René Westerholt

The Analysis of Spatially Superimposed and Heterogeneous Random Variables

March 2018

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

Acknowledgements

My special thanks go to **Professor Alexander Zipf**, who has supported and accompanied me over the past five years. He enabled me not only to do my doctoral thesis, but also to develop a number of other qualifications. This includes a considerable track record in teaching and a successful research proposal, among many other activities. A big thank-you goes to **Professor João Porto de Albuquerque**. Besides being my second adviser, he challenged me with inspiring and rigorous discussions from viewpoints well beyond my own technical expertise during his many research stays in Heidelberg. He further made it possible for me to convene a workshop with him at a major European conference, and provided a lot of support during the last stage of my thesis. My academic career so far would hardly have been possible without **Professor Bernd Resch**, who provided me mental support and motivation since he co-supervised my master's thesis in Osnabrück. Bernd also made it possible to go abroad for a research stay at Harvard University, which was clearly one of my personal highlights during my PhD project. Dear Bernd, I owe you a lot! A heartfelt thank you goes to **Dr Franz-Benjamin Mocnik**, who supported me with valuable thoughts and discussions and–more importantly–provided me with a considerable amount of patience and motivation, which I clearly needed during the last mile of my doctorate. **Dr Andreas Reimer**, **Dr Enrico Steiger** and **Dr Martin Hämmerle** accompanied me in a special way over long periods of time and were three of my most important colleagues. Dear Andi, my figures, charts and maps would not even look half as good without your guidance and support for good visualisations. My dear Enrico and Martin, I am grateful that we have successfully mastered the CrowdAnalyser together! Along with many other esteemed colleagues, I would like to express my special thanks to **Carolin Klonner**, **Sabrina Marx**, **Robin Peters**, **Dr Tessio Novack**, **Dr Adam Rousell**, **Dr Ming Li**, **Jan Kinne** and **Heinrich Lorei**, who have provided me with professional support and motivation over many years and have provided me with a lot of fun. I would like to thank **Professor Bernhard Höfle** for his strategic support and valuable discussions. A hearty thank you goes to **Bettina Knorr**. Bettina–I am sure that my doctorate would have lasted another five years without your competent, warm-hearted and well organized administrative and personal support! **Bastian Zimmer** has been my best and most valued friend for many years. Dear Bastian, even though you are geographically far away, you are always there for me when I need a helping hand and an open ear! There has been no greater fun and relaxation from PhD stress than taking a hike or city trip with you! Thank you for all your support over many years, even in difficult times! Last but not least, I would like to thank **Helga** (my mom) and **Pierre** (my brother) along with other **family members** who always believed in me, made everything possible and supported me whenever I needed it most!

My sincerest gratitude to all of you!

René Westerholt

Heidelberg, 30 March 2018

Abstract

Modern society produces spatiotemporal datasets at an unprecedented velocity and volume. People are leaving their digital traces when they tap their oyster cards to ride the London underground and when they share their brand preferences through paying with loyalty cards. These examples show how deeply digital services are now integrated into people's everyday lives. For this reason, geosocial media feeds, one type of everyday digital datasets, have recently gained considerable attention in academic research. The users who contribute data to these feeds collect subjective impressions about places and social events in a proactive yet often unconscious manner. Numerous examples are found in geography where social media data has been used to investigate human mobility, social livelihoods, or for other purposes, and a range of novel insights were achieved this way.

One step in the investigation of geosocial media content is the spatial analysis of the users' data records. It allows to assess spatial relations and the contextualization of the messages by relating them to covariates. However, the spatial analysis of this kind of data turns out to feature a number of formidable methodological challenges: the relationship between messages and places is often elusive, and issues like the self-selection bias and semantic ambiguities complicate the analysis. Probably the most problematic effect is that variegated phenomena are reflected in the data simultaneously, leading to the notion of spatially superimposed heterogeneous random variables. Available spatial analysis methods are not able to address these characteristics. We thus need a thorough understanding of their effects, and novel ways to investigate the fine-grained organization of places by, for instance, using geosocial media data.

The research presented in this thesis is located at the interface between spatial analysis methodology and the characteristics of spatially superimposed random variables. Three types of contributions are presented: (i) the interactions of spatial analysis techniques with spatially superimposed random variables are investigated; (ii) novel methods for their analysis and characterization are put forward; and (iii) the broader context of the discussed matters is explored, including a discussion of similar methodological issues in different fields. Three datasets are employed: two real-world Twitter samples covering different temporal scales, and one idealized synthetic dataset representing the characteristics of spatially superimposed random variables.

The empirical contribution focuses on estimators of spatial autocorrelation and hot-spot statistics. Thereby, the impacts of mixed geographic scales and adverse topological arrangements on spatial analysis results are investigated. It is found that geosocial media datasets contain a mixture of geometric scales, and that smaller scales dominate analysis results. This leads to biased hot-spot statistics which are inflated by false positives. Estimators of spatial autocorrelation show indications for false patterns that result from adverse topological arrangements caused by the representation of various phenomena in the data. The joint influences of scale and topology show a complex interplay leading to unpredictable behaviours, further complicating result interpretation.

The methodological contribution of this thesis is two-fold: First, a modified hot-spot statistic is proposed that takes account of the geometric characteristics of superimposed random variables. This statistic allows the detection of hot-spots when multiple scales are present in datasets. Second, a statistical test is derived that allows to investigate the local interactions between the arrangement of random variables in geographic space and their local variance. The latter test can be used to investigate how places, like those represented in geosocial media, are characterized in terms of their endogenous variability.

Exploring the broader context reveals that questions in the analysis of spatially superimposed random variables are not limited to geosocial media, but extend to other areas such as socio-ecological psychology. This latter field features novel techniques such as the so-called event sampling method, which allow the collection of geotagged *in situ* surveys under contextual conditions. These raise methodological questions similar to those around geosocial media, which renders a common research agenda meaningful. Furthermore, it is discussed how, despite of the outlined issues that are investigated in this thesis, some of the available spatial analysis methods have still been applied successfully in the empirical literature.

In summary, the presented findings will allow gaining a better understanding of the spatial organization of the digital representations of places on geosocial media feeds. Further, the obtained methodological results are an important step towards the notion of a place-based GIS, which is a long-term goal in GIScience. Ultimately, this will increase our understanding of human geographic everyday behaviours.

Kurzzusammenfassung

Unsere moderne Informationsgesellschaft produziert fortlaufend raumzeitliche Datensätze in einer nie dagewesenen Geschwindigkeit und Menge. Menschen hinterlassen ihre digitalen Spuren bei der Nutzung des öffentlichen Personennahverkehrs mit E-Tickets oder bei der Preisgabe ihrer Markenpräferenzen über den Einsatz von Kundenkarten. Diese Beispiele demonstrieren die flächendeckende und tiefe Verankerung digitaler Technologien im Alltag. Geosoziale Medien haben dabei jüngst eine besondere Aufmerksamkeit in der akademischen Forschung erfahren. Nutzer hinterlassen proaktiv, jedoch oft unbewusst, subjektive Eindrücke und Meinungen über Orte und soziale Ereignisse. In der Geographie erscheinen in jüngerer Zeit zahlreiche Arbeiten, die diese Daten für die Untersuchung menschlicher Mobilität, sozialer Aktivitätsräume, oder für andere Zwecke nutzen. Auf diese Weise wurden bereits zahlreiche neue Erkenntnisse über die räumliche Organisation des Alltagslebens unserer Gesellschaften gewonnen.

Ein wichtiger Schritt in der Untersuchung der Inhalte geosozialer Medien besteht in deren räumlicher Analyse. Diese erlaubt etwa die Untersuchung regionaler Verflechtungen und eine räumliche Kontextualisierung der Daten. Dabei treten jedoch veritable methodische Herausforderungen auf: die Beziehung zwischen einer Nachricht und dem zugehörigen Ort ist oft nicht eindeutig bestimmbar und Probleme wie Selbstselektivität oder die Ambiguität der semantischen Beiträge verkomplizieren räumliche Analysen. Das vermutlich gravierendste Problem besteht jedoch in der räumlich und zeitlich koinzidenten Repräsentation verschiedener Phänomene, was zum Begriff räumlich überlagerter und heterogener Zufallsvariablen führt. Vorhandene räumliche Analysemethodiken sind nicht in der Lage die Eigenschaften solcher Zufallsvariablen zu berücksichtigen. Erkenntnisse über deren Auswirkungen sowie neue methodische Ansätze sind deshalb notwendig, um detaillierte Erkenntnisse über die räumliche Organisation von Orten mittels Daten sozialer Medien zu erzielen.

Die in dieser Arbeit präsentierte Forschung befasst sich mit der Schnittstelle zwischen räumlicher Analysemethodik und den Eigenschaften der genannten räumlich überlagerten Zufallsvariablen. Drei Arten wissenschaftlicher Beiträge werden dabei gemacht: es werden (i) empirische Erkenntnisse hinsichtlich der Interaktion zwischen Methodik und Dateneigenschaften erzielt; (ii) methodische Beiträge zur Berücksichtigung und weitergehenden Charakterisierung dieser Eigenschaften präsentiert; und (iii) diese Erkenntnisse in einen breiteren Kontext eingebettet, der über die Geographie hinausgeht. Drei Datensätze kommen dabei zum Einsatz: zwei extrahierte Twitterdatensätze auf unterschiedlichen Zeitskalen, sowie ein synthetisch erzeugter Datensatz. Letzterer reflektiert die Eigenschaften räumlich überlagerter heterogener Zufallsvariablen in idealisierter Form.

Der empirische Beitrag hat die Untersuchung von Schätzern räumlicher Autokorrelation und Hot-Spot-Statistiken zum Thema. Dabei werden die Einflüsse überlagerter räumlicher Maßstäbe und durch Heterogenität verursachter topologischer Anordnungen auf die Ergebnisse dieser Methoden untersucht. Ein Resultat ist, dass in sozialen Medien eine Mischung unterschiedlicher Maßstäbe zu finden ist, wobei kleine Maßstäbe die Daten dominieren. Dies führt zu verzerrten Hot-Spot-Schätzungen, was wiederum eine erhöhte Zahl falsch-positiver Resultate nach sich zieht. Schätzer räumlicher Autokorrelation reagieren auf die Eigenschaften räumlicher überlagerter Zufallsvariablen mit Indikationen fälschlicher räumlicher

Strukturen, was wiederum auf die topologischen Anomalien der Überlagerung von Phänomenen zurückzuführen ist. Der gemeinsame Effekt von Topologie und Maßstabsvariation besteht in einem schwer vorhersagbaren und komplexen Verhalten in der Schätzung räumlicher Statistiken.

Die methodische Kontribution dieser Arbeit besteht aus zwei Teilen: Zum Einen wird ein modifizierter Schätzer für Hot-Spots entwickelt. Dieser ist in der Lage, verschiedene simultan vorhandene Maßstäbe zu behandeln und erlaubt deren getrennte Analyse. Zum Anderen wird eine Teststatistik hergeleitet, welche die Untersuchung von Wechselwirkungen zwischen geographischer Anordnung und lokaler Varianz erlaubt. Auf diese Weise ist es möglich, Orte, wie sie etwa in geosozialen Medien repräsentiert sind, über das Zusammenspiel ihrer Varianz und ihrer räumlichen Konfiguration zu beurteilen.

Die erzielten Ergebnisse deuten darauf hin, dass die aufgezeigten Problematiken von genereller Natur und somit nicht auf geosoziale Medien beschränkt sind. Ein Beispiel für einen verwandten Bereich mit ähnlichen Problematiken stellt die sozialökologische Psychologie dar. Neuere Methoden dieser Disziplin, wie zum Beispiel ereignisbasierte Umfragen die im realen Umgebungskontext in georeferenzierter Form erhoben werden, weisen ähnliche Eigenschaften wie Daten aus geosozialen Medien auf. Diese Ähnlichkeiten werden in der Arbeit herausgearbeitet. Ein Resultat hiervon ist, dass sich eine multidisziplinäre Betrachtungsweise und eine gemeinsame zukünftige Forschungsagenda zur effizienten Problemlösung anbieten. Darüber hinaus werden auch Fälle präsentiert, in denen die Anwendung räumlicher Analysemethoden erfolgreich zu sinnhaften Erkenntnissen geführt hat.

Zusammenfassend bietet diese Arbeit Erkenntnisse und Ansätze zum besseren Verständnis der räumlichen Organisation digital repräsentierter Orte. Außerdem stellen die entwickelten Methoden einen wesentlichen Beitrag hin zu einer ortsbasierten Analyse und einem ortsbasierten GIS dar—ein langfristiges und noch nicht erreichtes Ziel in der Geoinformatik. Die Ergebnisse dieser Arbeit werden somit nachhaltig zum besseren Verständnis menschlicher Raumnutzung im Alltagsverhalten beitragen.

Contents

List of Figures

List of Tables

List of Abbreviations

2D	Two-dimensional
3D	Three-dimensional
AAPOR	American Association for Public Opinion Research
AGI	Ambient geospatial information
API	Application programming interface
ATLS	Automatic terrestrial laser scanning station
BKG	Bundesamt für Kartographie und Geodäsie
CA	California
CBD	Central business district
CGI	Contributed geographic information
CCGI	Citizen-contributed geographic information
DBSCAN	Density-based spatial clustering of applications with noise
DGP	Data-generating process
ESM	Event sampling method
ESOMAR	European Society for Opinion and Market Research
ESS	European Social Survey
FDR	False-discovery rate
GESIS	Gesellschaft Sozialwissenschaftlicher Infrastruktureinrichtungen
GIS	Geographical information system
GIScience	Geographic information science
GPS	Global positioning system
GWR	Geographically weighted regression
HH	High-high interaction
HL	High-low interaction
i.i.d.	independent and identically distributed
iVGI	Involuntarily volunteered geographic information
LBSM	Location-based social media
LBSN	Location-based social networks
LDA	Latent dirichlet allocation
LiDAR	Light detection and ranging
LISA	Local indicator of spatial association
LOSH	Local spatial heteroscedasticity
LH	Low-high interaction
LL	Low-low interaction
LSD	Local spatial dispersion
LSI	Latent semantic indexing
MAUP	Modifiable areal unit problem
NCAA	National Collegiate Athletic Association
OSM	OpenStreetMap

RatSWD	Rat für Sozial- und Wirtschaftsdaten
RoI	Region of interest
RQ	Research question
SMS	Short message service
SVD	Singular value decomposition
TF–IDF	Term frequency–inverse document frequency
UGC	User-generated content
UK	United Kingdom
US	United States
VGI	Volunteered geographic information
VMAP	Vector map project
ZIP	Zone improvement plan

Part I

Summary and Conclusions

I.1 Introduction

A number of novel user-generated and georeferenced datasets have recently become available to geographical research, including those from geosocial media feeds. On the one hand, these datasets are a direct result of technological developments that have led to an omnipresence of mobile and networked devices. On the other hand, they result from a conflation of data usage and production (the *prosumer* and the *produser* paradigms, see Haklay et al. 2008; Coleman 2009; Ritzer et al. 2012). The related web services used to collect the data are characterized by a strong integration into everyday activities. Global communication and self-portrayal via geosocial media feeds have become routine behaviours, so that these feeds now collect geotagged and timestamped data reflecting various aspects of the lives of ordinary people.

Geosocial media feeds complement Al Gore's vision of a *digital earth* (Gore 1998; Craglia et al. 2012)—a 'digital skin' that mirrors information from and about our planet. It adds a layer depicting the everyday individual and collective social spheres. Recently and with increasing availability through accessible *application programming interfaces* (APIs), researchers are also using this kind of data for scientific studies. Examples are found across the empirical disciplines including geography, where user-generated content is leveraged in investigations of human mobility behaviours, to disclose geotemporal demographics and in numerous other fields (*cf.*, Mitchell et al. 2013; Steiger et al. 2016b; Longley and Adnan 2016). Although these novel data have provoked some criticism, especially with regard to their uncritical use (Lazer et al. 2014; Kitchin 2014; Rae and Singleton 2015), geosocial media data still provide information that would otherwise not be available for scientific exploitation.

Geosocial media datasets are collective and subjective in nature. People communicate their opinions and impressions about places, about activities happening therein, and about their personal feelings. However, people who are leaving their digital spatiotemporal trails are often not aware that their contributions are not only stored as individual data records, but form part of collective databases into which their contributions are fed through the underlying socio-technical applications. The collected data can then be used to derive detailed usage patterns from the users' everyday lives. The implied lack of user awareness clearly raises privacy concerns (Tasse et al. 2017). Yet, it also offers the compelling advantage of collecting data from and about people interacting in their daily environmental contexts in a rather natural way, though issues like self-selection bias or semantic ambiguities can still be reflected in the data (Sengstock and Gertz 2012; Tufekci 2014). In contrast to traditional forms of data acquisition like surveys or interviews where people are exposed to an unnatural setting, the everyday embedding of geosocial media leads to an inherent normality in using the outlined services, which reveals otherwise unavailable information making the related data appealing to academic research.

Data from geosocial media feeds allow detailed investigations of the everyday behaviours of people. The entirety of these behaviours forms the *everyday geography* of an individual which circumscribes the "sum total of a person's first-hand involvements with the geographical world in which he or she lives" (Seamon 1979, pp. 15-16). The concept of everyday geography describes the social practices that occur in so-called *activity spaces* (Horton and Reynolds 1971), which represent the fraction of earth's surface that is utilized by an individual. However, the analysis of everyday geographies and their associated places through geosocial media data is challenging for two main reasons: (i) The conditions under which the user-generated information is created are often intractable and remain elusive (the 'uncertain geographic context problem'; Vich et al. 2017); (ii) people contribute their data largely independent of each other, leading to a number of variegated subjective momentary assessments at similar times and places. These two issues indeed reflect the very nature of individual everyday geographies, but they also lead to technical challenges.

Geosocial media datasets thus represent different *data-generating processes* (DGPs, the 'true models'; Hammervold and Olsson 2012) simultaneously, each of which is associated with specific users and their idiosyncratic contexts. This type of data can therefore be seen as a collection of spatiotemporal slices, each describing a snapshot of the specific *in situ* and personalized geographical conditions in certain places at specific times. Because multiple people contribute their subjective impressions, these space-time slices are in themselves inhomogeneous in a potentially indivisible manner. This inherent diversity of geosocial media data hinders the achievement of substantial and rigorous geographical insights. However, it also provides surplus to geographical analysis, because it implies an ample amount of detailed information about places and situations from different perspectives.

The assessment of spatial relationships through spatial analysis techniques forms one important part in characterizing human everyday geographies. It allows the disclosure of geographical activity clusters and the identification of contextual influences (Unwin 1996; Fischer and Getis 2010b). One key assumption in spatial analysis is the notion of spatial uniformity of the observed spatial processes in null models of no spatial effects. This is formalized by the concept of *stationarity*, of which different forms exist. The notion of second-order stationarity is applied most frequently in spatial analysis, and it refers to constant means and variances in the observation area. This implies consistent covariance-based spatial relationships that are independent of their absolute locations and depend only on the interdependence structure between spatial units (Gaetan and Guyon 2010).

In the view of the above discussion, second-order stationarity is an unrealistic assumption with regard to geosocial media data. The observed processes vary depending on the individual, and are influenced by idiosyncratic and subjective local contexts. This inevitably leads to *spatial heterogeneity* (*i. e.*, unstable statistical parameters) (Anselin 1990), which is the most devastating disturbance to spatial analysis because it invalidates globally estimated parameters and null models used for drawing inferences (Griffith and Layne 1999). Further, because established stationarity concepts describe uniformity over spatially disjoint units, these are not capable to depict the internal inhomogeneity that is found within locations and time slots in a mixed manner like in geosocial media feeds. These conditions are summarized and discussed in this thesis under the term *spatially superimposed and heterogeneous random variables*.

The interface between spatially superimposed heterogeneous random variables and spatial analysis methodology sets the frame for this thesis, which is organized into three parts. Part (i) investigates in which ways the geographically superimposed manner of individual geosocial media contributions impacts spatial analysis results in the sense of invalidating or impeding drawn conclusions. The discussion is constrained to Twitter data and the focus is on measures of spatial autocorrelation and hot-spot techniques, two commonly applied types of methods that are relevant to empirical research. Part (ii) makes a methodological contribution by introducing two novel techniques: One is a modified hot-spot technique that takes account of the data characteristics of spatially superimposed random variables. The second proposed method is a statistical test of the links between the variance and the spatial layout in strictly local circumstances, allowing detailed investigations of the heterogeneity associated with local spatial superpositions of random variables. Finally, part (iii) discusses the implications of the achieved results with regard to their relevance beyond the field of spatial analysis. This includes a discussion of how spatial analysis techniques have been applied to geosocial media data in empirical studies. Further, a connection to recent methodological discussions in socio-ecological psychology is made. The latter underpins the relevance of the results presented in this thesis and is aimed to start a broader interdisciplinary discussion.

I.1.1 Presumptions

The research presented in this thesis and the formulation of viable research questions require some initial presumptions to be made. These are supported by the literature and the discussions in the introductory part above:

- *The data published on geosocial media capture snapshots of people's everyday behaviours.* Geolocated media and GIS technology have become deeply integrated with daily routines, and Sui and Goodchild (2011) have conjectured a continuation of this trend.
- *Geosocial media is influenced by contextual conditions.* The posting of messages depends on intrinsic contextual circumstances including physical and social conditions, activity contexts into which users are engaged and individual incentives (Zimmermann et al. 2007).
- *Geosocial media data provides information about the whereabouts of people.* Empirical research shows evidence that for instance Twitter data provides information about the whereabouts of people at the neighbourhood scale or larger (Steiger et al. 2015b; Ratnasari et al. 2016; Steiger et al. 2016a). It is thus reasonable to assume a relation between the whereabouts of users and the posted contents also at smaller geographic scales.
- *Users contribute their messages independent of each other.* People are largely independent individuals during the use of geosocial media. Beyond societal trends, the postings of other users usually have little influence on a user's decision to post contents on geosocial media.
- *The textual content of user contributions can be uncertain and vague.* People have different spatial perception characteristics (*cf.*, Wender et al. 2002), are prone to differing contextual settings (Zimmermann et al. 2007; Crampton et al. 2013) and follow individual incentives (Ames and Naaman 2007; Cuel et al. 2011; Oh and Syn 2015). This introduces uncertainty into geosocial media contents.
- *The geotags attached to geosocial media data are obtained by built-in smartphone GPS receivers or by network positioning techniques.* The contents are provided with distinct coordinates. However, because these are influenced by GPS inaccuracies, topography, and other factors (see Zandbergen 2009; Carrion et al. 2017), the coordinates contain modest measurement errors.
- *Geosocial media data is often pre-processed before spatial analysis is conducted.* Typical pre-processing steps include natural language processing and noise reduction, which is evident from the empirical literature (*e. g.*, Bakillah et al. 2015; Lansley and Longley 2016; Steiger et al. 2016b).
- *Geosocial media datasets are non-stationary in the traditional sense and, in addition, violate stationarity assumptions within the observed locations.* Geosocial media datasets provide spatially superimposed and variegated georeferenced random variables. Available spatial analysis methodologies do not take sufficient account of this circumstance.

In addition to these presumptions, the employed use cases in the following research are restricted to Twitter data. This means that many of the exemplary results reflect the characteristics of digital geolocated narratives, including the demographics of the users. Photo-sharing platforms like Flickr or check-in sites such as Foursquare are not considered in this thesis and might show slightly different data characteristics.

I.1.2 Research Questions

Two major and two minor research questions are formulated under the presumptions given above. The main questions RQ 1 and RQ 2 are too complex to be answered in a single step. Therefore, these are subdivided into more detailed sub-questions. The minor questions RQ 3 and RQ 4 offer a broader context and provide links to neighbouring fields.

RQ 1: How do spatially superimposed random variables affect spatial analysis results?

The analysis of spatially superimposed random variables often includes spatial analysis methods. Recent examples for their application to geosocial media data are found in Frank et al. (2013) (radius of gyration, Moran's I, Geary's c), Cheng and Wicks (2014) (spatiotemporal scan statistics) and Luo et al. (2016a) (space-time trajectory analysis). Traditional application areas of spatial analysis include census records and data collected in a controlled way for scientific purposes. However, geosocial media data is collected largely unsystematically, so that one questionable point in their analysis is the assumption of second-order stationarity (see I.3.3), as it is implied by many spatial analysis methods. Strategies are available to deal with the violation of this assumption, including the analysis of first and second-order derivatives (Lillesand et al. 2015; Gelfand and Banerjee 2015), spatially-constrained analysis approaches (Patil et al. 2006; Bravo and Weber 2011) or the use of adapted estimators (Ord and Getis 2001). However, these strategies assume processes to be heterogeneous in disjoint sub-regions, but to be homogeneous within the individual locations. The inherent heterogeneity of (non-identical) spatially superimposed random variables infringes this condition and prevents the application of the strategies outlined above to compensate for the violation of stationarity. One of the main objectives of this thesis is therefore to explore how the violation of stationarity assumptions affects spatial analysis results.

RQ 1.1: In what ways are hot-spot estimations affected by different overlaid spatial scales?

Hot-spot techniques allow to disclose regions where extremal values accumulate in geographic space. These are important to a range of theoretical and practical matters such as cluster detection or the identification of social, biological or similar activity centres. However, what complicates this kind of analysis with geosocial media data is that people contribute observations independently and that there are no restrictions on semantic subjects. In consequence, multiple phenomena at different and spatially mixed scales are reflected at the same times and locations. Further, people use different idiosyncratic concepts of scale when they contribute subjective impressions about the same whereabouts, leading to variegated spatial conceptualizations (Wender et al. 2002; Dangschat 2007). This is in stark contrast to data collected specifically for geographic research, where the acquisition scale is adjusted to a single coherent phenomenon and where no subjectively perceived spatial scales are included. Contrarily, when it comes to the interpretation of analysis results obtained from superimposed random variables and to inferences about these, it is important to be aware of potential conflicts and cross-scale interactions that can corrupt the obtained results. Question 1.1 thus explores how geosocial media datasets are composed with respect to contained geographic scales, and how that in turn impacts the estimation of hot-spots.

RQ 1.2: How do topological and statistical effects of spatial overlap influence spatial autocorrelation?

Data characteristics strongly influence the estimation of spatial autocorrelation. It is well-known that Moran's I (see I.3.2) converges to normality faster when random variables are approximately normal or at least follow a symmetric distribution (Griffith 2010). The estimator is also sensitive to topological imbalances in the geographic layout through unequal connectivity degrees (Tiefelsdorf and Boots 1997; Tiefelsdorf et al. 1999; Shortridge 2007) and requires a minimum variability within the investigated random variables to provide sufficient statistical power (Walter 1992a; Walter 1992b). By analogy, the variances of the underlying DGPs must be uniform to draw reliable inferences about Moran's *I* (Oden 1995; Waldhör 1996; Shen et al. 2016). These findings demonstrate strong indications that geosocial media data characteristics—like the outlined mixture of spatial scales and other implications caused by the heterogeneity inherent to geosocial media data—influence results from spatial analysis methods. Specifically, it is another goal of this thesis to investigate the influence of topological outliers on the disclosure of spatial structure by Moran's *I*. The liberal data acquisition schemes of geosocial media feeds are likely to cause an adverse placement of messages in geographic space, rendering the investigation of topology and connectedness relevant. Additionally, the influences of different statistical parameter values between overlapping DGPs is investigated. These cause a mixture distribution, but the role of the spatially overlapping geometric setup is yet unclear, and thus another objective of this research question.

RQ 1.3: Which influences do joint topological and scale effects have on spatial autocorrelation?

The combination of different effects can intensify their individual impacts on spatial analysis results. For instance, the effects of endogenous spatial dependence and exogenous spatial heterogeneity on regression analysis are problematic in terms of violating technical requirements (Anselin and Griffith 1988; Porojan 2001). However, their joint effects are even more severe and make it difficult to specify a correct model. It is possible in these cases that regression residuals might falsely appear to be spatially structured even though that might be a mere artefact caused by the joint effect of the two simultaneous issues. Another example is the influence of coexisting local and regional effects when assessing spatial structures (Ord and Getis 2001; Johnson et al. 2013). Neglecting their interdependence leads to an inflation of the variance of test statistics and to a wrong specification of null models, and thus to misleading conclusions about spatial patterns. These two examples make clear how important it is to take account of the interaction of different data characteristics. Thus, a third goal pertaining to main question 1 is to explore how scale-related (RQ 1.1) and topological issues (RQ 1.2) together influence the estimation of spatial autocorrelation by the example of Moran's *I*.

RQ 2: How to identify and characterize spatial structures in superimposed random variables?

Spatial structures are assessed and characterized by first and second-order spatial data characteristics. First-order characteristics describe the intensity, whereas second-order characteristics assess interactions within random variables (Fischer and Getis 2010b). Both are of interest to a range of scenarios like cluster detection or the characterization of stochastic processes (a list of possible applications is found in Getis 2007, p. 494). However, as discussed above, existing methods are not suitable for superimposed random variables. Therefore, the focus of this question is on the derivation of novel (i) hot-spot measures and (ii) measures that allow to investigate the relationships between geographic arrangements and the magnitude of local variances. The latter is related to spatial autocorrelation, but can additionally be used to characterize the variability within superimposed random variables in relation to their spatial organization, and thus to obtain a clearer characterization of places. A more detailed breakdown into specific sub-questions is introduced below.

RQ 2.1: How can hot-spots at different, geographically coexisting spatial scales be disclosed separately?

Matching the analysis and the phenomenon scale is crucial for identifying meaningful structures (Goodchild 2001). Hot-spot estimators are particularly prone to scale misspecifications, because these measures accumulate local data points in an additive manner. Especially when the analysis scale is coarse, results are then quickly biased by including unrelated samples. The scale of an analysis can be controlled by adjusting distance parameters (*e. g.*, in analyses of geographically continuous phenomena; Gneiting and Guttorp 2010), or by calibrating a so-called spatial weights matrix (when spatially discrete phenomena are analysed; Pace and LeSage 2010)). Because geosocial media feeds are indexed over discrete geographic units, it is required to adjust a matrix of pairwise relationships between the sampled locations. A multitude of different matrix designs is available (*cf.* Aldstadt and Getis 2006; Mawarni and Machdi 2016; Ermagun and Levinson 2017), but no approach exists that acknowledges the specific spatial characteristics of superimposed random variables. One aim of RQ 2.1 is therefore to establish a spatial weighting scheme that takes account of the representation of various spatially overlapping scales in data. Further, because geosocial media comprises multiple scales at the same time, a second objective is to adapt the popular hot-spot technique G_i^* (Getis and Ord 1992; Ord and Getis 1995) towards the disclosure of hot-spots at different scale levels in a separated manner.

RQ 2.2: How can the influence of spatial structures on local variance be tested in non-stationary superimposed random variables?

The investigation of the spatial heterogeneity of superimposed random variables is useful for a better understanding of this type of data. Spatial heterogeneity is a proxy of the spatial instability in random variables (Dutilleul and Legendre 1993). In traditional datasets this refers to instability in statistical moments over an observation area. In case of superimposed random variables, spatial heterogeneity should additionally be investigated locally within the locations, which allows detailed characterizations of how the local spatial arrangement of random variables influences their diversity. In turn, this helps to better understand the spatial organization of places. The analysis of spatial structure in variance can for instance be conducted by a recently published technique called *Local Spatial Heteroscedasticity* (LOSH; Ord and Getis 2012; Xu et al. 2014a) which allows to investigate how spatial structures affect the magnitude of the variance. Application areas of this method include investigations of the internal structure of clusters or the detection of geographic boundaries (*e. g.*, Getis (2015)). However, LOSH is not capable of dealing global spatial heterogeneity in the null model, as it cannot take account of strong differences in the dispersal behaviours of different random variables. Further, the measure cannot characterize the relation between structure and variance in superimposed scenarios, because LOSH identifies the globally dominating heterogeneous features instead of fine-grained local spatial variance structures. The latter is required for the aforementioned characterizations of places in geosocial media data. RQ 2.2 therefore objectifies to modify LOSH (including a suitable inference mechanism) to make it applicable to spatially superimposed random variables in a globally heterogeneous setting.

RQ 3: How are superimposed random variables analysed spatially in empirical studies?

Despite the methodological challenges outlined above, numerous empirical spatial analyses of superimposed random variables have been conducted. The question arises as to how these are carried out if the available methods are not tailored to the requirements of this type of data. This entails a discussion of typically used methods and ways how the high degree of heterogeneity of the data is considered. Beyond a review of methods and approaches, this research question also touches upon the interpretation and the validity of obtained results. These points are discussed at various scales and include the individual as well as the collective level in terms of the geosocial media users. Several selected case studies from human mobility research—a field in which geosocial media data is frequently used—are reviewed with regard to their spatial analysis approaches applied. Further, the potential of using this kind of data as well as future research topics are elaborated. Another focal aspect of this research question is a more general discussion of the application areas of geosocial media data in the spatial context. These largely include spatial analysis approaches, which connects them to the other research questions in this thesis.

RQ 4: To what extent are spatially superimposed random variables from other research fields related to geosocial media data?

Spatially superimposed heterogeneous random variables are not exclusively considered in geography and spatial analysis. A field with very similar types of data is socio-ecological psychology (Rentfrow 2013; Oishi 2014). This branch, which is also known as geographical psychology, explicitly considers geographical and other contextual effects in the design and interpretation of surveys and of other data acquisition techniques. One important methodology in this field is the so-called *event sampling method* (ESM; Reis and Gable 2000). Thereby, surveys are conducted under *in situ* conditions and the respondents answer questions while being engaged in everyday activities and exposed to contextual factors. This results in answers which are less affected by interviewer-induced effects, but on the other hand raises similar questions as with geosocial media data. ESM responses form superimposed random variables, even though the sample sizes are typically smaller than in case of geosocial media data. In answering this research question, parallels are drawn between the main results of this thesis (RQ 1 and RQ 2) and ESM responses. Further, it is discussed how contextual factors and subjective as well as cognitive influences impact the data, with an emphasis on their spatial characteristics. The achieved results for RQ 4 further strengthen the relevance of the other results achieved in this thesis, as they show the importance of these for methodological research areas beyond geography.

I.1.3 Structure and Context of the Research

As in statistics in general, contributions to spatial analysis must take account of methods, data characteristics, research fields and application scenarios simultaneously. The outlined research questions of this thesis cover the entire breadth of these elements through touching upon a range of related aspects. Figure I.1.1a illustrates that each of the four research questions is associated with one of the mentioned and interrelated components of spatial analysis. This links the research questions and shows that this work offers a comprehensive treatment of the presented topic. Further, I.1.1b makes clear that the thesis still has a clear focus through classifying the accompanying publications into four quadrants spanned by the research questions. The accumulation in the first quadrant shows that the main contribution is situated at the junction between methodology and data characteristics, whereas the other dimensions are treated too, but are covered less extensive. The obtained results therefore provide a multifaceted treatment of the analysis of spatially superimposed and heterogeneous random variables, while making a strong methodological contribution at the same time.

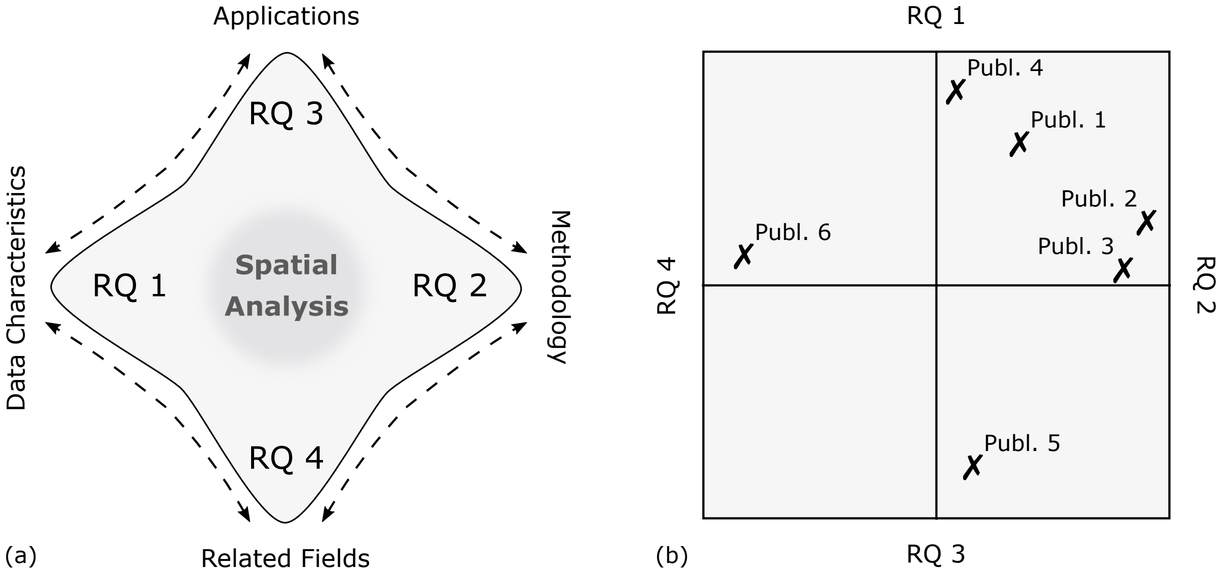


Figure I.1.1: Structure and context of the research in this thesis. (a) Placement in the subject-specific context. (b) Assignment of publications to research questions.

