilisation of transport infrastructure, and human mobility in general, have been extensively researched using spatial analysis methods (Y. Gao et al. 2019; Boss et al. 2018; Blazquez et al. 2018; Steiger et al. 2016). Similarly, green infrastructures, that is, those concerning the strategic deployment of urban greenery, have also been studied in terms of spatial statistics. Examples include studies on the influence of urban green on residential property values (Mei et al. 2018; Conway et al. 2010), accessibility of green spaces (Pearsall/Eller 2020; Dai 2011), and associations of urban green with mental well-being (Houlden et al. 2019; 2018). Also in focus, but less frequently studied with the methods presented in this chapter, are electricity and water supply infrastructures. For example, Ceci et al. (2019) exploit spatial autocorrelation to better understand the spatial properties of a network of photovoltaic plants. On a larger scale, Hong et al. (2020) use spatial regression models to investigate regional spillover effects in energy consumption. In terms of water supply, Zamenian et al. (2017) use hotspot statistics to reveal spatial clusters in certain characteristics of water pipes, while Abokifa and Sela (2019) disclose spatial patterns in pipe failures with Moran's I. These and other examples demonstrate the usefulness of spatial analysis in both infrastructure research as well as infrastructure management and monitoring.

2.3 Geosocial Media Analytics

A newer field of metropolitan research is the analysis of geosocial media. Despite concerns about skewed demographics of social media users (Jiang et al. 2019), these types of user-generated geographic information have been used in numerous fields, particularly in the social sciences. One such field that also ties in with Section 2.1 is using social media analytics as a proxy for geodemography. Spatial statistics have been used to construct daytime counterparts to the census, for example, in terms of peoples' whereabouts (Steiger et al. 2015), the ethnic composition of neighbourhoods (Longley/Adnan 2016; Longley et al. 2015), and for age and gender profiling (Lansley/Longley 2016). An-

other area of geosocial media analytics is the study of urban emotions. Frank et al. (2013) apply Geary's c and Moran's I to examine patterns of 'happiness' in tweets from across the US. Similarly, Rybarczyk et al. (2018) use exploratory spatial analysis and regression to establish links between tweet sentiments and travel modes. Examples of the use of spatial analysis of social media data in the field of crime research can be found in Ristea et al. (2020) and Kounadi et al. (2018). This selection shows the breadth of topics for which social media data has been spatially analysed, and numerous others exist (for overviews, see Steiger/Westerholt/Zipf 2016; Steiger et al. 2015). Methodological concerns have been raised about using established methods in the context of geosocial media analytics. Social media differs from scientific data in that it is generated without adherence to scientific protocols (Westerholt 2019a; 2019b). Studies have shown that using methods like Moran's I with these datasets can lead to variance inflation (due to mixtures of different phenomena, see Section 4.4) and scale-related problems (Westerholt et al. 2016). Accordingly, methods tailored to the spatial analysis of this type of data have been proposed (e.g., Westerholt 2021a; Westerholt et al. 2015).

3. Presumptions and Principles of Spatial Analysis

The application of spatial analysis methods is characterised and constrained by a number of principles and presumptions. These rest on two main circumstances: the georeferenced nature of spatial datasets and the fact that these are not independent samples. This section briefly presents some of the main resulting specificities with the aim of making the reader interested in applications aware of them when using spatial statistics. Easy-to-understand explanations, often in the form of footnotes, are given throughout to support the reader's rigorous understanding.

3.1 Spatial Processes

The primary goal of using spatial statistical measures is to understand the interaction behaviour of geographical phenomena. Making geographical phenomena including their spatial structure available to statistical analysis requires a formalisation,[1] which can be achieved via spatial processes (Cressie 1993, 8f.) of the form

$$\mathcal{Y} = \{Y_s : Y \in \Omega, s \in \mathcal{S} \subset \mathbb{R}^n\}\,, \tag{1}$$

where the Y denote random variables indexed over spatial units s (e.g. points, lines, or polygons). Set Ω is the sample space containing all possible outcomes of Y (e.g. $\mathbb{R}_{\geq 0}$ in case of precipitation or $\{1, 2, 3, 4, 5, 6\}$ when rolling a dice). This very general notion of a spatially indexed set of random variables can be endowed with various properties that lead to three specialisations of $\mathcal{Y}$. If we assume $\mathcal{S}$ to be fixed and with cardinality

1 In terms of notation, for this chapter, upper-case symbols mean random variables and lower-case symbols denote associated realisations of these in terms of concrete data, unless otherwise stated. Furthermore, bold upper-case letters stand for vectors or matrices. Superscript $\top$ is the vector or matrix transpose. Set-builder notation is used when introducing sets.

$|\mathcal{S}| = \infty$, we arrive at the notion of a geostatistical process that is used to analyse spatially continuous phenomena like precipitation or soil properties. These are defined at every location in a given study area, and thus at an infinite number of coordinates (e.g. by considering ever more precise decimal places of numeric coordinates). If we instead consider $\mathcal{S}$ as a finite but stochastic set of geometries, we can derive the notion of a point process. Here, not only the attributes but also the geometries are considered random, as is the case with locations of trees or crime sites. In other words, both the 'where' and the 'what' in these cases are subject to a certain degree of randomness. The notion used in exploratory spatial data analysis and spatial econometrics is the lattice (regular or irregular) based on a deterministic and finite set of locations with $|\mathcal{S}| < \infty$. Census variables are a common example of lattices, for which variables such as income or household sizes are only defined for the census units and any attempt to interpolate in between would be invalid. The remainder of this chapter will focus on the latter type of lattice processes and thus on the analysis of spatial structure in attributes mapped over fixed locations.

3.2 Stationarity Assumptions

Many statistical techniques rely on homogeneity assumptions. To ensure valid results and conclusions, it is often necessary that certain properties of the distribution of observations are constant. This property, called stationarity, is particularly challenging when geographical processes are involved. Geographical space is inherently heterogeneous, rendering the latter a candidate for a 'second law of geography' (Goodchild 2004). In less technical terms, this often observed lack of homogeneity in space means that there is no such thing as an average place on the Earth's surface (Goodchild 2009). Nevertheless, for technical reasons, many techniques of spatial analysis are bound to certain somewhat relaxed but still rigid stationarity assumptions. The strongest statistical homogeneity concept is that of strict stationarity. This concept implies that all properties of distributions remain constant regardless of where and when respective phenomena are observed. White noise is a non-spatial example of strict stationarity with mean $\mu = 0$ and a fixed variance σ^2, but this concept would be too restrictive and unrealistic to apply in geographical contexts. A slightly less strict form of stationarity is second-order or weak stationarity (Oliver 2010, 320f.). Here, the mean and the variance are assumed to be constant, while no assumptions are made for higher-order moments. Weak stationarity is the form of homogeneity required for many spatial analysis procedures. As a corollary, the properties of weak stationarity imply constant covariance. This is advantageous because it means that the spatial behaviour of the random variables under consideration is assumed as equally characterised everywhere on the map. We are therefore only dealing with one spatial process, not with a potentially complex mixture. One disadvantage is that strict conditions are still imposed that are not always fulfilled in practice.

3.3 Spatial Weights

A common way to incorporate spatial associations explicitly in statistical routines is to construct a spatial weights matrix. There are various forms of spatial weights for different purposes (Bavaud 2014; Harris et al. 2011), but all of them formalise the potential for either proactive interaction or passive relatedness between spatial units (Dray 2011). For example, if the process under study is based on physical contact or direct exchange as in the case of contagious diseases, the weights may be based on spatial contiguity (e.g., through immediate physical adjacency or flight connections facilitating movement). When distance plays a role, as is the case with the propagation of noise in cities, the weights pairwise connect locations based on some function of their joint physical distance. These examples illustrate two important classes of spatial weights: topological and distance-based (Getis 2009). A third way to construct weights is to derive them empirically from a given dataset, which can be helpful when no prior knowledge is available about the nature of the spatial mechanism under study. However, empirical weights carry a risk of introducing circular logic by deriving weights from the same dataset to which they are then applied, hence lowering the explanatory power of analyses. In still other cases, a third attribute can serve as a useful spatial proxy for connectedness. An example of this would be the study of the spread of invasive species using the global trade network as a proxy for their exchange.