The findings presented hereafter in Sections I.4 and I.5 are based on the following publications, which are attached in Part II:

Publication 1: Westerholt, R, Resch, B & A Zipf (2015). 'A Local Scale-Sensitive Indicator of Spatial Autocorrelation for Assessing High- and Low-Value Clusters in Multi-Scale Datasets'. *International Journal of Geographical Information Science*, 29 (5), 868-887. DOI: 10.1080/13658816.2014.1002499.

Publication 2: Westerholt, R, Steiger, E, Resch, B & A Zipf (2016). 'Abundant Topological Outliers in Social Media Data and Their Effect on Spatial Analysis'. *PLOS ONE*, 11 (9), e0162360. DOI: 10.1371/journal.pone.0162360.

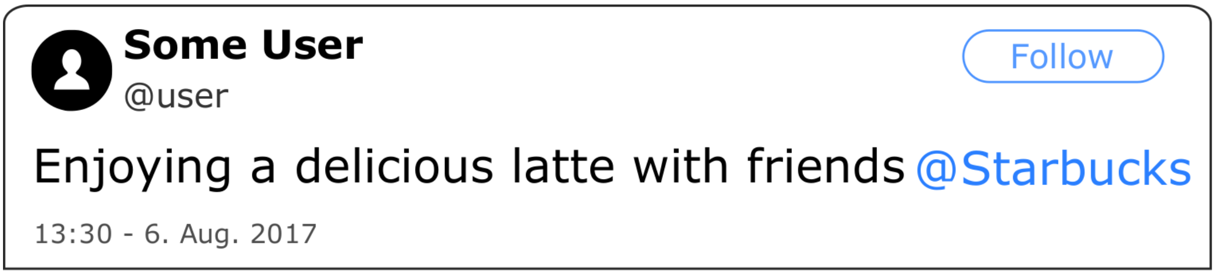
Publication 3: Westerholt, R (submitted, under review). 'The Impact of Different Statistical Parameter Values between Point Based Datasets when Assessing Spatial Relationships'. *Proceedings of the 21st AGILE Conference*.

Publication 4: Westerholt, R, Resch, B, Mocnik, F.-B. & D Hoffmeister (2018). 'A Statistical Test on the Local Effects of Spatially Structured Variance'. *International Journal of Geographical Information Science*, 32 (3), 571-600. DOI: 10.1080/13658816.2017.1402914.

Publication 5: Bluemke, M, Resch, B, Lechner, C, Westerholt, R, & JP Kolb (2017). 'Integrating Geographic Information into Survey Research: Current Applications, Challenges and Future Avenues'. *Survey Research Methods*, 11 (3), 307-327. DOI: 10.18148/srm/2017.v11i3.6733.

Publication 6: Steiger, E, Westerholt, R, & A Zipf (2016). 'Research on Social Media Feeds – A GIScience Perspective'. In: *European Handbook of Crowdsourced Geographic Information*. Ed. by Capineri, C, Haklay, M, Huang, H, Antoniou, V, Kettunen, J, Ostermann, F & R Purves. London: Ubiquity Press, 237-254. DOI: 10.5334/bax.r.

I.2 Geosocial Media


Some User
@user

Enjoying a delicious latte with friends @Starbucks
13:30 - 6. Aug. 2017

Figure I.2.1: Illustration of a constructed Twitter message.

Geosocial media content like the constructed Twitter message above (Figure I.2.1) is often colloquial. The tweet shown here tells of an apparently trivial situation that does not seem to be relevant for anyone beyond the author's private communication. Not surprising, researchers often consider messages like the one above irrelevant for their geosocial media investigations and frequently remove them as noise. However, a closer look beyond the ostensible triviality of this everyday message reveals that the tweet[1] contains more information than it seems at first glance. It features details about the author and about the related place where the message was likely generated. For example, the fact that the author meets with friends could describe their whereabouts as a so-called 'third place', *i. e.*, as a place "other than the home or workplace where people can [...] commune with friends, neighbo[u]rs, coworkers, and even strangers" (Mehta and Bosson 2010, p. 779). Further, latte macchiato consumption has been related to a "rootless cosmopolitanism" lifestyle (Wurgaft 2003). If this takes place in locations that appear intrusive and threatening to the local social fabric, like Starbucks chain coffeehouses, this may indicate a potential gentrification process in the surrounding neighbourhood, because, as Wurgaft (2003) puts it, "gentrification is the disease, [while] Starbucks is the symptom" (Wurgaft 2003, p. 74).

The simple example above shows that supposedly irrelevant information can describe phenomena of different social and geographical scales: it discloses individual-level information about the user, but also about the local geography in the nearby areas. Such information can characterize the material (a café), social (a sociable place) and economic (medium to high priced cafés) properties of whereabouts and individuals. However, the most profound difference between geosocial media messages and traditional geographical data sources like surveys is that the authors of such messages do most likely communicate this information unconsciously, and thus without any intent to provide researchers data for further exploitation. The message was instead composed for self-expression, for connecting with like-minded people, or simply for communicating with friends. The information about the related place is therefore ambient in nature and a mere by-product of the everyday communication, which makes geosocial media data appealing to the analysis of the subjective imaginary of routine places. The following sub-sections introduce key characteristics of geosocial media feeds and discuss differences to other forms of user-generated geographic information, as well as to established forms of geographic data.

[1] The term *tweet* is used synonymously with *Twitter message*.

I.2.1 A Technological Perspective

Two developments have taken place in recent years that made possible the emergence of user-generated geolocalizable content: an increasing penetration of our everyday lives with mobile technologies, and a changed attitude in using digital media. Both of these developments are intertwined and cannot be considered separately. The success of mobile technologies has only been possible because it was accompanied by changes in our societal norms and behaviours. It has become normal for internet users to not only consume static contents but to also contribute individual information to the Web. Users have literally become *produsers* (Coleman 2009; Sieber and Rahemtulla 2010), which reflects an increased willingness to share virtually all but the most private pieces of information online. These developments are in no way new, but form part of the larger societal process of the integration of technologies into the lives of people, starting in the 1940s with the advent of computers and continuing ever since (Poorthuis and Zook 2013). The Web 2.0 paradigm started to emerge at the end of the 1990s as one variant of this trend.

The Web 2.0 constitutes one of the backbones of modern mobile and participatory Internet technologies. The underlying stack of sociotechnical techniques and practices is characterized by an ubiquity of networking capabilities, the large-scale availability of heterogeneous datasets, and the willingness to share user-generated and often private content (O'Reilly 2010). These technological and societal developments fostered the proliferation of mobile and geolocalization technologies. It is thus the evolution of the Web 2.0 paradigm that has made possible the development of geographically annotated datasets like those provided by geosocial media feeds.

The outlined developments have increased the availability of geotagged datasets. Together with the required hardware and software components, as well as with positioning technologies like GPS receivers, geotagged datasets form the so-called *Geoweb* (Scharl and Tochtermann 2007). The Geoweb acts as an interface to the creation of, and access to, geoinformation in and from a variety of life circumstances (Haklay et al. 2008; Crampton 2009; Leszczynski 2015). For example, the fact that people can determine their momentary geolocations using GPS receivers installed in smartphones, and that these geotags can be attached to text messages like tweets, shows that it is the simultaneous interplay of hardware (smartphones and GPS) and software (social media platforms) components that makes the Geoweb thrive.

The umbrella term Geoweb subsumes all kinds of georeferenced contents on the Web. It is not limited to user-generated contents, requiring to distinguish further between the one-way Geoweb of georeferenced but static contents, from other datasets which are the result of a two-way interactive process involving the users' proactive participation (Johnson and Sieber 2012). These latter participatory types of Geoweb components can be further sub-divided into different modes of production of geospatial user-generated information (Rinner and Fast 2015):

- Crowd mapping (the digitizing of features from the material world)
- Citizen sensing (the passive collection of georeferenced data)
- Citizen reporting (the proactive collection of georeferenced data)
- Map-based discourse (the expression of opinions about real-world circumstances)
- Geosocial media (the active or passive collection of often unstructured opinions)

These modes of production relate to different kinds of datasets and social practices. Crowd mapping means the creation and sharing of maps in a collaborative manner, with OpenStreetMap[2] (OSM) being

[2] `http://www.openstreetmap.org`

the most notable example. This kind of data acquisition includes the explicit intent of users to share geographic information. In contrast, citizen sensing and citizen reporting refer to the collection of ambient information about particular whereabouts of people. People provide geolocalizable information, but the practice of mapping is typically not part of that. Geographic coordinates are therefore acquired *en passant*. The latter is also the case with map-based discourse platforms. These include portals like TripAdvisor[3] where people discuss about real-world places that exist in the material world. Finally, geosocial media feeds are inherently different from all other types of Geoweb applications in that the related applications are based on subjective opinions about circumstances ranging from material places to the sphere of socially-constructed spaces.

I.2.2 Disambiguation of Geosocial Media Feeds

Geosocial media feeds are inherently diverse and various concurrent terms exist to capture their characteristics. However, these terms are not equivalent, as they have subtle connotations that can change the meaning of the underlying software and data. Some terms are of technical nature (*e. g.*, *location-based social media (LBSM)*; Evans and Saker 2017), others highlight certain functionalities of the platforms, such as the ability of social networking (*location-based social networks (LBSN)*; Roick and Heuser 2013). See et al. (2016) recently provided an overview of the most important terminologies around geosocial media and how these are used across different domains. They proposed the following taxonomy (the list is sorted by decreasing generality):

Pervasive user-generated content (Pervasive UGC). This term was introduced by Krumm et al. (2008). It is very general and covers all sorts of digital artefacts that are made available by users on the Web (*e. g.*, pictures and videos). The adjective ‘pervasive’ expresses that these artefacts are entrenched in the daily routines of users through the use of mobile devices. Users literally carry the content with them and use it in local contexts where the digital files become part of and influence social practices. However, since the data does neither have to be geotagged nor must it be related to any geographic place, pervasive UGC describes a high-level concept, of which geosocial media feeds form a special case.

Citizen-contributed geographic information (CCGI). Introduced by Spyratos et al. (2014), CCGI covers volunteered data from crowd mapping platforms and geosocial media data. In contrast to pervasive UGC, the concept of CCGI only includes data that has specifically been created by the users themselves. It also only refers to explicitly georeferenced content. Still, because its focus is very broad including projects like OSM, the term does not permit further subdivision of different sorts of geosocial media feeds into subcategories. It is thus treated here as another overarching category that specializes further the term pervasive UGC, but is superordinate to geosocial media feeds.

Contributed geographic information (CGI). Harvey (2013) puts forward the term CGI, which emphasizes the role of the users in collecting localizable content. Thereby, this may or may not happen without the users’ knowledge and explicit decision to contribute geographical information. The focus is thus on the fact that data is contributed by lay persons as a by-product of their device usage. People contribute geographic data in a largely unconscious manner through sensors installed in smartphones, geosocial media, or by contributing to citizen sensing initiatives. The latter is a category of typically specialized projects requiring a minimum level of domain knowledge (see Boulos et al. 2011), and does typically not reflect well peoples’ everyday life behaviours.

[3] `http://www.tripadvisor.com`

Involuntarily volunteered geographic (iVGI), and ambient geospatial information (AGI). The terms iVGI (Fischer 2012) and AGI (Stefanidis et al. 2013) are largely congruent and describe two important facets of geosocial media data: The term AGI makes very explicit the ambient nature of the collected information by partly representing the characteristics of local contexts in which people use geosocial media. Further, the term iVGI describes the unconsciousness and opportunism in this data collection (Kelley 2013; Graham et al. 2013; Sester et al. 2014; Kitchin et al. 2017). People use geosocial media feeds not for data collection but for following their personal intents. Stefanidis et al. (2013) emphasize another important characteristic, which is that users may still have intrinsic motivations to proactively communicate either a description or the coordinates of a location, for instance when they share photos of popular places to improve their own reputation. This is relevant for characterizing the spatial everyday behaviours of people in a realistic manner. It is this intrinsic motivation of users to capture and disclose information about places that leads to a wealth of collective and individual-level geographic information.

Geosocial media data can appear in different forms. For instance, a tweet behaves like a subjective narrative about a place and may not refer to any material entity. In contrast, a user-generated photo may reflect a material geographical entity, while the choice of taking the photo and the way it is taken are still subjective decisions. Geosocial media messages are therefore to be understood as reflections of the subjective perception of places, rather than as precise representations of objective geographical facts. Technically, this is reflected by various formats of geosocial media (Coleman 2009; Rinner and Fast 2015):

- plain coordinate locations (like a whereabout posted on Facebook[4]),
- categorical or numerical values (for instance age information in a user profile),
- attribute tags (like Twitter[5] hashtags),
- content ratings (as those awarded to Foursquare/Swarm[6] venues),
- multimedia items (like Flickr[7] photos),
- complex narratives (such as microblog content posted on Twitter).

From a technical standpoint, this thesis focuses on content from microblogging services like Twitter that deliver textual data in an unstructured manner. The focus is thus on the spatial analysis of complex narratives describing places. Because of the strong integration of geosocial media into the everyday lives of the users, their interpretation—including their geospatial characteristics—requires to also understand the social processes that accompany the production of the associated data.

I.2.3 Geosocial Media as a Form of Social Practice

The use of geosocial media entails new forms of social practices. Due to the ubiquity of geosocial media feeds and because of their high degree of embedding in routine situations, they also imply a strong societal dimension. This is conceptualized by the term *Neogeography* (Turner 2007; Haklay et al. 2008; Hudson-Smith et al. 2009; Haklay 2013; Leszczynski 2014), which was introduced by Turner (2007) in his seminal book. He defines Neogeography as a collective term for practices involving “people using and creating their own maps” and “about sharing location information with friends and visitors, helping

[4] `http://www.facebook.com`

[5] `http://www.twitter.com`

[6] `http://www.swarmapp.com`

[7] `http://www.flickr.com`

shape context, and conveying understanding through knowledge of place" (Turner 2007, p. 3). While the term Geoweb emphasizes the technological perspective on user-generated geographic content, the study of Neogeography thus focuses on the processes how, and the circumstances under which these data are being created and shared (Elwood et al. 2012).

The social practices involved in the everyday production of geodata impact social norms and behaviours beyond the appearance of new technologies and data (Kitchin et al. 2017). Their impact reaches out into the material world (Sui and Goodchild 2011), making the geographical dimensions of cyberspace and the material world evolve in a coexisting manner (Poorthuis and Zook 2013; Crampton et al. 2013). Adopting Leibniz's notion of relative space as the sum total of human geospatial interactions anchored in the material reality (Quesnot and Roche 2015), this suggests the interaction between material and virtual spaces. For instance, when geosocial media users read negative formulated messages about a place, this can impact an individual's mobility behaviour in that people may try to avoid the corresponding region. By analogy, when Facebook users can see the whereabouts of their friends on a map, this may trigger spontaneous meetings and can thus support the strengthening of social cohesion (Sutko and de Souza e Silva 2011; de Souza Silva 2013). These examples demonstrate the dualism between material and virtual spaces and illustrates the feedback mechanisms between the physical behaviours of people in the material world and the digital image of places.

Interactions between the virtual and the material geographic space are influenced by a range of endogenous and exogenous factors. For instance, the motivation for sharing whereabouts may either be prompted by pragmatic purposes or social-driven (Tang et al. 2010; Lindqvist et al. 2011). Purpose-based location-sharing thereby includes the communication of a location for the purpose of making contact with other people, whereas social-driven location-sharing includes incentives such as self-expression or the enhancement of ones' digital self-representation (Barkhuus et al. 2008; Evans 2011; Quesnot and Roche 2015). Clearly, the second type of incentives is more selective in the sense that people would not communicate places that would run counter these goals. The latter has recently been supported by results from an investigation of 22 qualitative interviews with Foursquare users (Saker 2017). The answers given suggest that users who are sharing locations are well aware that their behaviours are observed by others. Respondents also indicated that they carefully select their communicated locations, including those they do not wish to disclose. Clearly, the act of sharing spatial and place-based information on geosocial media is far from being unconscious, even though the motivation for doing it is not the purpose of data collection.

The selectivity in communicating whereabouts bears a strong influence on the social imaginary of places (Taylor 2004; Kelley 2013). It determines which places are mirrored into digital space and how positive or negative this image appears. In addition, the underlying stack of technologies has an influence on the representation of places through the designs of the geosocial media platforms, which constrain how people communicate their whereabouts. For example, because a platform like Twitter is targeted at a specific group of people and focused on specific purposes, that influences the types of places which are represented accordingly. Platform design also partially shapes user behaviours through the expectations that are imposed from the expected target audience. For example, an expectation to present oneself in a professional, casual or other manner could ultimately lead to Michael Focault's 'technology of self', where the technology becomes absorbed as part of the 'self' but, also shapes the representation of the 'self' in the material world in a reciprocal manner (Rzeszewski and Beluch 2017). Closely related is the so-called 'code/space' metaphor (Kitchin and Dodge 2011), whereby code generates space and space generates code. These arguments lead to the outlined blending of material and digital spaces.

I.2.4 Technical and Demographic Challenges

The outlined interactions between the design of geosocial media feeds and their influence on behavioural patterns show how strong the technical sphere shapes social processes. Geosocial media data are therefore subject to complex challenges making their scientific use difficult. Some of these challenges are of technical nature. For instance, it is often difficult to disclose the semantics of messages because people write messages in a colloquial and ambiguous manner (Zappavigna 2012; Sester et al. 2014). Technical length restrictions and highly sophisticated semantics make it difficult to reveal information by natural language processing. Further, because geographic regions are not populated evenly, the amount of information is volatile in space and time (Sester et al. 2014). This leads to a spatially varying number of messages and may in turn lead to bias in the assessed importance of events and places. Similarly, this bias may also create the impression of highly redundant data in places where many people post about similar topics. Some messages might further be of a noisy character and thus of little use for scientific investigation (Sengstock and Gertz 2012).

Geospatial characteristics also have an influence on the data from geosocial media platforms. Because most data is available in the form of points, these are technically scale-free (Sester et al. 2014). Therefore, even though social media data does not provide explicit scales it often conveys implicit ones, especially when it refers to social or material circumstances. However, these implicit scales are not fixed over time. Crampton et al. (2013) investigated a Twitter hashtag indicating the victory of the University of Kentucky's men's basketball team in the National Collegiate Athletic Association (NCAA) championship. Over the course of one evening they found that the scale of the digital representation of this event changed from a local to the national level, and later moved back to a local level. Thus, the scale at which conclusions could be drawn about the investigated phenomenon changed drastically, which clearly influenced the spatial analysis of the respective messages.

In a similar vein, the degree of localness of users and phenomena represented on geosocial media can differ significantly. Johnson et al. (2016) report that the number of distinctly local users varies between 75% and 88% across the platforms Swarm, Twitter and Flickr. The number of local users is thereby determined by the size of the white population, youth and the degree of urbanization. The least degree of localness is observed for contributors of Flickr photos, which seems reasonable because photo-sharing services are frequently used by tourists. This is supported by findings from Li et al. (2013) who report that photo-sharing services are more uncertain with respect to the localness of the shared photos. However, Rzeszewski and Beluch (2017) found that the degree of localness is not consistent across different cities, whereby some cities are more affected by importing lifestyles from other places, which has an impact on the interpretation of analysis results obtained from geosocial media data.

The varying degree of localness is related to biases in the types of sampled places and their demographic compositions. Research on geosocial media feeds is strongly skewed towards data from world cities like London or New York. These are cultural and social melting pots that, by their very nature, act in an averaging way (Rzeszewski and Beluch 2017). Heterogeneous groups of people like tourists mix-up temporarily with the domestic population, making it difficult to disclose local behavioural patterns. Another related kind of digital divide is the under-representation of rural areas on geosocial media (Mislove et al. 2011; Hecht and Stephens 2014). The literature on findings from conurbations is vast, but little is known about how people use geosocial media in rural contexts. There are further dividing lines along demographic characteristics. A majority of geosocial media content is produced by only a small number of individuals (Haklay 2012) that typically belong to the young, wealthy and educated parts of the society, especially to Caucasian and Asian ethnic communities (Li et al. 2013; Longley et al. 2015). Further, a majority of the contributors is male (Mislove et al. 2011; Longley et al. 2015), although this bias was reported to be flattening out recently (Mislove et al. 2011). Relevant geographic factors include landuse classes (Longley and Adnan 2016) and variations at the national level (Mislove et al. 2011). While these results are largely based on intrinsic investigations that do not take account of external data for validation, Sloan (2017) report findings from a comparison of a British Twitter dataset to a representative socio-demographic panel. These results largely confirm the already mentioned results, while the bias towards younger cohorts seems to be less severe than was suggested by other studies.

While some of the outlined characteristics are valuable for the investigation of subjective momentary experiences, they clearly have an impact on the data, and thus on obtained analysis results. This includes spatial analysis, which offers a useful set of techniques for the geospatial characterization of places and locations. This field is introduced in the next Chapter.

I.3 Spatial Analysis

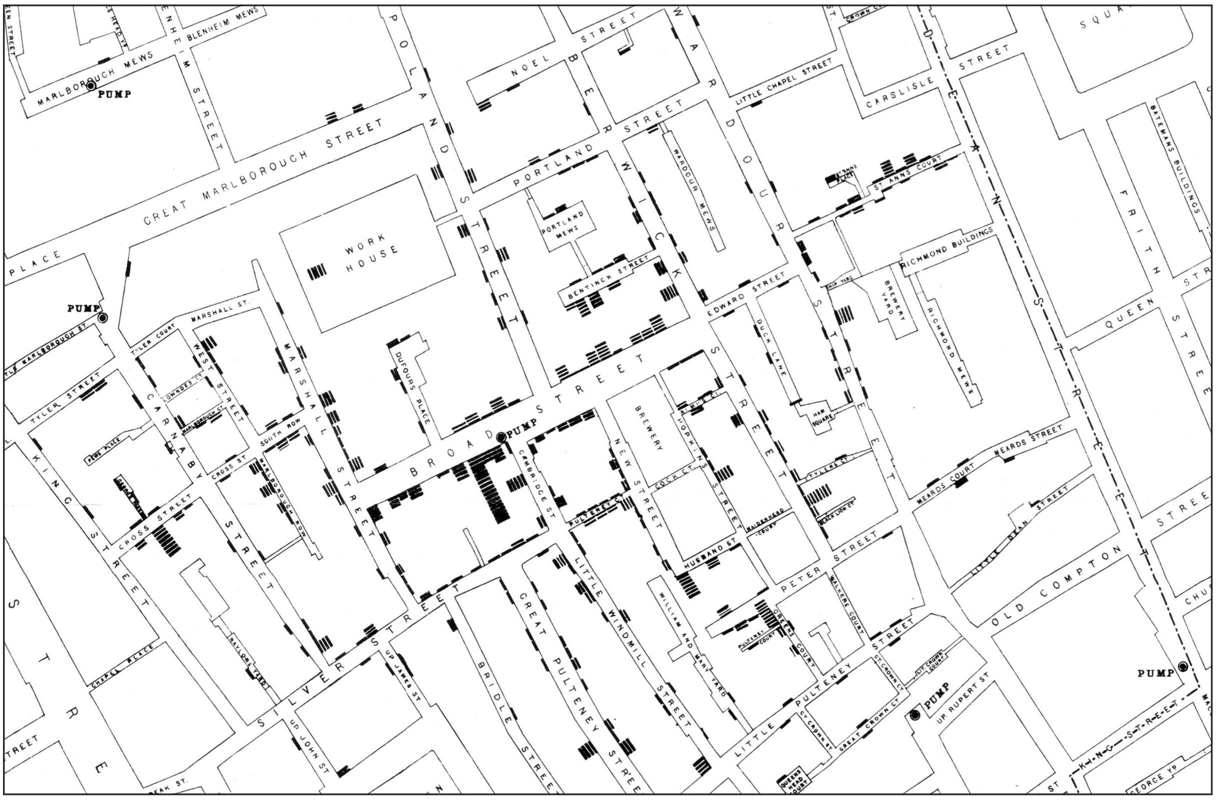


Figure I.3.1: Dr Snow's London cholera map—the result of an early example of spatial analysis. The piled black rectangles indicate the numbers of cholera cases. The figure is reproduced from a publicly available map from the Department of Epidemiology[1], University of California, Los Angeles, CA.

Dr Snow's well-known map of cholera cases (Figure I.3.1) is a classic example for motivating the statistical field of spatial analysis. In 1854, the Soho district of London was hit by a cholera epidemic. It was caused by contaminated drinking water, which came from public pumps operated by different utility companies. Dr Snow identified the source of the epidemic by recording all cholera cases on a topographic map. That map illustrates a range of important topics from spatial analysis, which are also relevant for this thesis. One of these is the notion of a spatial random variable. The variation in the numbers of cholera cases on the map (*i. e.*, the piled black rectangles) is influenced by stochastic factors, such as the numbers of people living in the houses, cholera-related health indicators like population overcrowding, malnutrition, unhygienic conditions and access to basic medical services (Soto 2009). These random influences show that Dr Snow dealt with a random quantity generated by a spatial stochastic process, rather than with a fixed (*i. e.*, deterministic) characteristic of geographic space. Such spatial random variables interact with

[1] `http://www.ph.ucla.edu/epi/snow/snowmap1.pdf`, last accessed on 19 September 2017

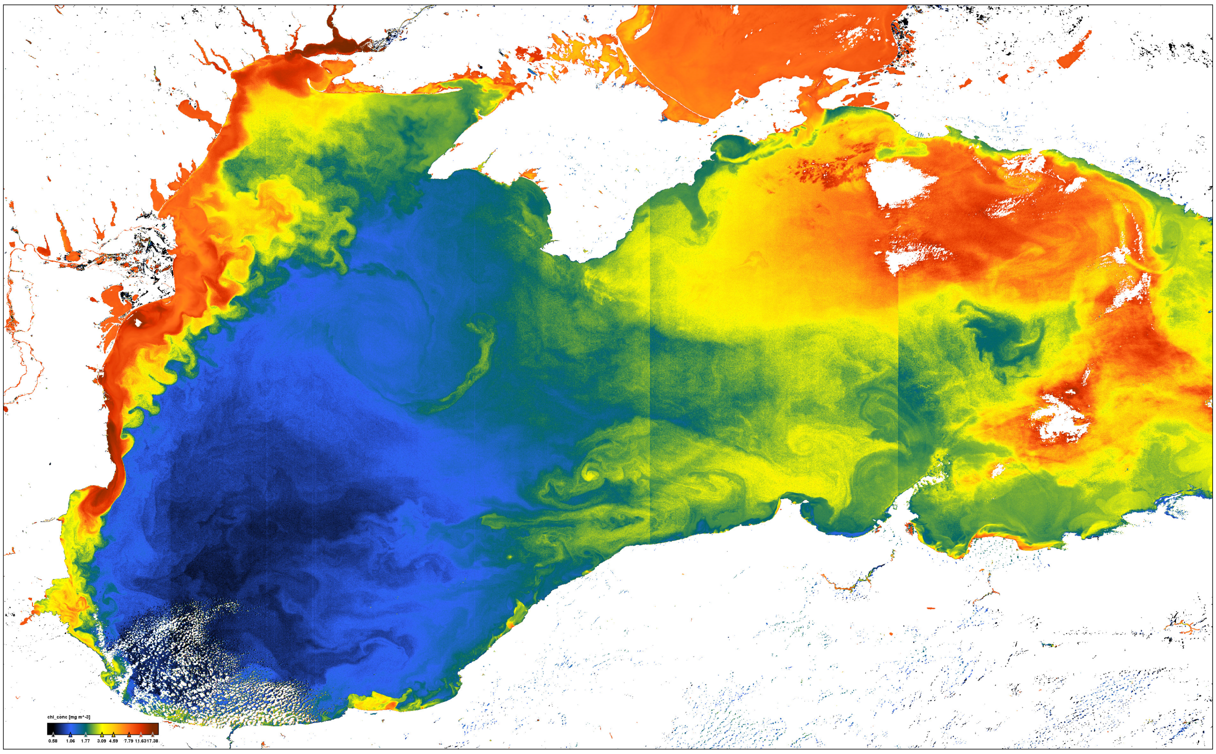


Figure I.3.2: Chlorophyll *a* concentration in the Black Sea as an example of a spatial random field. The figure is provided by the Plymouth Marine Laboratory[2].

their immediate geographic vicinity, which causes them to have an inherent interaction behaviour. This is evident from the distinctive clustering that is notable on the map in Figure I.3.1. The clustering shows that the geographic distribution of the cholera cases is not completely random, but follows systematic yet probabilistic rules. Modern statistics characterizes these rules by the estimation of spatial autocorrelation, another important spatial analysis concept. Dr Snow did not have such well-defined tools available at his time. Nevertheless, it was the same characteristic that allowed him to finally disclose the source of the cholera epidemic, which turned out to be a pump located in Broad Street.

Dr Snow's map provides an intuitive sense of some key ideas of spatial analysis, the core of which is the determination of spatial behaviours of random variables by taking account of geographic circumstances. The following sub-sections introduce some spatial analysis concepts that are important for this thesis in a concise and technical manner. Further, a connection between this statistical field and geosocial media data is made.

I.3.1 Spatial Stochastic Processes

Deterministic variables like the positions of a swinging pendulum allow the precise prediction of future states. However, this is not possible with random variables that are subject to unpredictable influences. Still, random variables are not entirely arbitrary but described by probabilistic rules. Their behaviours are therefore oftentimes structured, which allows a prediction of future states with some degree of certainty

[2] http://www.coastcolour.org/site03_mediterranean_blacksea/MER_FRS_1PNMAP20050405_082012_000003012036_00107_16193_0001_c2r_chl_conc.jpg, last accessed on 22 September 2017

(Everitt and Skrondal 2010, p. 356). Technically speaking, random variables are described by probability spaces:

Definition. ***(Probability spaces, random variables)*** *Let $(\Omega, \mathcal{A}, P)$ be a probability space where Ω is a sample space of possible outcomes, $\mathcal{A}$ denotes a σ-algebra of events and P is a probability measure over Ω. Any mapping $X\colon \Omega \mapsto E$ is called a random variable, whereby $E = (\Omega, \mathcal{A})$ is a measurable space with respect to P.*

The sample space Ω contains all theoretically possible outcomes of a random phenomenon, *e. g.*, the positive real numbers in case of temperatures measured in Kelvin. The event space $\mathcal{A}$ is a so-called σ-algebra that combines single outcomes into subsets of Ω that can be assigned probabilities. For instance, in case of temperatures measured in Kelvin, this would include events like "$X \leq 10\,\mathrm{K}$" forming a half-open interval on the real numbers. The probability measure $P\colon \omega \mapsto [0, 1]$ assigns probabilities to individual outcomes $\omega \in \Omega$ and also allows to assess the probability of events $A \in \mathcal{A}$.

When a family of random variables is referenced over an additional structure that allows to sort the sequence of random outcomes in some way, we call that collection a *stochastic process*. The structure over which the variables are referenced (or *indexed*) is called its index set (Cox and Miller 1977). The latter can be any arbitrary set endowed with a metric, but in a narrower sense the index set is typically defined to be a time interval:

Definition. ***(Stochastic processes)*** *Let (T, d) be a metric space with T being an index set and d being a metric. Then, a collection of random variables $X = \{X_t : t \in T, X \in \Omega\}$ is called a stochastic process. Each random variable $X_t \equiv X(t)$ is bound to one index from the set T. In practice, T often denotes a time interval $[t_0, t_N] \subset \mathbb{R}$.*

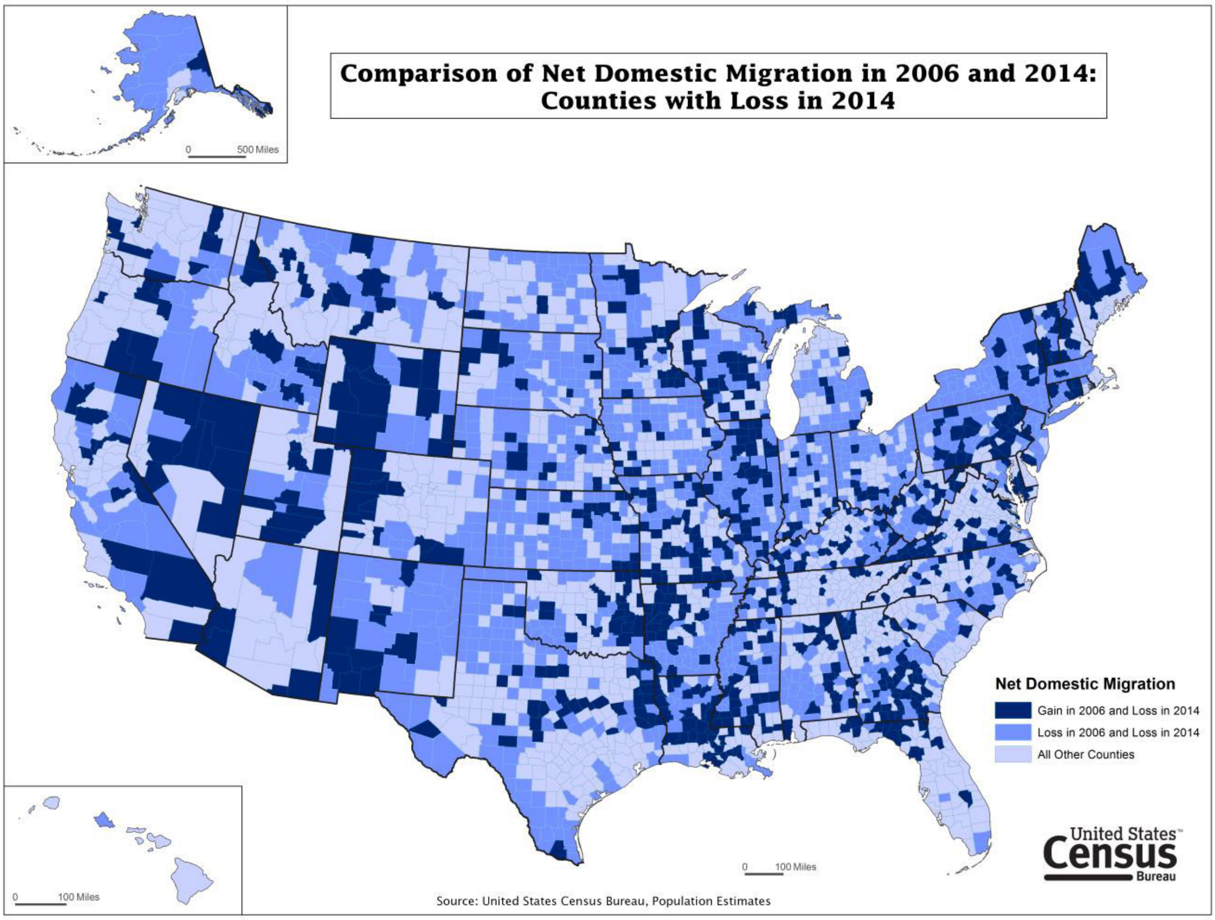
The definition given above is operational in time series analysis, but it is possible to extend the concept to the spatial case. Let $T = \mathbb{R}^2$ be a two-dimensional real-valued index set and let $\mathcal{S} \subset T$ be a set of so-called spatial units. Three different kinds of *spatial processes* can be derived from this (see Cressie 1993; Gneiting and Guttorp 2010): spatial random fields, lattices and spatial point patterns.

Definition. ***(Spatial random fields)*** *A collection of random variables $X = \{X_s : s \in \mathcal{S}, X \in \Omega\}$ is called a spatial random field, iff the spatial index set is continuous (i. e., $|\mathcal{S}| = \infty$) and fixed (i. e., if it is not subject to randomness).*

Spatial random fields form the conceptual basis for geostatistical investigations, a branch that deals with spatially continuous phenomena like soil properties or water temperatures. Figure I.3.2 shows the Chlorophyll *a* concentration in the Black Sea and thus gives an example of a spatially continuous phenomenon. This variable is defined at each point of the water surface. The term 'continuous' thus refers to the notion of having a random variable defined at any location of the underlying spatial index. The figure also demonstrates that the spatial index may be bounded as the Chlorophyll *a* concentration is only defined within the bounds of the water body.