Regardless of the underlying semantics of spatial weights, they have a profound technical impact on the results of spatial-statistical methods. Spatial weights matrices are composed of positive coefficients whereby, by convention, the diagonal is filled with zeros (Bavaud 1998). Self-interactions are thus deliberately ignored (though they would technically be includable in many cases) in order to concentrate on spatial effects between units. In addition, the weights are used to adjust the geographical scale of an analysis. Scale is an important characteristic because geographical phenomena typically operate at specific scale ranges (Dungan et al. 2002). Incorrect scale adjustment through spatial weights is a common source of error. This can have far-reaching effects on the validity of obtained results, including difficult interpretation and, in the worst case, wrong conclusions on which further theory-building could then in turn be based. Another reason why spatial weights are important for spatial analysis is a more technical and less obvious one. Many spatial statistics are given as so-called ratios of quadratic forms[2] $\boldsymbol{Y}^\top \boldsymbol{W} \boldsymbol{Y}$, with $\boldsymbol{Y}$ being a sample of n observed values for Y, and $\boldsymbol{W}$ denoting a spatial weights matrix. The range and shape of the distributions of such statistics are determined by the eigenvalue spectrum of the (in this case) spatial weights matrix (de Jong et al. 1984), a property explained in more detail in Section 4. Spatial weights therefore determine, to some extent, how we should correctly interpret spatial-statistical results.

2 Quadratic forms are polynomials in n variables with terms whose sum of powers is not greater than two (Lam 2005, 1). $f(X_1, X_2) = a_1 X_1^2 + a_2 X_2^2 + X_1 X_2$ is an example.

3.4 Modifiable Areal Unit Problem

Metropolitan research often involves the analysis of secondary data. These have most likely been collected for purposes other than studying the process we, as analysts, want to study. Furthermore, many such datasets, including census data, information divided into raster cells, or anonymised datasets to preserve geoprivacy, involve the use of aggregation units. This frequently results in the so-called Modifiable Areal Unit Problem (MAUP, Bluemke et al. 2017; Openshaw 1983), which is caused by an at least partial arbitrariness of the location, shape, and scale of aggregation units used. A consequence can then be a wrongly specified scale leading to a possible discrepancy between the scale of the process under investigation and that of the data collected. The phenomenon of interest may then not be optimally represented in the data. One problem with arbitrarily shaped aggregation units is that they may combine smaller-scale units that should possibly not be combined from a geographical point of view and with regard to the stationarity assumptions discussed above. Another concern can be the creation of possibly meaningless boundaries, which then lead to problematic spatial weights. Furthermore, the position of the aggregation units is beyond our control when using secondary data, which adds to the uncontrolled geographical mixing effects. A common example of the occurrence of the MAUP is the use of polygonal census units. These are intended for demographic purposes, but not necessarily for other types of geographical study. Another typical example of MAUP occurs in the use of grid cells. While convenient to use, these often do not reflect well underlying spatial structures. Unfortunately, the MAUP remains unsolved and challenging, mainly because it is a theoretical rather than an empirical problem (Wolf et al. 2020). It is therefore essential for any spatial analyst to be cautious when interpreting results obtained from secondary data.

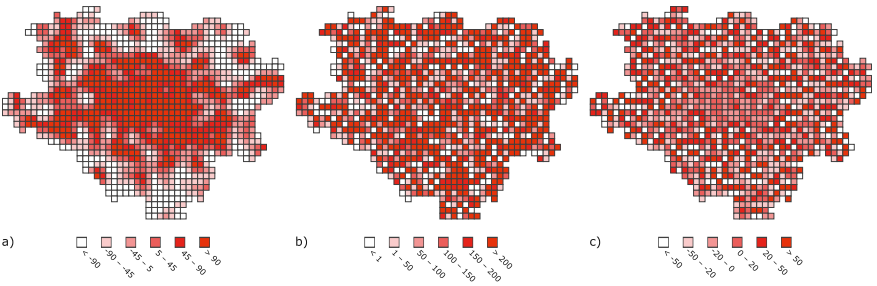
4. Spatial Autocorrelation and Moran's I

A very essential part of the statistical analysis of spatial structures deals with the estimation of spatial autocorrelation (see fig. 1). In this section, this concept is first discussed in general before the estimator Moran's I is considered in detail. An overview of other measures is given in Section 5.

4.1 Conceptual Remarks

The so-called First Law of Geography, which states that "everything is related to everything else, but near things are more related than distant things" (Tobler 1970, 236) is one of the major conceptual underpinnings of spatial analysis. Spatial autocorrelation is a way to formalise and quantify this idea in statistical terms. It describes the tendency of mapped attributes to be significantly spatially clustered (positive), dispersed (negative), or random (no significant spatial autocorrelation). Fig. 1 illustrates all three characteristics using a population grid from Dortmund, Germany. Fig. 1a shows a positively autocorrelated version of the population data resulting in more similar values occurring together than by chance. Fig. 1c shows a negatively spatially autocorrelated

Fig. 1: Illustration of positive and negative spatial autocorrelation as well as spatial randomness based on a 500-m population grid from Dortmund, Germany. Partial graphics a) and c) are based on spatially filtered variables using the method presented in Westerholt (2021a). a) Positive spatial autocorrelation; b) spatial randomness; c) negative spatial autocorrelation. The maps are based on data from the 2011 German Census.



counterpart with fewer similar values sticking together than would be expected at random. Fig. 1b shows a randomised version of the original population data with clustered areas being compensated by negatively autocorrelated parts of the map. Measures of spatial autocorrelation are thus a way to characterise different types of spatial structures in datasets.

Spatial autocorrelation is encountered in a number of situations and is often methodologically useful. Geostatistics, for example, is based on assuming positive spatial autocorrelation (Getis 2008). Natural phenomena such as precipitation and temperatures do not normally show sudden jumps in neighbouring areas unless barriers like rivers or rock walls are present. Data about such phenomena exhibit smooth spatial variation that is exploited in geostatistics for Kriging, a statistical interpolation procedure that takes into account the spatial correlations estimated from data (Calder/Cressie 2009). Exploratory spatial data analysis uses spatial autocorrelation to explore the *a priori* unknown spatial nature of phenomena and generate hypotheses. Inferences from measurements of spatial autocorrelation combined with different spatial weights, each reflecting different possible spatial interaction mechanisms, are then a way to investigate possible spatial functioning mechanisms and geographical relationships. Spatial autocorrelation also gives rise to specific forms of spatial regression modelling including the spatial autoregressive (long-range spatial effects; global spillovers) and the spatial error model (small-scale, limited spatial effects; local spillovers) (LeSage/Pace 2014; Anselin 2003). Further uses of the concept of spatial autocorrelation include testing for model misspecification, testing spatial stationarity assumptions, uncovering spatial relationships, examining the influences of geometric units and aggregation, revealing the roles of time and space, and supporting the study of spatial outliers (Getis 2007).

4.2 Definition of Moran's I

Particular attention has been given to estimators of spatial autocorrelation for interval-scaled random variables. Due to its widespread availability in software packages and statistical computing environments, Moran's I is one of the most widely used such methods. Other popular methods include the G-statistics (Ord/Getis 2001; 1995) for hotspots and Geary's c (Cliff/Ord 1981; Geary 1954), but Moran's I was shown to be better behaved than c with respect to statistical power[3] and sensitivity to spatial weights (Chun/Griffith 2013). Another reason for its popularity may be that Moran's I resembles the non-spatial Pearson correlation coefficient r. This similarity is appealing, but it may also tempt researchers to misinterpret Moran's I in spirit of Pearson's r, which is sometimes justified but much more often is not. Global and local Moran's I are given as

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

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

whereby the w_{ij} terms denote spatial weights connecting locations i and j, and y_i are n numeric variates with mean $\bar{y}$. Note that alternative notions of I exist for the analysis of regression residuals that account for the additional exogenous variation contributed by regressors (Tiefelsdorf 2000).

4.3 Interpretation

Interpreting Moran's I is more complex than interpreting Pearson's r. Pearson's r has a straightforward interpretation where values on $[-1, 0)$ indicate a negative and values on $(0, 1]$ indicate a positive correlation, and where the mean 0 signifies decorrelation. Significance of r depends on whether r is far away enough from 0, and on sample distribution and size. The mean of Moran's I also goes towards 0 as n increases but is generally given by $-1/(n-1)$ (Cliff/Ord 1981, 44), whose deviation from 0 is important especially for smaller samples. For better understanding the feasible range of Moran's I, it is useful to first look at Pearson's r. Let $\boldsymbol{E} = (e_{ij})$ be the $n \times n$ identity matrix.[4]

3 The ability of a statistical test to successfully identify significant effects when they are present.

4 Identity matrices have ones on the diagonal and zeros elsewhere. They are the neutral element for vectors and matrices, since multiplication with them leaves the latter unchanged. Thus, identity matrices are the vector and matrix equivalent of the scalar 1.