Definition. ***(Lattice data)*** *A collection of random variables $X = \{X_s : s \in \mathcal{S}, X \in \Omega\}$ is called lattice data, iff the spatial index set is discrete (i. e., $|\mathcal{S}| < \infty$) and fixed (i. e., not random). The spatial units in the index set can be of any regular or irregular shape.*

Lattice data is used to model spatially discrete phenomena like election results at the constituency level and census variables. These are only defined in specific locations in a possibly aggregated form. For instance,


Comparison of Net Domestic Migration in 2006 and 2014:
Counties with Loss in 2014
0 500 Miles
Net Domestic Migration
Gain in 2006 and Loss in 2014
Loss in 2006 and Loss in 2014
All Other Counties
0 100 Miles
0 100 Miles
United States Census Bureau

Source: United States Census Bureau, Population Estimates

Figure I.3.3: Net domestic migration within the US counties as an example of lattice data. The figure is provided by the United States Census Bureau[3].

the map in Figure I.3.3 shows the exchange in domestic migration within the American counties. Each attribute value is bound to a specific county, which are of arbitrary shape and therefore form an irregular lattice of non-random administrative units. It would not be admissible to interpolate such attribute values from their lattices to the entire Euclidean geographic space. The migration net balance is only defined for whole counties, but not for the individual locations within these counties. The analysis of lattice data is therefore related to the analysis of networks, as each county can be represented by a point and neighbourhood relations between these counties define a network of regions. Lattices prevail in spatial regression scenarios and in exploratory/confirmatory spatial data analysis.

Definition. ***(Marked and unmarked spatial point patterns)*** *A set of spatial units $\mathcal{S} = \{S_i \in \mathbb{R}^2 : i \in \mathbb{N}\}$ is called a spatial point pattern, iff the spatial units are random variables. In addition, if any arbitrary random variables $X_i \in \Omega$ are assigned to these random locations, then the collection of ordered pairs $\mathcal{S}_X = \{(S_i, X_i) : S_i \in \mathcal{S}, X_i \in \Omega\}$ is called a marked spatial point pattern. Like with lattices, the spatial units can be either regular or irregular in shape.*

[3] `https://www.census.gov/content/dam/Census/newsroom/blogs/2015/03/moving-in-the-usa-domestic-migration-before-and-after-the-recession/blog-graphic-2.jpg`, last accessed on 22 September 2017

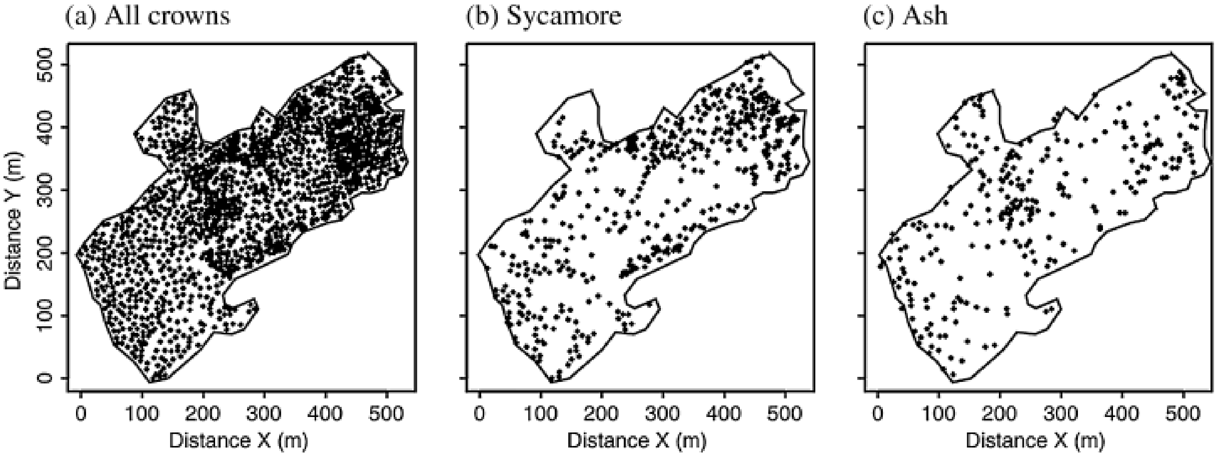


Figure I.3.4: Crown locations of sycamore and ash trees in an English woodland, an example of spatial point patterns (taken from Atkinson et al. 2007).

Spatial point patterns are used in the investigation of geometric arrangements of random spatial units. The pattern of trees in an ancient semi-natural woodland (Atkinson et al. 2007) as shown in Figure I.3.4 is one example of a spatial point pattern. The locations of the trees are themselves considered outcomes of a random process which is influenced by ecological factors, climatic conditions, the seed dispersal mechanism, and other factors. Further, labels indicating the tree species (ashes and sycamore trees) may be treated as additional random outcomes assigned to the geometric points as so-called marks (not shown in the figure). Spatial point pattern analysis thus allows to investigate the spatial configuration of random geometries and, if marks are available, the interactions between geometric patterns and those in the attached attributes.

The introduced types of spatial processes conceptualize different types of geographic phenomena. This thesis focuses on the relationship between geosocial media data and spatial analysis techniques. Times and locations of the corresponding messages are random, because they are influenced by unpredictable user behaviours. In combination with attributes assigned to the messages, this gives rise to the investigation of marked spatial point patterns. Technically, the analysis of spatial relationships within the marks of a spatial point pattern boils down to the analysis of a lattice (Shimatani 2002). That is, the spatial analysis of the attribute values is conducted by letting the stochastic marks vary over the sampled locations, which are held fixed accordingly. The methods that are used to analyze the spatial second-order[4] behaviour of such random variables are summarized by the term *mark correlation function*, of which different versions exist (a comprehensive list is found in Illian et al. 2008). All of these assess the statistical associations between geographically adjacent units, and thus the characteristic of *spatial autocorrelation*. The next section introduces this concept as well as those variants of the mark correlation function which are relevant for the remainder of this document.

I.3.2 Spatial Autocorrelation

Conventional and non-spatial statistics are largely based on the assumption of *i.i.d.* (independent and identically distributed) random variables (Durbin 1973; Gaenssler and Stute 1979). This assumption is

[4] The term *second-order* refers to the second statistical moment (variance or covariance) (see Illian et al. 2008, 223 ff.). *First-order* is used analogously.

reasonable when samples are taken from controlled experiments without interactions within the samples. In such cases, the i.i.d. assumption is convenient because it allows to disregard potentially complex relationships between random variables when drawing inference. The occurrence of dependency structures, in turn, introduces redundancy and thus reduces the degrees of freedom of statistical estimators. If not taken into account, this causes an overestimation of statistical effects and, in addition, an obscuration of characteristics of individual random variables when samples are not identically distributed. These challenges render the drawing of meaningful and generalizable conclusions difficult in case of systematically structured and heterogeneous random variables.

Geographical data is not collected in a fully controlled manner but taken from *in situ* physical and social contextual conditions (Goodchild 2009). The data is also often aggregated into arbitrary units that are defined for purposes other than spatial analysis, *e. g.*, in case of census data. Adjacent geographic units may then be subject to similar contextual influences affecting the phenomenon of interest (*e. g.*, political conditions or climatic factors). These characteristics inevitably cause redundancy within the related random variables, which in turn violates the i.i.d. assumption. Even though it is possible to adopt an appropriate spatial sampling scheme when collecting data (a review is found in Wang et al. 2012b), spatial dependencies still occur through proactive interaction or other forms of spatial diffusion mechanisms that are endemic to the underlying data-generating processes. These circumstances motivate the *first law of geography* (Tobler 1970; Sui 2004), which states that “everything is related to everything else, but near things are more related than distant things” (Tobler 1970, p. 236). This heuristic rather than rigorous law is statistically quantified by the concept of *spatial autocorrelation*, which features prominently in the field of spatial analysis.

Spatial autocorrelation is a second-order characteristic that describes the spatial interaction behaviour within random variables (Fischer and Getis 2010b). A range of different estimators exist (see Getis 2007; Getis 2008): the covariance-based Moran’s *I* and Geary’s *c* (Cliff and Ord 1969), the spatial autoregressive coefficients ρ and λ (Anselin 2001), or G_i^* (Getis and Ord 1992; Ord and Getis 1995) which emphasizes structures within extremal values. These estimators evaluate the spatial interaction behaviours within random variables but are used in different application scenarios. These include the assessment of influences of distance effects, of the roles of geometry and topology, or of the impact that individual geographic features have on spatial processes (Getis 2007). In this thesis, two estimators of particularly high academic and practical interest are investigated in more detail: Moran’s *I* and G_i^*.

Moran’s *I* (Moran 1950; Cliff and Ord 1969) is one of the most frequently applied measures of spatial autocorrelation. It estimates the normalized spatially-weighted covariance within random variables, whereby it takes account of geographic structures by incorporating a so-called spatial weights matrix. Spatial weights define the fixed geographic structure connecting those spatial units $s_i \in S$ upon which the investigated phenomenon is believed to operate (Bavaud 1998; Getis 2009; Harris et al. 2011). For instance, in an investigation of commuting processes it may be useful to utilize the numbers of major traffic routes between administrative units, while in an epidemiological analysis it might be more useful to adopt physical contiguity as an appropriate spatial weighting scheme. Using Moran’s *I* as a test statistic allows the investigation of whether the modeled geographic layout plays a significant role in the structure of an attribute, and thus if geographic factors are major drivers of interactions within the analysed random variables. Moran’s *I* is widely used because it has several advantages over other test statistics. For example, it has higher statistical power, is less affected by attribute outliers, and is more robust against configurational outliers in the spatial layout than Geary’s *c* (Chun and Griffith 2013).

Definition. ***(Moran's I)*** *Let $cov(X_i, X_j)$ be the covariance between the spatial random variables $X = \{X_i\}$ at locations i and j. Moran's I is an estimator of the normalized spatially-weighted covariance (i. e., of spatial autocorrelation). It takes account of pairwise relationships between spatial units by using a spatial weights matrix $W = (w_{ij})$. The global form of Moran's I which averages the overall spatial autocorrelation in a region is defined as*

$$I = \frac{n}{\sum_{i,j}^{n} w_{ij}} \cdot \frac{\sum_{i,j}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i}^{n}(x_i - \bar{x})^2}, \tag{I.3.1}$$

where n is the number of spatial units and $\bar{x}$ denotes the average of the observed attribute values. By analogy, the local version of Moran's I allows to disaggregate the global measure into its local constituents, and makes it possible to estimate and map local structures. It is defined as (Anselin 1995)

$$I_i = \frac{x_i - \bar{x}}{\frac{1}{n} \cdot \sum_{j}^{n}(x_j - \bar{x})^2} \cdot \sum_{j \neq i}^{n} w_{ij} \cdot (x_j - \bar{x}). \tag{I.3.2}$$

A second estimator used in this thesis is G_i^* (Getis and Ord 1992; Ord and Getis 1995), which assesses spatial structures within the tails of an attribute value distribution. Significant structures in the left tail (*i. e.*, within low values) are called *cold spots*, whereas structures in the right tail are called *hot spots*, which is why G_i^* is often referred to as a hot-spot statistic. Its distribution is asymptotically normal in the null hypothesis, regardless of the distribution of the investigated random variables (Zhang 2008).

Definition. ***(Hot spot statistic G_i^*)*** *The measure G_i^* is an additive indicator of spatial autocorrelation in the tails of the attribute value distribution of spatial random variables $X = \{X_i\}$. Similar to Moran's I, it takes account of geospatial relationships between spatial units i and j by using a spatial weights matrix $W = (w_{ij})$. The global measure that averages the extreme-value accumulation is defined as*

$$G = \frac{\sum_{i,j}^{n} w_{ij} \cdot x_i x_j}{\sum_{i,j}^{n} x_i x_j}, j \neq i, \tag{I.3.3}$$

with n being the number of spatial units. Two local versions of Equation I.3.3 exist. These are called G_i and G_i^* and they are almost identical with the only difference between them being that G_i omits the local observation i in the estimation. Therefore, only the definition of G_i^* is given here:

$$G_i^* = \frac{\sum_{j}^{n} w_{ij} \cdot x_j}{\sum_{j}^{n} x_j}. \tag{I.3.4}$$

The considered measures of spatial autocorrelation assume uniformity in the investigated random variables in their null models (Tiefelsdorf 1998). This statistical uniformity is called stationarity, of which there are various types of different strengths. An additional and related spatial concept is spatial heterogeneity, which refers to varying statistical parameters.

I.3.3 Spatial Heterogeneity and Stationarity

Spatial heterogeneity describes geographic non-uniformity within spatial random variables (Dutilleul and Legendre 1993). In the empirical sciences, it refers to mixtures of different processes like they are found in transitional ecological habitats (*e. g.*, in-between aquatic and terrestrial habitable conditions; Turner

1989), or to an uneven distribution of a single process (*e. g.*, in case of malaria infectivity; Santos-Vega et al. 2016). Spatial heterogeneity then describes the coexistence and interaction of variegated processes and is often used synonymously with general diversity. Different types of heterogeneity are characterized by their causal origins, maintenance mechanisms, and dynamics (Strayer et al. 2003), as well as by their structural types (*e. g.*, fuzzy vs. crisp boundaries; Jacquez et al. 2000). Apparently, no consensus exists on the exact meaning of spatial heterogeneity in the empirical literature.

Spatial statistics distinguishes different orders of spatial heterogeneity in statistical parameters (Kolasa and Rollo 1991; Dutilleul and Legendre 1993). First-order heterogeneous processes have a spatially varying mean and second-order heterogeneity relates to unstable variance. First-order heterogeneity appears as a trend if the observation scale does not match the scale of the observed phenomenon. In contrast, if multiple processes are observed at small scales, the region appears patchy forming spatial regimes (Anselin 1988b; Camara et al. 2004). The local versions of spatial autocorrelation measures introduced in Section I.3.2 can be used to disclose heterogeneity. While G_i^* allows to investigate varying means, local Moran's *I* is a measure of second-order heterogeneity in the covariance. In addition, Ord and Getis (2012) put forward a measure called *Local Spatial Heteroscedasticity* (LOSH) which allows to characterize heterogeneity in the variance, and is used in this thesis alongside G_i^* and Moran's *I*.

Definition. ***(LOSH)*** *Let $\bar{x}_j$ be a local spatially weighted mean and e_j be a residual about $\bar{x}_j$. h_1 is then the global average of the local residuals e_j and LOSH is given by*

$$\begin{aligned} H_i &= \frac{\sum_j^n w_{ij} \cdot |e_j|^a}{h_1 \cdot \sum_j^n w_{ij}}, \qquad & e_j &= x_j - \bar{x}_j, \\ \bar{x}_j &= \frac{\sum_k^n w_{jk} \cdot x_k}{\sum_k^n w_{jk}}, \qquad & h_1 &= \frac{\sum_j^n |e_j|^a}{n}. \end{aligned} \tag{I.3.5}$$

One reason for spatial heterogeneity is the absence of stationarity, which describes the uniformity of statistical parameters in data generating processes. Different types of the concept exist which are discriminated by their degrees of restrictiveness (Cressie 1993; Oliver 2010). The strictest variant is strong stationarity, which requires all random variables to be drawn from the same distribution regardless of their absolute locations. This is equivalent to a static world with homogeneous contextual conditions and thus a too strong and unrealistic assumption for geographic data. The statistical measures considered in this thesis presume so-called second-order stationarity (Cliff and Ord 1981; Zimmermann and Stein 2010, p. 42). This concept requires all moments up to the second order to be constant, while skew, kurtosis and higher-order moments are allowed to vary geographically. The covariance reduces then to a function of distance, assuring a consistent spatial autocorrelation function in null models of spatial statistical tests.

Definition. ***(Second-order spatial stationarity over finite spatial indexes)*** *Let $\mathcal{S}$ be a finite set of n discrete spatial units. Further, consider the set $X = \{X_s : s \in \mathcal{S}, X \in \Omega\}$ of spatial random variables. Let $s, s_1, s_2 \in \mathcal{S}$ and $E[X_s^2] < \infty$ for all s. Then, the process is second-order stationary if*

$$\begin{aligned} E[X_s] = \mu, \quad Var[X_s] = \sigma^2, \\ Cov[X_{s_1}, X_{s_2}] = \sigma^2 \cdot C(\|s_1 - s_2\|), \end{aligned} \tag{I.3.6}$$

where $C(\cdot)$ is a correlation function and $\|\cdot\|$ denotes the Euclidean norm. Second-order stationarity thus requires a constant mean and variance. It follows that the covariance between any two spatial units is only a function of (not necessarily Euclidean) distance (Zimmermann and Stein 2010).

I.4 Scientific Contributions

The main results to answer the research questions from Section I.1.2 are given in the following and their relevance and limitations are discussed. All results discussed are laid out in detail in Part II, where the accompanying publications are attached as individual chapters. In all following subsections, the achieved key findings are highlighted in bold font before they are briefly discussed.

I.4.1 Research Question 1

Research question 1 examines how scale and topological characteristics of spatially superimposed heterogeneous random variables influence the hot-spot estimator G_i^* and the autocorrelation measure Moran's I. The results given here are based on Chapters II.1, II.2 and II.6, which correspond to Westerholt et al. (2015), Westerholt et al. (2016) and Westerholt (2018).

RQ 1.1: Impact of co-occurring spatial scales on hot-spot estimation

Two objectives are pursued to investigate the influence of scale-related characteristics of spatially superimposed and heterogeneous random variables on hot-spot estimation: (i) a better understanding of the scales included and of related statistical data characteristics is achieved by analysing two Twitter datasets; (ii) the impact of overlapping scales on results of G_i^* and on conclusions drawn on hot-spots is examined.

Geosocial media data contain different scales in a spatially commingled way. The results obtained show that spatial lags[1] of geosocial media data are heterogeneous. Five different analytical scales are applied and on each of these, 70–90% of the included observations interact on scales beyond the respective range of interest. Additionally, small spatial scales dominate the constructed spatial lags in such a way that the smallest two of the investigated scales account for more than 40% of all comprised observations. Taking account of the attribute values (here: semantic similarities) further reveals that the proportion of attribute values contributed by the smallest-scale observations exceeds their quantitative share by up to 80%, suggesting that small scales are strongly overvalued. Therefore, it is difficult to recognize significant hot-spot patterns on larger scales from the geosocial media data examined without being influenced by the smallest contained scales.

Hot-spots are frequently misinterpreted or remain undetected when scales co-occur in spatially superimposed data. Application of G_i^* to Twitter data confirms the observed bias towards small scales. In the obtained results the number of significant hot-spots increases strongly with the analytical scale (see Figure I.4.1). This is caused by pronounced small-scale observations, many of which are falsely included on coarser analytical scales due to the additive nature of G_i^*. An investigation of the means of standardized G_i^* values on different scales supports this observation by a positive trend. As a result, the null hypothesis is rejected too often at large scales because significant hot-spots from smaller scales are propagated to larger-scale inferences, causing type I error inflation (*i. e.*, false-positives). Type II errors (*i. e.*, false-negatives) are in turn observed on small analytical scales when large-scale observations with

[1] A spatial lag quantifies the spatial neighbourhood of a random variable.

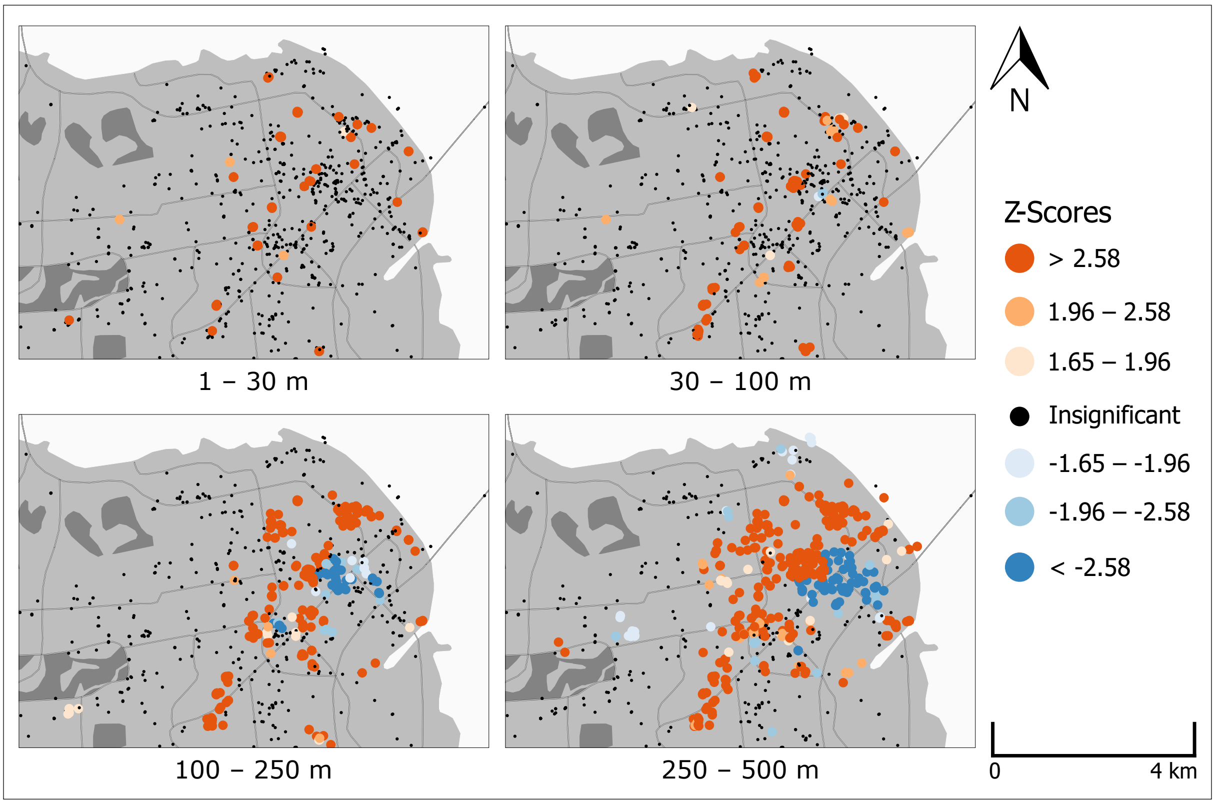


Figure I.4.1: A series of z-scores from applying G_i^* on different scales; based on Figure II.1.6.

low attribute values are located next to small-scale observations. This then diminishes potential hot-spot patterns, resulting in the overlooking of effects.

Small-scale hot and cold-spots constrain large-scale inferences. The applied G_i^* estimator overestimates either significant hot-spots strongly, or the effects remain undetected. One reason for this is that the G_i^* values obtained on different scales are not mutually independent. Similar to the well-known multiple hypothesis testing problem that occurs with local statistics (Caldas de Castro and Singer 2006; Nelson 2012), superposition introduces spurious dependencies between scale levels. These dependencies constrain the possible interpretations of hot and cold-spots, because smaller-scale information is propagated into larger-scale inferences. Just like multiple hypothesis testing requires correction to avoid type I error inflation, the obtained results suggest that the scale-related dependency structures introduced by superimposition must be additionally controlled to ensure correct hot-spot interpretations. Another difficulty in interpreting hot-spots is the complex interaction of type I and type II errors. Some local arrangements make existing patterns appear weaker than they are, while effects are overestimated in other situations. However, there is a non-trivial interaction between these situations that impacts the statistical power of G_i^* and deserves further investigation in future research.

RQ 1.2: Effects of falsely but strongly connected units on estimating spatial autocorrelation from spatially superimposed data

The estimation of spatial autocorrelation from random variables is influenced by their topological arrangement and how they are connected to each other. Research question 1.2 addresses the impact of unfavourable arrangements caused by superimposition on the spatial autocorrelation measure Moran's *I*.

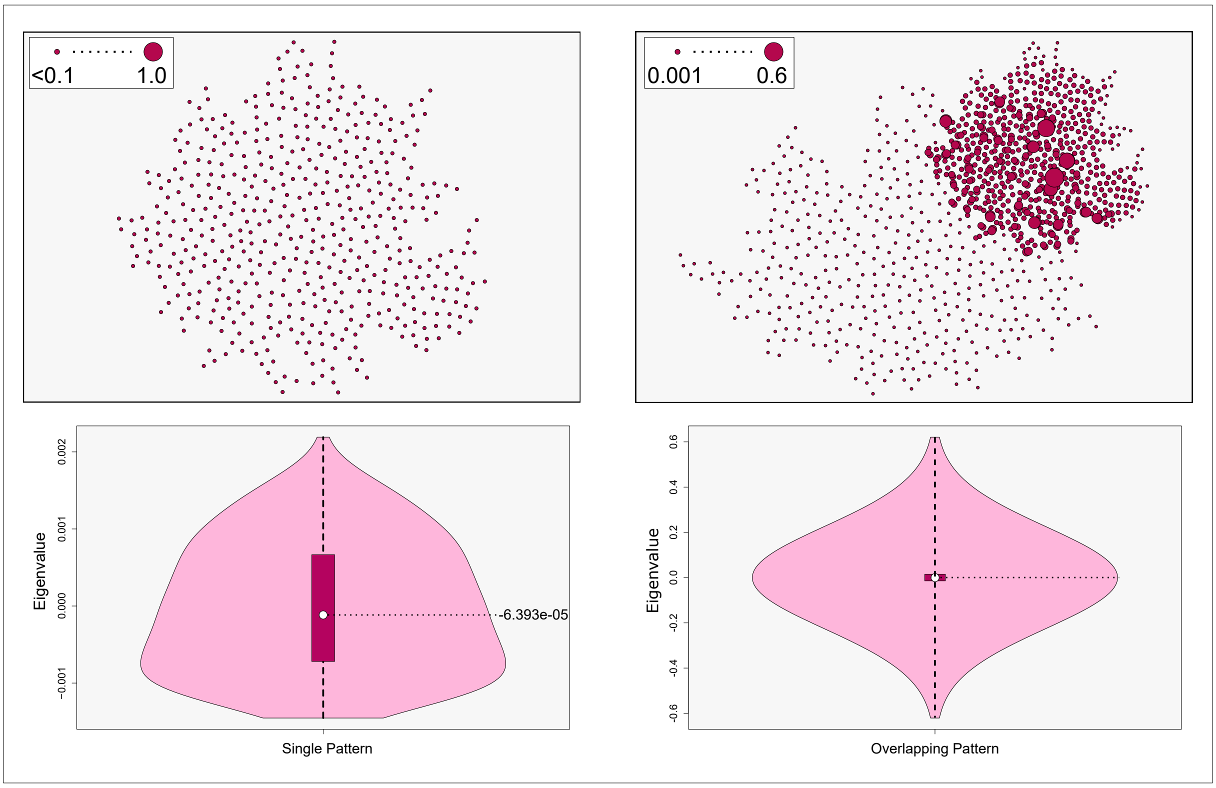


Figure I.4.2: Two maps of local eigenvalues and their spectra obtained for a single and a superimposed pattern; based on Figures II.2.4 and II.2.5.

Three analyses are performed with a Twitter and a synthetic dataset: (i) semivariances and autocovariances are estimated to investigate cumulative variability effects; (ii) the eigenvalues of different spatial weighting schemes are analysed to assess the influence of superimposed geographic layouts on Moran's I; and (iii) Moran scatter plots are used to decompose Moran's I for a clear identification of topological effects. In addition, the effect of statistical differences between superimposed processes is investigated by analysing 20.000 configurations of normal attributes with differing means and variances.

Estimating spatial autocorrelation with spatial superimposition suggests non-existent patterns. Semivariance describes the variance on distance bands and allows to investigate cumulative topological effects. The estimated semivariances show a maximum over short spatial distances, pointing to a high degree of diversity in geographically close random variables. Further, the trend in the semivariances has an initial negative slope, which quickly levels out to the general overall variance. In non-superimposed datasets, that slope is typically positive (the 'first law of geography', *cf.* section I.3.2). To better understand their unexpected shape, the semivariances are broken down into pairwise spatial autocovariance terms. This reveals an accumulation of autocovariances close to zero over short distances, which confirms the local heterogeneity from the semivariances. Two further peaks are noteworthy: one reaches into the positive values (clustering), and another goes into the negative values (repulsion). Positive, negative and neutral spatial associations thus occur together, rendering an intuitive ad hoc interpretation of the trend of the estimated semivariances in the sense of negative spatial autocorrelation problematic.

Strongly interacting spatial units existing in the superposition area of different processes change the distribution of spatial autocorrelation statistics. The eigenvalues of spatial weights matrices (which determine the potential for interaction between units) provide a detailed understanding of how the connectedness in a spatial layout influences spatial autocorrelation. Figure I.4.2 gives the eigenvalues for spatial weights obtained for the synthetic data. The non-superimposed points show a homogeneous eigenvalue pattern, meaning that each spatial unit contributes equally to the estimation of spatial autocorrelation. In contrast, the superimposed eigenvalue pattern is diverse with simultaneous high and low values, whereby strong eigenvalues occur for those units that interact across the two superimposed patterns. These units increase the variability in the eigenvalue spectrum, which influences the range of possible Moran's *I* values and stretches and changes their distribution (Tiefelsdorf and Boots 1997; Tiefelsdorf et al. 1999). The interaction between the superimposed patterns thus has a strong influence on spatial pattern disclosure. Inferences drawn about Moran's *I* are then biased and have lower statistical power, in particular when utilizing normal theory as derived under the central limit and the Pitman-Koopman theorems (Cliff and Ord 1973; Cliff and Ord 1981).

Several possibly contrary spatial processes are incorrectly identified in superimposed data. Moran scatter plots allow the decomposition of Moran's *I* into its parts. Even though only one spatial pattern exists in the synthetic superimposed data, the scatter plot reveals three different spatial processes. One of these reflects the actual spatial interaction in the data and shows a positive slope in the regression line (both combined patterns are autocorrelated at $I = 0.81$). In addition, another spurious positive and a negative sloping line appear, representing interactions between the two processes involved. Each of these lines is associated with one of the scales of the two processes. However, if not sorted out explicitly, they are included in the characterization of the overall spatial interaction behaviour of the analysed data. Setting these components into relation with the eigenvalues outlined above further shows that the two components behave in opposite directions when the eigenvalues increase. However, in both cases their influences become stronger, showing the strong effect of spatial superimposition on the interpretation of spatial patterns.

Superimposed mean values of different intensities change the interpretation of Moran's *I* estimates. With regard to statistical differences, the determined strength of spatial autocorrelation is underestimated with strongly differing mean values if the overlapping patterns are spatially random, *i. e.*, if these are themselves not spatially autocorrelated. In contrast, the degree of underestimation of spatially structured overlaid patterns additionally depends on the geometric scale associated with the stronger of the mean values involved. Therefore, dominant large-scale patterns cause stronger underestimation, and the rate at which these effects become operational is faster than with dominating small-scale patterns. The effect of differing mean values thus leads to a misinterpretation of the magnitude of spatial patterning in data.

Different attribute variances in superimposed random variables increase the uncertainty of spatial autocorrelation estimates. The uncertainty in the estimation of Moran's *I* increases, when the variances of overlaid attributes differ. This effect is similar for both, spatially random and spatially structured overlapping patterns. Also, with respect to the scales of the involved patterns, the influence of differing variances is symmetric. It makes no difference whether the larger or the smaller-scale pattern dominates in terms of dispersion. The variance impact becomes effective quickly, meaning that even small variations cause rather strong increases in the range of Moran's *I* values. What is further noteworthy is that variance deviations in general lead to a prevalence towards larger Moran's *I* estimates. That is, while mean deviations cause underestimation, differences in the involved variances may counterbalance the impact of the means. This finding adds to the counterbalancing effects detected for the different jointly occurring spatial processes in the Moran scatter plot. It suggests that, when both effects appear simultaneously, their impacts might become even stronger.

RQ 1.3: Joint impacts of scale and misspecified spatial weights on spatial autocorrelation

Research question 1.3 focuses on the joint influences of the systematic scale and topological effects revealed in RQ 1.1 and RQ 1.2. Nine-thousand random point patterns, each containing two differently scaled sub-patterns, have been generated that mimic different interaction scales, various types of superpositions, and two different attribute dispersal mechanisms. These patterns are investigated in two regards: (i) the numbers of interactions between differently scaled patterns are analysed to determine the extent to which these interact; (ii) the impact of scale differences and related topological effects on Moran's *I* is assessed.

The number of spatial interactions possible between overlapping patterns is constrained by their scale differences and the adjusted analytical scale. Two scenarios are investigated in this regard: one using a fixed degree of geometric overlap, with this condition being abandoned in the second case studied. The adjusted analytical scale turned out to be unimportant when the degree of pattern overlap is kept constant. Small scale differences between the jointly analysed patterns lead to high numbers of interactions in this case. This number declines at an exponential rate and levels out to a low figure as the differences in scale become greater. The latter convergence happens because few points from one sub-pattern interact with only few of the other. In contrast, letting the degree of overlap vary causes the interaction behavior to be highly dependent on the analytical scale used. On small analytical scales, the frequency of interactions remains high even when the sub-pattern scales differ strongly. In this case, many observations from a small-scale process interact with few large scale points, meaning that the disclosure of spatial structures is then dominated by relatively few and eventually extreme cross-pattern interactions. Contrarily and by analogy to the case of fixed geometric overlap, the frequency drops exponentially on large analytical scales. The spatial analysis of superimposed random variables is hence very sensitive to the combination of differently scaled spatial patterns and how these are arranged relative to each other. Scale differences between overlapping patterns thus determine to what extent the other disruptive factors (*e. g.*, statistical influences) described above can become effective.

Scale differences influence the characteristics of the processes incorrectly identified in spatially superimposed data. Interactions reflecting the actual pattern in the data do only provide reliable estimates of Moran's *I* when the involved scales of contained sub-patterns are almost similar. Very quickly, the autocorrelation drops notably causing Moran's *I* to be underestimated. This is an important finding, because the scales of combined patterns are very likely to vary marginally in practical scenarios. However, over large parts of medium scale differences, spatial autocorrelation tends to be overestimated, before it drops again and finally converges to a state of underestimation. In contrast, the superimposition-related false interactions between the involved patterns behave different. These contribute a positive additive component to Moran's *I* on small and moderate scale differences. When the scale differences are stronger, their contribution turns to chaotic behaviour making interpretations of disclosed patterns difficult. Estimating Moran's *I* from superimposed data thus leads to unreliable and often highly unpredictable outcomes. In summary, Moran's *I* is estimated close to correct if the involved scales are almost similar. In contrast, Moran's *I* is underestimated at small scale differences, overestimated if scales differ moderately, and it shows unpredictable behaviours at large scale differences through chaotic inter-scale effects.

I.4.2 Research Question 2

Research question 2 puts forward novel methodological routines for the spatial analysis and characterization of spatially superimposed and heterogeneous random variables. One proposed method is a novel spatial hot-spot estimator based on G_i^*, which explicitly considers the spatial coincidence of different scales in local neighbourhoods. The second contribution is a local statistical test about the influence of spatial arrangements on the variance under non-stationary conditions. These methods are laid out in detail in Chapters II.1 and II.3, which conform to Westerholt et al. (2015) and Westerholt et al. (2018).

RQ 2.1: A hot-spot estimator for spatially superimposed random variables

Research question 2.1 addresses hot-spot estimation from superimposed random variables by proposing two contributions: (i) A novel approach to spatial weighting is presented that differs from available ones in stratifying neighbourhoods with respect to the contained interaction scales. In addition, (ii) a novel hot-spot estimator allowing to disclose hot-spots on different scales in a separated manner is derived.

Geometric and topological criteria allow to extract information on relevant scales from spatial neighbourhoods. A two-step procedure is put forward to derive spatial weights for superimposed random variables: a circular boundary is drawn first around each spatial unit, whose distance threshold corresponds to the geometric range of the analysed process. All pairwise relations between the contained random variables are then examined to assess whether these are located on the scale at which the analysed process is assumed to interact. The derived weighting scheme thus forms a hybrid approach incorporating geometric (the circular boundary) and topological principles (the relative placement of spatial units) that allows to stratify local neighbourhoods into various distinct but geometrically overlapping parts. Certain scale ranges can be switched on and off which can then be evaluated separately. In summary, whereas conventional available spatial weighting schemes assume neighbourhoods to be internally coherent without the need to further sub-stratify them, the introduced approach contributes a scheme for the case of spatially non-exclusive geographic random variables.

Hot-spots can be disclosed separately on different scales by consistently limiting all stages of the statistic to relevant scales. A modified version of G_i^* called GS_i^* is derived and its first two moments are determined. In GS_i^*, the normalizing denominator takes account of the existence of different spatial scales through evaluating a binary vector indicating scale fit. The original method takes no account of different scales and uses a constant denominator which is based on all data available. In addition, the numerator of GS_i^* is integrated with the proposed weighting scheme. This allows to sort out irrelevant information, but requires a correction in the degrees of freedom. Finally, expressions for the mean and variance of GS_i^* are derived which are constrained to relevant scales, too. The finally proposed measure is presented as an asymptotically normal z-score, which facilitates convenient inference and interpretation. The presented measure makes it possible to evaluate hot-spots on different scales in an isolated manner and thus to disclose otherwise undetectable phenomena.

Hot-spot detection is more reliable when it is constrained to relevant scales. The application of G_i^* and GS_i^* to a Twitter dataset demonstrates the usefulness of GS_i^* for this type of data. While the mean value of G_i^* shows a strong positive trend on larger scales, the trend line for GS_i^* is flat and remains close to zero, which is the expected behaviour for z-scores. The maps in Figure I.4.3 further show that GS_i^* allows a better identification and separation of spatial hot and cold-spots on different scales. The diverse central business district (CBD) of San Francisco appears as a strong cold-spot[2], while the Asian quarter in the northern part features a prominent hot-spot representing Chinese New Year celebrations. In addition, small hot-spots appear in central neighbourhoods (*e. g.*, a college campus) on the largest analysed scale, while others are only present on smaller scales (*e. g.*, a secondary school in the north). These phenomena

[2] The attribute studied is semantic similarity.

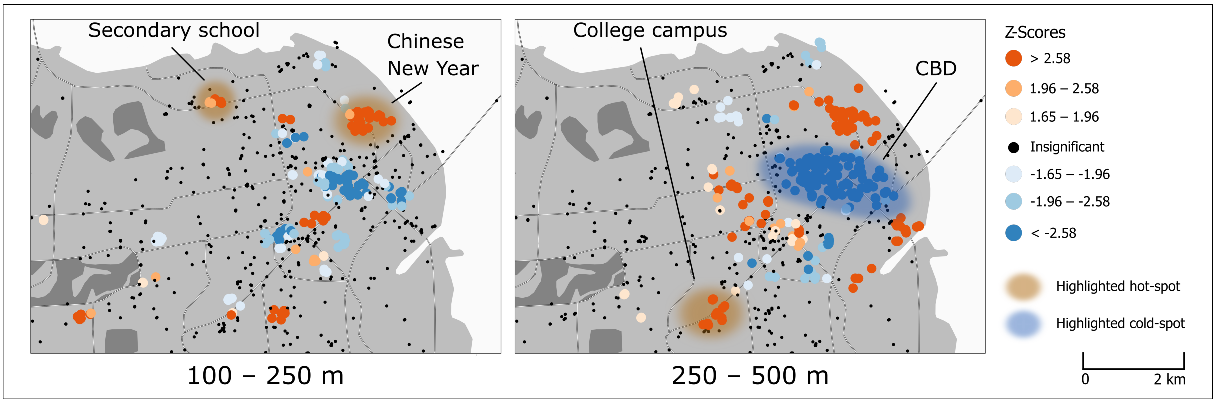


Figure I.4.3: GS_i^* scores obtained on different scales; based on Figure II.1.6.

are either undetectable with G_i^* (type II errors), or remain dominant features throughout most analytical scales (type I errors). G_i^* thus shows a high number of false positives on large analysis scales with 33.56% of all random variables being flagged significant. This is not the case with GS_i^*, which evaluates 3.77% of all observations as significant and thus comes close to the specified error probability of 5%.

RQ 2.2: Testing relationships between spatial arrangements and the local variance

Spatial superposition affects the geospatial arrangement of random variables, which in turn influences their variability. To better understand geospatial mixtures of processes and how places are organized spatially, a statistical test about the relationship between spatial arrangement and the variance has been proposed. This test statistic is based on two principles: (i) it only makes use of local information to permit global spatial heterogeneity; and (ii) it includes a strictly local inferential framework. These principles allow for assessing whether the way how random variables are arranged geospatially reduces or increases the variance in a certain location, or whether these two characteristics are unrelated.

Local dispersion analysis allows to test the impact of spatial arrangement on the local variability. The proposed measure called Local Spatial Dispersion (LSD) makes it possible to test the influence of a spatial pattern on the local on-site variability without being influenced by the overall geographic variance distribution. In contrast, Local Spatial Heteroscedasticity (LOSH), a recently proposed technique from which LSD is derived, reveals hot-spots of variability that stand out in a global comparison, but fails to detect entirely local variance structures that do not appear to be outstanding globally. The key technical difference to LOSH is that LSD compares estimated residuals about local, spatially-weighted mean values to their own local averages, whereas LOSH incorporates the global average of these residuals in local comparisons. This way, LSD assesses the entirely local impact of spatial arrangement on the variance without taking reference to the dispersal behaviour in other locations. It is thus possible to detect and characterize how the local geospatial layout affects the variance even in locations that are identified as globally non-significant by LOSH.

Local inferences and the prediction of additional synthetic data increase the reliability in testing for local spatial dispersion. The inferential framework introduced for LSD permits second-order spatial heterogeneity by using local randomization. However, drawing inferences locally comes at the cost of basing decisions on sparse information from potentially small neighbourhoods leading to unreliable reference distributions. To overcome this issue, the proposed solution includes a Bayesian framework for predicting additional synthetic local mean values. Additional local residuals can then be estimated about these means, making it possible to calculate a local bootstrap from any number of Monte Carlo replications. In a first step, the global statistical information is assessed from all available local, spatially-weighted mean values to derive their averaged prior distribution. The local information from the neighbourhood in question is omitted in this step to prevent a double use of data. The prior is then combined with local information to adapt the initially constructed distribution to local conditions. This approach has two advantages: the global prior reduces the risk of local overadaptation to eventually extreme observed situations, whereas the use of local information avoids strong global averaging and leads to a more realistic depiction of local dispersal behaviours.

Joint evaluation of LSD and LOSH reveals a detailed spatial variance characterization. The joint application of LSD and LOSH to a LiDAR-derived dataset of height differences in an Alpine meadow makes it possible to disclose features that cannot be detected by using either measure alone. For instance, a haystack is discriminated from another area of high variability generated from fence posts through taking account of locally (LSD) and globally relevant (LOSH) variance characteristics simultaneously. Both are outstanding global features, but their internal diversities differ: the haystack is locally homogeneous, whereas the fence posts cause fluctuation. Further, the joint assessment of both measures also allows the variance structure of a global dividing line between a mown and an unmown part of the meadow to be characterized in great detail. The internal structure of this boundary is homogeneous at its centre, but gets more diverse and fuzzy towards its edges. This finding cannot be obtained by using either measure alone, because the local internal structures are not remarkable globally and the solely local variance behaviour does not provide a comprehensive characterization. By analogy, further features could be revealed and characterized. This shows two things: Evaluating both measures jointly allows to characterize the local structures of global features. It is also possible to identify local features that would otherwise go unnoticed.

I.4.3 Research Question 3

Research question 3 reviews the role of spatial analysis in studies that use geosocial media data. Two contributions are made: (i) The areas of application and the general strategies applied in the spatial analysis of geosocial media data are explored. Further, (ii) the impact of conceptual and institutional shortcomings on empirical spatial studies are identified and discussed. The outcomes summarized here are based on Chapter II.5 which corresponds to Steiger et al. (2016c).

Geosocial media data are commonly aggregated geographically before their spatial analysis. Most available studies that use geosocial media data are conducted on regional or coarser geographic scales. Individual social media messages are thereby aggregated into areal units like census tracts, administrative areas or grid cells. Fields where this is prevalent are the analysis of collective human dynamics (*e. g.*, commuting behaviours), delineations of socially coherent living environments, and the detection of physical urban structures. Oftentimes, the choice for geographic aggregation is made for pragmatic reasons. For instance, when using additional information like census data or socio-economic indicators, these are usually only available in aggregated form or measured at coarse scales. Such additional covariates are frequently used for finding similarities with established datasets, which is done because the majority of studies still investigate the potential of geosocial media data for empirical research. This goal partially explains the need for aggregation. Another reason is that, at smaller scales, required knowledge about the scales of phenomena or their form of associations in space is often lacking. An extensive literature therefore exists on findings on coarse scales, while very little is known about non-aggregated geosocial media data on local scales.

Analysis of unaggregated geosocial media data is rarely conducted explicitly spatial. Some studies do analyse unaggregated geosocial media data. Typical domains for this are event detection, geo-social network analysis or the investigation of sentiments and emotions. However, the processes investigated are oftentimes unknown and analyses are exploratory in nature. Knowledge about appropriate spatial parametrization is then lacking and the technical issues discussed for RQ 1 and RQ 2 hamper the achievement of thorough insights about the spatial behaviour within individual messages. As a result, the individual messages are treated in a spatially isolated fashion—for instance, when assigning the messages emotional scores without taking into account their geographic context and relations. Therefore, spatial analysis is often reduced to the mapping of non-spatial results, whereas more complex spatial patterns remain undetected.

Analysis of geosocial media is often carried out in non-spatial disciplines. Beyond the technical issues discussed above, another noteworthy observation is that most geosocial media analyses are performed by scholars from non-geographic backgrounds. For instance, computer scientists have a strong record in event detection and the spatial analysis of linguistic patterns is a major focus of computational linguists. However, these researchers are often unaware of the importance of accounting for spatial associations (with respect to both technical and substantial implications) when it comes to the analysis of spatial random variables. A stronger interdisciplinary effort between geographers and researchers from other empirical disciplines is therefore needed and could help overcome this discrepancy.

I.4.4 Research Question 4

The appearance of superimposed and heterogeneous random variables is not limited to user-generated geographic content. They also appear in survey research where a novel paradigm called *event sampling method* (ESM) provides an event-driven approach to collect surveys *in situ*. The following paragraphs draw parallels and reveal differences between the characteristics of geosocial media data and ESM survey responses. A detailed elaboration is found in Chapter II.4 according to Bluemke et al. (2017).

User-generated and scientifically collected *in situ* random variables share technical challenges. Like geosocial media data, ESM responses are collected from contextual conditions. The latter are partly reflected in the samples and influence the collected contents. In addition, the contributions express the idiosyncratic spatial concepts used by the respondents. The resulting data are therefore inherently heterogeneous and of limited intersubjectivity. In addition, people assign different meanings to similar places and processes, leading to the collection of phenomena that appear more diverse in the data than they really are. This means that, even though ESM responses are more structured through the use of predefined questions, many of the issues identified in this thesis for geosocial media do also apply for this kind of data. ESM responses and geosocial media feeds are thus very similar and give rise to a joint methodological effort by geographers and psychologists.

Different forms of superimposed, heterogeneous random variables exist. The data collection of ESM responses and the goals in their analysis differ from those of geosocial media data. The latter is collected in the vein of the humans-as-sensors concept, whereby the premise is that people sense their immediate environments and publish this information. In contrast, ESM responses are used to collect information about individuals while they are influenced by their momentary spatial contexts. Geosocial media analysis thus focuses on space and place, whereas these are treated as contextual covariates in case of ESM responses. For this reason, the data acquisition schemes used are different: Geosocial media uses a liberal scheme of unprompted messages and makes only few constraints about the collected contents. In contrary, ESM is based on a structured survey approach with pre-defined questions and spatial triggers to ensure a controlled contribution of the responses. ESM is therefore less affected by the self-selection bias, which is a serious problem with geosocial media data, as the collected samples then constitute an insufficient representation of the analysed processes. However, ESM-based responses are in turn affected stronger by the individual spatial capabilities of the surveyed persons. These include potentially biased distance estimates, insufficient local spatial knowledge, or disturbing external influences. In summary, while many technical challenges are similar, the issues in terms of content and interpretation are different. This shows that the notion of spatially superimposed and heterogeneous random variables is a largely technical one in the first place and that different kinds of these variables exist.

I.4.5 Limitations

Even though the obtained findings are generalizable, limitations exist in the applied research design that narrow the scope of the drawn conclusions. These restrictions are organized contentwise:

Spatial and temporal limitations

One limitation of the results is the spatial focus of the work and that *time is not considered*. This deliberate decision is made for the sake of clarity in the drawn conclusions. The impacts of superposition on either dimension—space and time—are both not yet well understood. Their joint investigation would therefore hamper the interpretation of disclosed findings, making it difficult to separate spatial from temporal superposition effects. For this reason, all obtained results are interpreted in a solely spatial manner. In addition, the applied concept of geographic space is constrained to a *Euclidean notion*. Non-Euclidean concepts exist in geography, but the choice of an appropriate concept requires considerable knowledge of the investigated phenomena. The everyday phenomena represented in geosocial media are not well understood, which is why the more conventional Euclidean approach is chosen here. However, while the idea of straight-line distance is intuitive, other types of geographic associations might be reasonable too and could lead to additional complementary insights. Based on the decision to work with the Euclidean vector space, the modeling of geographic relations is limited to *distance decay effects*. This weighting scheme reflects the first law of geography in an intelligible way, but a plethora of other spatial weighting schemes is available that could unveil differing results, though the general characteristics revealed should remain similar.

Statistical limitations

With regards to statistical configurations, it is of note that only *certain statistical characteristics* have been investigated. For instance, the synthetic data used is populated with normally distributed attribute values. These are then tested for interactions with geometric and topological factors, as well as with regard to different mean-variance combinations. However, different statistical populations are likely to occur in reality, including their combinations in the form of mixture distributions. All obtained conclusions are therefore interpreted conditional on the chosen statistical setups. Similarly, with respect to the generated statistical point patterns used for answering RQ 2, the *generative mechanism for constructing the employed patterns* is based on a random walk procedure. However, different kinds of geometric point dispersal appear and these could reveal further interesting outcomes that could not be investigated in this work. In the same vein, the *ways to overlay different point patterns* could also be varied further. Most of these limitations are caused by the necessity to control statistical parameters in order to facilitate the clear interpretation of the experimental results. Varying all parameters at the same time would hamper this endeavour, which is why the laboratory-like and idealized synthetic data was used.

Methodological limitations

This thesis focuses on the investigation of the impact of spatial superposition on *selected methodologies*. These methods include Moran's *I*, G_i^* and LOSH, which were chosen for their relevance to practical scenarios and empirical research. Even though most other estimators of spatial autocorrelation (*e. g.*, Geary's *c*) should behave similar (these form a common family of special cases of the Mantel test; Hubert et al. 1981; Getis 2010), the likeness of such results is not guaranteed. Separate investigations including a comparison with the results obtained here could shed light on the actual transferability of the achieved scientific contributions. Another constraint related to the choice of methods is the *limitation to real-valued attributes*. Spatial patterns can also be assessed within other statistical data types, such as categorical and integer random variables (*e. g.*, dichotomous variables like gender, or counts). The set of methods and their distributional characteristics are different in these cases, and the impacts of topological outliers, scale differences and other effects might differ from those disclosed here. Further, in terms of the presented novel methodological approaches, it is noteworthy that the derived GS_i^* hot-spot estimator is based on the *assumption that scales are geometrically separable* within superimposed point sets. This is clearly a strong and restrictive assumption, partly based on the choice of spatial weights outlined above, which literally moves the problem of inseparability from the attribute space to the geometric domain. Even though the experimental results are promising, there might be cases where GS_i^* might not perform well. Similarly, the presented LSD method uses *global prior knowledge and local information both in equal parts*, which may not be generally suitable in all possible use cases.

Scope restrictions

The most profound limitation in the scope of the obtained results is their restriction to user-generated data collected from *liberal and unmoderated acquisition schemes*. This implies that users are free in the choice of locations, times and contents of their contributions. Other forms of user-generated geodata that also represent the everyday behaviours of people include check-ins or shared photos. These are more constraint and likely to show other statistical and spatial characteristics. Therefore, the results presented are not trivially on-by-one transferable to these types of data. By analogy the discussion of the role of spatial analysis techniques in previously conducted analyses of geosocial media content is limited to *certain selected fields* (human mobility, event detection and few related areas). While it is still assumed that the discussion is representative, other application domains might exist which are not fully covered by the drawn conclusions. Similarly, the comparison of geosocial media content with ESM responses is *only one possible way of connecting the results obtained with other fields* by reviewing methodological and data similarities. This non-exhaustive treatment has been undertaken to keep the discussions tractable and for starting an interdisciplinary dialogue between related fields.

I.5 Synopsis and Conclusions

This thesis studied the spatial analysis of spatially superimposed random variables. These variables appear in user-generated datasets like those extracted from geosocial media feeds and they partially represent externalizations of everyday spatial practices of people. The findings obtained provide basic knowledge necessary for the spatial-statistical characterization and understanding of these types of data and related social phenomena. In order to enable a holistic view of the topic, four aspects have been addressed: (i) the spatial data characteristics of superimposed random variables were investigated, (ii) suitable methodological approaches were derived, (iii) a contribution was made to the exploration of how spatial analysis with superimposed random variables is carried out in empirical studies, and (iv) conceptual and methodological commonalities with neighbouring fields and related types of statistical data were determined. The following paragraphs establish connections between these individual parts and conclusions are drawn from this synthesis. The chapter closes with the major key conclusions of this thesis.

I.5.1 Synoptic Integration

Geosocial media and related ambient user-generated datasets represent information about human experiences with and within places. Places are experienced individually and are defined as locations infused with human meaning (Tuan 1977; Agnew and Livingstone 2011). Their representations thus appear multi-layered in collective spatial datasets, making their spatial analysis a challenging task. This is reflected in the mainly aggregated type of spatial analysis found in the literature and discussed in Chapter II.5 and Steiger et al. (2016c) for RQ 3. Apart from intended large-scale analysis, this strategy is a workaround to avoid the technical challenges widely discussed in this thesis. For instance, in Bakillah et al. (2015), one of the studies discussed, Twitter messages about disaster-related damage sightings are investigated by applying an algorithm treating the spatial point data in an aggregated way. This is done to mitigate the influence of the low intersubjectivity of the messages and to instead find a spatial consensus. However, the descriptions provided by the people are related directly to the concept of place, since they reflect individually perceived impressions of damaged places. Their combination and the aim of reaching a consensus is therefore questionable and, at best, a simplification. What these findings reveal is that the widely adopted humans-as-sensors concept (Goodchild 2007) needs to be extended beyond space by including stronger the notion of place.

Similar arguments hold true for the analysis of *in situ* survey responses discussed in Chapter II.4 and Bluemke et al. (2017) for RQ 4. The contextual conditions of the responders make them provide subjective representations of perceived geographical, mental and other situations. These are further strongly influenced by personal traits and the application of idiosyncratic distance estimations and scale assessments. Therefore, both geosocial media data and *in situ* survey responses suggest that the spatial analysis of superimposed random variables corresponds to the analysis of spatial representations of actually subjective platial[1] data. Given the evident differences between geosocial media and *in situ* surveys, this finding demonstrates the broad relevance of the research conducted in this thesis. A range of user-generated datasets face similar technical and conceptual challenges and their joint methodological treatment will benefit a better understanding of how people interact geographically with digital media.

The platial nature of geosocial media data is confirmed by the analytical results obtained for RQ 1. The diverse spectrum of the contained scales discovered in Chapter II.1 and Westerholt et al. (2015), gives hints that the contributed information describe individual encounters with places. These are based on different phenomena and how these are perceived by different social media contributors, but also on the idiosyncratic cognitions discussed in the previous paragraph. Geosocial media and related data are thus externalizations of how people perceive different yet simultaneous phenomena occurring in geographic locations. Different users have different preferences, which, in turn, depend on context and individual characteristics. The users therefore communicate different aspects of a local geography and their behaviour therein to geosocial media feeds. This creates a rich digital mirror image of the platial characteristics of locations. The spectrum of overlapping scales obtained from Twitter messages is therefore a reflection of the complexity of the (social) geography of an area.

The distortion of spatial patterns caused by simultaneously observed local clustering and repulsion behaviours (Chapter II.2 and Westerholt et al. 2016) further corroborates the platial nature of geosocial media data. Sometimes people communicate related phenomena (clustering), whereas they may also utilize places in completely unrelated ways (repulsion). The spatial pattern detected from such data cannot be interpreted unambiguously, which is why aggregation was found to be the preferred way of anaylsis with this kind of data (RQ 3). However, the simulation experiments (given in Chapters II.1, II.2 and II.6 and in Westerholt et al. 2015; Westerholt et al. 2016; Westerholt 2018) revealed that the technical issues that occur with individual-level analysis are in fact also transferred to aggregate-level results. For example, it was shown by a semivariogram how analysing Twitter messages in a combined way (though still using the individual messages) leads to erroneous conclusions about spatial patterns. This was further corroborated by showing that complex interactions exist between scales and other characteristics and that these do not vanish when treating data in an aggregated manner. Aggregation (either prior to or during an analysis) is thus a rather simple remedy to circumvent place-induced technical issues. The obtained analytical results therefore highlight the problems of aggregated analyses and provide a detailed understanding of the spatial characteristics of platial data. These achieved insights called for methodological contributions to better incorporate the spatial characteristics of platial data, and for gaining an enhanced understanding of these in empirical scenarios.

In his seminal paper on the cornerstones of *GIScience* Mike Goodchild suggested that in the future methods will be needed to investigate overlapping spatial continuities reflecting the complex geographic nature of the human-comprehensible world (Goodchild 1992). This forward-looking statement anticipates the idea of place and platial analysis. However, GIS and GIScience mostly remained in the vein of crisp vector units or field based notions for representing spatial phenomena. A slow shift towards the notion of place and place-based GIS is only taking place recently. Goodchild further conjectured that spatial statistics will play a pivotal role in these developments. This thesis provides insights that support and advance the early visionary views of Goodchild. The results obtained for RQ 1, RQ 3 and RQ 4 (outlined and connected in the previous paragraphs) provide findings that reveal a strong connection between the spatial-statistical characteristics of platial data and drawn conclusions about spatial patterns related to these. Methodologically, this work is thus in the vein of the platial analysis movement and it

[1] The term *platial* is used synonymously with the term *place-based*.

contributes to an enhanced understanding of the complex set of overlapping continua representing the human-comprehensible world.

The methodological contributions obtained for RQ 2 in Chapters II.1 and II.3 (corresponding to Westerholt et al. 2015; Westerholt et al. 2018) contribute to the advancement of platial analysis. They explicitly incorporate the characteristics of spatially superimposed random variables disclosed in the empirical parts of this thesis. For instance, the spatial weighting scheme proposed in Chapter II.3 takes account of the mixture of scales found in spatial representations of platial data, as was investigated in Chapter II.1. Similarly, the method LSD derived in Westerholt et al. (2018) puts forward the investigation of spatial patterning of spatial variance. This characteristic is related to the strong, spatially overlapping heterogeneity evident from Chapters II.2 and II.6. Thus, this connection shows that LSD allows to investigate heterogeneity caused by topological and statistical characteristics of superimposed random variables, and thus a spatial characteristic of places. The methodological achievements in this thesis therefore contribute to the recent advancement of a theory on platial analysis and platial GIS, both of which have recently been named some of the foremost research topics in GIScience (Duncan 2011; Goodchild and Li 2011; Goodchild 2015).

The findings obtained represent a major step towards place-based analysis and will influence a number of empirical research areas within and beyond geography. Findings from the behavioural sciences and from psychology suggest a strong link between personal, demographic and other characteristics and the subjective cognition of places (*e. g.*, Weiss et al. 2003; Dangschat 2007; Witt et al. 2010; Zadra and Clore 2011; Sugovic and Witt 2013). Thus, the obtained findings are of relevance to all types of analyses using user-generated geographic data in the AGI sense. These have gained momentum recently and include human mobility investigations (*e. g.*, Wu et al. 2015; Steiger et al. 2016b; Steiger et al. 2016a), natural hazard analysis (*e. g.*, Thomson et al. 2012; Crooks et al. 2013; Albuquerque et al. 2015) and event detection (*e. g.*, Hiruta et al. 2012; Sakaki et al. 2013; Cheng and Wicks 2014), among numerous other fields. However, making sense of these data quantitatively beyond the purely spatial domain requires a mature GIScience theory of places and their analysis, to which this thesis contributes important findings. Beyond place, the thesis additionally allows better ways to find spatial solutions to enhance the spatial analysis of these kinds of data. This was for instance demonstrated by the refined hot-spot detection presented in Chapter II.1, where taking account of the spatial organization of places significantly increased the meaningfulness of the obtained spatial analysis results. It is thus expected that the findings obtained provide geographers and related spatial scholars the means to make sense of user-generated georeferenced and ambient data, and to provide a profound impetus to the ongoing advancement of platial analysis.

I.5.2 Main Conclusions

Overall, the findings indicate that spatially superimposed and heterogeneous random variables are spatial-statistical representations of platial information. This thesis therefore confirms recent discussions conjecturing geosocial media data to be of largely platial nature (Quesnot and Roche 2015; Roche 2016; Jenkins et al. 2016; Mckenzie and Adams 2017). On this basis, and on the basis of further evidence presented in this thesis, the following main conclusions are drawn:

Everything is related to everything else. This well-known quote from Waldo Tobler's first law of geography (Tobler 1970) summarizes one of the most important analytical findings of this thesis, which is the interrelatedness of most of the investigated statistical characteristics. The investigations conducted revealed strong connections between scale, statistical parameters and topological effects. For instance, changing the scales of overlapping processes influenced the effect of mean values on spatial analysis results (Westerholt 2018). The mean values, in turn, impacted the strength of the falsely disclosed non-existing processes that were found to interfere with spatial autocorrelation results (Westerholt et al. 2016). The meaning of these concatenations of joint effects is two-fold: On the one hand, they complicate the traceability of the influence of individual statistical characteristics on spatial analysis results. Figuring out the influence of all of them together in a complex real-world empirical scenario is thus difficult, and eventually intractable. On the other hand, the disclosed connections between the effects of characteristics show the sensitivity of spatial analysis methods regarding data characteristics when platial information is analysed in a spatial way. This is partly contrary to the results obtained by Griffith (2010), who investigated Moran's *I* and found a certain robustness of this measure against non-normal distributional characteristics of random variables. The discrepancy shows that varying a single characteristic (like distributional assumptions) is not sufficient for a holistic understanding of statistical variations when it comes to the analysis of complex platial data like it was considered in this thesis.

Superimposed random variables are spatial representations of platial phenomena. The results obtained indicate that the random variables analysed in this thesis reflect subjective human interactions with geographic space. The discussion of existing empirical studies in Steiger et al. (2016c) revealed a strong focus on user subjectivity. In some cases, that subjectivity is treated as a nuisance calling for treatment prior to further analysis (*e. g.*, Sengstock and Gertz 2012). In other cases, researchers exploit the wealth of personal information it offers (*e. g.*, Steiger et al. 2014a; Steiger et al. 2016b). In almost all cases, however, the focus of the analyses is on how people perceive and interact with geographic space. For instance, the reviewed analyses of human mobility behaviours investigate individual activity spaces. By analogy, the discussed studies on detecting urban structures typically reveal information about how people perceive the topography of a city. Additionally, the derived information are often enriched by further qualitative dimensions like emotions and sentiments (*e. g.*, Mitchell et al. 2013; Resch et al. 2015d) or social network properties (Croitoru et al. 2015). Although the derived information referred to above is technically treated as spatial random variables, subjectivity combined with a number of additional content dimensions gives a strong indication of the actually platial character of this data. This impression is confirmed by the analytical results obtained in this thesis and makes spatially-superimposed random variables spatial representatives of platial phenomena.

Analysis of superimposed random variables provides clues to the conjectured container property of places. A location together with its characteristics is considered a place when it is perceived as a unique whole that can be contrasted with the literal "everything else". Places have been conjectured to be containers that provide a context for the emergence and development of phenomena (Johnson 1987; Winter and Freksa 2012). This idea has mostly been discussed conceptually so far. However, the results obtained in this thesis contribute empirical evidence for this characteristic. For instance, the fact that multiple scales co-occur (Westerholt et al. 2015) shows that each of the underlying causal processes is evolving in its own container, largely detached from the others and only connected through certain shared contextual parameters. Beyond these, each of the phenomena instantiates its own space in which it evolves. By analogy, the simultaneous observation of clustering and repulsion behaviours (Westerholt et al. 2016) indicates that different types of spatial processes utilize similar spaces, but require different platial affordances. One implication of this is that statistical characterizations of superimposed random variables enable important insights into the spatial organization of places. The spatial analysis of this kind of random variables can thus be considered an important tool for getting a better understanding of one of the dimensions of places. This thesis thus gives strong empirical indications for the *places as containers* argument and makes a statistical connection between space and place.

Some issues caused by superimposed random variables are related to known statistical problems. It was demonstrated in this thesis that the spatial analysis of superimposed random variables faces statistical problems. These were identified in the experiments conducted and it was shown that these may lead to erroneous conclusions. One of these is the dependencies between scales which are induced by their spatial overlap. This issue is structurally similar to the well-known *multiple hypothesis testing* problem that appears with local statistical tests (Caldas de Castro and Singer 2006; Nelson 2012). The spatial co-incidence of different scales leads to the repeated testing of hypotheses on different scales, whereby the same data is used each time. Like with local hypothesis tests, the significance level requires adjustment because there is a risk of making at least one error in each test that has been carried out, instead of just once. Another issue mimicking an established problem is the inherent heterogeneity of the investigated random variables, which is related to a lack of stationarity. This makes the corresponding data appear noisy and has led to a considerable methodical effort in eliminating this effect (Sengstock and Gertz 2012; Lovelace et al. 2016). Some of the methodological work conducted in this thesis can thus be considered noise reduction approaches, for instance, the geometric stratification approach presented in Westerholt et al. (2015). In this sense, the proposed technique $\mathrm{GS}_{\mathrm{i}}^*$ complements noise removal strategies from other fields like *minimum noise fraction* (remote sensing) (see Luo et al. 2016b) or *stopword removal* (natural-language processing) (see Saif et al. 2014) by a geometric *spatial noise* approach sorting out wrongly scaled information. However, though heterogeneity is a technical issue violating assumptions of spatial analysis techniques, the term noise is not appropriate in this context. Noise is a disturbing effect, but the heterogeneity within superimposed random variables is a reflection of the diversity attached to the represented phenomena. These two examples show that many issues are related to known problems. Existing solutions for solving these may thus allow bridging technical issues in the spatial with those in the platial world.

Spatial analysis results obtained with superimposed random variables shall be carefully revisited. The insights gained in this work could, after conducting a critical review, lead to the conclusion that many available empirical results obtained from geosocial media data may prove to be incorrect. These results were largely obtained by using spatial analysis methodology without taking account of the effects discovered in this thesis. A comparable situation has occurred in ecology in the early 2000s, after researchers had found that spatial autocorrelation was almost never taken into account in statistical analyses in that field (Diniz-Filho et al. 2003). It is possible that some of the findings from fields for which geosocial media data has been considered particularly valuable (*e. g.*, human disaster response, human mobility or spatial communication behaviour) are flawed in a similar direction in that the place-based nature of the analysed data and the respective spatial implications may change the nature of the already obtained findings. For example, the disclosed impacts of spatially overlapping scales on hot-spot detection are severe. If these are not considered in analyses, the results are likely to indicate wrong centres of social activity and other processes. Further, quantitative results on spatial relationships (*e. g.*, spatial communication behaviour) may be equally distorted by varying statistical parameter values and their effect on the magnitudes and signs of spatial measures. The findings obtained in this thesis may therefore not only provide an impetus for future work towards platial analysis, but could also provoke a critical discussion of existing geographical findings regarding geosocial media and related sources of information.

I.6 Future Research

The conclusions drawn and the knowledge gained from this work permit a wide range of further research. Therein, the following recommendations have been identified as particularly valuable. These focus on the fields of spatial analysis, place-based analysis, and additional adjacent areas.

I.6.1 Spatial Analysis

Interesting future opportunities can be found by integrating superimposed random variables further with the field of spatial analysis. It was demonstrated in this thesis that many established concepts are not directly applicable to superimposed random variables. However, many of these concepts are still important for obtaining reasonable results with these kinds of information, and substantial effort is thus needed to integrate the issues discovered in this thesis with available concepts from the spatial analysis framework.

One particularly important concept that requires adaptation in future research is stationarity. This concept is related to the notion of equilibrium and it describes a set of assumptions about the stability of statistical parameters within an observation area (see Section I.3.3). Stationarity is relevant to ensure the validity of auxiliary parameters, and for drawing reasonable conclusions based on appropriate null models. However, available spatially-exclusive notions like second-order or intrinsic stationarity are not useful when phenomena appear spatially intertwined. Analysing the related spatially superimposed random variables requires an operational adaptation of the stationarity concept towards the case of spatial (and temporal) simultaneity, allowing variation within local subsets of observation areas. One possible approach could be to develop a multivariate notion of stationarity by treating mixed phenomena with different parameters separately. Yet, this would require prior separation, which can be difficult if the information available on phenomena is scarce, as it is typically the case with geosocial media and related datasets. Another possible strand might be the spectral analysis of variances and other parameters (Fuentes and Reich 2010) by means of decomposing the mixture of superimposed parameter values into their constituent parts. Regardless of how, a suitable stationarity notion will be required in the future in order to make geographical sense of the wealth of collected georeferenced user-generated and ambient datasets.

Changing the concept of stationarity has broad implications on closely related topics such as asymptotics. For instance, second-order stationarity is required for a well-behaved asymptotic convergence behaviour of statistical estimators like Moran's I, by means of operational central limit theorems. In spatial statistics, asymptotics have been derived under two different premises: increasing domain, and infill asymptotics (Anselin 2001). The first of these is conceptualized by an infinite extension of the investigated spatial domain through adding observations to its edges. In contrast, infill asymptotics approaches the limit by dividing the observation area into ever smaller units. Both approaches are not feasible with superimposed random variables and novel concepts regarding asymptotics are thus needed. Infill asymptotics cannot be established because that would conceptually lead to non-stationarity in case of superimposed random variables. Very fine-grained subsets may then no longer be filled up with the mixture of processes that characterizes a region, which implies qualitative differences. Similarly, increasing the domain would simply mean to add more noisy mixture observations, which helps only marginally. Thus, a spatial stationarity notion for superimposed random variables requires to rethink a range of further concepts like the discussed asymptotics. Obtaining stable and reliable spatial-statistical estimators relies heavily on their successful future derivation.

Future methodological research should focus on finding generalized forms of measures of spatial autocorrelation like Moran's I, Geary's c and related statistics. This would enhance the understanding of how superimposed random variables are related to, and fit into, the conventional spatial analysis theory. Further, if generalized forms covering both these types of information exist, that would allow to elaborate a joint framework in which theoretical results for both forms could be obtained simultaneously. Such a joint framework would be in accordance with existing generalizations available in spatial analysis. For instance, it has been shown that most available measures of spatial autocorrelation are mutually related and that they form special cases of the Mantel test (Hubert et al. 1981). Further common frameworks have been presented for global (Lee 2004) and for local hypothesis testing (Lee 2009), both unifying different kinds of spatial hypothesis tests. Integrating spatial random variables with the superimposed random variables considered in this thesis could lead to two future results: It might either demonstrate that the superimposed case itself forms a generalization, subsuming the conventional variables as a special case. Another possible outcome could be the discovery of an even more general mechanism connecting superimposed and spatially-exclusive forms of analysis. The latter could serve as an interesting bridge between spatial and place-based analysis.

Another important future research topic is that of appropriate spatial weights. A large number of different spatial weighting schemes is available (see Getis and Aldstadt 2004; Getis 2009). However, as was shown in this thesis, spatially superimposed random variables require novel strategies beyond the available geometric, topological and empirical approaches. Future research should consider further properties of spatially mixed phenomena beyond the geometric domain that was considered in this thesis. This could include qualitative dimensions such as evaluated topic associations or other semantic features to maximize the number of spatial units correctly related. Time and cross-phenomenal relationships should also be considered to allow spatiotemporal and cross-correlation analysis. The temporal case, however, implies further conceptual difficulties. For instance, it is not always clear how space and time shall be treated together mathematically. Temporal scales of everyday phenomena depicted in geosocial media feeds can further be complex both in their own individual regard and with respect to their mutual entanglement. Elaborating on these aspects will enhance the plausibility and level of detail of future research results obtained from superimposed random variables.