Instead of using the usual expression, we can write r as:

$$r = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} e_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} e_{ij}(x_i - \bar{x})(y_j - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}. \tag{3}$$

The left-hand summation over elements e_{ij} evaluates to n and the double sum in the numerator on the right-hand side adds a number of zero terms, since all off-diagonal elements of $\boldsymbol{E}$ are zero (since Pearson's r does not contain pair-wise weights). Mathematically, nothing has changed by this more complicated notation. Writing r this way, however, one can see the structural similarity to Moran's I including the ranges of the two statistics. Determined by the only non-zero eigenvalue of $\boldsymbol{E}$, which is $\lambda = 1$, Pearson's r ranges on $[-\lambda, \lambda]$ and thus on $[-1, 1]$. This range no longer holds if we replace $\boldsymbol{E}$ by some spatial weights matrix $\boldsymbol{W}$ with a different eigenvalue spectrum.

It is possible to substitute $\boldsymbol{W}$ in Equations 2a and 2b by its eigenvalue representation[5] $\boldsymbol{HWH} = \boldsymbol{U\Lambda U}^\top$ with $\boldsymbol{H}$ being the centring operator[6] (Dray 2011; Dray et al. 2006). Here, $\boldsymbol{\Lambda}$ is a diagonal matrix of eigenvalues associated with eigenvectors in the columns of $\boldsymbol{U}$. It can be shown that for symmetric[7] matrices $\boldsymbol{W}$, the feasible domain of Moran's I is given as $[a \cdot \lambda_{\text{min}},\ a \cdot \lambda_{\text{max}}]$ with $a = n/\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}$ and $\lambda_{\text{min}}, \lambda_{\text{max}}$ denoting the smallest and largest eigenvalues of $\boldsymbol{HWH}$ (de Jong et al. 1984). This shows that for quadratic forms such as Pearson's r and Moran's I there is a close relationship between the eigenvalues of (implicit or explicit) weighting structures and the feasible value ranges of the corresponding statistical measures.

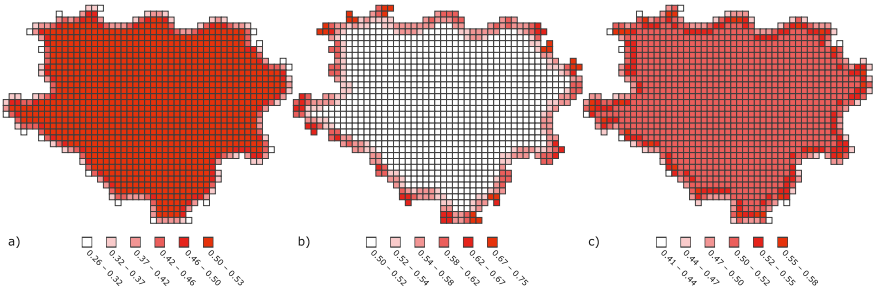
Not only the bounds of I depend on the eigenvalues of $\boldsymbol{HWH}$, but also the shape of the distribution and resulting map patterns. The locations of the eigenvalues on the spectrum are important for the shape of I's distribution, and some eigenvalues can mark inflection points (Tiefelsdorf/Boots 1995) especially for smaller lattices (for large-sample asymptotics, see Section 4.4). In addition, the spatial weights are sometimes normalised, for example, to make them comparable across different study areas. Common normalisations include the W-coding (each row sums to 1), and the C-coding scheme (each weight represents its global share; for an overview, see Bavaud 2014). However, these normalisation schemes affect the topology-induced variance and change the influence of spatial units on a spatial analysis. The W-coding scheme gives excessive weight to low-connected units (see fig. 2b) that are typically found along the boundary of a study area, but can also occur elsewhere, for example, when spatial units vary strongly in size. In contrast, the C-coding scheme favours highly connected units (fig. 2a). Tiefelsdorf et al. (1999) have presented the S-coding scheme to balance the

5 An eigenvalue representation of the spatial weights matrix can intuitively be thought of as a decomposition and reorganisation of the initial matrix into all possible spatial substructures represented by the weights. The eigenvalues then reflect the strengths of these substructures.

6 A centring operator is a matrix that subtracts the mean value either by column or by row.

7 The symmetric part $1/2\left(\boldsymbol{W} + \boldsymbol{W}^\top\right)$ of $\boldsymbol{W}$ can safely be computed in the present case of Moran's I since the antisymmetric part leads to quadratic forms evaluating to zero.

Fig. 2: Effect of different kinds of normalisations of spatial weights represented by the eigenvalues of local spatial weights matrices. a) C-coding scheme. b) W-coding scheme. c) S-coding scheme. The maps are based on data from the 2011 German Census.



effects of C and W-coding (fig. 2c). Shortridge (2007) has further found for grid configurations that both positive and negative autocorrelation are overestimated when using rook[8] instead of queen[9] weights, an effect that is more pronounced in the case of negative spatial autocorrelation. In summary, interpreting Moran's I and related measures depends strongly on the spatial weights and the way neighbourhood relations are specified. Reporting results without elaborating on the spatial weighting scheme used is therefore of limited informative value.

A useful graphical tool to understand Moran's I results beyond distributional concerns is the Moran scatterplot (Anselin 1996). The plot maps the standardised[10] attribute values y_i on the x-axis against their also standardised[11] spatial lags $\sum_j w_{ij} y_j$ (i.e. the spatially weighted sum of neighbours) on the y-axis. This shows the relationship of Moran's I to the regression of the lags on the variates. Fig. 3a shows an example of a Moran scatterplot for the filtered population data used in fig. 1a. The data are strongly positively autocorrelated, which is manifested in the diagram by the clustering of data points in the first (high values surrounded by other high values) and third quadrants (low values surrounded by other low values) and by a positively sloping trendline. In contrast, the negatively spatially autocorrelated data from fig. 1c result in an arrangement centred in the second and fourth quadrants (fig. 3c). Spatial randomness is characterised by the absence of a discernible trend as visualised in fig. 3b. The plot can be used to examine different features of the spatial autocorrelation structure present in a dataset. For example, it is possible to identify deviant data points that exhibit unusual behaviour. These are often of particular geographical interest because they do not fit into their spatial surroundings. In addition, the scatterplot can be used to identify structural breaks, that is, possible non-stationarities with respect to the spatial process.

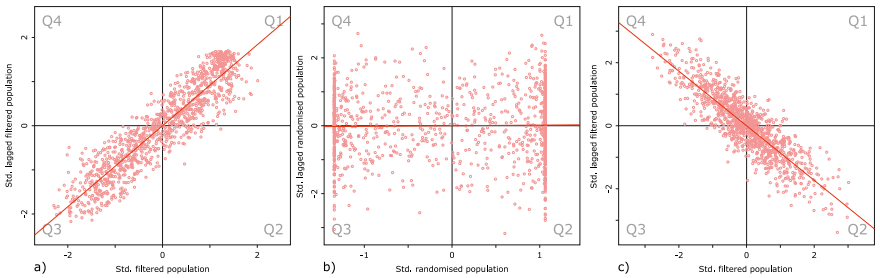
8 Rook weights connect grid cells along the four cardinal directions.

9 Queen weights connect grid cells along the four cardinal directions and the diagonals.

10 Standardisation means to centre the variables by subtracting their mean and then dividing by their standard deviation.

11 Standardisation means to centre the variables by subtracting their mean and then dividing by their standard deviation.

Fig. 3: Moran scatterplots for the positively and negatively spatially autocorrelated population variable from fig. 1 and for its spatially randomised version. Q1–Q4 denote quadrants 1 to 4 as defined by the dashed lines. The red trend lines indicate the regressions of the lags on the standardised variables. Moran scatterplots for a) a positively spatially autocorrelated variable; b) a spatially randomised variable; and c) a negatively spatially autocorrelated variable.



The Moran scatterplot is thus a very helpful tool especially for the initial exploration of spatial structures.

The main takeaway from these technical considerations is that the range of possible values for Moran's I and its distribution depend strongly on the spatial weights. This is analogously the case with Pearson's r but is not of practical relevance there, since the (implicit) weighting structure is always the same. Very similar results apply for other spatial measures like Geary's c, which also depend on given exogenous spatial linkages. Moreover, the spatial weights determine to some extent how much influence individual spatial units exert on the overall spatial analysis; a property that can sometimes affect analyses in unexpected ways, for example, when normalisations are involved. Empirical researchers should be mindful of these aspects and take them into account when interpreting corresponding results.