Spatial data analysis evaluates spatial pattern in attributes by conditioning on fixed geometric layouts. In contrast, the related field of spatial point pattern analysis deals with stochastic geometries whereby the locations of points, networks and other geometries are themselves considered outcomes of random processes (an overview is found in Illian et al. 2008). Considering attributes in case of stochastic geometries (where attributes are called 'marks') is done by evaluating the so-called 'typical point', which results in fixing the geometric layout and letting the attribute vary upon this. One major technique to investigate these cases is the so-called mark correlation function, the different variants of which are mathematically equivalent to Moran's I, G_i^* and other measures from spatial data analysis (Shimatani 2002). However, spatially superimposed random variables are outcomes of two spatial processes: One being related to the decision of a user to post content in a specific location, and the other referring to the posted content itself. Based upon this presumption, analysing this kind of information spatially should consider both these random processes simultaneously. This requires models and measures that take account of the joint probability of both geometric layout and attribute at the same time. The result of that would be a spatial autocorrelation approach that allows evaluating the magnitude of spatial patterning within attributes and within the geometries simultaneously. Such a doubly-stochastic notion of spatial autocorrelation would be useful to a range of different fields, ranging from botany to the social sciences, and could bridge stochastic geometry with traditional spatial analysis.

I.6.2 Place-based Analysis

The importance of the relationship between platial analysis and spatially superimposed random variables was outlined and emphasized in the conclusions section (I.5). Future research should thus focus on the integration of place-based concepts with those from spatial analysis. The GIScience literature on places still pertains a strongly spatial viewpoint. Prevalent attempts to project places onto spatial maps illustrate this. Place is therefore often condensed into a mere attribute which is then treated as a function of space. Clearly, as shown by the literature from spatial cognition, social geography and related fields, the notion of place is more complex and powerful than its current analytical treatment in GIScience suggests. For instance, beyond their relationship to space, places have also been considered relationships between individual persons and locations (Mennis and Mason 2016). At a conceptual level, GIScience should therefore work towards better ways to formalize places and to make them available to place-based GIS and platial analysis. This may build upon available concepts like activity spaces (describing the everyday geography of individuals; Horton and Reynolds 1971; Golledge and Stimson 1997) and conceptual spaces (quality dimensions among which space is included as one dimension; Gärdenfors and Williams 2001), both of which could serve as valuable instruments towards formal GIScience place representations. The debate around places would also benefit from a more fundamental discussion of the ontological and epistemological implications of switching from spatial to platial viewpoints. Scheider and Janowicz (2014) provide some initial ideas, and future research should follow these strands and related discussions.

Viewed from the statistical perspective of this work, platial analysis in the sense of an analogue to spatial analysis is still a largely undefined field. Some initial attempts were made towards platial counterparts of spatial GIS functionalities like buffers and joins (Gao et al. 2013) as well as on qualitative spatial reasoning (which largely relies on the notion of place; Freksa 1991; Wolter and Lee 2010). However, a statistical platial pattern detection and characterization in analogy to spatial analysis is not yet in sight and many essential concepts are still missing. For example, we do not yet have a notion of platial units available. Spatial units provide containers used to model those parts of geographic space that are utilized by spatial processes for their dynamic diffusion. This thesis has underpinned the container notion of places in the sense that these provide the contextual conditions for processes to develop. However, it is yet unclear how a place-based counterpart to spatial units would technically and conceptually look like. Similarly, it is also unclear what kinds of statistical analyses would be possible and what a useful hypothesis testing framework would be for them. Is it possible to define a concept of platial autocorrelation, and, in case it is, what would a revised place-basesd first law of geography look like? These questions are largely unaddressed to date and GIScience should focus on the development of platial analysis in future research to enhance the understanding of places and to support quantitative (social) geography.

I.6.3 Related Research Areas

The findings obtained in this thesis encourage further conceptual work beyond spatial analysis and the space/place dualism elaborated above. As the discussions have shown, geosocial media are only one way to obtain spatially superimposed and heterogeneous random variables. For instance, the investigated *in situ* survey responses are structurally similar but differ with respect to various characteristics. Many further types of user-generated datasets may also fall into the same category. For example, user-collected air pollution sensor data (Sîrbu et al. 2015) or crowd-sourced physiological conditions and emotions (Resch et al. 2015b) may be affected by the same technical issues identified in this thesis. Although these types of information are very different from geosocial media feeds, the related datasets are also collected by users in a largely uncontrolled way. User behaviour patterns are thus likely to be contained, leading to similar kinds of heterogeneity. However, these additional types of spatially superimposed random variables may vary with respect to certain statistical and other characteristics. For instance, calibrated devices like carbon sensors carried by people in their everyday lives exhibit lower degrees of uncontrolled heterogeneity in their recording of content than human textual inputs provided by geosocial media. Still, the data collected is prone to contextual conditions and to the activities the contributors are engaged in during data collection. Forthcoming research should identify how and to what degree the statistical analysis of these types of data coincides with the results obtained in this work and the extent to which there are systematic differences. This will reveal a better understanding and disambiguation of different types of user-generated superimposed random variables—and thus ultimately of different facets of human everyday behaviours.

Another interesting relationship that should be investigated in the future is that with statistical mixing. Statistical mixing refers to the property of different gases or fluids mixed into each other to be indistinguishable after a certain period of time (Lebowitz and Penrose 1973; Frigg and Werndl 2018). For instance, if red and blue colour are blended into each other, that mixing leads to the colour violet. The formation of violet takes a while, but red and blue pigments can no longer be detected visually. The superimposition process found in the types of information considered in this thesis might also be analyzed and explained well in terms of statistical mixing. Different processes depicted in geosocial media or other datasets can be viewed as analogues to the colours and the superposition corresponds to the blend-over of these as was illustrated above. Viewed over time, this idea corresponds to a non-equilibrium starting point where different social processes start to develop in a certain region. After some time, because people utilize the same places in different ways, these evolve into a collective equilibrium-like state in which the entire geographical area appears like a noisy but almost uniform surface. However, even this theory from statistical mechanics is not directly applicable because the different social processes do not necessarily scatter evenly in a region, but, instead, the superposition process (*i. e.*, the mix of processes) becomes equalized. Further, statistical mixing refers to the asymptotic independence of a stochastic process. Thus, application of this theory is inevitably linked to future work on asymptotics outlined earlier in this section. It is only possible to investigate how processes mix over time and space, if it is clear what the term *limit* means conceptually with the kinds of information treated in this work. Therefore, this suggestion of adopting the well-studied theory of statistical mixing requires other future work to be solved first and is formulated as a long-term goal towards a better understanding of superimposed random variables, the composition and characterization of places, and, ultimately, of how people organize their everyday life geographically.

References (Summary and Conclusions)

Agnew, J and D Livingstone (2011). 'Space and Place'. In: *Handbook of Geographical Knowledge*. Ed. by J Agnew and D Livingstone. London, UK: SAGE, pp. 316–330. DOI: 10.4135/9781446201091.n24.

Albuquerque, J de, B Herfort, A Brenning and A Zipf (2015). 'A Geographic Approach for Combining Social Media and Authoritative Data Towards Identifying Useful Information for Disaster Management'. *International Journal of Geographical Information Science* 29 (4), pp. 667–689. DOI: 10.1080/13658816.2014.996567.

Aldstadt, J and A Getis (2006). 'Using AMOEBA to Create a Spatial Weights Matrix and Identify Spatial Clusters'. *Geographical Analysis* 38 (4), pp. 327–343. DOI: 10.1111/j.1538-4632.2006.00689.x.

Ames, M and M Naaman (2007). 'Why we Tag: Motivations for Annotation in Mobile and Online Media'. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Ed. by D Gilmore. San Jose, CA: ACM Press, p. 971. DOI: 10.1145/1240624.1240772.

Anselin, L (1988b). *Spatial Econometrics: Methods and Models*. Vol. 4. Studies in Operational Regional Science. Dordrecht: Springer Netherlands. DOI: 10.1007/978-94-015-7799-1.

— (1990). 'Spatial Dependence and Spatial Structural Instability in Applied Regression Analysis'. *Journal of Regional Science* 30 (2), pp. 185–207.

— (1995). 'Local Indicators of Spatial Association - LISA'. *Geographical Analysis* 27 (2), pp. 93–115. DOI: 10.1111/j.1538-4632.1995.tb00338.x.

— (2001). 'Spatial Econometrics'. In: *A Companion to Theoretical Econometrics*. Ed. by B Baltagi. Hoboken, NJ: Wiley-Blackwell, pp. 310–330. DOI: 10.1002/9780470996249.

Anselin, L and D Griffith (1988). 'Do Spatial Effects Really Matter in Regression Analysis?' *Papers in Regional Science* 65 (1), pp. 11–34. DOI: 10.1111/j.1435-5597.1988.tb01155.x.

Assuncao, R and E Reis (1999). 'A New Proposal to Adjust Moran's I for Population Density'. *Statistics in Medicine* 18 (16), pp. 2147–2162. DOI: 10.1002/(SICI)1097-0258(19990830)18:16<2147::AID-SIM179>3.0.CO;2-I.

Atkinson, P, G Foody, P Gething, A Mathur and C Kelly (2007). 'Investigating Spatial Structure in Specific Tree Species in Ancient Semi-Natural Woodland Using Remote Sensing and Marked Point Pattern Analysis'. *Ecography* 30 (1), pp. 88–104. DOI: 10.1111/j.0906-7590.2007.04726.x.

Bakillah, M, R Li and S Liang (2015). 'Geo-Located Community Detection in Twitter with Enhanced Fast-Greedy Optimization of Modularity: The Case Study of Typhoon Haiyan'. *International Journal of Geographical Information Science* 29 (2), pp. 258–279. DOI: 10.1080/13658816.2014.964247.

Barkhuus, L, B Brown, M Bell, S Sherwood, M Hall and M Chalmers (2008). 'From Awareness to Repartee: Sharing Location within Social Groups'. In: *Proceeding of the 26th Annual CHI Conference on Human Factors in Computing Systems - CHI '08*. New York, NY: ACM Press, pp. 497–506. DOI: 10.1145/1357054.1357134.

Basile, R, M Durbán, R Mínguez, J María Montero and J Mur (2014). 'Modeling Regional Economic Dynamics: Spatial Dependence, Spatial Heterogeneity and Nonlinearities'. *Journal of Economic Dynamics and Control* 48, pp. 229–245. DOI: `10.1016/j.jedc.2014.06.011`.

Bavaud, F (1998). 'Models for Spatial Weights: A Systematic Look'. *Geographical Analysis* 30 (2), pp. 153–171. DOI: `10.1111/j.1538-4632.1998.tb00394.x`.

Birenboim, A (2017). 'The Influence of Urban Environments on our Subjective Momentary Experiences'. *Environment and Planning B: Urban Analytics and City Science* forthcomin, p. 239980831769014. DOI: `10.1177/2399808317690149`.

Blei, D, A Ng and M Jordan (2003). 'Latent Dirichlet Allocation'. *The Journal of Machine Learning Research* 3, pp. 993–1022.

Bolin, R and D Klenow (1983). 'Response of the Elderly to Disaster: An Age-Stratified Analysis'. *The International Journal of Aging and Human Development* 16 (4), pp. 283–296. DOI: `10.2190/MQEG-YN39-8D5V-WKMP`.

Boulos, M, B Resch, D Crowley, J Breslin, G Sohn, R Burtner, W Pike, E Jezierski and K Chuang (2011). 'Crowdsourcing, Citizen Sensing and Sensor Web Technologies for Public and Environmental Health Surveillance and Crisis Management: Trends, OGC Standards and Application Examples'. *International Journal of Health Geographics* 10 (1), p. 67. DOI: `10.1186/1476-072X-10-67`.

Bouten, W, E Baaij, J Shamoun-Baranes and K Camphuysen (2013). 'A Flexible GPS Tracking System for Studying Bird Behaviour at Multiple Scales'. *Journal of Ornithology* 154 (2), pp. 571–580. DOI: `10.1007/s10336-012-0908-1`.

Bravo, C and R Weber (2011). 'Semi-Supervised Constrained Clustering with Cluster Outlier Filtering'. In: *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*. Ed. by C San Martin and S Kim. Heidelberg: Springer, pp. 347–354. DOI: `10.1007/978-3-642-25085-9_41`.

Caldas de Castro, M and B Singer (2006). 'Controlling the False Discovery Rate: A New Application to Account for Multiple and Dependent Tests in Local Statistics of Spatial Association'. *Geographical Analysis* 38 (2), pp. 180–208. DOI: `10.1111/j.0016-7363.2006.00682.x`.

Camara, G, A Sposati, D Koga, A Monteiro, F Ramos, E Camargo and S Fuks (2004). 'Mapping Social Exclusion and Inclusion in Developing Countries'. In: *Spatially Integrated Social Science*. Ed. by M Goodchild and D Janelle. Oxford, UK: Oxford University Press. Chap. 11, pp. 223–238.

Capannelli, G, J Lee and P Petri (2010). 'Economic Interdependence in Asis: Developing Indicators for Regional Integration and Cooperation'. *The Singapore Economic Review* 55 (01), pp. 125–161. DOI: `10.1142/S021759081000364X`.

Carrion, D, F Migliaccio and D Pagliari (2017). 'Exploring Geolocation Issues in Social Media Analytics: A Case Study with Tweet Messages'. In: *The 6th International Virtual Scientific Conference on Informatics and Management Sciences*, pp. 100–103. DOI: `10.18638/ictic.2017.6.1.316`.

Cheng, T and T Wicks (2014). 'Event Detection using Twitter: A Spatio-Temporal Approach'. *PLOS ONE* 9 (6), e97807. DOI: `10.1371/journal.pone.0097807`.

Chun, Y and D Griffith (2013). *Spatial Statistics and Geostatistics*. London, UK: SAGE.

Cliff, A and J Ord (1969). 'The Problem of Spatial Autocorrelation'. In: *London Papers in Regional Science (1), Studies in Regional Science*. Ed. by A Scott. London: Pion, pp. 25–55.

— (1973). *Spatial Autocorrelation*. London, UK: Pion.

— (1981). *Spatial Processes: Models & Applications*. London, UK: Pion.

Coleman, D (2009). ‘Volunteered Geographic Information in Spatial Data Infrastructure: An Early Look at Opportunities and Constraints’. In: *Spatially Enabling Society: Research, Emerging Trends and Critical Assessment*. Ed. by A Rajabifard, J Crompvoets, M Kanantari and B Kok. Leuven: Leuven University Press. Chap. 10, pp. 131–148.

Cox, D and H Miller (1977). *The Theory of Stochastic Processes*. Boca Raton, FL: Chapmann & Hall.

Craglia, M, K de Bie, D Jackson, M Pesaresi, G Remetey-Fülöpp, C Wang, A Annoni, L Bian, F Campbell, M Ehlers, J van Genderen, M Goodchild, H Guo, A Lewis, R Simpson, A Skidmore and P Woodgate (2012). ‘Digital Earth 2020: Towards the Vision for the Next Decade’. *International Journal of Digital Earth* 5 (1), pp. 4–21. DOI: 10.1080/17538947.2011.638500.

Crampton, J (2009). ‘Cartography: Maps 2.0’. *Progress in Human Geography* 33 (1), pp. 91–100. DOI: 10.1177/0309132508094074.

Crampton, J, M Graham, A Poorthuis, T Shelton, M Stephens, M Wilson and M Zook (2013). ‘Beyond the Geotag: Situating ’Big Data’ and Leveraging the Potential of the Geoweb’. *Cartography and Geographic Information Science* 40 (2), pp. 130–139. DOI: 10.1080/15230406.2013.777137.

Cressie, N (1993). *Statistics for Spatial Data*. Wiley Series in Probability and Statistics. Hoboken, NJ: John Wiley & Sons. DOI: 10.1002/9781119115151.

Croitoru, A, N Wayant, A Crooks, J Radzikowski and A Stefanidis (2015). ‘Linking Cyber and Physical Spaces through Community Detection and Clustering in Social Media Feeds’. *Computers, Environment and Urban Systems* 53, pp. 47–64. DOI: 10.1016/j.compenvurbsys.2014.11.002.

Crooks, A, A Croitoru, A Stefanidis and J Radzikowski (2013). ‘#Earthquake: Twitter as a Distributed Sensor System’. *Transactions in GIS* 17 (1), pp. 124–147. DOI: 10.1111/j.1467-9671.2012.01359.x.

Csikszentmihalyi, M and J Hunter (2003). ‘Happiness in Everyday Life: The Uses of Experience Sampling’. *Journal of Happiness Studies* 4 (2), pp. 185–199. DOI: 10.1023/A:1024409732742.

Cuel, R, O Morozova, M Rohde, E Simperl, K Siorpaes, O Tokarchuk, T Wiedenhoefer, F Yetim and M Zamarian (2011). ‘Motivation Mechanisms for Participation in Human-Driven Semantic Content Creation’. *International Journal of Knowledge Engineering and Data Mining* 1 (4), p. 331. DOI: 10.1504/IJKEDM.2011.040653.

Dangschat, J (2007). ‘Raumkonzept zwischen struktureller Produktion und individueller Konstruktion’. *Ethnologie und Raum* 9 (1), pp. 24–44.

de Souza Silva, A (2013). ‘Location-aware Mobile Technologies: Historical, Social and Spatial Approaches’. *Mobile Media & Communication* 1 (1), pp. 116–121. DOI: 10.1177/2050157912459492.

Dever, J, A Rafferty and R Valliant (2008). ‘Internet Surveys: Can Statistical Adjustments Eliminate Coverage Bias?’ *Survey Research Methods* 2 (2), pp. 47–62. DOI: 10.18148/srm/2008.v2i2.128.

Diniz-Filho, J, L Bini and B Hawkins (2003). ‘Spatial Autocorrelation and Red Herrings in Geographical Ecology’. *Global Ecology and Biogeography* 12 (1), pp. 53–64. DOI: 10.1046/j.1466-822X.2003.00322.x.

Dörry, S and A Decoville (2016). ‘Governance and Transportation Policy Networks in the Cross-Border Metropolitan Region of Luxembourg: A Social Network Analysis’. *European Urban and Regional Studies* 23 (1), pp. 69–85. DOI: 10.1177/0969776413490528.

Duncan, J (2011). ‘Space, Place, and Population Research: Challenges and Future Directions’. In: *2011 Specialist Meetingâ€”Future Directions in Spatial Demography*.

Durbin, J (1973). *Distribution Theory for Tests Based on the Sample Distribution Function*. 9th ed. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics. DOI: 10.1137/1.9781611970586.

Dutilleul, Pierre and Pierre Legendre (1993). 'Spatial Heterogeneity Against Heteroscedasticity: An Ecological Paradigm Versus a Statistical Concept'. *Oikos* 66 (1), pp. 152–171. DOI: 10.2307/3545210.

Elwood, S (2008). 'Volunteered Geographic Information: Future Research Directions Motivated by Critical, Participatory, and Feminist GIS'. *GeoJournal* 72 (3-4), pp. 173–183. DOI: 10.1007/s10708-008-9186-0.

Elwood, S, M Goodchild and D Sui (2012). 'Researching Volunteered Geographic Information: Spatial Data, Geographic Research, and New Social Practice'. *Annals of the Association of American Geographers* 102 (3), pp. 571–590. DOI: 10.1080/00045608.2011.595657.

Ermagun, A and D Levinson (2017). 'An Introduction to the Network Weight Matrix'. *Geographical Analysis* forthcomin. DOI: 10.1111/gean.12134.

Evans, L (2011). 'Location-Based Services: Transformation of the Experience of Space'. *Journal of Location Based Services* 5 (3-4), pp. 242–260. DOI: 10.1080/17489725.2011.637968.

Evans, L and M Saker (2017). *Location-Based Social Media*. Cham: Springer. DOI: 10.1007/978-3-319-49472-2.

Everitt, BS and A Skrondal (2010). *The Cambridge Dictionary of Statistics*. Cambridge, UK: Cambridge University Press.

Fischer, F (2012). 'VGI as Big Data - A New but Delicate Geographic Data-Source'. *GEOInformatics* 3, pp. 46–47.

Fischer, M and A Getis (2010b). 'Introduction'. In: *Handbook of Applied Spatial Analysis*. Ed. by M Fischer and A Getis. Heidelberg: Springer, pp. 1–24. DOI: 10.1007/978-3-642-03647-7_1.

Frank, M, L Mitchell, P Dodds and C Danforth (2013). 'Happiness and the Patterns of Life: A Study of Geolocated Tweets'. *Scientific reports* 3, p. 2625. DOI: 10.1038/srep02625.

Freksa, C (1991). 'Qualitative Spatial Reasoning'. In: *Cognitive and Linguistic Aspects of Geographic Space*. Ed. by D Mark and A Frank. Dordrecht: Springer, pp. 361–372. DOI: 10.1007/978-94-011-2606-9_20.

Frigg, R and C Werndl (2018). 'Equilibrium in Gibbsian Statistical Mechanics'. In: *Routledge Companion to Philosophy of Physics*. Ed. by E Knox and A Wilson. London, UK: Routledge, in press.

Fuentes, M and B Reich (2010). 'Spectral Domain'. In: *Handbook of Spatial Statistics*. Ed. by A Gelfand, P Diggle, M Fuentes and P Guttorp. Boca Raton, FL: CRC Press, pp. 58–77.

Gabler, S and A Quatember (2013). 'Repräsentativität von Subgruppen bei Geschichteten Zufallsstichproben'. *AStA Wirtschafts- und Sozialstatistisches Archiv* 7 (3-4), pp. 105–119. DOI: 10.1007/s11943-013-0132-3.

Gaenssler, P and W Stute (1979). 'Empirical Processes: A Survey of Results for Independent and Identically Distributed Random Variables'. *The Annals of Probability* 7 (2), pp. 193–243.

Gaetan, C and X Guyon (2010). *Spatial Statistics and Modeling*. Springer Series in Statistics. New York, NY: Springer. DOI: 10.1007/978-0-387-92257-7.

Gao, S, K Janowicz, G McKenzie and L Li (2013). 'Towards Platial Joins and Buffers in Place-Based GIS'. In: *Proceedings of The First ACM SIGSPATIAL International Workshop on Computational Models of Place - COMP '13*. Orlando, FL, pp. 42–49. DOI: 10.1145/2534848.2534856.

Gärdenfors, P and M Williams (2001). 'Reasoning About Categories in Conceptual Spaces'. In: *International Joint Conference on Artificial Intelligence*. Seattle, WA, pp. 385–392.

Gelfand, A and S Banerjee (2015). 'Bayesian Wombling: Finding Rapid Change in Spatial Maps'. *Wiley Interdisciplinary Reviews: Computational Statistics* 7 (October), pp. 307–315. DOI: 10.1002/wics.1360.

Gentles, S, C Charles, J Ploeg and K McKibbon (2015). 'Sampling in Qualitative Research: Insights from an Overview of the Methods Literature'. *The Qualitative Report* 20 (11), pp. 1772–1789.

Getis, A (2007). 'Reflections on Spatial Autocorrelation'. *Regional Science and Urban Economics* 37 (4), pp. 491–496. DOI: 10.1016/j.regsciurbeco.2007.04.005.

— (2008). 'A History of the Concept of Spatial Autocrrelation: A Geographer's Perspective'. *Geographical Analysis* 40 (3), pp. 297–309. DOI: 10.1111/j.1538-4632.2008.00727.x.

— (2009). 'Spatial Weights Matrices'. *Geographical Analysis* 41 (4), pp. 404–410. DOI: 10.1111/j.1538-4632.2009.00768.x.

— (2015). 'Analytically Derived Neighborhoods in a Rapidly Growing West African City: The Case of Accra, Ghana'. *Habitat International* 45 (Part 2), pp. 126–134. DOI: 10.1016/j.habitatint.2014.06.021.

Getis, A and J Aldstadt (2004). 'Constructing the Spatial Weights Matrix Using a Local Statistic'. *Geographical Analysis* 34 (2), pp. 130–140. DOI: 10.1353/geo.2004.0002.

Getis, A and J Ord (1992). 'The Analysis of Spatial Association by Use of Distance Statistics'. *Geographical Analysis* 24 (3), pp. 189–206. DOI: 10.1111/j.1538-4632.1992.tb00261.x.

Gneiting, T and P Guttorp (2010). 'Continuous Parameter Stochastic Process Theory'. In: *Handbook of Spatial Statistics*. Ed. by A Gelfand, P Diggle, M Fuentes and P Guttorp. Boca Raton, FL: CRC Press, pp. 17–28.

Golledge, R and R Stimson (1997). *Spatial Behavior: A Geographic Perspective*. New York, NY: Guilford Press.

Goodchild, M (2001). 'Models of Scale and Scales of Modeling'. In: *Modelling Scale in Geographical Information Science*. Ed. by N Tate and P Atkinson. Chichester, UK: John Wiley & Sons, pp. 3–10.

— (2009). 'What Problem? Spatial Autocorrelation and Geographic Information Science'. *Geographical Analysis* 41 (4), pp. 411–417. DOI: 10.1111/j.1538-4632.2009.00769.x.

Goodchild, M and L Li (2011). 'Formalizing space and place To cite this version :' in: *Proceedings du 1er colloque international du CIST*. Paris, pp. 177–183.

Gore, A (1998). 'The Digital Earth'. *Australian Surveyor* 43 (2), pp. 89–91. DOI: 10.1080/00050326.1998.10441850.

Graham, M (2011). 'Wiki Space: Palimpsests and the Politics of Exclusion'. In: *A Wikipedia Reader*. Ed. by G Lovink and N Tkacz. Amsterdam: Institute of Network Cultures, pp. 269–282.

Graham, M, M Zook and A Boulton (2013). 'Augmented Reality in Urban Places: Contested Content and the Duplicity of Code'. *Transactions of the Institute of British Geographers* 38 (3), pp. 464–479. DOI: 10.1111/j.1475-5661.2012.00539.x.

Griffith, D (1988). *Advanced Spatial Statistics*. Dordrecht, NL: Springer, p. 292.

— (2010). 'The Moran Coefficient for Non-Normal Data'. *Journal of Statistical Planning and Inference* 140 (11), pp. 2980–2990. DOI: 10.1016/j.jspi.2010.03.045.

Griffith, D and L Layne (1999). *A Casebook for Spatial Statistical Data Analysis*. Oxford, UK: Oxford University Press.

Haklay, M (2012). ‘Nobody Wants to Do Council Estates: Digital Divide, Spatial Justice and Outliers’. In: *Annual Meeting of the American Association of Geographers 2012*. New York, NY: American Association of Geographers.

— (2013). ‘Neogeography and the Delusion of Democratisation’. *Environment and Planning A* 45 (1), pp. 55–69. DOI: 10.1068/a45184.

Haklay, M, A Singleton and C Parker (2008). ‘Web Mapping 2.0: The Neogeography of the GeoWeb’. *Geography Compass* 2 (6), pp. 2011–2039. DOI: 10.1111/j.1749-8198.2008.00167.x.

Hammervold, R and U Olsson (2012). ‘Testing Structural Equation Models: The Impact of Error Variances in the Data Generating Process’. *Quality & Quantity* 46 (5), pp. 1547–1570. DOI: 10.1007/s11135-011-9466-5.

Harris, R, J Moffat and V Kravtsova (2011). ‘In Search of ‘W ’’. *Spatial Economic Analysis* 6 (3), pp. 249–270. DOI: 10.1080/17421772.2011.586721.

Harvey, F (2013). ‘To Volunteer or to Contribute Locational Information? Towards Truth in Labeling for Crowdsourced Geographic Information? Towards Truth in Labeling for Crowdsourced Geographic Information’. In: *Crowdsourcing Geographic Knowledge*. Ed. by D Sui, S Elwood and M Goodchild. Dordrecht: Springer Netherlands, pp. 31–42. DOI: 10.1007/978-94-007-4587-2_3.

Hecht, B and M Stephens (2014). ‘A Tale of Cities: Urban Biases in Volunteered Geographic Information’. In: *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*. Ann Arbor, MI: AAAI Press, pp. 197–205.

Hewings, G, Y Okuyama and M Sonis (2001). ‘Economic Interdependence Within the Chicago Metropolitan Area: A Miyazawa Analysis’. *Journal of Regional Science* 41 (2), pp. 195–217. DOI: 10.1111/0022-4146.00214.

Hiruta, S, T Yonezawa, M Jurmu and H Tokuda (2012). ‘Detection, Classification and Visualization of Place-Triggered Geotagged Tweets’. In: *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. Pittsburgh, PA: ACM Press, pp. 956–963. DOI: 10.1145/2370216.2370427.

Horton, F and D Reynolds (1971). ‘Effects of Urban Spatial Structure on Individual Behavior’. *Economic Geography* 47 (1), pp. 36–48. DOI: 10.2307/143224.

Hubert, L, R Golledge and C Costanzo (1981). ‘Generalized Procedures for Evaluating Spatial Autocorrelation’. *Geographical Analysis* 13 (3), pp. 224–233. DOI: 10.1111/j.1538-4632.1981.tb00731.x.

Hudson-Smith, A, A Crooks, M Gibin, R Milton and M Batty (2009). ‘NeoGeography and Web 2.0: Concepts, Tools and Applications’. *Journal of Location Based Services* 3 (2), pp. 118–145. DOI: 10.1080/17489720902950366.

Illian, J, A Penttinen, H Stoyan and D Stoyan (2008). *Statistical Analysis and Modelling of Spatial Point Patterns*. Hoboken, NJ: John Wiley & Sons, p. 534.

Jacquez, G, S Maruca and M Fortin (2000). ‘From Fields to Objects: A Review of Geographic Boundary Analysis’. *Journal of Geographical Systems* 2 (3), pp. 221–241. DOI: 10.1007/PL00011456.

Jenkins, A, A Croitoru, A Crooks and A Stefanidis (2016). ‘Crowdsourcing a Collective Sense of Place’. *PLOS ONE* 11 (4), e0152932. DOI: 10.1371/journal.pone.0152932.

Johnson, I, S Sengupta, J Schöning and B Hecht (2016). ‘The Geography and Importance of Localness in Geotagged Social Media’. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 515–526. DOI: 10.1145/2858036.2858122.

Johnson, M (1987). *The Body in the Mind: The Bodily Basis of Meaning, Imagination, and Reason*. Chicago, IL: University of Chicago Press.

Johnson, P, J Hoverman, V McKenzie, A Blaustein and K Richgels (2013). 'Urbanization and Wetland Communities: Applying Metacommunity Theory to Understand the Local and Landscape Effects'. *Journal of Applied Ecology* 50 (1), pp. 34–42. DOI: 10.1111/1365-2664.12022.

Johnson, P and R Sieber (2012). 'Motivations Driving Government Adoption of the Geoweb'. *GeoJournal* 77 (5), pp. 667–680. DOI: 10.1007/s10708-011-9416-8.

Kaplan, M, B McFarland, N Huguet, K Conner, R Caetano, N Giesbrecht and K Nolte (2013). 'Acute Alcohol Intoxication and Suicide: A Gender-Stratified Analysis of the National Violent Death Reporting System'. *Injury Prevention* 19 (1), pp. 38–43. DOI: 10.1136/injuryprev-2012-040317.

Kelley, M (2013). 'The Emergent Urban Imaginaries of Geosocial Media'. *GeoJournal* 78 (1), pp. 181–203. DOI: 10.1007/s10708-011-9439-1.

Khoo, T, F Fu and S Pauls (2016). 'Coevolution of Cooperation and Partner Rewiring Range in Spatial Social Networks'. *Nature Scientific Reports* 6, p. 36293. DOI: 10.1038/srep36293.

Kitchin, R (2014). *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. London, UK: SAGE.

Kitchin, R and M Dodge (2011). *Code/Space*. Cambridge, MA: The MIT Press. DOI: 10.7551/mitpress/9780262042482.001.0001.

Kitchin, R, T Lauriault and M Wilson (2017). *Understanding Spatial Media*.

Kolasa, J and C Rollo (1991). 'The Heterogeneity of Heterogeneity: A Glossary'. In: *Ecological Heterogeneity*. Ed. by J Kolasa and S Pickett. Heidelberg: Springer. Chap. 1, pp. 1–23. DOI: 10.1007/978-1-4612-3062-5_1.

Krumm, J, N Davies and C Narayanaswami (2008). 'User-Generated Content'. *IEEE Pervasive Computing* 7 (4), pp. 10–11. DOI: 10.1109/MPRV.2008.85.

Lansley, G and P Longley (2016). 'The Geography of Twitter Topics in London'. *Computers, Environment and Urban Systems* 58, pp. 85–96. DOI: 10.1016/j.compenvurbsys.2016.04.002.

Lazer, D, R Kennedy, G King and A Vespignani (2014). 'The Parable of Google Flu: Traps in Big Data Analysis'. *Science* 343 (6176b), pp. 1203–1205. DOI: 10.1126/science.1248506.

Lebowitz, J and O Penrose (1973). 'Modern Ergodic Theory'. *Physics Today* 26, pp. 155–175.

Lee, S (2004). 'A Generalized Significance Testing Method for Global Measures of Spatial Association: An Extension of the Mantel Test'. *Environment and Planning A* 36 (9), pp. 1687–1703. DOI: 10.1068/a34143.

— (2009). 'A Generalized Randomization Approach to Local Measures of Spatial Association'. *Geographical Analysis* 41 (2), pp. 221–248. DOI: 10.1111/j.1538-4632.2009.00749.x.

Leszczynski, A (2014). 'On the Neo in Neogeography'. *Annals of the Association of American Geographers* 104 (1), pp. 60–79. DOI: 10.1080/00045608.2013.846159.

— (2015). 'Spatial Media/tion'. *Progress in Human Geography* 39 (6), pp. 729–751. DOI: 10.1177/0309132514558443.

Lillesand, T, R Kiefer and J Chipman (2015). *Remote Sensing and Image Interpretation*. 7th ed. Hoboken, NJ: Wiley & Sons.

Lindqvist, J, J Cranshaw, J Wiese, J Hong and J Zimmerman (2011). 'I'm the Mayor of my House: Examining why People Use Foursquare - a Social-driven Location Sharing Application'. In: *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11*. New York, NY: ACM Press, pp. 2409–2418. DOI: 10.1145/1978942.1979295.

Longley, P and M Adnan (2016). 'Geo-temporal Twitter Demographics'. *International Journal of Geographical Information Science* 30 (2), pp. 369–389. DOI: 10.1080/13658816.2015.1089441.

Lovelace, R, M Birkin, P Cross and M Clarke (2016). 'From Big Noise to Big Data: Toward the Verification of Large Data Sets for Understanding Regional Retail Flows'. *Geographical Analysis* 48 (1), pp. 59–81. DOI: 10.1111/gean.12081.

Luo, F, G Cao, K Mulligan and X Li (2016a). 'Explore Spatiotemporal and Demographic Characteristics of Human Mobility via Twitter: A Case Study of Chicago'. *Applied Geography* 70, pp. 11–25. DOI: 10.1016/j.apgeog.2016.03.001.

Luo, G, G Chen, L Tian, K Qin and S Qian (2016b). 'Minimum Noise Fraction versus Principal Component Analysis as a Preprocessing Step for Hyperspectral Imagery Denoising'. *Canadian Journal of Remote Sensing* 42 (2), pp. 106–116. DOI: 10.1080/07038992.2016.1160772.

MacKerron, G and S Mourato (2013). 'Happiness is Greater in Natural Environments'. *Global Environmental Change* 23 (5), pp. 992–1000. DOI: 10.1016/j.gloenvcha.2013.03.010.

Mawarni, M and I Machdi (2016). 'Dynamic Nearest Neighbours for Generating Spatial Weight Matrix'. In: *2016 International Conference on Advanced Computer Science and Information Systems*. Malang, Indonesia: IEEE, pp. 257–262. DOI: 10.1109/ICACSIS.2016.7872771.

Mckenzie, G and B Adams (2017). 'Juxtaposing Thematic Regions Derived from Spatial and Platial User-Generated Content'. In: *Proceedings of the 13th International Conference on Spatial Information Theory (COSIT 2017)*. Ed. by E Clementini, M Donnelly, M Yuan, C Kray, P Fogliaroni and A Ballatore. L'Aquila. DOI: 10.4230/LIPIcs.COSIT.2017.20.

McLeod, K (2000). 'Our Sense of Snow: The Myth of John Snow in Medical Geography'. *Social Science & Medicine* 50 (7-8), pp. 923–935. DOI: 10.1016/S0277-9536(99)00345-7.

Mehta, V and J Bosson (2010). 'Third Places and the Social Life of Streets'. *Environment and Behavior* 42 (6), pp. 779–805. DOI: 10.1177/0013916509344677.

Mennis, J and M Mason (2016). 'Modeling Place as a Relationship between a Person and a Location'. In: *Proceedings of the 9th International Conference on GIScience*. Montréal, CA, pp. 2014–2017. DOI: 10.21433/B3119W316472.

Mitchell, L, M Frank, K Harris, P Dodds and C Danforth (2013). 'The Geography of Happiness: Connecting Twitter Sentiment and Expression, Demographics, and Objective Characteristics of Place'. *PLoS ONE* 8 (5), e64417. DOI: 10.1371/journal.pone.0064417.

Mooney, P (2009). 'The Physical and Social Aspects of Place Attachment'. *Topos* 2, pp. 28–31.

Moran, P (1950). 'Notes on Continuous Stochastic Phenomena'. *Biometrika* 37 (1/2), pp. 17–23. DOI: 10.2307/2332142.

Nelson, T (2012). 'Trends in Spatial Statistics'. *The Professional Geographer* 64 (1), pp. 83–94. DOI: 10.1080/00330124.2011.578540.

Nilsson, L (1972). 'Habitat Selection, Food Choice, and Feeding Habits of Diving Ducks in Coastal Waters of South Sweden during the Non-Breeding Season'. *Ornis Scandinavica (Scandinavian Journal of Ornithology)* 3 (1), pp. 55–78.

Oden, N (1995). 'Adjusting Moran's I for Population Density'. *Statistics in Medicine* 14 (1), pp. 17–26. DOI: 10.1002/sim.4780140104.

Oh, S and S Syn (2015). 'Motivations for Sharing Information and Social Support in Social Media: A Comparative Analysis of Facebook, Twitter, Delicious, YouTube, and Flickr'. *Journal of the Association for Information Science and Technology* 66 (10), pp. 2045–2060. DOI: 10.1002/asi.23320.

Oishi, S (2014). 'Socioecological Psychology'. *Annual Review of Psychology* 65 (1), pp. 581–609. DOI: 10.1146/annurev-psych-030413-152156.

Oliver, M (2010). 'The Variogram and Kriging'. In: *Handbook of Applied Spatial Analysis*. Ed. by M Fischer and A Getis. Heidelberg: Springer, pp. 319–352. DOI: 10.1007/978-3-642-03647-7_17.

Ord, J and A Getis (1995). 'Local Spatial Autocorrelation Statistics: Distributional Issues and an Application'. *Geographical Analysis* 27 (4), pp. 286–306. DOI: 10.1111/j.1538-4632.1995.tb00912.x.

— (2001). 'Testing for Local Spatial Autocorrelation in the Presence of Global Autocorrelation'. *Journal of Regional Science* 41 (3), pp. 411–432. DOI: 10.1111/0022-4146.00224.

— (2012). 'Local Spatial Heteroscedasticity (LOSH)'. *The Annals of Regional Science* 48 (2), pp. 529–539. DOI: 10.1007/s00168-011-0492-y.

O'Reilly, T (2010). 'What is Web 2.0? Design Patterns and Business Models for the Next Generation of Software'. In: *Online Communication and Collaboration: A Reader*. Ed. by H Donelan, K Kear and M Ramage. London, UK: Taylor & Francis, pp. 225–235.

Pace, K and J LeSage (2010). 'Spatial Econometrics'. In: *Handbook of Spatial Statistics*. Ed. by A Gelfand, P Diggle, M Fuentes and P Guttorp. Boca Raton, FL: CRC Press, pp. 245–260.

Patil, G, R Modarres, W Myers and P Patankar (2006). 'Spatially Constrained Clustering and Upper Level Set Scan Hotspot Detection in Surveillance Geoinformatics'. *Environmental and Ecological Statistics* 13 (4), pp. 365–377. DOI: 10.1007/s10651-006-0017-5.

Poorthuis, B and M Zook (2013). 'Spaces of Volunteered Geographic Information'. In: *Ashgate Research Companion on Geographies of Media*. Ed. by P Adams and J Craine. Oxford, UK: Taylor & Francis, pp. 311–328. DOI: 10.2139/ssrn.2259845.

Porojan, A (2001). 'Trade Flows and Spatial Effects: The Gravity Model Revisited'. *Open Economies Review* 12 (3), pp. 265–280. DOI: 10.1023/A:1011129422190.

Quesnot, T and S Roche (2015). 'Platial or Locational Data? Toward the Characterization of Social Location Sharing'. *Proceedings of the Annual Hawaii International Conference on System Sciences*, pp. 1973–1982. DOI: 10.1109/HICSS.2015.236.

Rae, A and A Singleton (2015). 'Putting big data in its place: a Regional Studies and Regional Science perspective'. *Regional Studies, Regional Science* 2 (1), pp. 1–5. DOI: 10.1080/21681376.2014.990678.

Rajaram, S, L Heinrich, J Gordan, J Avva, K Bonness, A Witkiewicz, J Malter, C Atreya, R Warren, L Wu and S Altschuler (2017). 'Sampling Strategies to Capture Single-Cell Heterogeneity'. *Nature Methods*. DOI: 10.1038/nmeth.4427.

Ratnasari, N, E Candra, D Saputra and A Perdana (2016). 'Urban Spatial Pattern and Interaction based on Analysis of Nighttime Remote Sensing Data and Geo-social Media Information'. In: *IOP Conference Series: Earth and Environmental Science*. Vol. 47. 1. IOP Publishing, p. 012038. DOI: 10.1088/1755-1315/47/1/012038.

Reis, H and S Gable (2000). 'Event-Sampling and other Methods for Studying Everyday Experience'. In: *Handbook of Research Methods in Social and Personality Psychology*. Ed. by H Reis and C Judd. Cambridge, UK: Cambridge University Press, pp. 190–222.

Rentfrow, P (2013). *Geographical Psychology: Exploring the Interaction of Environment and Behavior*. Washington, DC: American Psychological Association.

Resch, B, M Sudmanns, G Sagl, A Summa, P Zeile and J Exner (2015b). 'Crowdsourcing Physiological Conditions and Subjective Emotions by Coupling Technical and Human Mobile Sensors'.

GI_Forum â€' Journal for Geographic Information Science 1, pp. 514–524. DOI: 10.1553/giscience2015s514.

Resch, B, A Summa, G Sagl, P Zeile and J Exner (2015d). 'Urban Emotionsâ€"Geo-Semantic Emotion Extraction from Technical Sensors, Human Sensors and Crowdsourced Data'. In: *Progress in Location-Based Services 2014*. Ed. by Gartner G and Huang H. Cham: Springer, pp. 199–212. DOI: 10.1007/978-3-319-11879-6_14.

Richardson, D, N Volkow, M Kwan, R Kaplan, M Goodchild and R Croyle (2013). 'Spatial Turn in Health Research'. *Science* 339 (6126), pp. 1390–1392. DOI: 10.1126/science.1232257.

Rinner, C and V Fast (2015). 'A Classification of User Contributions on the Participatory Geoweb'. In: *Advances in Spatial Data Handling and Analysis*. Ed. by F Harvey and Y Leung. Cham: Springer, pp. 35–49. DOI: 10.1007/978-3-319-19950-4_3.

Ritzer, G, P Dean and N Jurgenson (2012). 'The Coming of Age of the Prosumer'. *American Behavioral Scientist* 56 (4), pp. 379–398. DOI: 10.1177/0002764211429368.

Roche, S (2016). 'Geographic Information Science II'. *Progress in Human Geography* 40 (4), pp. 565–573. DOI: 10.1177/0309132515586296.

Roick, O and S Heuser (2013). 'Location Based Social Networks - Definition, Current State of the Art and Research Agenda'. *Transactions in GIS* 17 (5), pp. 763–784. DOI: 10.1111/tgis.12032.

Rollinson, P (1998). 'The Everyday Geography of the Homeless in Kansas City'. *Geografiska Annaler, Series B: Human Geography* 80 (2), pp. 101–115. DOI: 10.1111/j.0435-3684.1998.00032.x.

Rzeszewski, M and L Beluch (2017). 'Spatial Characteristics of Twitter Users - Toward the Understanding of Geosocial Media Production'. *ISPRS International Journal of Geo-Information* 6 (8). DOI: 10.3390/ijgi6080236.

Saif, H, M Fernández, Y He and H Alani (2014). 'On stopwords, filtering and data sparsity for sentiment analysis of Twitter'. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation*. Ed. by N Calzolari, K Choukri, T Declerck, H Loftsson, B Maegaard, J Mariani, A Moreno, J Odijk and S Piperidis. Reykjavik: European Language Resources Association, pp. 810–817.

Sakaki, T, M Okazaki and Y Matsuo (2013). 'Tweet Analysis for Real-Time Event Detection and Earthquake Reporting System Development'. *IEEE Transactions on Knowledge and Data Engineering* 25 (4), pp. 919–931. DOI: 10.1109/TKDE.2012.29.

Saker, M (2017). 'Foursquare and Identity: Checking-in and Presenting the Self through Location'. *New Media & Society* 19 (6), pp. 934–949. DOI: 10.1177/1461444815625936.

Santos-Vega, M, M Bouma, V Kohli and M Pascual (2016). 'Population Density, Climate Variables and Poverty Synergistically Structure Spatial Risk in Urban Malaria in India'. *PLOS Neglected Tropical Diseases* 10 (12), e0005155. DOI: 10.1371/journal.pntd.0005155.

Scharl, A and K Tochtermann (2007). *The Geospatial Web : How Geobrowsers, Social Software and the Web 2.0 are Shaping the Network Society*. Dordrecht: Springer.

Scheider, S and K Janowicz (2014). 'Place Reference Systems'. *Applied Ontology* 9 (2), pp. 97–127. DOI: 10.3233/AO-140134.

Seamon, D (1979). *A Geography of the Lifeworld*. London, UK: Croom Helm.

See, L, P Mooney, G Foody, L Bastin, A Comber, J Estima, S Fritz, N Kerle, B Jiang, M Laakso, H Liu, G Milcinski, M Niksic, M Painho, A Podör, A Olteanu-Raimond and M Rutzinger (2016). 'Crowdsourcing, Citizen Science or Volunteered Geographic Information? The Current State of

Crowdsourced Geographic Information'. *ISPRS International Journal of Geo-Information* 5 (5), p. 55. DOI: 10.3390/ijgi5050055.

Sengstock, C and M Gertz (2012). 'Latent Geographic Feature Extraction from Social Media'. In: *Proceedings of the 20th International Conference on Advances in Geographic Information Systems*. New York, NY: ACM Press, pp. 149–158. DOI: 10.1145/2424321.2424342.

Sester, M, J Arsanjani, R Klammer, D Burghardt and J-H Haunert (2014). 'Integrating and Generalizing Volunteered Geographic Information'. In: *Abstracting Geographic Information in a Data Rich World*. Ed. by D Burghardt, C Duchêne and W Mackaness. Cham: Springer. Chap. 5, pp. 119–155. DOI: 10.1007/978-3-319-00203-3.

Shen, C, C Li and Y Si (2016). 'Spatio-Temporal Autocorrelation Measures for Nonstationary Series: A New Temporally Detrended Spatio-Temporal Moran's Index'. *Physics Letters, Section A: General, Atomic and Solid State Physics* 380 (1-2), pp. 106–116. DOI: 10.1016/j.physleta.2015.09.039.

Shi, F, E Petriu and R Laganiere (2013). 'Sampling Strategies for Real-Time Action Recognition'. In: *2013 IEEE Conference on Computer Vision and Pattern Recognition*. Portland, OR, USA: IEEE, pp. 2595–2602. DOI: 10.1109/CVPR.2013.335.

Shimatani, K (2002). 'Point Processes for Fine-Scale Spatial Genetics and Molecular Ecology'. *Biometrical Journal* 44 (3), pp. 325–352. DOI: 10.1002/1521-4036(200204)44:3<325::AID-BIMJ325>3.0.CO;2-B.

Shortridge, A (2007). 'Practical Limits of Moran's Autocorrelation Index for Raster Class Maps'. *Computers, Environment and Urban Systems* 31 (3), pp. 362–371. DOI: 10.1016/j.compenvurbsys.2006.07.001.

Sieber, R and M Haklay (2015). 'The Epistemology(s) of Volunteered Geographic Information: A Critique'. *Geo: Geography and Environment* 2 (2), pp. 122–136. DOI: 10.1002/geo2.10.

Sieber, R and H Rahemtulla (2010). 'Model of Public Participation on the Geoweb'. In: *Proceedings of the 6th International Conference on Geographic Information Science*. Zürich.

Sila-Nowicka, K, J Vandrol, T Oshan, J Long, U Demsar and A Fotheringham (2016). 'Analysis of Human Mobility Patterns from GPS Trajectories and Contextual Information'. *International Journal of Geographical Information Science* 30 (5), pp. 881–906. DOI: 10.1080/13658816.2015.1100731.

Sîrbu, A, M Becker, S Caminiti, B De Baets, B Elen, L Francis, P Gravino, A Hotho, S Ingarra, V Loreto, A Molino, J Mueller, J Peters, F Ricchiuti, F Saracino, V Servedio, G Stumme, J Theunis, F Tria and J Van den Bossche (2015). 'Participatory Patterns in an International Air Quality Monitoring Initiative'. *PLOS ONE* 10 (8), e0136763. DOI: 10.1371/journal.pone.0136763.

Sloan, L (2017). 'Who Tweets in the United Kingdom? Profiling the Twitter Population Using the British Social Attitudes Survey 2015'. *Social Media + Society* 3 (1), p. 205630511769898. DOI: 10.1177/2056305117698981.

Snow, J (1855). *On the Mode of Communication of Cholera*. London: John Churchill.

Sonnentag, S, C Binnewies and S Ohly (2012). 'Event-Sampling in Occupational Health Psychology'. In: *Research Methods in Occupational Health Psychology : Measurement, Design and Data Analysis*. Ed. by R Sinclair, M Wang and L Tetrick. New York, NY: Routledge, pp. 208–228.

Soto, SM (2009). 'Human Migration and Infectious Diseases'. *Clinical Microbiology and Infection* 15 (s1), pp. 26–28. DOI: 10.1111/j.1469-0691.2008.02694.x.

Spyratos, S, M Lutz and F Pantisano (2014). 'Characteristics of Citizen-Contributed Geographic Information'. In: *Proceedings of the AGILE'2014 International Conference on Geographic Information Science*. Ed. by J Huerta, S Schade and C Granell. Castellón: AGILE.

Stefanidis, A, A Crooks and J Radzikowski (2013). 'Harvesting Ambient Geospatial Information from Social Media Feeds'. *GeoJournal* 78 (2), pp. 319–338. DOI: 10.1007/s10708-011-9438-2.

Steiger, E, B Resch, J de Albuquerque and A Zipf (2016a). 'Mining and Correlating Traffic Events from Human Sensor Observations with Official Transport Data Using Self-Organizing-Maps'. *Transportation Research Part C: Emerging Technologies* 73, pp. 91–104. DOI: 10.1016/j.trc.2016.10.010.

Steiger, E, B Resch and A Zipf (2016b). 'Exploration of Spatiotemporal and Semantic Clusters of Twitter Data Using Unsupervised Neural Networks'. *International Journal of Geographical Information Science* 30 (9), pp. 1694–1716. DOI: 10.1080/13658816.2015.1099658.

Steiger, E, R Westerholt, B Resch and A Zipf (2015b). 'Twitter as an Indicator for Whereabouts of People? Correlating Twitter with UK Census Data'. *Computers, Environment and Urban Systems* 54, pp. 255–265. DOI: 10.1016/j.compenvurbsys.2015.09.007.

Steiger, E, R Westerholt and A Zipf (2016c). 'Research on Social Media Feeds â€" A GIScience Perspective'. In: *European Handbook of Crowdsourced Geographic Information*. Ed. by C Capineri, M Haklay, H Huang, V Antoniou, J Kettunen, F Ostermann and R Purves. London: Ubiquity Press. Chap. 18, pp. 237–254. DOI: 10.5334/bax.r.

Strayer, D, M Power, W Fagan, S Pickett and J Belnap (2003). 'A Classification of Ecological Boundaries'. *BioScience* 53 (8), pp. 723–729. DOI: 10.1641/0006-3568(2003)053[0723:ACOEB]2.0.CO;2.

Sugovic, M and J Witt (2013). 'An Older View on Distance Perception: Older Adults Perceive Walkable Extents as Farther'. *Experimental Brain Research* 226 (3), pp. 383–391. DOI: 10.1007/s00221-013-3447-y.

Sui, D (2004). 'Tobler's First Law of Geography: A Big Idea for a Small World?' *Annals of the Association of American Geographers* 94 (2), pp. 269–277. DOI: 10.1111/j.1467-8306.2004.09402003.x.

Sui, D and M Goodchild (2011). 'The Convergence of GIS and Social Media: Challenges for GIScience'. *International Journal of Geographical Information Science* 25 (11), pp. 1737–1748. DOI: 10.1080/13658816.2011.604636.

Sutko, D and A de Souza e Silva (2011). 'Location-aware Mobile Media and Urban Sociability'. *New Media & Society* 13 (5), pp. 807–823. DOI: 10.1177/1461444810385202.

Tang, K, J Lin, J Hong, D Siewiorek and N Sadeh (2010). 'Rethinking Location Sharing: Exploring the Implications of Social-driven vs. Purpose-driven Location Sharing'. In: *Proceedings of the 12th ACM International Conference on Ubiquitous Computing - Ubicomp '10*. New York, NY: ACM Press, pp. 85–94. DOI: 10.1145/1864349.1864363.

Tasse, D, Z Liu, A Sciuto and J Hong (2017). 'State of the Geotags: Motivations and Recent Changes'. In: *Proceedings of the Eleventh International AAAI Conference on Web and Social Media*. International Conference on Web and Social Media (ICWSM). Montréal, CA: AAAI Press, pp. 250–259.

Taylor, C (2004). *Modern Social Imaginaries*. Durham, NC: Duke University Press.

Thielmann, T (2010). 'Locative Media and Mediated Localities'. *Aether: The Journal of Media Geography* 5, pp. 1–17.

Thomson, R, N Ito, H Suda, F Lin, Y Liu, R Hayasaka, R Isochi and Z Wang (2012). ‘Trusting Tweets: The Fukushima Disaster and Information Source Credibility on Twitter’. In: *Proceedings of the 9th International ISCRAM Conference*. Vancouver.

Tiefelsdorf, M (1998). ‘Some Practical Applications of Moran’s I’s Exact Conditional Distribution’. *Papers in Regional Science* 77 (2), pp. 101–129. DOI: 10.1111/j.1435-5597.1998.tb00710.x.

Tiefelsdorf, M and B Boots (1997). ‘A Note on the Extremities of Local Moran’s Iis and Their Impact on Global Moran’s I’. *Geographical Analysis* 29 (3), pp. 248–257. DOI: 10.1111/j.1538-4632.1997.tb00960.x.

Tiefelsdorf, M, D Griffith and B Boots (1999). ‘A Variance-Stabilizing Coding Scheme for Spatial Link Matrices’. *Environment and Planning A* 31 (1), pp. 165–180. DOI: 10.1068/a310165.

Tobler, W (1970). ‘A Computer Movie Simulating Urban Growth in the Detroit Region’. *Economic Geography* 46, pp. 234–240. DOI: 10.2307/143141.

Tuan, Y (1977). *Space and Place: The Perspective of Experience*. Minneapolis, MN: University of Minnesota Press.

Tufekci, Z (2014). ‘Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls’. In: *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*. Ed. by E Adar and P Resnick. Ann Arbor, MI: The AAAI Press, pp. 505–514.

Turner, A (2007). *An Introduction to Neogeography*. Santa Clara, CA: O’Reilly Media.

Turner, M (1989). ‘Landscape Ecology: the Effect of Pattern on Process’. *Annual Review of Ecology and Systematics* 20 (1), pp. 171–197. DOI: 10.1146/annurev.es.20.110189.001131.

Unwin, D (1996). ‘GIS, Spatial Analysis and Spatial Statistics’. *Progress in Human Geography* 20 (4), pp. 540–551. DOI: 10.1177/030913259602000408.

Vich, G, O Marquet and C Miralles-Guasch (2017). ‘Suburban Commuting and Activity Spaces: Using Smartphone Tracking Data to Understand the Spatial Extent of Travel Behaviour’. *The Geographical Journal* 183 (4), pp. 426–439. DOI: 10.1111/geoj.12220.

Volkart, E (1951). *Social Behavior and Personality*. New York, NY: Social Science Research Council.

Waldhör, T (1996). ‘The Spatial Autocorrelation Coefficient Moran’s I Under Heteroscedasticity’. *Statistics in Medicine* 15 (7-9), pp. 887–892. DOI: 10.1002/(SICI)1097-0258(19960415)15:7/9<887::AID-SIM257>3.0.CO;2-E.

Walmsley, D and G Lewis (2014). *People and Environment: Behavioural Approaches in Human Geography*. London, UK: Routledge.

Walter, S (1992a). ‘The Analysis of Regional Patterns in Health Data. I. Distributional Considerations’. *American Journal of Epidemiology* 136 (6), pp. 730–741.

— (1992b). ‘The Analysis of Regional Patterns in Health Data. II. II. The Power to Detect Environmental Effects’. *American Journal of Epidemiology* 136 (6), pp. 742–759.

Wang, J, A Stein, B Gao and Y Ge (2012b). ‘A Review of Spatial Sampling’. *Spatial Statistics* 2 (1), pp. 1–14. DOI: 10.1016/j.spasta.2012.08.001.

Wang, X, J Ge, W Wei, H Li, C Wu and G Zhu (2016b). ‘Spatial Dynamics of the Communities and the Role of Major Countries in the International Rare Earths Trade: A Complex Network Analysis’. *PLOS ONE* 11 (5). Ed. by G Sun, e0154575. DOI: 10.1371/journal.pone.0154575.

Weiss, E, G Kemmler, E Deisenhammer, W Fleischhacker and M Delazer (2003). ‘Sex Differences in Cognitive Functions’. *Personality and Individual Differences* 35 (4), pp. 863–875. DOI: 10.1016/S0191-8869(02)00288-X.

Wender, K, D Haun, B Rasch and M Blümke (2002). 'Context Effects in Memory for Routes'. In: *Spatial Cognition III*. Ed. by C Freksa, W Brauer, C Habel and K Wender. Tutzing: Springer, pp. 209–231. DOI: 10.1007/3-540-45004-1_13.

Westerholt, R (2018). 'The Impact of Different Statistical Parameter Values between Point Based Datasets when Assessing Spatial Relationships'. In: *Proceedings of the AGILE'2018 International Conference on Geographic Information Science*. AGILE, submitted.

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 (3), pp. 571–600. DOI: 10.1080/13658816.2017.1402914.

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 (5), pp. 868–887. DOI: 10.1080/13658816.2014.1002499.

Westerholt, R, E Steiger, B Resch and A Zipf (2016). 'Abundant Topological Outliers in Social Media Data and Their Effect on Spatial Analysis'. *PLOS ONE* 11 (9), e0162360. DOI: 10.1371/journal.pone.0162360.

Winter, S and C Freksa (2012). 'Approaching the Notion of Place by Contrast'. *Journal of Spatial Information Science* 5 (5), pp. 31–50. DOI: 10.5311/JOSIS.2012.5.90.

Witt, J, D Proffitt and W Epstein (2010). 'When and How are Spatial Perceptions Scaled?' *Journal of Experimental Psychology: Human Perception and Performance* 36 (5), pp. 1153–1160. DOI: 10.1037/a0019947.

Wolter, D and J Lee (2010). 'Qualitative Reasoning with Directional Relations'. *Artificial Intelligence* 174 (18), pp. 1498–1507. DOI: 10.1016/J.ARTINT.2010.09.004.

Wu, F, Z Li, W Lee, H Wang and Z Huang (2015). 'Semantic Annotation of Mobility Data using Social Media'. In: *Proceedings of the 24th International Conference on World Wide Web*. Ed. by A Gangemi, S Leonardi and A Panconesi. New York, NY: ACM Press, pp. 1253–1263. DOI: 10.1145/2736277.2741675.

Wurgaft, B (2003). 'Starbucks and Rootless Cosmopolitanism'. *Gastronomica* 3 (4), pp. 71–75. DOI: 10.1525/gfc.2003.3.4.71.

Xu, M, C Mei and N Yan (2014a). 'A Note on the Null Distribution of the Local Spatial Heteroscedasticity (LOSH) Statistic'. *The Annals of Regional Science* 52 (3), pp. 697–710. DOI: 10.1007/s00168-014-0605-5.

Zadra, J and G Clore (2011). 'Emotion and Perception: The Role of Affective Information'. *Wiley Interdisciplinary Reviews: Cognitive Science* 2 (6), pp. 676–685. DOI: 10.1002/wcs.147.

Zandbergen, P (2009). 'Accuracy of iPhone Locations: A Comparison of Assisted GPS, WiFi and Cellular Positioning'. *Transactions in GIS* 13 (s1), pp. 5–25. DOI: 10.1111/j.1467-9671.2009.01152.x.

Zappavigna, M (2012). *Discourse of Twitter and Social Media: How We Use Language to Create Affiliation on the Web*. London, UK: Continuum International Publishing.

Zhang, T (2008). 'Limiting Distribution of the G Statistics'. *Statistics and Probability Letters* 78 (12), pp. 1656–1661. DOI: 10.1016/j.spl.2008.01.023.

Zhang, T and G Lin (2016). 'On Moran's I Coefficient Under Heterogeneity'. *Computational Statistics and Data Analysis* 95, pp. 83–94. DOI: 10.1016/j.csda.2015.09.010.

Zimmermann, A, A Lorenz and R Oppermann (2007). 'An Operational Definition of Context'. In: *Modeling and Using Context*. Ed. by B Kokinov, D Richardson, T Roth-Berghofer and L Vieu. Heidelberg: Springer, pp. 558–571. DOI: 10.1007/978-3-540-74255-5_42.

Zimmermann, D and M Stein (2010). 'Classical Geostatistical Methods'. In: *Handbook of Spatial Statistics*. Ed. by A Gelfand, P Diggle, M Fuentes and P Guttorp. Boca Raton, FL: CRC Press, pp. 29–44.

Part II

Publications

II.1 A Local Scale-Sensitive Indicator of Spatial Autocorrelation for Assessing High- and Low-Value Clusters in Multi-Scale Datasets

Abstract

Georeferenced user-generated datasets like those extracted from Twitter are increasingly gaining the interest of spatial analysts. Such datasets oftentimes reflect a wide array of real-world phenomena. However, each of these phenomena takes place at a certain spatial scale. Therefore, user-generated datasets are of multi-scale nature. Such datasets cannot be properly dealt with using the most common analysis methods, because these are typically designed for single-scale datasets where all observations are expected to reflect one single phenomenon (e. g., crime incidents). In this paper, we focus on the popular local G statistics. We propose a modified scale-sensitive version of a local G statistic. Furthermore, our approach comprises an alternative neighbourhood definition that is enables to extract certain scales of interest. We compared our method with the original one on a real-world Twitter dataset. Our experiments show that our approach is able to better detect spatial autocorrelation at specific scales, as opposed to the original method. Based on the findings of our research, we identified a number of scale-related issues that our approach is able to overcome. Thus, we demonstrate the multi-scale suitability of the proposed solution.

Keywords: Scale, Spatial Autocorrelation, User-Generated Data, Social Media, Twitter

II.1.1 Introduction

Spatial patterns of geographic phenomena can be explored using indicators of spatial autocorrelation. Such indicators express the degree of dependence among different observations of some spatial variable (Cliff and Ord 1969). In more general terms, spatial autocorrelation can be described as the correlation between a matrix of spatial relations (usually referred to as “spatial weights matrix”) and an attribute value matrix. Corresponding indices are often designed as test statistics. In such circumstances, their goal is to find unusually high degrees of spatial dependence by testing against the null hypothesis of spatial independence (Getis 2010). Typical fields where this kind of statistic is particularly helpful are human geography, epidemiology or criminology. In such fields, spatial autocorrelation statistics can, for instance, be used for finding areas of high economic prosperity, regions of elevated infectivity or crime hot spots.

One recurring problem with spatial autocorrelation statistics is their sensitivity to spatial scale effects. Most geographic phenomena operate on a specific scale range. This typically includes both an upper and a lower distance bound. Some processes occur globally, while others are limited to small regions (Dungan et al. 2002). Therefore, geographic data acquisition requires adjusting the measuring scale to the phenomenon of interest. This is achievable with little effort in controlled experiments that rely on automated measuring devices. Appropriate geographic deployment of such devices leads to a correctly scaled dataset. However, adjusting the measurement scale becomes more difficult (or even impossible) when employing uncontrolled data acquisition methods, for instance when observing social activities through georeferenced human reports in social media feeds like Twitter. Such uncontrolled data acquisition does not allow for a priori scale adjusting and thus causes a potential misfit of the measuring scale. In addition, user-generated data often represents more than one phenomenon. Observations originating from such data sources reflect a wide array of underlying phenomena. Moreover, single contributors reporting about these phenomena typically do not interact directly. Thus, their contributions appear in a geometrically superimposed manner. A similar effect can be observed in census data, where processes operating at different scales are interacting crosswise and are aggregated to the respective datasets (Manley et al. 2006). The analysis of user-generated data in general is of ever increasing interest. Recently, social media in particular has been leveraged in diverse fields such as human mobility analysis (*e. g.*, Hawelka et al. (2014)), event detection (*e. g.*, Crooks et al. (2013)) or sentiment analysis (*e. g.*, Mitchell et al. (2013)).

However, most of the available spatial autocorrelation statistics have been developed in the context of controlled data acquisition processes. They assume some spatial variable to represent only one phenomenon, measured at a best fitting scale. In such case, it is possible to adopt a region-oriented point of view by asking the question *"What region of a dataset is out of the ordinary?"* Here, one just has to properly model the size and shape of the focal neighbourhoods. However, multi-topic and thus multi-scale datasets like those extracted from social media are of heterogeneous nature. Every sub-region can contain observations at small scales being situated next to others at larger scales. These observations appear to be crosswise and overlapping. In fact, one region cannot be regarded as one coherent spatial unit in such cases. The question here changes to *"Which observation at a certain scale in what region of a dataset is out of the ordinary?"* Thus, the focus changes from being purely region-oriented towards a phenomenon-oriented viewpoint. The question is how to separate the extraordinary from the ordinary without drawing wrong conclusions from such heterogeneous mixed-scale regions.

Existing spatial autocorrelation approaches apply various strategies for coping with scale issues. One of these is to vary the spatial weights matrix in size, shape or topological configuration. A broad range of different approaches was developed over the last decades. Getis and Aldstadt (2004) figured out eleven different general schemes, without claiming completeness. A well-known scale-related issue that is related to neighbourhood definition is that of topological invariance. Different topological configurations might comprise the same spatial weights matrix when being modelled by simple binary contiguity. This effect even appears across different scales (Dacey 1965). One can avoid this kind of problem by recognizing topology in the neighbourhood definition via applying an appropriate weighting scheme (Cliff and Ord 1969). Another way of dealing with scale is to use local statistics instead of global measures. These can account for non-stationary spatial processes and exogenous factors causing heterogeneity (such as topography). Thus, they can model local-scale characteristics more realistically (Fotheringham 2009).

In this paper, we propose a modified version of a local G statistic, which we call *GS statistic*. The "S" in the name reflects our emphasis on scale. Our version of the local G statistic is able to deal with multi-scale datasets. Spatial autocorrelation can be assessed by following a two-step approach: First, the scale range of interest is extracted by relying on a new neighborhood definition. Our neighborhood definition differs from common approaches in that all tuples of observations within the local focus are examined with respect to their scale. Furthermore, the principle of the statistic itself is modified towards operating at a certain scale, instead of mixing up different ones. This allows for unraveling the autocorrelation structure of all locally available scales separately. We further develop equations for assessing the variance and the expectation and we present a standardized version of our statistic. Finally, we test our approach by comparing it to the original method. We apply both the original and our method to a Twitter dataset consisting of a snapshot of an urban setting from the city of San Francisco and we discuss some scale-related issues.

We start the remainder of this article by giving background information on the ambiguous term of geographic scale in Section II.1.2. Afterwards, in Section II.1.3, we present a literature review on the field of spatial autocorrelation statistics, with special focus on scale. In Section II.1.4, we define our modified statistic, which is being tested in Section II.1.5. We end our paper with some concluding remarks in Section II.1.6.

II.1.2 Background: Some Notes on Geographic Scale

The concept of geographic scale is central to this paper. Spatial phenomena are supposed to operate at a certain scale. Therefore, accounting for this property is crucial for obtaining realistic results from spatial autocorrelation analysis. However, scale is an ambiguous term. While the concept is of interest to several disciplines, each adopted a different meaning (see Gibson et al. (2000) for a multi-disciplinary overview). Ecologists use the term for describing levels in the hierarchical system of biological taxonomy or in the hierarchy of a food chain (Allen and Hoekstra 1992). Sociologists classify their research according to the scale of human relationships, *i. e.*, into micro-, meso-, macro- and global-sociology (Smelser 1995). Scholars from political sciences or from urban planning use the term “scale” less from a quantitative than a conceptual point of view. In analogy to political jurisdictions, they classify their research into studies at the local, regional, national or international scale (Turner 1989; Gibson et al. 2000).

Different notions of scale are also common even within the single discipline of Geography. Cartographic scale, for instance, refers to a ratio between model and reality. It is a proxy for the degree of spatial reduction during the process of reality abstraction (Turner 1989). In contrast, phenomenon scale (or operational scale) describes the areal magnitude that a phenomenon covers in the real world (Lam and Quattrochi 1992; Montello 2001). Its counterpart is analysis scale (or methodological scale), which denotes the unit size used for aggregation (Lam and Quattrochi 1992; Montello 2001). The concurrent term “resolution” basically describes the same concept in remote sensing, where it is used to specify the width of equally sized grid cells. Another more general description of the concept of resolution/analysis scale has been given by Waldo Tobler. He describes this concept as the representation of the smallest distinguishable parts (Tobler 1988).

Throughout the remainder of this paper, we use the term “scale” to refer either to phenomenon or analysis scale. Both are interrelated. If one is analyzing a spatial phenomenon at a wrongly adjusted analysis scale, the analyst misses out the essential information (*i. e.*, spatial variation) (Goodchild 2001). Thus, it is crucial to harmonize the phenomenon scale (or the “real-world” scale) and the analysis scale.

II.1.3 Literature Review

A broad range of indicators for measuring spatial autocorrelation has been developed over the last decades. Many of them are of global nature and describe the average spatial autocorrelation across a given region. Popular examples include the autocovariance-based Moran’s I (Moran 1950), the semivariance-based Geary’s c (Geary 1954) or Tango’s C (Tango 1995) and Rogerson’s R (Rogerson 1998), the two latter being both related to the χ^2-goodness-of-fit test. A statistic that moreover allows statements about the characteristics of the involved observations is Getis & Ord’s G (Getis and Ord 1992; Ord and Getis 1995).

Zhang and Lin (2006) modified G for overcoming the problem whereby high and low values might cancel each other out. These authors also presented an alternative approach to G by decomposing Moran's *I* into three separate statistics (Zhang and Lin 2007). These are respectively capable of finding either high-value, medium-value or low-value accumulation.

The indicators presented above are designed for dealing with numerical attribute values. However, more recently, some research has also taken place around indicating spatial autocorrelation in the context of categorical data. This kind of spatial association is indeed beyond the focus of this paper. However, some recent examples can be found in (Boots 2003; Ruiz et al. 2010; Leibovici et al. 2014). Most of these indicators are based on entropy measures.

Approaches to the treatment of scale and related issues can be distinguished into two general but complementary strategies: The use of local statistics and the design of spatial weight matrices. Local statistics are better suited for taking into account the local context than global ones (Fotheringham 2009). These measures assess the autocorrelation of a given local sub-region instead of subsuming the whole spatial autocorrelation structure by just one number. This category of statistics is relatively recent and is often designed to complement some corresponding and already available global measure. Examples of such statistic include $\mathrm{G_i}$ and $\mathrm{G_i}^*$ (Getis and Ord 1992), LISA statistics (the local versions of Moran's *I* and Geary's *c*) (Anselin 1995), U (Tango 1995) or local R (Rogerson 1998). The general principle of these statistics is to compare a local neighbourhood to some overall dataset. However, this is problematic when considering the potential heterogeneity of spatial regions with respect to underlying covariates. A recent approach that has been presented by Ord and Getis (2001) tries to overcome this issue by comparing contiguous regions instead.