4.4 Asymptotic Distribution of Moran's I

Drawing inferences about Moran's I requires knowledge of the null distribution assuming no spatial autocorrelation. This knowledge allows to assess how likely or unlikely an observed I value would occur by chance given the connectivity defined by the spatial weights. Resorting to a Monte Carlo-style approach with estimation of an empirical null distribution is possible but may sometimes be impractical, for example, if the sample size is large. Moran's I can also be evaluated analytically by approximating its asymptotic distribution in two different ways. One possible approach uses a randomisation argument implying a conditional viewpoint holding the observed values fixed. The null hypothesis then assumes that the observed values could occur randomly anywhere on the lattice with equal chance (Besag/Newell 1991). An alternative approach is based on the assumption that the observed values were drawn from a joint normal distribution and is thus not limited to certain already observed values. In this case, the distribution of I in the null hypothesis corresponds to repeated and independent sampling from a

normal distribution for each spatial unit. Both approaches allow testing for complete spatial randomness using normal approximations. Which of these viewpoints to adopt depends largely on the nature of the process under investigation.

Approximating the null distribution of I using normal distributions is possible for both inference approaches outlined above. They are subject to only mild regularity conditions (Cliff/Ord 1981, 46ff.). One condition is a sufficiently large dataset. Cliff and Ord (1972) suggest using a Beta approximation for small samples. A second condition is that the number of non-zero connections per spatial unit and established through the weights is not a function of the size of the lattice. A third condition is that no geographical subregion should dominate the lattice too much. Indeed, unfavourable spatial configurations exist that may prohibit the use of the normal approximation, but in many practical cases the latter will be possible. Normal approximation requires estimates of the mean and variance of I. The mean is given in Section 4.3 and does not differ between the two types of null hypotheses concerned. The variance terms for the two cases, however, are not identical and, due to the complex weighting structures involved, are quite cumbersome. For the latter reason, I refrain from replicating these terms here and refer the reader to Cliff and Ord (1981) and Griffith (2010). The analyst may sometimes encounter underlying, exogenous spatial processes that interfere with the analysis. Also, as described in Section 4.3, spatial configurations may impact the null distribution. Both situations can result in skewness invalidating the normal approximation. Tiefelsdorf (2002) suggests using a saddlepoint approximation in these cases that can accommodate such circumstances better than the normal approximation.

Data in metropolitan research often come in the form of counts, rates, or other forms of non-normal observations. It is therefore of practical interest to consider inferences about Moran's I with non-normally distributed samples. Griffith (2010) has shown that the normal approximations introduced above can be extended to a range of random variables that mimic the normal distribution. This is the case, for example, for counts drawn from a Poisson distribution, provided their mean is sufficiently large. A similar argument applies to binomial variables under the restriction of a large number of trials. For these types of variables, the equation for the mean of I holds if the random variables are reasonably symmetric about their mean. In the cases of skewness or non-symmetry, the mean estimator will be asymptotically valid if n is large enough. The term for the variance of Moran's I under the normality assumption is asymptotically valid as long as independence and identical distribution hold, and when n is roughly larger than 25. These encouraging results allow the extension of the normal approximation to a number of distributions. To some extent, this even applies to mixtures of differently distributed random variables, although this would require even larger sample sizes. For the latter case, however, it has been shown that the spatial arrangement of the random variables involved has an influence on both the mean and the variance of I, especially when the underlying means and variances of the distributions that enter mixtures differ greatly (Westerholt 2018; Westerholt et al. 2016). Therefore, caution is still required when drawing conclusions about Moran's I using non-normal data.

5. Specialised Measures of Spatial Association

In metropolitan research, non-interval scaled variables are often considered. Specialised estimators have been developed, of which the following subsections provide an overview.

5.1 Rate Variables

Rates like disease incidences and unemployment shares often violate the stationarity assumptions of spatial-statistical tests. High rates are more likely to occur when the underlying base population is small. Depending on the composition of the underlying populations, the resulting heteroscedasticity leads to either Type-I (false positive) or Type-II (false negative) error inflation (Walter 1992a; 1992b). Various approaches have been proposed to deal with rate variables. Acknowledging that the variance of normal variables depends on the sample size, Waldhör (1996) proposes to use inverse local population sizes as approximators of the local variances in the estimator for the variance of I. Oden (1995) instead traces regional rates back to individual cases. This, however, would incur a high downstream computational effort. The analysis is thus brought back to spatial units by using global comparison values based on the same base population everywhere. Assuncao and Reis (1999) propose an empirical Bayesian solution considering rates as conditional on local propensities. Rates are standardised with a constant global mean estimated from raw counts (instead of averaging the rates) and a variance estimate taking into account local numbers of cases. A similar but improved method has been proposed recently by Jung et al. (2019b). Jackson et al. (2010) propose to include the spatial weights matrix in the variance estimator used in the denominator of Moran's I. In this way, explanatory power is borrowed from exploiting spatial redundancy in nearby populations. In a similar vein, Bucher et al. (2020) incorporate additional uncertainty weights in I in a mobile sensor measurement context. Zhang and Lin (2016) develop an adjustment factor to account for heteroscedasticity as well as spatial structure in the variance. This factor can be used in the variance estimator for the randomisation-based hypothesis testing framework. The range of approaches introduced shows that heteroscedasticity has received considerable attention.

5.2 Categorical Variables

Categorical variables can mean either the analysis of binary outcomes like the presence or absence of a species, or multi-categorical data such as the Index of Multiple Deprivation in the UK censuses. The traditional and still widely used methods for assessing spatial structure in these types of variables are the join-count statistics, either in a binary way (Cliff/Ord 1981; Moran 1948; 1947) or for so-called k-colour maps (Cliff/Ord 1981; Krishna Iyer 1949). These measures count the numbers of ties of certain, spatially neighboured attribute values. In addition to using a Monte Carlo permutation approach, two different types of analytical evaluation based on normal approximations are available. These are in principle analogous to the normal and randomisation assumptions for the evaluation of I: free sampling with replacement and unfree sampling. Boots (2003) de-

velops local tests for categorical value clustering. These tests are conditional on local compositions of classes and thus address the problem that the spatial configuration of categorical variables is not independent of their class composition. To address this compositional issue, Ruiz et al. (2010), Matilla-García et al. (2012), and Farber et al. (2015) develop global tests based on the entropy of the different locally occurring attribute value compositions. These ideas are taken further towards local testing by Naimi et al. (2019). Also focusing on local tests, Anselin and Li (2019) and Anselin (2019b) have proposed categorical counterparts to local Moran's I, which are an amalgamation of the latter with the join count statistics.

5.3 Vectors and Flows

Many phenomena including traffic flows, commuter patterns, and migration can be represented in networks or origin–destination matrices. Analysing these is possible from various perspectives (Chun 2008): spatial dependence in the origins, clustering of destinations, or combinations of these. Additional modelling steps are thus required for the spatial weights in addition to adapted models. Liu et al. (2015) have modified Moran's I towards considering origin and destination geometries as attribute values. Tao and Thill (2020) have extended this geometric idea to include attribute values (e.g. exchange intensity) and bivariate cases (in their example taxi trips and use of ride-hailing services). Analogous approaches exist for the analysis of the tails of a distribution (hotspot analysis). Berglund and Karlström (1999) have presented an approach to the G-statistics that allows clusters of high and low fluxes to be identified, again taking into account different perspectives through respectively modelled spatial weights. Another approach to hotspot analysis is developed by Tao and Thill (2019b). They extend an incremental method called AMOEBA (Aldstadt/Getis 2006), which grows spatially connected clusters from the bottom up, towards detecting coherent 'ecotopes' of flows. Another perspective on flows is a geometric one in the sense of unmarked point pattern analysis. Tao and Thill (2016) introduce a method based on the widely used K-function that considers flows in a joint four-dimensional space, and thus without separating the origin and destination geometries. This idea was later extended to the multivariate case (Tao/Thill 2019a). Similarly, Shu et al. (2021) present an analogous technique based on the L-function, a variance-stabilised version of K.