The compilation of spatial weights matrices is another strategy for dealing with scale issues. Getis and Aldstadt (2004) revealed at least eleven different schemes for this purpose. Getis (2009) categorized them into three categories according to their respective nature. Following this, spatial weight matrices can be constructed by following a theoretical, empirical or topological point of view. Theoretical approaches are based on some underlying distance theory such as Zipf's law (Zipf 1949). They assume the spatial weights to be exogenous to any system. The most frequently applied approach of this kind is using some sort of inverse distance. Scale is typically modelled by inducing an upper distance bound. The opposite of the theoretical approach to constructing weight matrices is constructing them in an empirical manner. Here, the analyst tries to estimate the neighbourhood structure by extracting it from some reference region of a given dataset. However, this reference region is also the limiting factor for the explanatory power of such matrices. A third approach to matrix construction is trying to depict the topology as realistically as possible. These approaches are motivated by the well-known issue of topological invariance (Dacey 1965), which leads to similar matrices across different topological settings when using binary contiguity indicators. An issue related to scale here is that differently sized spatial units are nevertheless treated similarly. Cliff and Ord (1969) suggested using suitable weighting schemes to overcome this problem. Examples of recent approaches for matrix construction include that of Getis and Aldstadt (2004) (utilization of a local statistic for assessing a proper matrix) or LeSage (2003) (Gaussian distance). Two interesting approaches with specific focus on scale are presented by Aldstadt and Getis (2006) and Rogerson and Kedron (2012). Both of them are based on successive expansions of the neighbourhood size until a maximum value of a given local statistic (*e. g.*, local Moran's *I*) is reached. Another approach for finding a suitable scale is leveraging the range of local semivariograms (Lloyd 2011). However, this is more common with geostatistical scenarios such as kriging.

In summary, research on indicators for measuring spatial autocorrelation has a long-standing tradition. Indicators can be found for different types of data and originate from different domains. The same is true for scale problems, which have indeed always been important to geographic problems. However, dealing with scale remains a challenging and yet unsolved task (Getis 2006). It is interesting to note that even today, after decades of research, modeling scale remains one of the biggest challenges in spatial analysis (Fotheringham 2009). With the rise of mixed-scale datasets like those extracted from social media, this issue is becoming even more challenging. None of the available approaches focuses on this specific problem. Thus, this is the motivation for our research.

II.1.4 A Scale-Sensitive Local G Statistic

Before defining our scale-sensitive local G statistic, we first introduce the original method (Getis and Ord 1992; Ord and Getis 1995). This statistic aims to assess not only spatial autocorrelation but also the character of the observations that are involved. More specifically, it shows whether any local accumulation primarily consists of high, medium or low attribute values. Two slightly different versions of the local G statistics are available. One of them (called $\mathrm{G_i}^*$) includes the current observation under investigation. Its counterpart (called $\mathrm{G_i}$) neglects the observation being examined and only accounts for its neighbours. Equations II.1.1 and II.1.2 define both measures.

$$G_i{}^* = \frac{\sum_j \omega_{i,j} \cdot x_j}{\sum_j x_j} \tag{II.1.1}$$

$$G_i = \frac{\sum_{j \neq i} \omega_{i,j} \cdot x_j}{\sum_j x_j} \tag{II.1.2}$$

The variable x represents the attribute values. The matrix ω denotes a binary spatial weights matrix, where values of one indicate adjacency to observation i. However, non-binary matrices are also allowed. The index j iterates over the adjacent observations.

II.1.4.1 Issues Regarding Scale

The problem that is addressed in this paper is the issue of inadequate scale treatment when it comes to multi-scale datasets. One issue that arises is related to the different scales involved in the nominator and denominator of the local G statistic. In Equations II.1.1 and II.1.2, the nominators represent the sum of the accumulated attribute values contained in a given local neighbourhood. That neighbourhood may be defined by any given distance threshold. This sum is being compared against the overall sum of the attributes' values throughout the entire dataset (represented by the denominators). Now, if one changes the distance threshold used to define the neighbourhood, it will clearly result in a scale change in the nominators. However, there is no effect on the values they are being compared to, for the denominators remain unchanged. This fact causes a serious issue when it comes to multi-scale datasets, whereby phenomena occurring at different scales are compared with each other.

While the nominator represents spatial relations within a given distance range, the denominator comprises spatial relations across all scales that are present in the dataset. This is indeed not an issue with single-scale datasets, since only one scale is of interest under such circumstances. However, it becomes a problem when analysing multi-scale datasets. In such cases, different scales are being mixed up, although they might represent different phenomena. Another problem is the way in which neighbourhoods are

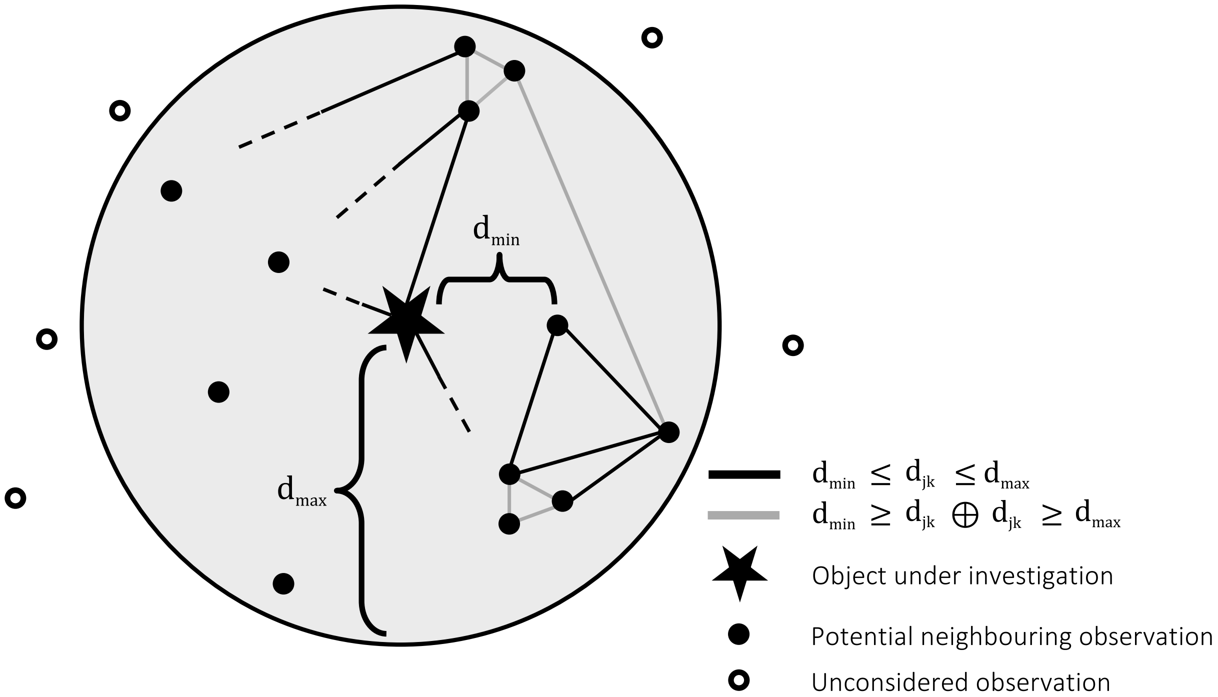

d_min
d_max
d_min ≤ d_jk ≤ d_max
d_min ≥ d_jk ⊕ d_jk ≥ d_max
Object under investigation
Potential neighbouring observation
Unconsidered observation

Figure II.1.1: Schematic sketch of the proposed scale-adjusted neighbourhoods. d = distance; $j, k \in \mathbb{N}$ = indices of observations; $\oplus$ = 'exclusive-or'.

typically defined. As mentioned in Section II.1.3, many different approaches exist. However, they typically model the neighbourhood as a fixed-size area around some observation. Furthermore, they assume to include single-scale observations. This is inappropriate for multi-scale datasets, since phenomena at different scales might be situated in close proximity to each other and overlap. Thus, prior to redefining the original statistic, we need to introduce an alternative neighbourhood definition.

II.1.4.2 Scale-Adjusted Neighbourhoods

The first step of our proposed solution for overcoming the problems with multi-scale datasets is the use of scale-adjusted neighbourhoods. Common approaches for neighbourhood definition specify their shape, size or topological ordering (Getis 2009). The focal scale is usually modelled by choosing a sufficient neighbourhood size. All instances being situated closer than a defined distance threshold are taken into account. The threshold's value is set based on the phenomenon being studied. However, in case of multi-scale datasets, one is implicitly dealing with observations at scales that are smaller (or even larger) than the intended one. Therefore, we suggest using an upper and a lower distance threshold. Moreover, these thresholds are then used for evaluating the pairwise distances between all features in the vicinity of the examined observation. If the distance between two of these features exceeds the upper bound or is shorter than the lower one, their relationship is neglected and excluded from the neighbourhood. Figure II.1.1 illustrates this approach.

II.1.4.3 Development of the Proposed GS Statistic

In this sub-section, we define our approach to defining a local scale-sensitive high/low value autocorrelation statistic. This measure is derived by adapting the local G statistic, as stated above. We call our statistic "GS

Table II.1.1: Preliminary variable definitions.

n	Total number of point features
ϕ_{jk}	Binary variable, indicating scale fit (1) or misfit (0)
ω_{jk}	Spatial weights, indicating adjacency of k to j
$f(x_j, x_k) \mathrel{\widehat{=}} f_{jk} := f : D \times D \mapsto \mathbb{R}$	A function that maps two input attributes associated with points j and k to a real-valued variable
$\Gamma = \sum\limits_{j}^{n} \sum\limits_{k \neq j}^{j-1} f(x_j, x_k)$	The attribute value sum of all scale-fitting relationships shared by points j and k
$\Phi = \sum\limits_{j}^{n} \sum\limits_{k \neq j}^{j-1} \phi_{jk}$	The total number of relationships fitting the analysis scale
$W = \sum\limits_{m}^{n} \sum\limits_{j}^{n} \sum\limits_{k \neq j}^{j-1} \omega_{mj}\, \omega_{mk}\, \phi_{jk}$	The cumulative number of relationships across all neighbourhoods fitting the analysis scale
$W_i = \sum\limits_{j}^{n} \sum\limits_{k \neq j}^{j-1} \omega_{ij}\, \omega_{ik}\, \phi_{jk}$	The number of scale-fitting relationships adjacent to observation i
$A = \sum\limits_{m}^{n} \sum\limits_{j}^{n} \sum\limits_{k \neq j}^{j-1} \omega_{mj}\, \omega_{mk}\, \phi_{jk}\, f(x_j, x_k)$	The cumulative attribute value sum across all neighbourhoods at the given analysis scale

statistic", where the added "S" reflects the emphasis on scale. It should be noted that our definition given below focuses on pairwise relationships among observations. This kind of analysis is of broad interest for the analysis of data extracted from social media, where analysts are often interested in collective processes that occur within some geographic region. Thus, one might want to consider relationships among observations instead of focusing on single occurrences. Our tests, which are presented in Section II.1.5, deal with one such example (where semantic similarities are used to establish relationships). However, it would also be of interest to generalize our basic principles to other geometric configurations. Since this is beyond the scope of this paper, we leave that open to future research.

It is necessary to introduce some preliminary definitions, which are presented in Table II.1.1. These are used throughout the remainder of this paper. We define them at this early stage for the sake of readability of our equations. In addition, please note that we are using reduced designator notations (*i. e.*, GS_i^* instead of $GS_{i\;d_{min}}^{*d_{max}}$ and f_{jk} instead of $f(x_j, x_k)$) for notational convenience.

The definition of the proposed statistic is based on the original statistic as given by Equation II.1.1. Most formulas in the text are given without derivation. More detailed derivations can be found in Appendices II.1.7.1 to II.1.7.5. Equation II.1.3 shows our modified version of a scale-sensitive $\mathrm{G_i}^*$ statistic:

$$GS_i{}^* = \frac{\sum_j^n \sum_{k \neq j}^{j-1} \omega_{ij}\, \omega_{ik}\, \phi_{jk}\, f_{jk}}{\sum_j^n \sum_k^n \sum_{m \neq k}^{k-1} \omega_{jk}\, \omega_{jm}\, \phi_{km}\, f_{km}} \tag{II.1.3}$$

As we are operating on pairwise relationships between tuples of observations, the indices j and k represent the two observations being involved in that relationship. Thereby, the indices j and k have to be different. Otherwise, a single point would be set into a relationship to itself. An additional indicator variable denoted ϕ_{jk} has also been included. Its value is 1 if the distance between two contiguous features j and k is within the interval $[d_{min}, d_{max}]$, and 0 otherwise. Furthermore, the spatial weights matrix ω is evaluated twice. This is necessary because both observations j and k must be adjacent to observation i. These modifications allow the inclusion of scale-adjusted neighbourhoods as described in Section II.1.4.2 and lead to a match between nominator and denominator scales.

Under the null hypothesis (H_0) of spatial independence, each outcome of function f is supposed to be occurring equally likely (*i. e.*, $P(f_{jk}) = 1/n$). Furthermore, we suppose pairwise independence between those outcomes. It follows that the expectation for f is estimated by:

$$\hat{E}[f] = \frac{\sum_j^n \sum_{k \neq j}^{j-1} \phi_{jk} \, f_{jk}}{\Phi} \tag{II.1.4}$$

By using Equation II.1.4, we can define the empirical expectation of the $\mathrm{GS_i}^*$ statistic under H_0 (Equation II.1.5). The first factor and the denominator in Equation II.1.5 are constant across all neighbourhoods. Therefore, these can be ignored and the equation reduces to W_i/W. It follows that the statistic's local expectation is supposed to be proportional to the respective neighbourhood's fraction among all neighbourhoods at the given scale. This is analogous to the original method.

$$\hat{E}[GS_i{}^*] = \frac{\hat{E}[f] \cdot W_i}{A} \sim \frac{W_i}{W} \tag{II.1.5}$$

In equations II.1.6 and II.1.7, we develop equations for the variance of the $\mathrm{GS_i}^*$ statistic. Therefore, we first need an equation for the estimate of the expectation of the squared test statistic in Equation II.1.6. This is then used to estimate the empirical variance in II.1.7 by applying the so called one-pass algorithm (Chan et al. 1983).

$$\hat{E}[GS_i^{*2}] = \frac{\frac{W_i \cdot \sum_j^n \sum_{k \neq j}^{j-1} \phi_{jk} \, f_{jk}^2}{\Phi} + \frac{W_i(W_i - 1)(\Gamma^2 - \sum_j^n \sum_{k \neq j}^{j-1} (\phi_{jk} \, f_{jk})^2)}{\Phi(\Phi - 1)}}{A^2} \tag{II.1.6}$$

$$V\hat{a}r[GS_i^{*2}] = \frac{\frac{W_i \cdot \sum_j^n \sum_{k \neq j}^{j-1} \phi_{jk} \, f_{jk}^2}{\Phi} - \frac{W_i^2 \cdot \Gamma^2}{\Phi^2} + \frac{2W_i(\Gamma^2 - \sum_j^n \sum_{k \neq j}^{j-1} (\phi_{jk} \, f_{jk})^2)}{\Phi(\Phi - 1)}}{A^2} \tag{II.1.7}$$

As expected, the variance under H_0 becomes 0 if there are no neighbours in the vicinity of observation i (*i. e.*, $W_i = 0$). The same applies if no corresponding scale-fitting relationships are located in the neighbourhood ($\Phi = 0$) or if the total attribute value sum (A) of all those features equals zero. Similarly, the variance estimation also becomes zero if the overall neighbourhood sum equals zero. In contrast, the variance is greater than zero if all observations are contained in the neighbourhood of the current feature. This is a difference from the original method. However, this becomes clear when recalling the statistic's principle: One neighbourhood is compared against all other neighbourhoods. Thus, the denominator is always greater than the nominator, resulting in a non-zero variance.

The maximum value of our statistic is reached when all neighbourhoods mutually contain each other. In such circumstances, the aggregation of all ϕ_{jk} for any tuple across the whole neighbourhood forms an all-ones matrix. It follows that the maximum value of the $\mathrm{GS_i}^*$ statistic is given as:

$$\max GS_i^* = \frac{1}{n} \tag{II.1.8}$$

Accordingly, the minimum value is reached if no values except the investigated observation itself are contained in some neighbourhood. It follows that the minimum value is given by:

$$\min GS_i^* = 0 \tag{II.1.9}$$

Equations II.1.8 and II.1.9 show that the range of the $\mathrm{GS_i^*}$ statistic is not fixed. This is a major difference compared to the original G statistics, which range is the interval $[0, 1]$. In contrast, the GSi* statistic depends on the number of input features. Thus, two $\mathrm{GS_i^*}$ values should not be compared with each other directly. A comparison is only meaningful after standardization. The standardized version of $\mathrm{GS_i^*}$ is given in Equation II.1.10. Applying this equation produces standard deviates (*i. e.*, z-scores), which appear to be on the interval $[-\infty, \infty]$. Furthermore, following the well-known central limit theorem, these scores tend to be approximately normal, given a sufficiently large sample size. Therefore, these scores can be evaluated by means of normal theory.

$$Z_{GS_i^*} = \frac{\sum_j^n \sum_{k\neq j}^{j-1} \omega_{ij}\, \omega_{ik}\, \phi_{jk}\, f_{jk} - \frac{W_i \cdot \sum_j^n \sum_{k\neq j}^{j-1} \phi_{jk}\, f_{jk}}{\Phi}}{\sqrt{\frac{W_i \cdot \sum_j^n \sum_{k\neq j}^{j-1} \phi_{jk}\, f_{jk}^2}{\Phi} + \frac{W_i (W_i - 1)(\Gamma^2 - \sum_j^n \sum_{k\neq j}^{j-1} (\phi_{jk}\, f_{jk})^2)}{\Phi(\Phi-1)} - \frac{W_i^2 (\sum_j^n \sum_{k\neq j}^{j-1} \phi_{jk}\, f_{jk})^2}{\Phi^2}}} \qquad \text{(II.1.10)}$$

II.1.5 Empirical Comparison Between $\mathrm{GS_i}$ and $\mathrm{G_i}$

We now empirically illustrate the problems that occur when applying the original Gi* statistic to multi-scale datasets such as those extracted from social media. Furthermore, we also show that our approach overcomes these problems. Before this is done, we explain the datasets that we used and all the necessary preprocessing. Please note that we do not aim to analyze the regions that we sampled with respect to the qualitative properties of the underlying phenomena. All following steps are merely illustrative for testing our suggested approach with respect to the scale issues that are mentioned in Section II.1.5.3.

II.1.5.1 Dataset Description

The datasets we used were extracted from the social media service Twitter. They originate from an urban setting in the city of San Francisco, CA. We used two randomly chosen time slots. One of them covers the time period of January 30, 2014, from 8 p.m. until 10 p.m.; the second slot covers a whole week from the 20th of January until the 26th of January 2014.

Our automated crawler leveraged the public Twitter Streaming API. Since we are interested in applying methods from spatial statistics, we restricted our query to georeferenced tweets only. We crawled all tweets from a bounding box covering the city of San Francisco and its immediate surroundings. The bounding box had a size of approximately 15×15 km. We did not restrict our data collection by using keywords or any other type of filter. The subsets of our dataset that we used for this paper sum up to a size of 1,291 tweets (for the two-hour slot) and 69,345 tweets (for the one-week slot). Figure II.1.2 provides an overview of this subset and shows its distribution over the city.

II.1.5.2 Preprocessing and Data Preparation

The crawled datasets consist of textual tweets. However, our approach as well as the original $\mathrm{G_i}^*$ statistic are designed for dealing with numerical values. Thus, we first have to transform the textual tweets into some numerical representation. We have chosen to use similarities among the tweets for our test and comparative study. In a realistic scenario, high similarity scores might be interpreted as indicators of coherent social activity (*i. e.*, people might be reporting about similar topics). In order to obtain meaningful similarities, several steps are conducted.

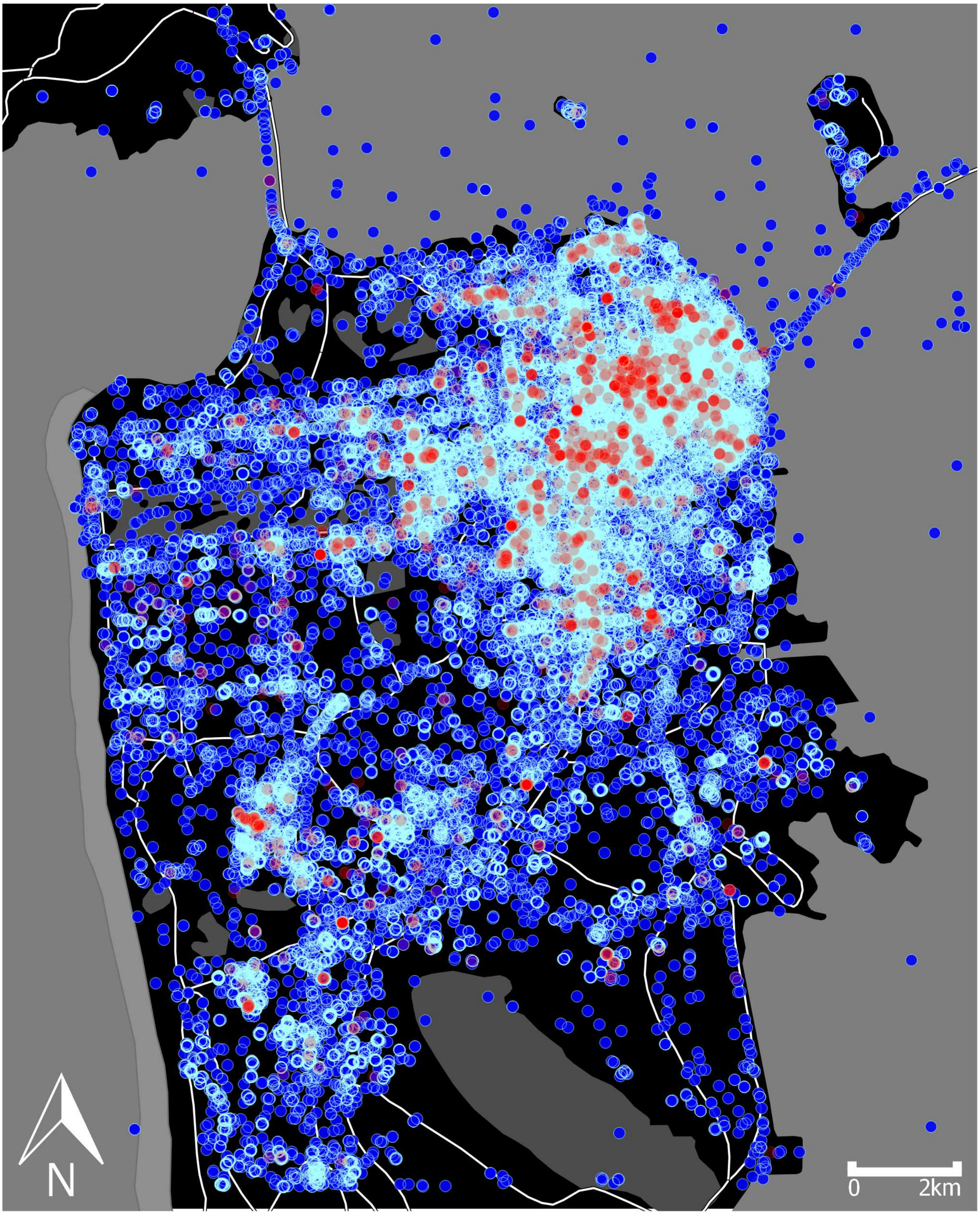


Figure II.1.2: Overview of our test datasets originating from San Francisco, CA. Blue = 20th of January until 26th of January; Red = 30th of January, 8 p.m. until 10 p.m. More intense colours indicate higher numbers of superimposed Tweets. Base data: VMAP, National Geospatial-Intelligence Agency, USA.

The first step is to split up the cohesive strings of words into single tokens. The tokenisation process that we used follows some rules that have been adapted from the recent literature: The texts are split up at case changes, except if they occur at the beginning of a word (Metke-Jimenez et al. 2011); Twitter's specific symbols (*e. g.*, #, @) are kept (O'Connor et al. 2010), and short forms or contractions of English words (*e. g.*, I'm) are retained (Pak and Paroubek 2010). Moreover, we split the tweets at white-spaces and punctuation marks. A large portion of the resulting tokens occurs frequently, but adds little meaning (*e. g.*, "to", "or"). Therefore, these so-called stop words are removed from the corpus in the second step. For this purpose, we relied on the English stop word list provided by the database system PostgreSQL.

The actual similarity assessment is based on the method of Latent Semantic Indexing (LSI) (Deerwester et al. 1990). The core principle of this method is based on a singular value decomposition (SVD). First of all, the tokens are transformed into normalised frequencies (called Term Frequency-–Inverse Document Frequency (TF–IDF) scores). These are then used for extracting inherent components, based on word co-occurrence. LSI works in an unsupervised manner. Thus, no a priori knowledge about the text corpus is needed. However, a criterion for maintaining a reasonable number of components is required. In our experiments, we used a broken stick model for this purpose. This approach is usually used for modelling resource allocation in ecology. However, it has also proven to be useful for application of the SVD (Cangelosi and Goriely 2007).

Again, note that our approach for assessing similarities has been chosen for the sake of producing numerical tweet representations. Neither similarity assessment itself nor analysing our test site is the focus of this paper. Thus, the chosen approach is appropriate for our experiments regarding the proposed statistic. We point out that more accurate semantic similarity approaches might be available (*e. g.*, Latent Dirichlet Allocation (Blei et al. 2003) or probabilistic LSI (Hofmann and Thomas 1999)). However, these are more sophisticated and require more detailed a priori knowledge about the composition of the text corpus. Whenever realistic conclusions are to be drawn from any dataset, careful consideration should be given to the choice of an appropriate semantic similarity approach.

II.1.5.3 Comparison Between $\mathrm{GS_i^*}$ and $\mathrm{G_i}^*$

Our comparison focuses on three central problems that occur when the $\mathrm{G_i}^*$ statistic is applied to multi-scale datasets. All these problems occur due to the issues highlighted in Section II.1.4.1. Moreover, we also demonstrate that these issues are solved by our proposed solution.

Overemphasis of Dominant Scales

Recall the property of scale mixing within social media data. Figure II.1.3 illustrates the average composition of five differently scaled neighbourhoods. These neighbourhoods are heterogeneous. In most cases, the actual scale of interest contributes only approximately 30 % of the total attribute value sum. This means that approximately 70 % of all variation is contributed by scales other than the one of interest. Accordingly, when applying standard (*i. e.*, single-scale) approaches for neighbourhood definition, all these scales are considered together.

However, if 70 % of the total variation is contributed by phenomena beyond interest, it is likely to create some bias in autocorrelation results. This is particularly the case when one or more of these non-relevant scales are dominating a dataset. Figure II.1.4 shows to what extent the respective scales are under- or overrated. It illustrates the ratio between the share in the attribute value sum and the share in the quantitative composition of the neighbourhoods. It can be seen that the small scales (1–30 m and

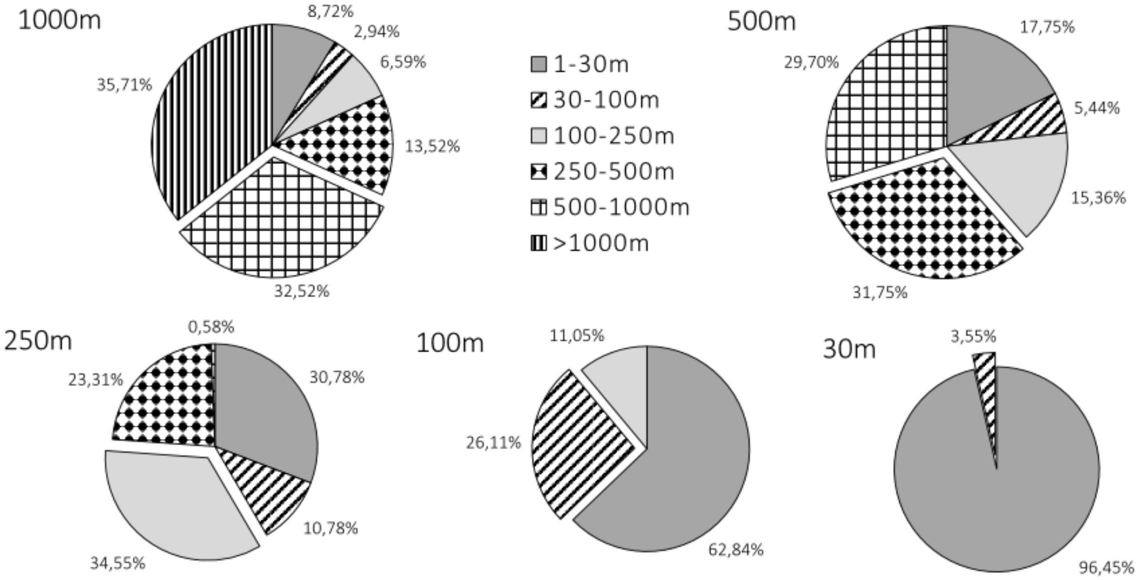


Figure II.1.3: Average composition of the attribute value sum for five classes of neighbourhood sizes. The respective scales of interest are highlighted by displacement. Dataset: Twitter, 30th of January 2014, 8 p.m. until 10 p.m.

30–100 m) are overrepresented in most neighbourhoods. Thus, phenomena operating at such scales are excessively biasing the results at other scales.

The problems described above affect the original $G_i{}^*$ statistic in two ways: On the one hand, scales are superimposed in the focal neighbourhoods. On the other hand, these are then compared against an overall mixture of scales (*i. e.*, the denominator of the statistic). The larger the scale, the more different scales are potentially being mixed up. Figure II.1.5 shows one of the effects caused by that behaviour. The mean of the z-values obtained through the $G_i{}^*$ statistic shows a strong trend with increasing scale. However, we are dealing with a standardized version of the statistic. Following the central limit theorem, the resulting standard variates are expected to be approximately normal. Thus, the mean is expected to be an unbiased estimator of the expectation, which should be close to zero in the present case. That is obviously not true for $G_i{}^*$ when it is applied to social media datasets. It is very likely that this effect is caused by the scale mixture described in the previous paragraph. That mixture implies different underlying populations, since different phenomena might be operating at the different scales. Thus, there are also different means present in the mixture. The mean of the z-values is influenced by that variety of means, which in turn leads to the observed bias.

These effects are diminished with our suggested scale-sensitive approach. Our method only extracts those scales from the vicinity of observations that are relevant for the current analysis scale. Thus, each diagram shown in Figure II.1.3 would only consist of one pie slice, each representing the respective scale of interest. The composition of the attribute value sum of the neighbourhoods is completely made up of observations fitting the scale of interest. Moreover, the same applies to the comparative size. The modified statistic only includes those observations in any calculation that are fitting the current scale of interest. Therefore, the estimated means obtained through our modified statistic (see Figure II.1.5) remain close to zero across all investigated scales.

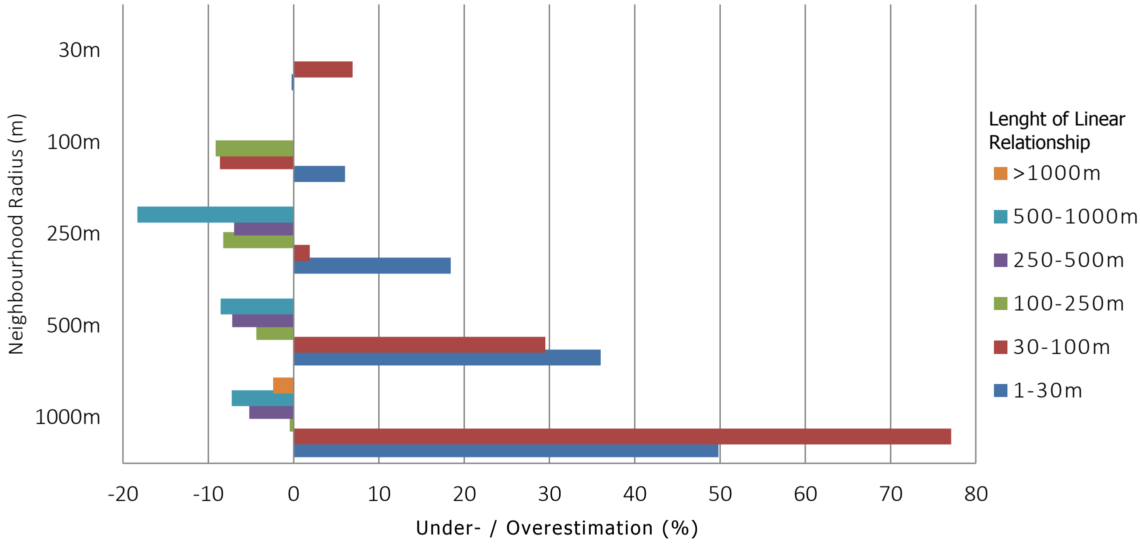


Figure II.1.4: Under-/overestimation of various scales within neighbourhoods at different analysis scales. Dataset: Twitter, 30th of January 2014, 8 p.m. until 10 p.m.

Type I/Type II Errors

The problem of overemphasising dominant scales leads to another closely related problem, which is the occurrence of type I/II errors. This is a well-known general issue of all local statistics (Nelson 2012). It is usually caused by missing strategies for facing multiple testing problems. However, when dealing with multi-scale datasets, this problem is further exacerbated by an additional problem. When some scales are dominating a dataset, they also hide weaker phenomena at less dominant scales. However, these less pronounced phenomena are not necessarily less important. Some analyst might indeed be interested in analysing these weaker phenomena. Now, several different configurations are possible: Some weaker phenomenon might, for instance, consist of some high-value accumulation. These values might, however, only be high according to their own respective scale. Some contiguous and more dominant scale might comprise even higher values. In such situations, the dominance of the other scale with high values leads

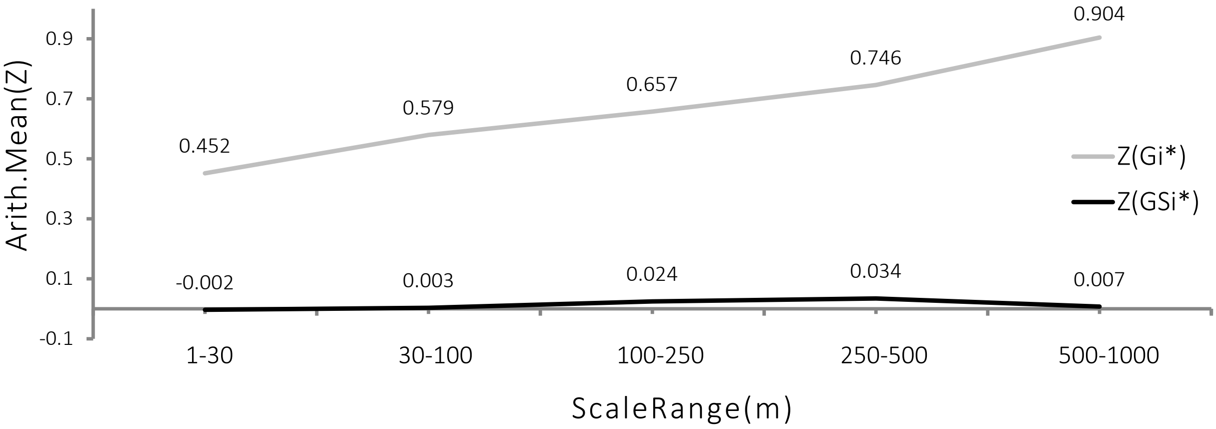


Figure II.1.5: Arithmetic means of $Z(G_i^*)$ and $Z(GS_i^*)$ across all tested scales. Dataset: Twitter, 20th of January until 26th of January.

to type II errors. H_1 is rejected although high values are present at the adjusted scale of interest. These values just appear to be quite low in comparison to the more dominant adjacent scale that is present in the same neighbourhood. The same situation occurs if a phenomenon of interest shows low-value accumulation. Higher values at another scale are again artificially raising the neighbourhood score, leading to H_1 rejection. In contrast, type I errors occur whenever a scale of interest is actually not out of the ordinary, but gets interfered by a more dominant scale. This situation might occur in both directions, either toward low values (cold spots) or high values (hot spots). In such cases, the neighbourhood score is artificially raised (or lowered) to a level that leads to a wrong acceptance of H_1.