5.4 Multivariate Analysis

Sometimes it is of interest to analyse joint spatial patterns and linkages between different processes. The use of standard correlation measures like Pearson's r with spatially autocorrelated variables is problematic due to Type-I error inflation (Dutilleul/Legendre 1993; Clifford et al. 1989; Bivand 1980). A number of alternative approaches have been discussed. Wartenberg (1985) proposes to expand data vectors to matrices with different variables per spatial unit. These $n \times m$ matrices can be used in a cross product with the spatial weights matrix, and the eigenvectors of the resulting matrix can then be evaluated in the sense of a principal component analysis. A local version of this method has recently been proposed (Lin 2020). The relationship between spatial and non-spatial

correlation is not unique, however, and different spatial configurations can be found producing the same Pearson's r values. Lee (2001) therefore combines Pearson's r and Moran's I into a common measure that captures point-to-point and spatial correlations. The latter proved to be more suitable for small sample sizes than the Wartenberg (1985) approach (Khamis et al. 2010). Anselin et al. (2002) focus solely on spatial arrangement and present a multivariate version of local Moran's I, including a generalised Moran scatterplot, in which the spatial lag of one variable is regressed on observations of another. In the same vein, Anselin (2019a) extends Geary's c for multivariate datasets and points out that for the multivariate case, measures like Geary's c based on differences in attribute space offer conceptual advantages over cross products of mean deviations such as Moran's I. Also modifying Geary's c and Moran's I, Eckardt and Mateu (2021) propose partial versions of these statistics.

5.5 Spatial Heterogeneity

Spatial heterogeneity describes geographical instabilities of statistical parameters (Dutilleul/Legendre 1993). This property can be caused either by endogenous non-stationarity or by exogenous contextual variation. Often regarded as a nuisance, spatial heterogeneity can be an interesting feature for scientific enquiry. One way to investigate heterogeneity is to use local statistics such as the local version of Moran's I (see equation 2b) or other so-called Local Indicators of Spatial Association (Anselin 1995). Mapping such measures and inspecting visualisations like the Moran scatterplot enables the exploration of local pockets of heterogeneity. More recently, specialised measures of spatial heterogeneity have been developed. Ord and Getis (2001) propose a measure of spatial concentration and thus of spatially varying means, whereby the method controls for possibly interfering global autocorrelation structures. Focusing on variance, Ord and Getis (2012) propose a measure of spatial heteroscedasticity. This method called LOSH (*Local Spatial Heteroscedasticity*) calculates variances about locally estimated means and allows for the detection of irregular clusters and spatial boundaries. Xu et al. (2014) have investigated the distributional properties of LOSH and recommend a Monte Carlo strategy for inference instead of the parametric chi-square test originally proposed. Based on LOSH, Westerholt et al. (2018) develop a test for strictly local spatial heteroscedasticity to characterise spatial variance in subregions regardless of other locations. Background is the detection of pronounced variances that may only stand out within small subregions but not in a global comparison. On a more practical note, Aldstadt et al. (2012) have suggested using LOSH and the G-statistics in tandem to investigate internal cluster heterogeneity. An alternative approach to this with improved discriminability between cluster boundaries and interiors is to use a spatial filtering approach (Westerholt 2021a; 2021b).

6. Summary and Outlook

In this chapter, basic principles and methods of statistical spatial analysis were presented. The topic was motivated through outlining selected application areas. These

give the reader an impression of the breadth of metropolitan research for which spatial analysis has been applied. Presumptions and possible pitfalls were then outlined before Moran's I was presented in detail. The latter included not only the definition and interpretation of the measure, but also associated inference mechanisms. After the main methodological part, further spatial statistical estimators were discussed. However, this chapter is not exhaustive. Many of the principles presented do also apply to the various spatial regression approaches. Examples include spatial error and spatial lag models (Anselin 2001), geographically weighted regression (Wheeler/Páez 2010), and spatial filtering (Getis/Griffith 2002; Griffith 2000). Another area that is largely left out in the chapter is the topic of spatiotemporal analysis. Whilst broadly similar to what is discussed in this chapter, the latter differs conceptually, particularly in terms of modelling spatial weights (e.g., Gao 2015). In terms of future trends, one current research direction is towards a deeper integration of spatial analysis with computer science, leading to the notion of 'spatial' or 'geographical' data science (Bacao et al. 2020; Singleton/Arribas-Bel 2021). A related direction deals with a stronger integration of spatial analysis and machine learning (Klemmer/Neill 2020; Klemmer et al. 2019). Another current trend is the deeper integration of spatial analysis with human geography, manifesting itself in human-centred, place-based approaches (Westerholt et al. 2020; Purves et al. 2019). These areas will complement the traditional directions of spatial statistics in interesting ways and open up new pathways, both theoretically and in terms of practical integration with new fields of application in metropolitan research.

References

Abokifa, Ahmed A., Lina Sela. "Identification of Spatial Patterns in Water Distribution Pipe Failure Data Using Spatial Autocorrelation Analysis." Journal of Water Resources Planning and Management 145.12 (2019): 04019057.

Aldstadt, Jared, Arthur Getis. "Using AMOEBA to Create a Spatial Weights Matrix and Identify Spatial Clusters." Geographical Analysis 38.4 (2006): 327–343.

Aldstadt, Jared, Michael Widener, Neal Crago. "Detecting Irregular Clusters in Big Spatial Data." In: Proceedings of the 7th International Conference on Geographic Information Science (GIScience 2012). Eds. Ningchuan Xiao, Mei-Po Kwan, Michael F. Goodchild, Shashi Shekhar. Columbus, OH, USA, 2012.

Andrews, Marcus R., et al. "Geospatial Analysis of Neighborhood Deprivation Index (NDI) for the United States by County." Journal of Maps 16.1 (2020): 101–112.

Anselin, Luc. "Local Indicators of Spatial Association – LISA." Geographical Analysis 27.2 (1995): 93–115.

Anselin, Luc. "The Moran Scatterplot as an ESDA Tool to Assess Local Instability in Spatial Association." In: Spatial Analytical Perspectives on GIS. Eds. Manfred M. Fischer, Henk J. Scholten, David Unwin. London, UK: Taylor & Francis, 1996. 111–125.

Anselin, Luc. "Spatial Econometrics." In: A Companion to Theoretical Econometrics. Ed. Badi H. Baltagi. Malden, MA, USA: Blackwell, 2001. 310–330.

Anselin, Luc. "Spatial Externalities, Spatial Multipliers, and Spatial Econometrics." International Regional Science Review 26.2 (2003): 153–166.

Anselin, Luc. "A Local Indicator of Multivariate Spatial Association: Extending Geary's c." Geographical Analysis 51.2 (2019a): 133–150.

Anselin, Luc. "Quantile Local Spatial Autocorrelation." Letters in Spatial and Resource Sciences 12.2 (2019b): 155–166.

Anselin, Luc, Xun Li. "Operational Local Join Count Statistics for Cluster Detection." Journal of Geographical Systems 21.2 (2019): 189–210.

Anselin, Luc, Ibnu Syabri, Oleg Smirnov. "Visualizing Multivariate Spatial Correlation with Dynamically Linked Windows." In: Proceedings of the CSISS Workshop on New Tools for Spatial Data Analysis. Santa Barbara, CA, USA, 2002.

Assunção, Renato M., Edna A. Reis. "A New Proposal to Adjust Moran's I for Population Density." Statistics in Medicine 18.16 (1999): 2147–2162.

Bacao, Fernando, Maribel Yasmina Santos, Martin Behnisch. "Spatial Data Science." ISPRS International Journal of Geo-Information 9.7 (2020): 428.

Barreca, Alice, Rocco Curto, Diana Rolando. "Housing Vulnerability and Property Prices: Spatial Analyses in the Turin Real Estate Market." Sustainability 10.9 (2018): 3068.

Bavaud, François. "Models for Spatial Weights: A Systematic Look." Geographical Analysis 30.2 (1998): 153–171.

Bavaud, François. "Spatial Weights: Constructing Weight-Compatible Exchange Matrices from Proximity Matrices." In: Proceedings of the 8th International Conference on Geographic Information Science (GIScience 2014), Vienna, Austria, September 24–26, 2014. Eds. Matt Duckham, Edzer Pebesma, Kathleen Stewart, Andrew U. Frank. Cham, Switzerland: Springer, 2014. 81–96.

Berglund, Svante, Anders Karlström. "Identifying Local Spatial Association in Flow Data." Journal of Geographical Systems 1.3 (1999): 219–236.

Besag, Julian, James Newell. "The Detection of Clusters in Rare Diseases." Journal of the Royal Statistical Society: Series A (Statistics in Society) 154.1 (1991): 143–155.

Bivand, Roger. "A Monte Carlo Study of Correlation Coefficient Estimation with Spatially Autocorrelated Observations." Quaestiones Geographicae 6 (1980): 5–10.

Blazquez, Carola A., Barbara Picarte, Juan Felipe Calderón, Fernando Losada. "Spatial Autocorrelation Analysis of Cargo Trucks on Highway Crashes in Chile." Accident Analysis & Prevention 120 (2018): 195–210.

Bluemke, Matthias, Bernd Resch, Clemens Lechner, René Westerholt, Jan-Philipp Kolb. "Integrating Geographic Information into Survey Research: Current Applications, Challenges and Future Avenues." Survey Research Methods 11.3 (2017): 307–327.

Boots, Barry. "Developing Local Measures of Spatial Association for Categorical Data." Journal of Geographical Systems 5.2 (2003): 139–160.

Boss, Darren, Trisalyn Nelson, Meghan Winters, Colin J. Ferster. "Using Crowdsourced Data to Monitor Change in Spatial Patterns of Bicycle Ridership." Journal of Transport & Health 9 (2018): 226–233.

Brachman, Micah L. "Don't Forget About Geography." Journal of Spatial Information Science 21 (2020): 263–266.

Brenner, Neil, Christian Schmid. "The 'Urban Age' in Question." International Journal of Urban and Regional Research 38.3 (2014): 731–755.

Bucher, Dominik, Henry Martin, David Jonietz, Martin Raubal, René Westerholt. "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), September 27–30, 2021, Poznań, Poland; Part I. Eds. Krzysztof Janowicz, Judith A. Verstegen. Saarbrücken/Wadern, Germany: Dagstuhl Publishing, 2020. 2:1–2:15.

Calder, Catherine, Noel A. C. Cressie. "Kriging and Variogram Models." In: International Encyclopedia of Human Geography. Eds. Rob Kitchin, Nigel Thrift. Vol. 1. Oxford, UK: Elsevier, 2009. 49–55.

Castells, Manuel. From the Urban Question to the City and the Grassroots. Brighton, UK: Urban and Regional Studies, University of Sussex, 1985.

Cebrecos, Alba, et al. "Geographic and Statistic Stability of Deprivation Aggregated Measures at Different Spatial Units in Health Research." Applied Geography 95 (2018): 9–18.

Ceci, Michelangelo, Roberto Corizzo, Donato Malerba, Aleksandra Rashkovska. "Spatial Autocorrelation and Entropy for Renewable Energy Forecasting." Data Mining and Knowledge Discovery 33.3 (2019): 698–729.

Chun, Yongwan. "Modeling Network Autocorrelation within Migration Flows by Eigenvector Spatial Filtering." Journal of Geographical Systems 10.4 (2008): 317–344.

Chun, Yongwan, Daniel A. Griffith. Spatial Statistics and Geostatistics: Theory and Applications for Geographic Information Science and Technology. London, UK: SAGE, 2013.

Cliff, Andrew David, J. Keith Ord. "Testing for Spatial Autocorrelation among Regression Residuals." Geographical Analysis 4.3 (1972): 267–284.

Cliff, Andrew David, J. Keith Ord. Spatial Processes: Models and Applications. London, UK: Pion, 1981.

Clifford, Peter, Sylvia Richardson, Denis Hémon. "Assessing the Significance of the Correlation Between Two Spatial Processes." Biometrics 45.1 (1989): 123–134.

Conway, Delores, et al. "A Spatial Autocorrelation Approach for Examining the Effects of Urban Greenspace on Residential Property Values." The Journal of Real Estate Finance and Economics 41.2 (2010): 150–169.

Cressie, Noel A. C. Statistics for Spatial Data. New York, NY, USA: John Wiley & Sons, 1993.

Dai, Dajun. "Racial/Ethnic and Socioeconomic Disparities in Urban Green Space Accessibility: Where to Intervene?" Landscape and Urban Planning 102.4 (2011): 234–244.

de Jong, Peter, C. Sprenger, Frans van Veen. "On Extreme Values of Moran's I and Geary's c." Geographical Analysis 16.1 (1984): 17–24.

Dray, Stéphane. "A New Perspective about Moran's Coefficient: Spatial Autocorrelation as a Linear Regression Problem." Geographical Analysis 43.2 (2011): 127–141.

Dray, Stéphane, Pierre Legendre, Pedro R. Peres-Neto. "Spatial Modelling: A Comprehensive Framework for Principal Coordinate Analysis of Neighbour Matrices (PCNM)." Ecological Modelling 196.3–4 (2006): 483–493.

Dungan, Jennifer L., et al. "A Balanced View of Scale in Spatial Statistical Analysis." Ecography 25.5 (2002): 626–640.

Dutilleul, Pierre, Pierre Legendre. "Spatial Heterogeneity against Heteroscedasticity: An Ecological Paradigm versus a Statistical Concept." Oikos 66.1 (1993): 152–171.

Eckardt, Matthias, Jorge Mateu. "Partial and Semi-Partial Statistics of Spatial Associations for Multivariate Areal Data." Geographical Analysis 53.4 (2021): 818–835.

Farber, Steven, Manuel Ruiz Marín, Antonio Páez. "Testing for Spatial Independence Using Similarity Relations." Geographical Analysis 47.2 (2015): 97–120.

Fingleton, Bernard, Daniel Olner, Gwilym Pryce. "Estimating the Local Employment Impacts of Immigration: A Dynamic Spatial Panel Model." Urban Studies 57.13 (2020): 2646–2662.

Fischer, Manfred M. "Spatial Analysis: Retrospect and Prospect." In: Geographical Information Systems: Principles, Technical Issues, Management Issues and Applications. Eds. Paul A. Longley, Michael F. Goodchild, David J. Maguire, David W. Rhind. Vol. 1. New York, NY, USA: Wiley, ^{2}2005. 283–292.

Fischer, Manfred M., Arthur Getis. "Introduction." In: Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications. Eds. Manfred M. Fischer, Arthur Getis. Berlin/Heidelberg, Germany: Springer, 2010. 1–24.

Frank, Morgan R., Lewis Mitchell, Peter Sheridan Dodds, Christopher M. Danforth. "Happiness and the Patterns of Life: A Study of Geolocated Tweets." Scientific Reports 3.1 (2013): 2625.

Gao, Song. "Spatio-Temporal Analytics for Exploring Human Mobility Patterns and Urban Dynamics in the Mobile Age." Spatial Cognition & Computation 15.2 (2015): 86–114.

Gao, Xingchuan, Tao Li, Xiaoshu Cao. "Spatial Fairness and Changes in Transport Infrastructure in the Qinghai-Tibet Plateau Area from 1976 to 2016." Sustainability 11.3 (2019a): 589.

Gao, Yong, Jing Cheng, Haohan Meng, Yu Liu. "Measuring Spatio-Temporal Autocorrelation in Time Series Data of Collective Human Mobility." Geo-spatial Information Science 22.3 (2019b): 166–173.

Geary, Robert C. "The Contiguity Ratio and Statistical Mapping." The Incorporated Statistician 5.3 (1954): 115–127 + 129–146.

Getis, Arthur. "Reflections on Spatial Autocorrelation." Regional Science and Urban Economics 37.4 (2007): 491–496.

Getis, Arthur. "A History of the Concept of Spatial Autocorrelation: A Geographer's Perspective." Geographical Analysis 40.3 (2008): 297–309.

Getis, Arthur. "Spatial Weights Matrices." Geographical Analysis 41.4 (2009): 404–410.

Getis, Arthur, Daniel A. Griffith. "Comparative Spatial Filtering in Regression Analysis." Geographical Analysis 34.2 (2002): 130–140.

Goodchild, Michael F. "The Validity and Usefulness of Laws in Geographic Information Science and Geography." Annals of the Association of American Geographers 94.2 (2004): 300–303.

Goodchild, Michael F. "What Problem? Spatial Autocorrelation and Geographic Information Science." Geographical Analysis 41.4 (2009): 411–417.

Griffith, Daniel A. "A Linear Regression Solution to the Spatial Autocorrelation Problem." Journal of Geographical Systems 2.2 (2000): 141–156.

Griffith, Daniel A. "The Moran Coefficient for Non-Normal Data." Journal of Statistical Planning and Inference 140.11 (2010): 2980–2990.

Harris, Richard, John Moffat, Victoria Kravtsova. "In Search of 'W'." Spatial Economic Analysis 6.3 (2011): 249–270.

Hong, Jingke, et al. "Unfolding the Spatial Spillover Effects of Urbanization on Interregional Energy Connectivity: Evidence from Province-Level Data." Energy 196 (2020): 116990.

Houlden, Victoria, João Porto de Albuquerque, Scott Weich, Stephen Jarvis. "A Spatial Analysis of Proximate Greenspace and Mental Wellbeing in London." Applied Geography 109 (2019): 102036.

Houlden, Victoria, et al. "The Relationship between Greenspace and the Mental Wellbeing of Adults: A Systematic Review." PLoS ONE 13.9 (2018): e0203000.

Jackson, Monica C., Lan Huang, Qian Xie, Ram C. Tiwari. "A Modified Version of Moran's *I*." International Journal of Health Geographics 9.1 (2010): 33.

Jiang, Yuqin, Zhenlong Li, Xinyue Ye. "Understanding Demographic and Socioeconomic Biases of Geotagged Twitter Users at the County Level." Cartography and Geographic Information Science 46.3 (2019): 228–242.

Jung, Paul H., Jean-Claude Thill, Michele Issel. "Spatial Autocorrelation and Data Uncertainty in the American Community Survey: A Critique." International Journal of Geographical Information Science 33.6 (2019a): 1155–1175.

Jung, Paul H., Jean-Claude Thill, Michele Issel. "Spatial Autocorrelation Statistics of Areal Prevalence Rates under High Uncertainty in Denominator Data." Geographical Analysis 51.3 (2019b): 354–380.

Kandt, Jens, Paul A. Longley. "Ethnicity Estimation Using Family Naming Practices." PLoS ONE 13.8 (2018): e0201774.

Khamis, Faisal G., Abdul Aziz Jemain, Kamarulzaman Ibrahim. "On a Comparison between Two Measures of Spatial Association." Journal of Modern Applied Statistical Methods 9.1 (2010): 13.

Kim, Byoungjun et al. "COVID-19 Testing, Case, and Death Rates and Spatial Socio-Demographics in New York City: An Ecological Analysis as of June 2020." Health & Place 68 (2021): 102539.

Klemmer, Konstantin, Adriano Koshiyama, Sebastian Flennerhag. "Augmenting Correlation Structures in Spatial Data Using Deep Generative Models." arXiv (2019): arXiv:1905.09796 [preprint].

Klemmer, Konstantin, Daniel B. Neill. "SXL: Spatially Explicit Learning of Geographic Processes with Auxiliary Tasks." arXiv (2020): arXiv:2006.10461 [preprint].

Kounadi, Ourania, Alina Ristea, Michael Leitner, Chad Langford. "Population at Risk: Using Areal Interpolation and Twitter Messages to Create Population Models for Burglaries and Robberies." Cartography and Geographic Information Science 45.3 (2018): 205–220.

Krishna Iyer, P.V. "The First and Second Moments of some Probability Distributions Arising from Points on a Lattice and their Application." Biometrika 36.1–2 (1949): 135–141.

Lam, Tsit-Yuen. Introduction to Quadratic Forms over Fields. Providence: American Mathematical Society, 2005.

Lansley, Guy, Paul A. Longley. "Deriving Age and Gender from Forenames for Consumer Analytics." Journal of Retailing and Consumer Services 30 (2016): 271–278.

Lee, Sang-Il. "Developing a Bivariate Spatial Association Measure: An Integration of Pearson's r and Moran's I." Journal of Geographical Systems 3.4 (2001): 369–385.

LeSage, James P., R. Kelley Pace. "Interpreting Spatial Econometric Models." In: Handbook of Regional Science. Eds. Manfred M. Fischer, Peter Nijkamp. Berlin/Heidelberg, Germany: Springer, 2014. 1535–1552.

Lin, Jie. "A Local Model for Multivariate Analysis: Extending Wartenberg's Multivariate Spatial Correlation." Geographical Analysis 52.2 (2020): 190–210.

Lin, Liyue, Yu Zhu, Pengfei Liang, Baoyu Xiao. "The Spatial Patterns of Housing Conditions of the Floating Population in China Based on the Sixth Census Data." Geographical Research 33.5 (2014): 887–898.

Liu, Yu, Daoqin Tong, Xi Liu. "Measuring Spatial Autocorrelation of Vectors." Geographical Analysis 47.3 (2015): 300–319.

Liu, Yunzhe, Tao Cheng. "Understanding Public Transit Patterns with Open Geodemographics to Facilitate Public Transport Planning." Transportmetrica A: Transport Science 16.1 (2020): 76–103.

Longley, Paul A., Muhammad Adnan. "Geo-Temporal Twitter Demographics." International Journal of Geographical Information Science 30.2 (2016): 369–389.

Longley, Paul A., Muhammad Adnan, Guy Lansley. "The Geotemporal Demographics of Twitter Usage." Environment and Planning A: Economy and Space 47.2 (2015): 465–484.

Manley, David, Robin Flowerdew, David Steel. "Scales, Levels and Processes: Studying Spatial Patterns of British Census Variables." Computers, Environment and Urban Systems 30.2 (2006): 143–160.

Martín-Román, Ángel L., Jaime Cuéllar-Martín, Alfonso Moral. "Labor Supply and the Business Cycle: The 'Bandwagon Worker Effect'." Papers in Regional Science 99.6 (2020): 1607–1642.

Massey, Doreen. For Space. London, UK: SAGE, 2005.

Matilla-García, Mariano, Julián Rodríguez Ruiz, Manuel Ruiz Marín. "Detecting the Order of Spatial Dependence via Symbolic Analysis." International Journal of Geographical Information Science 26.6 (2012): 1015–1029.

Mei, Yingdan, Xiaoli Zhao, Lu Lin, Li Gao. "Capitalization of Urban Green Vegetation in a Housing Market with Poor Environmental Quality: Evidence from Beijing." Journal of Urban Planning and Development 144.3 (2018): 05018011.

Moran, Patrick A. P. "Random Associations on a Lattice." Mathematical Proceedings of the Cambridge Philosophical Society 43.3 (1947): 321–328.

Moran, Patrick A. P. "The Interpretation of Statistical Maps." Journal of the Royal Statistical Society: Series B (Methodological) 10.2 (1948): 243–251.

Naimi, Babak, et al. "ELSA: Entropy-Based Local Indicator of Spatial Association." Spatial Statistics 29 (2019): 66–88.

Oden, Neal. "Adjusting Moran's I for Population Density." Statistics in Medicine 14.1 (1995): 17–26.

Oliver, Margaret A. "The Variogram and Kriging." In: Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications. Eds. Manfred M. Fischer, Arthur Getis. Berlin/Heidelberg; Germany: Springer, 2010. 319–352.

Openshaw, Stan. The Modifiable Areal Unit Problem. Norwich, UK: Geo Books, 1983.

Ord, J. Keith, Arthur Getis. "Local Spatial Autocorrelation Statistics: Distributional Issues and an Application." Geographical Analysis 27.4 (1995): 286–306.

Ord, J. Keith, Arthur Getis. "Testing for Local Spatial Autocorrelation in the Presence of Global Autocorrelation." Journal of Regional Science 41.3 (2001): 411–432.

Ord, J. Keith, Arthur Getis. "Local Spatial Heteroscedasticity (LOSH)." The Annals of Regional Science 48.2 (2012): 529–539.

Pearsall, Hamil, Jillian K. Eller. "Locating the Green Space Paradox: A Study of Gentrification and Public Green Space Accessibility in Philadelphia, Pennsylvania." Landscape and Urban Planning 195 (2020): 103708.

Potoglou, Dimitris, Hanna Maoh, Yiming Wang, Scott Orford. "The Impact of Public Transport Infrastructure on Residential Land Value: Using Spatial Analysis to Uncover Policy-Relevant Processes." In: The Practice of Spatial Analysis. Eds. Helen Briassoulis, Dimitris Kavroudakis, Nikolaos Soulakellis. Cham, Switzerland: Springer, 2019. 275–293.

Purves, Ross S., Stephan Winter, Werner Kuhn. "Places in Information Science." Journal of the Association for Information Science and Technology 70.11 (2019): 1173–1182.

Ristea, Alina, et al. "Spatial Crime Distribution and Prediction for Sporting Events Using Social Media." International Journal of Geographical Information Science 34.9 (2020): 1708–1739.

Ruiz, Manuel, Fernando López, Antonio Páez. "Testing for Spatial Association of Qualitative Data Using Symbolic Dynamics." Journal of Geographical Systems 12.3 (2010): 281–309.

Rybarczyk, Greg, Syagnik Banerjee, Melissa D. Starking-Szymanski, Richard R. Shaker. "Travel and Us: The Impact of Mode Share on Sentiment Using Geo-Social Media and GIS." Journal of Location Based Services 12.1 (2018): 40–62.

Shortridge, Ashton. "Practical Limits of Moran's Autocorrelation Index for Raster Class Maps." Computers, Environment and Urban Systems 31.3 (2007): 362–371.

Shu, Hua, et al. "L-Function of Geographical Flows." International Journal of Geographical Information Science 35.4 (2021): 689–716.

Singleton, Alex, Alexandros Alexiou, Rahul Savani. "Mapping the Geodemographics of Digital Inequality in Great Britain: An Integration of Machine Learning into Small Area Estimation." Computers, Environment and Urban Systems 82 (2020): 101486.

Singleton, Alex, Daniel Arribas-Bel. "Geographic Data Science." Geographical Analysis 53.1 (2021): 61–75.

Steiger, Enrico, João Porto de Albuquerque, Alexander Zipf. "An Advanced Systematic Literature Review on Spatiotemporal Analyses of Twitter Data." Transactions in GIS 19.6 (2015): 809–834.

Steiger, Enrico, Bernd Resch, João Porto de Albuquerque, Alexander Zipf. "Mining and Correlating Traffic Events from Human Sensor Observations with Official Transport Data Using Self-Organizing-Maps." Transportation Research Part C: Emerging Technologies 73 (2016): 91–104.

Steiger, Enrico, René Westerholt, Bernd Resch, Alexander Zipf. "Twitter as an Indicator for Whereabouts of People? Correlating Twitter with UK Census Data." Computers, Environment and Urban Systems 54 (2015): 255–265.

Steiger, Enrico, René Westerholt, Alexander Zipf. "Research on Social Media Feeds – A GIScience Perspective." In: European Handbook of Crowdsourced Geographic Information. Eds. Cristina Capineri et al. London, UK: Ubiquity Press, 2016. 237–254.

Tao, Ran, Jean-Claude Thill. "Spatial Cluster Detection in Spatial Flow Data." Geographical Analysis 48.4 (2016): 355–372.

Tao, Ran, Jean-Claude Thill. "Flow Cross K-Function: A Bivariate Flow Analytical Method." International Journal of Geographical Information Science 33.10 (2019a): 2055–2071.

Tao, Ran, Jean-Claude Thill. "FlowAMOEBA: Identifying Regions of Anomalous Spatial Interactions." Geographical Analysis 51.1 (2019b): 111–130.

Tao, Ran, Jean-Claude Thill. "BiFlowLISA: Measuring Spatial Association for Bivariate Flow Data." Computers, Environment and Urban Systems 83 (2020): 101519.

Tiefelsdorf, Michael. Modelling Spatial Processes: The Identification and Analysis of Spatial Relationships in Regression Residuals by Means of Moran's *I*. Berlin/Heidelberg, Germany: Springer, 2000.

Tiefelsdorf, Michael. "The Saddlepoint Approximation of Moran's *I*'s and Local Moran's I_i's Reference Distributions and their Numerical Evaluation." Geographical Analysis 34.3 (2002): 187–206.

Tiefelsdorf, Michael, Barry Boots. "The Exact Distribution of Moran's *I*." Environment and Planning A: Environment and Space 27.6 (1995): 985–999.

Tiefelsdorf, Michael, Daniel A. Griffith, Barry Boots. "A Variance-Stabilizing Coding Scheme for Spatial Link Matrices." Environment and Planning A: Environment and Space 31.1 (1999): 165–180.

Tobler, Waldo R. "A Computer Movie Simulating Urban Growth in the Detroit Region." Economic Geography 46.sup1 (1970): 234–240.

van Dijk, Justin, Paul A. Longley. "Platial Geo-Temporal Demographics Using Family Names." In: Proceedings of the 2nd International Symposium on Platial Information Science (PLATIAL'19). Eds. Franz-Benjamin Mocnik, René Westerholt. Coventry, UK: 2020a. 23–31.

van Dijk, Justin, Paul A. Longley. "Interactive Display of Surnames Distributions in Historic and Contemporary Great Britain." Journal of Maps 16.1 (2020b): 68–76.

Waldhör, Thomas. "The Spatial Autocorrelation Coefficient Moran's *I* under Heteroscedasticity." Statistics in Medicine 15.7–9 (1996): 887–892.

Walter, Stephen D. "The Analysis of Regional Patterns in Health Data: I. Distributional Considerations." American Journal of Epidemiology 136.6 (1992a): 730–741.

Walter, Stephen. D. "The Analysis of Regional Patterns in Health Data: II. The Power to Detect Environmental Effects." American Journal of Epidemiology 136.6 (1992b): 742–759.

Wang, Chao, et al. "Railway and Road Infrastructure in the Belt and Road Initiative Countries: Estimating the Impact of Transport Infrastructure on Economic Growth." Transportation Research Part A: Policy and Practice 134 (2020): 288–307.

Wartenberg, Daniel. "Multivariate Spatial Correlation: A Method for Exploratory Geographical Analysis." Geographical Analysis 17.4 (1985): 263–283.

Westerholt, René. "The Impact of the Spatial Superimposition of Point Based Statistical Configurations on Assessing Spatial Autocorrelation." In: Geospatial Technologies

for All: Short Papers, Posters and Poster Abstracts of the 21st AGILE Conference on Geographic Information Science. Eds. Ali Mansourian, Petter Pilesjö, Lars Harrie, Ron von Lammeren. Lund, Sweden: 2018.

Westerholt, René. "Methodological Aspects of the Spatial Analysis of Geosocial Media Feeds: From Locations towards Places." gis.Science: Die Zeitschrift für Geoinformatik 31.2 (2019a): 65–76.

Westerholt, René. "Statistische räumliche Analyse in der Digitalen Transformation: Das Beispiel Geosozialer Medien." In: Geoinformationssysteme 2019: Beiträge zur 6. Münchner GI-Runde. Eds. Thomas H. Kolbe, Ralf Bill, Andreas Donaubauer. Munich, Germany: 2019b. 29–35.

Westerholt, René. "Emphasising Spatial Structure in Geosocial Media Data Using Spatial Amplifier Filtering." Environment and Planning B: Urban Analytics and City Science 48.9 (2021a): 2842–2861.

Westerholt, René. "Exploring and Characterising Irregular Spatial Clusters Using Eigenvector Filtering." In: Proceedings of the 29th Annual GIS Research UK Conference (GISRUK). Cardiff, UK, 2021b.

Westerholt, René, Franz-Benjamin Mocnik, Alexis Comber. "A Place for Place: Modelling and Analysing Platial Representations." Transactions in GIS 24.4 (2020): 811–818.

Westerholt, René, Bernd Resch, Franz-Benjamin Mocnik, Dirk Hoffmeister. "A Statistical Test on the Local Effects of Spatially Structured Variance." International Journal of Geographical Information Science 32.3 (2018): 571–600.

Westerholt, René, Bernd Resch, Alexander Zipf. "A Local Scale-Sensitive Indicator of Spatial Autocorrelation for Assessing High- and Low-Value Clusters in Multiscale Datasets." International Journal of Geographical Information Science 29.5 (2015): 868–887.

Westerholt, René, Enrico Steiger, Bernd Resch, Alexander Zipf. "Abundant Topological Outliers in Social Media Data and their Effect on Spatial Analysis." PLoS ONE 11.9 (2016): e0162360.

Wheeler, David C., Antonio Páez. "Geographically Weighted Regression." In: Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications. Eds. Manfred M. Fischer, Arthur Getis. Berlin/Heidelberg, Germany: Springer, 2010. 461–486.

Wolf, Levi John, et al. "Quantitative Geography III: Future Challenges and Challenging Futures." Progress in Human Geography 45.3 (2020): 596–608.

Xu, Min, Chang-Lin Mei, Na Yan. "A Note on the Null Distribution of the Local Spatial Heteroscedasticity (LOSH) Statistic." The Annals of Regional Science 52.3 (2014): 697–710.

Yazgi Walsh, Burcin, Chris Brunsdon, Martin Charlton. "Open Geodemographics: Classification of Small Areas, Ireland 2016." Applied Spatial Analysis and Policy 14.1 (2021): 51–79.

Zamenian, Hamed, Juyeong Choi, Seyed Amir Sadeghi, Nader Naderpajouh. "Systematic Approach for Asset Management of Urban Water Pipeline Infrastructure Systems." Built Environment Project and Asset Management 7.5 (2017): 506–517.

Zhang, Tonglin, Ge Lin. "On Moran's *I* Coefficient under Heterogeneity." Computational Statistics & Data Analysis 95 (2016): 83–94.