One example from our dataset is depicted in Figure II.1.6, which is showing two series of maps. Each of those series comprises four different scales of interest in ascending order. Those series illustrate both issues described so far. On the one hand, one can see the overemphasis of dominant scales. The results obtained through the original $G_i{}^*$ at the two smallest scales show a large number of statistically significant high-value accumulations. In fact, 33.56 % of all tweets of the dataset are identified to be statistically significant with $G_i{}^*$ (scale = 1–30 m; two-sided test; α = 0.1 each). In other words, every third tweet is considered to be part of a neighbourhood that comprises high-value accumulation higher than 90 % of the other tweets. This is obviously an upwards biased value, due to the dominance of that scale compared to larger ones. In comparison, the results obtained through our modified approach show a considerably lower number of extraordinary observations. Since the dominance of scales is not affecting the results, that method only evaluates 3.77 % of the tweets to be somehow abnormal. Another issue that can be seen in Figure II.1.6 is the existence of type I errors. Because of the dominance effect described above, H_0 is rejected too often. This does not only appear at the dominant scales, but is transferred onto all larger levels as well. Hot or cold spots occurring at small scales appear to be acting like "seeds" that are being enlarged at the next larger scale. Thus, the type I errors can be found with increasing frequency by enlarging the analysis scale. This effect also does not occur in the results obtained through our proposed statistic. Every scale is only analysed against observations at the same scale. Thus, there is no dominance to be transferred, resulting into a lower number of type I errors. However, the effect of "seed" locations with $G_i{}^*$ leads to another issue that is described in the following subsection.

Loss of Statistical Independence Between Scales

We already mentioned the spill-over effect of dominant scales that are transferred onto all larger ones. We can also observe that this effect results into "seed" locations that appear to be growing as the scale is getting enlarged. However, this phenomenon leads to another much more serious problem, which is the loss of independence between spatial autocorrelation results obtained for different scales. We assume all possible outcomes of spatial autocorrelation statistics to be equally likely. That is, we assume the probabilities to be $P(x_a) = \sum x_a/n$. If different scales are being admixed, however, this assumption is no longer verified; this occurs, for instance, when assessing a non-zero spatial autocorrelation at some small scale. If the scale is adjusted to some larger value, these small-scale instances are again included. The problem is that now, the outcome of zero has become impossible. The effect of the non-zero spatial autocorrelation at the smaller scale might be blurred (due to mixing) or be changed in nature (from negative to positive or vice versa) because other observations are included in the neighbourhood. However, the result of having no autocorrelation is no longer possible at any larger scale. In other words, the independence requirement $P(x_b|x_a) = P(x_b)$ is no longer met. Since we are dealing with multi-scale datasets that reflect potentially unrelated phenomena, this is an inappropriate property.

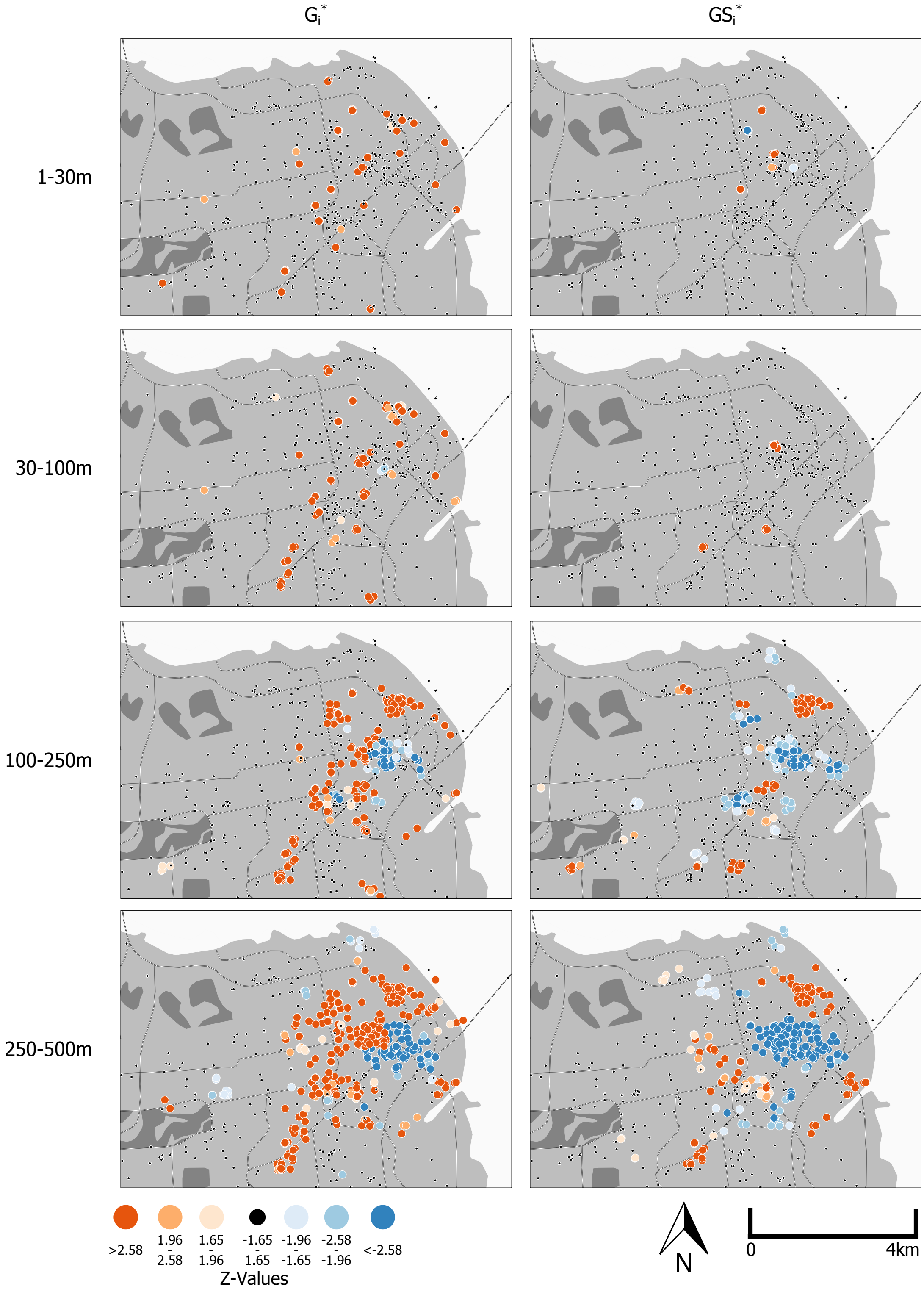


Figure II.1.6: Two series of analysis results. The left-hand side was obtained by applying the original G_i^* statistic, the right-hand side originates from our proposed GS_i^* statistic; Dataset: Twitter, 30th of January 2014, 8 p.m. until 10 p.m.; Base data: VMAP, National Geospatial-Intelligence Agency, USA.

II.1.6 Conclusions

The arising interest in analysing social media feeds and other kinds of human-generated datasets compels us to address the specific problems of such data. One of those problems is their multi-scale nature that is due to the uncontrolled data acquisition process. However, most spatial statistics are designed for single-scale datasets that result from controlled experiments. This paper introduced a scale-sensitive version of the popular G_i^* statistic. The proposed approach comprises an alternative approach for neighbourhood definition and a scale-adjustment of the statistic itself. Moreover, some scale-related issues that arise when dealing with multi-scale datasets are highlighted by comparing the results obtain through the original and the proposed statistics. These comparisons are carried out on a Twitter dataset for the city of San Francisco, CA. The results demonstrate that the suggested approach is better suited for dealing with multi-scale datasets, because it allows analysing certain scales without cross-scale interferences. Thus, it can be used in real-world scenarios whenever social media or other human-generated datasets are analysed.

However, scale-related effects affecting social media datasets are not yet fully understood. The list of issues mentioned in Section II.1.5 is given without claiming completeness. There might be many more effects that are still to be discovered. Moreover, the effects we listed and observed have not yet been fully investigated. Thus, future research should focus on getting a better understanding of the multi-scale nature of user-generated datasets. In addition, there are many more methods from spatial statistics and other fields that are not yet sufficiently capable of dealing with multi-scale datasets. Our suggested approach might serve as a starting point for initiating methodological research towards multi-scale enablement.

With respect to local autocorrelation statistics in general, more emphasis should be put on the definition of the null hypothesis. Geographic space imposes uncontrolled variance, due to varying local environmental conditions (Goodchild 2009; Anselin 1989). Local statistics such as G_i^* and our proposed solution already account for heterogeneity with respect to the spatial distribution of observations. In contrast, they usually include constant expectations of the observed variable. However, the outcomes of those variables might also be influenced by non-stationary environmental conditions. One way of overcoming this problem might be to use location-dependent expectation functions instead of constant values. Corresponding local values might be determined by methods such as Geographically Weighted Regression (Brunsdon et al. 1996). However, a specific problem to social media data is that the underlying driving forces are not yet fully understood.

Acknowledgements

Many fruitful discussions and valuable thoughts came to our minds in the course of discussions with colleagues. Therefore, we'd like to thank all staff and research affiliates from the *GIScience research group at Heidelberg University* and the *Harvard Center for Geographic Analysis*. You really helped a lot by steering us into the right direction. Special thanks go to Andreas Reimer (*GIScience Heidelberg*) for his substantive contributions and for his help in creating appealing maps. We also thank Shih-Pei Chen (*Max Planck Institute for the History of Science, Berlin*) for proof-reading this article. Furthermore, this work has partially been funded by the *Klaus Tschira Stiftung gGmbH* as well as by the *Graduate Funding Programme of the state of Baden-Württemberg*.

References (Chapter II.1)

Aldstadt, J and A Getis (2006). 'Using AMOEBA to Create a Spatial Weights Matrix and Identify Spatial Clusters'. *Geographical Analysis* 38 (4), pp. 327–343. DOI: 10.1111/j.1538-4632.2006.00689.x.

Allen, T and T Hoekstra (1992). *Toward a Unified Ecology*. New York: Columbia University Press, p. 505.

Anselin, L (1989). *What is Special About Spatial Data? Alternative Perspectives on Spatial Data Analysis*. Tech. rep. Santa Barbara, CA: National Center for Geographic Information and Analysis.

— (1995). 'Local Indicators of Spatial Association - LISA'. *Geographical Analysis* 27 (2), pp. 93–115. DOI: 10.1111/j.1538-4632.1995.tb00338.x.

Blei, D, A Ng and M Jordan (2003). 'Latent Dirichlet Allocation'. *The Journal of Machine Learning Research* 3, pp. 993–1022.

Boots, B (2003). 'Developing Local Measures of Spatial Association for Categorical Data'. *Journal of Geographical Systems* 5 (2), pp. 139–160. DOI: 10.1007/s10109-003-0110-3.

Brunsdon, C, A Fotheringham and M Charlton (1996). 'Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity'. *Geographical Analysis* 28 (4), pp. 281–298. DOI: 10.1111/j.1538-4632.1996.tb00936.x.

Cangelosi, R and A Goriely (2007). 'Component Retention in Principal Component Analysis with Application to cDNA Microarray Data'. *Biology Direct* 2 (2). DOI: 10.1186/1745-6150-2-2.

Chan, T, G Golub and R Leveque (1983). 'Algorithms for Computing the Sample Variance: Analysis and Recommendations'. *The American Statistician* 37 (3), pp. 242–247. DOI: 10.1080/00031305.1983.10483115.

Cliff, A and J Ord (1969). 'The Problem of Spatial Autocorrelation'. In: *London Papers in Regional Science (1), Studies in Regional Science*. Ed. by A Scott. London: Pion, pp. 25–55.

Crooks, A, A Croitoru, A Stefanidis and J Radzikowski (2013). '#Earthquake: Twitter as a Distributed Sensor System'. *Transactions in GIS* 17 (1), pp. 124–147. DOI: 10.1111/j.1467-9671.2012.01359.x.

Dacey, M (1965). *A Review on Measures of Contiguity for Two and k-Color Maps (Spatial Diffusion Project)*. Tech. rep. Evanston: Department of Geography, Northwestern University.

Deerwester, S, S Dumais, G Furnas, T Landauer and R Harshman (1990). 'Indexing by Latent Semantic Analysis'. *Journal of the American Society for Information Science* 41 (6), pp. 391–407. DOI: 10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9.

Dungan, J, J Perry, M Dale, P Legendre, J Fortin, A Jakomulska, M Miriti and M Rosenberg (2002). 'A Balanced View of Scale in Spatial Statistical Analysis'. *Ecography* 25 (2), pp. 626–640. DOI: 10.1034/j.1600-0587.2002.250510.x.

Fotheringham, S (2009). '"The Problem of Spatial Autocorrelation" and Local Spatial Statistics'. *Geographical Analysis* 41 (4), pp. 398–403. DOI: 10.1111/j.1538-4632.2009.00767.x.

Geary, R (1954). 'The Contiguity Ratio and Statistical Mapping'. *The Incorporated Statistician* 5 (3), pp. 115–127 + 129–146. DOI: 10.2307/2986645.

Getis, A (2006). 'Spatial Statistics'. In: *Geographical Information Systems: Principles, Techniques, Management and Applications*. Ed. by P Longley, M Goodchild, D Maguire and D Rhind. Hoboken, NJ: Wiley & Sons, pp. 239–251.

— (2009). 'Spatial Weights Matrices'. *Geographical Analysis* 41 (4), pp. 404–410. DOI: 10.1111/j.1538-4632.2009.00768.x.

Getis, A (2010). ‘Spatial Autocorrelation’. In: *Handbook of Applied Spatial Analysis*. Ed. by M Fischer and A Getis. Heidelberg: Springer, pp. 255–278.

Getis, A and J Aldstadt (2004). ‘Constructing the Spatial Weights Matrix Using a Local Statistic’. *Geographical Analysis* 34 (2), pp. 130–140. DOI: 10.1353/geo.2004.0002.

Getis, A and J Ord (1992). ‘The Analysis of Spatial Association by Use of Distance Statistics’. *Geographical Analysis* 24 (3), pp. 189–206. DOI: 10.1111/j.1538-4632.1992.tb00261.x.

Gibson, C, E Ostrom and T Ahn (2000). ‘The Concept of Scale and the Human Dimensions of Global Change’. *Ecological Economics* 32 (2), pp. 217–239. DOI: 10.1016/S0921-8009(99)00092-0.

Goodchild, M (2001). ‘Models of Scale and Scales of Modeling’. In: *Modelling Scale in Geographical Information Science*. Ed. by N Tate and P Atkinson. Chichester, UK: John Wiley & Sons, pp. 3–10.

— (2009). ‘What Problem? Spatial Autocorrelation and Geographic Information Science’. *Geographical Analysis* 41 (4), pp. 411–417. DOI: 10.1111/j.1538-4632.2009.00769.x.

Hawelka, B, I Sitko, E Beinat, S Sobolevsky, P Kazakopoulos and C Ratti (2014). ‘Geo-Located Twitter as Proxy for Global Mobility Patterns’. *Cartography and Geographic Information Science* 41 (3), pp. 260–271. DOI: 10.1080/15230406.2014.890072.

Hofmann, Thomas and Thomas (1999). ‘Probabilistic Latent Semantic Indexing’. In: *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Ed. by F Gey, M Hearst and R Tong. New York, NY: ACM Press, pp. 50–57. DOI: 10.1145/312624.312649.

Lam, N and D Quattrochi (1992). ‘On the Issues of Scale, Resolution, and Fractal Analysis in the Mapping Sciences’. *Professional Geographer* 44 (1), pp. 88–98. DOI: 10.1111/j.0033-0124.1992.00088.x.

Leibovici, D, C Claramunt, D Le Guyader and D Brosset (2014). ‘Local and Global Spatio-Temporal Entropy Indices Based on Distance-Ratios and Co-Occurrences Distributions’. *International Journal of Geographical Information Science* 28 (5), pp. 1061–1084. DOI: 10.1080/13658816.2013.871284.

LeSage, J (2003). ‘A Family of Geographically Weighted Regression Models’. In: *Advances in Spatial Econometrics: Methodology, Tools and Applications*. Ed. by L Anselin, J Florax and S Rey. Heidelberg: Springer, pp. 241–264. DOI: 10.1007/978-3-662-05617-2_11.

Lloyd, C (2011). *Local Models for Spatial Analysis*. London, UK: CRC Press, p. 336.

Manley, D, R Flowerdew and D Steel (2006). ‘Scales, Levels and Processes: Studying Spatial Patterns of British Census Variables’. *Computers, Environment and Urban Systems* 30 (2), pp. 143–160. DOI: 10.1016/j.compenvurbsys.2005.08.005.

Metke-Jimenez, A, K Raymond and I MacColl (2011). ‘Information Extraction from Web Services: A Comparison of Tokenisation Algorithms’. In: *Proceedings of the International Workshop on Software Knowledge*. Paris: SciTePress, pp. 12–23. DOI: 10.5220/0003698000120023.

Mitchell, L, M Frank, K Harris, P Dodds and C Danforth (2013). ‘The Geography of Happiness: Connecting Twitter Sentiment and Expression, Demographics, and Objective Characteristics of Place’. *PLoS ONE* 8 (5), e64417. DOI: 10.1371/journal.pone.0064417.

Montello, D (2001). ‘Scale in Geography’. In: *International Encyclopedia of the Social and Behavioural Sciences*. Ed. by N Smelser and P Baltes. Oxford, UK: Pergamon Press, pp. 13501–13504.

Moran, P (1950). ‘Notes on Continuous Stochastic Phenomena’. *Biometrika* 37 (1/2), pp. 17–23. DOI: 10.2307/2332142.

Nelson, T (2012). 'Trends in Spatial Statistics'. *The Professional Geographer* 64 (1), pp. 83–94. DOI: 10.1080/00330124.2011.578540.

O'Connor, M, M Krieger and D Ahn (2010). 'TweetMotif: Exploratory Search and Topic Summarization for Twitter'. In: *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*. Ed. by M Hearst. Menlo Park, CA: The AAAI Press, pp. 384–385.

Ord, J and A Getis (1995). 'Local Spatial Autocorrelation Statistics: Distributional Issues and an Application'. *Geographical Analysis* 27 (4), pp. 286–306. DOI: 10.1111/j.1538-4632.1995.tb00912.x.

— (2001). 'Testing for Local Spatial Autocorrelation in the Presence of Global Autocorrelation'. *Journal of Regional Science* 41 (3), pp. 411–432. DOI: 10.1111/0022-4146.00224.

Pak, A and P Paroubek (2010). 'Twitter as a Corpus for Sentiment Analysis and Opinion Mining'. In: *Proceedings of the Seventh International Conference on Language Resources and Evaluation*. Ed. by N Calzolari and K Choukri. Paris: European Language Resources Association, pp. 1320–1326.

Rogerson, P (1998). 'The Detection of Clusters Using a Spatial Version of the Chi-Square Goodness-Of-Fit Statistic'. *Geographical Analysis* 31 (2), pp. 130–147. DOI: 10.1111/j.1538-4632.1999.tb00973.x.

Rogerson, P and P Kedron (2012). 'Optimal Weights for Focused Tests of Clustering Using the Local Moran Statistic'. *Geographical Analysis* 44 (2), pp. 121–133. DOI: 10.1111/j.1538-4632.2012.00840.x.

Ruiz, M, F López and A Páez (2010). 'Testing for Spatial Association of Qualitative Data Using Symbolic Dynamics'. *Journal of Geographical Systems* 12 (3), pp. 281–309. DOI: 10.1007/s10109-009-0100-1.

Smelser, N (1995). *Problematics of Sociology : the Georg Simmel Lectures*. Berkeley, CA: University of California Press.

Tango, T (1995). 'A Class of Tests for Detecting "General" and "Focused" Clustering of Rare Diseases'. *Statistics in Medicine* 14 (21-22), pp. 2323–2334. DOI: 10.1002/sim.4780142105.

Tobler, W (1988). 'Resolution, Resampling, and all That'. In: *Building Database for Global Science*. Ed. by H Mounsey and R Tomlinson. London, UK: Taylor & Francis, pp. 129–137.

Turner, M (1989). 'Landscape Ecology: the Effect of Pattern on Process'. *Annual Review of Ecology and Systematics* 20 (1), pp. 171–197. DOI: 10.1146/annurev.es.20.110189.001131.

Zhang, T and G Lin (2006). 'A Supplemental Indicator of High-Value or Low-Value Spatial Clustering'. *Geographical Analysis* 38 (2), pp. 209–225. DOI: 10.1111/j.0016-7363.2006.00683.x.

— (2007). 'A Decomposition of Moran's I for Clustering Detection'. *Computational Statistics and Data Analysis* 51 (12), pp. 6123–6137. DOI: 10.1016/j.csda.2006.12.032.

Zipf, G (1949). *Human Behaviour and the Principle of Least Effort*. Cambridge, MA: Addison-Wesley.

II.1.7 Appendix

II.1.7.1 A1. Derivation of the Empirical Expectation of $\mathrm{GS_i}^*$

$$\begin{aligned}\hat{E}[GS_i^*] &= \frac{\sum_j^n \sum_{k\neq j}^{j-1} \omega_{ij}\, \omega_{ik}\, \phi_{jk}\, \hat{E}[f]}{\sum_j^n \sum_k^n \sum_{m\neq k}^{k-1} \omega_{jk}\, \omega_{jm}\, \phi_{km}\, f_{km}} \\ &= \frac{\hat{E}[f]\cdot \sum_j^n \sum_{k\neq j}^{j-1} \omega_{ij}\, \omega_{ik}\, \phi_{jk}}{\sum_j^n \sum_k^n \sum_{m\neq k}^{k-1} \omega_{jk}\, \omega_{jm}\, \phi_{km}\, f_{km}} \\ &= \frac{\hat{E}[f]\dot{W}_i}{A}\end{aligned} \tag{II.1.11}$$

Since $\hat{E}[f]$ and A are constant, we can infer that the expectation is proportional to the share of the neighbourhood's size among the overall sum of relationship outcomes:

$$\hat{E}[GS_i^*] \sim \frac{W_i}{W}$$

II.1.7.2 A2. Derivation of the Expectation of the Squared GS Statistic

$$GS_i^{*2} = \frac{\sum_j^n \sum_{k\neq j}^{j-1} \sum_m^n \sum_{p\neq m}^{m-1} \omega_{ij}\, \omega_{ik}\, \omega_{im}\, \omega_{ip}\, \phi_{jk}\, \phi_{mp}\, f_{jk}\, f_{mp}}{\sum_j^n \sum_k^n \sum_{m\neq k}^{k-1} \sum_p^n \sum_q^n \sum_{s\neq q}^{q-1} \omega_{jk}\, \omega_{jm}\, \omega_{pq}\, \omega_{ps}\, \phi_{km}\, \phi_{qs}\, f_{km}\, f_{qs}} \tag{II.1.12}$$

$$\begin{aligned}\hat{E}[f1, f2] &= \frac{\sum_j^n \sum_{k\neq j}^{j-1} \sum_m^n \sum_{p\neq m}^{m-1} \phi_{jk}\, f_{jk}\, \phi_{mp}\, f_{mp} - \sum_j^n \sum_{k\neq j}^{j-1} (\phi_{jk}\, f_{jk})^2}{\Phi(\Phi-1)} \\ &= \frac{\Gamma^2 - \sum_j^n \sum_{k\neq j}^{j-1} (\phi_{jk}\, f_{jk})^2}{\Phi(\Phi-1)}\end{aligned} \tag{II.1.13}$$

$$\hat{E}[f^2] = \frac{\sum_j^n \sum_{k\neq j}^{j-1} \phi_{jk}\, f_{jk}^2}{\Phi} \tag{II.1.14}$$

Solving Equation II.1.12 leads to quadratic and non-quadratic terms. Thus, we need $\hat{E}[f^2]$ and $\hat{E}[f_1, f_2]$ for inferring the expectation of the squared GS statistic. Both of these values are constant. Therefore, we can extract them from the sums. Furthermore, ω and ϕ are binary and $\omega_{ij}^2 = \omega_{ij}$, $\phi_{jk}^2 = \phi_{jk}$. Accordingly we can write:

$$\begin{aligned}\hat{E}[GS_i^{*2}] &= \frac{W_i \cdot \hat{E}[f^2] + W_i(W_i-1)\cdot \hat{E}[f_1, f_2]}{A^2} \\ &= \frac{\frac{W_i \sum_j^n \sum_{k\neq j}^{j-1} \phi_{jk}\, f_{jk}^2}{\Phi} + \frac{W_i(W_i-1)(\Gamma^2 - \sum_j^n \sum_{k\neq j}^{j-1} (\phi_{jk}\, f_{jk})^2)}{\Phi(\Phi-1)}}{A^2}\end{aligned} \tag{II.1.15}$$

II.1.7.3 A3. Derivation of the Empirical Variance of the Local $\mathrm{GS_i}^*$ Statistic

Applying the Steiner translation theorem leads to the variance of the statistic:

$$\begin{aligned}\hat{Var}[GS_i^*] &= \hat{E}[GS_i^{*2}] - (\hat{E}[GS_i^*])^2 \\ &= \frac{\frac{W_i \sum_j^n \sum_{k\neq j}^{j-1} \phi_{jk}\, f_{jk}^2}{\Phi} + \frac{(W_i^2 - W_i)(\Gamma^2 - \sum_j^n \sum_{k\neq j}^{j-1} (\phi_{jk}\, f_{jk})^2)}{\Phi(\Phi-1)} - \frac{W_i^2(\sum_j^n \sum_{k\neq j}^{j-1} \phi_{jk}\, f_{jk})^2}{\Phi^2}}{A^2}\end{aligned} \tag{II.1.16}$$

II.1.7.4 A4. Derivation of the Maximum of the GS$_{\mathrm{i}}$* Statistic

$$\begin{aligned}\max GS_i^* &= \frac{\sum_j^n \sum_{k=j+1}^n \omega_{ij}\,\omega_{ik}\,\phi_{jk}\,f_{jk}}{n\sum_j^n \sum_{k=j+1}^n \omega_{ij}\,\omega_{ik}\,\phi_{jk}\,f_{jk}} \\ &= \frac{\sum_j^n \sum_{k=j+1}^n \omega_{ij}\,\omega_{ik}\,f_{jk}}{n\sum_j^n \sum_{k=j+1}^n \omega_{ij}\,\omega_{ik}\,f_{jk}} = \frac{1}{n}\end{aligned} \tag{II.1.17}$$

II.1.7.5 A5. Derivation of the Standardised GS$_{\mathrm{i}}$* Statistic

$$\begin{aligned}Z_{GS_i^*} &= \frac{GS_i^* - \hat{E}[GS_i^*]}{\sqrt{\hat{Var}[GS_i^*]}} \\ &= \frac{\dfrac{\sum_j^n \sum_{k\neq j}^{j-1} \omega_{ij}\,\omega_{ik}\,\phi_{jk}\,f_{jk} - \frac{W_i \sum_j^n \sum_{k\neq j}^{j-1} \phi_{jk}\,f_{jk}}{\Phi}}{A}}{\sqrt{\dfrac{\frac{W_i \sum_j^n \sum_{k\neq j}^{j-1} \phi_{jk}\,f_{jk}^2}{\Phi} + \frac{(W_i^2 - W_i)(\Gamma^2 - \sum_j^n \sum_{k\neq j}^{j-1}(\phi_{jk}\,f_{jk})^2)}{\Phi(\Phi-1)} - \frac{W_i^2(\sum_j^n \sum_{k\neq j}^{j-1} \phi_{jk}\,f_{jk})^2}{\Phi^2}}{A^2}}} \\ &= \frac{\dfrac{\sum_j^n \sum_{k\neq j}^{j-1} \omega_{ij}\,\omega_{ik}\,\phi_{jk}\,f_{jk} - \frac{W_i \sum_j^n \sum_{k\neq j}^{j-1} \phi_{jk}\,f_{jk}}{\Phi}}{A}}{\dfrac{\sqrt{\frac{W_i \sum_j^n \sum_{k\neq j}^{j-1} \phi_{jk}\,f_{jk}^2}{\Phi} + \frac{(W_i^2 - W_i)(\Gamma^2 - \sum_j^n \sum_{k\neq j}^{j-1}(\phi_{jk}\,f_{jk})^2)}{\Phi(\Phi-1)} - \frac{W_i^2(\sum_j^n \sum_{k\neq j}^{j-1} \phi_{jk}\,f_{jk})^2}{\Phi^2}}}{A}} \\ &= \frac{\sum_j^n \sum_{k\neq j}^{j-1} \omega_{ij}\,\omega_{ik}\,\phi_{jk}\,f_{jk} - \frac{W_i \sum_j^n \sum_{k\neq j}^{j-1} \phi_{jk}\,f_{jk}}{\Phi}}{\sqrt{\frac{W_i \sum_j^n \sum_{k\neq j}^{j-1} \phi_{jk}\,f_{jk}^2}{\Phi} + \frac{(W_i^2 - W_i)(\Gamma^2 - \sum_j^n \sum_{k\neq j}^{j-1}(\phi_{jk}\,f_{jk})^2)}{\Phi(\Phi-1)} - \frac{W_i^2(\sum_j^n \sum_{k\neq j}^{j-1} \phi_{jk}\,f_{jk})^2}{\Phi^2}}}\end{aligned} \tag{II.1.18}$$

II.2 Abundant Topological Outliers in Social Media Data and Their Effect on Spatial Analysis

Abstract

Twitter and related social media feeds have become valuable data sources to many fields of research. Numerous researchers have thereby used social media posts for spatial analysis, since many of them contain explicit geographic locations. However, despite its widespread use within applied research, a thorough understanding of the underlying spatial characteristics of these data is still lacking. In this paper, we investigate how topological outliers influence the outcomes of spatial analyses of social media data. These outliers appear when different users contribute heterogeneous information about different phenomena simultaneously from similar locations. As a consequence, various messages representing different spatial phenomena are captured closely to each other, and are at risk to be falsely related in a spatial analysis. Our results reveal indications for corresponding spurious effects when analysing Twitter data. Further, we show how the outliers distort the range of outcomes of spatial analysis methods. This has significant influence on the power of spatial inferential techniques, and, more generally, on the validity and interpretability of spatial analysis results. We further investigate how the issues caused by topological outliers are composed in detail. We unveil that multiple disturbing effects are acting simultaneously and that these are related to the geographic scales of the involved overlapping patterns. Our results show that at some scale configurations, the disturbances added through overlap are more severe than at others. Further, their behaviour turns into a volatile and almost chaotic fluctuation when the scales of the involved patterns become too different. Overall, our results highlight the critical importance of thoroughly considering the specific characteristics of social media data when analysing them spatially.

Keywords: Spatial Analysis, Spatial Autocorrelation, Eigenvalues, Social Media, Twitter

II.2.1 Introduction

One aspect in the analysis of social phenomena is the search for spatial structures and patterns. The aim thereby is to explain the organization of complex spaces such as urban areas (Cranshaw et al. 2012; Lee et al. 2013) as well as social behaviour patterns (Newsome et al. 1998; Rai et al. 2007). Twitter and related social media feeds have recently become promising data sources in this regard. These online social networks capture a vast amount of georeferenced data from the everyday life of users, and are thus expected to represent a fraction of social happenings in geographic space. However, user-generated datasets have some unique shortcomings, such as their potential lack of trustworthiness, missing representativeness with respect to demographics, and self-selection bias (Gayo-Avello 2012). The authors of (Sengstock and Gertz 2012) further describe this from a technical perspective by highlighting potential spatial, temporal and semantic inaccuracies. Nevertheless, Twitter provides a high temporal and spatial resolution, offering a unique opportunity to gain novel insights into the spatiotemporal behaviour of humans.

A large body of literature dealing with social media analysis from geographic and social sciences has evolved throughout the last years. Examples span across a broad variety of fields such as the investigation of human mobility (Hawelka et al. 2014; Lenormand et al. 2014), natural hazards and disaster management (Crooks et al. 2013; Albuquerque et al. 2015), and geodemographics (Mislove et al. 2011; Longley et al. 2015). These research efforts are summarized by (Steiger et al. 2015a) through providing a systematic literature review emphasizing spatial analyses of social media feeds. These authors note that one important but prevailing shortcoming is the naïve application of existing spatial methods when conducting social media analysis. Similar critiques including a lack of theory have recently also been raised elsewhere (Rae and Singleton 2015), though from a less technical perspective. Most established spatial methods were designed for datasets with different characteristics, *i. e.*, data generated in some well-defined acquisition processes. It is therefore questionable whether existing methodological approaches produce reliable results. Although a majority of applied and empirical research on social media has been carried out, the scientific community is still lacking a thorough understanding of the interplay between applied spatial analysis methods and the specific characteristics that come with social media data.

One of the main differences between social media feeds and more traditional datasets is the data collection process, which appears highly unstructured. Mutually independent social media users contribute information about numerous real-world as well as fictional phenomena. To further stress the heterogeneity argument, issues arise even within representations of single phenomena. Due to varying spatial cognition and perception skills, user-generated data face the problem of user-induced heterogeneity (*cf.* Hegarty et al. 2006; Iosa et al. 2012). Different phenomenon representations thus occur simultaneously, and their geometric overlap leads to a disrupted topology, whereby we refer to topology as the spatial arrangement of tweets. The result is a number of topological outliers that would not occur when only one phenomenon would be reflected in a clear manner in the data. Intuitively, the data acquisition process of social media thus causes spatial analysis methods to combine different actually unrelated tweets. Established density-based clustering techniques like DBSCAN (Ester et al. 1996), for instance, include tweets that represent different underlying phenomena. The result then is an averaged density being too high for some and too low for other reflected phenomena. Similarly, covariance-based techniques incorporating attribute values like Kriging (Oliver 2010) infer their spatial relationships from misleading tweet comparisons when incorporating different phenomenon representations. In all these cases, the analysis results might lead to wrong conclusions. Clearly, such techniques are designed for mono-categorical and spatially exclusive datasets. The following Section ‘A motivating example’ provides an example from a London twitter dataset indicating the abovementioned problem statement.

In this paper we investigate how topological outliers caused by the abovementioned heterogeneities influence spatial analysis methodology in a general sense. The problem outlined above is not restricted to any specific method, but prevalent across a range of spatial analysis techniques when applied on highly uncertain user-generated datasets. Therefore, instead of studying any specific spatial method, we rather investigate the underlying characteristic called spatial autocorrelation. This second-order data characteristic drives spatial patterning and is the conceptual basis for spatial analysis (see Fischer and Getis (2010b)). Analysing spatial autocorrelation hence guarantees a high degree of generalizability of our results beyond the specificities of any particular technique. Table II.2.1 lists all investigations that we conduct within this paper, including associated methodology. These tasks cover a broad range of issues around topological outliers and the way how these influence spatial analyses.

The remainder of this paper starts out with further motivating our research (Section ’A motivating example’) and putting it into context (’Spatial analysis and spatial heterogeneity’). We then outline

Table II.2.1: Overview of the investigations conducted in this paper.

Scientific objectives		Methods
1)	• Determination of the interplay between tweets and spatial analysis methodology. • Illustration of unexpected behaviour when spatially analysing tweets.	Semivariogram, autocovariance
2)	• Calculation and mapping of increased topological heterogeneity caused by pattern overlap. • Demonstration of an additionally induced topological outlier region, which controls spatial patterning.	Eigenvalue analysis of local spatial weight matrices
3)	• Influence of topological heterogeneity on the distribution of Moran's $\mathcal{I}$ (a measure of spatial autocorrelation). • Determining consequences for drawing inference about spatial patterns.	Eigenvalue analysis of a global spatial weight matrix, violin plot
4)	• Discovery of effects of topological outliers on spatial pattern quantification. • Identification of disturbing spatial components induced by increased topological heterogeneity.	Moran's $\mathcal{I}$, Moran scatterplots
5)	• Determination of the role of scale differences between overlapping patterns on spatial analysis. • Detection of regularity and chaotic behaviour within the disturbing components from row 4.	Serial correlation, correlograms

some problematic covariation-based characteristics that emerge when analysing Twitter messages with established spatial analysis methods ('Indications from the Twitter dataset'). Afterwards, we investigate these characteristics within a simulated dataset, the latter allowing us to control different parameters such as spatial scale and attributes. We analyse how geometric overlap influences the power characteristics of conclusions drawn by spatial methods and how the topological arrangement pre-determines the range of expected results ('Increased topological variability'). Afterwards, we identify interfering components that lead to spurious and misleading analysis results (Section 'Influences on spatial autocorrelation'). This includes an analysis of their interdependencies with scale-differences among overlapping patterns (Section 'The roles of scale differences and the degree of overlap'). The article closes with a discussion of the achieved results and concluding remarks, the latter including future research prospects and practical hints for scholars employing georeferenced social media data.

II.2.1.1 A Motivating Example

Figure II.2.1 provides an example by showing geotagged tweets that occurred at the 'Trades Union Congress House', an umbrella organization of British labour unions headquartered in London. The tweets and their attributes are drawn from another study that we conducted earlier (see Section 'Datasets' for an

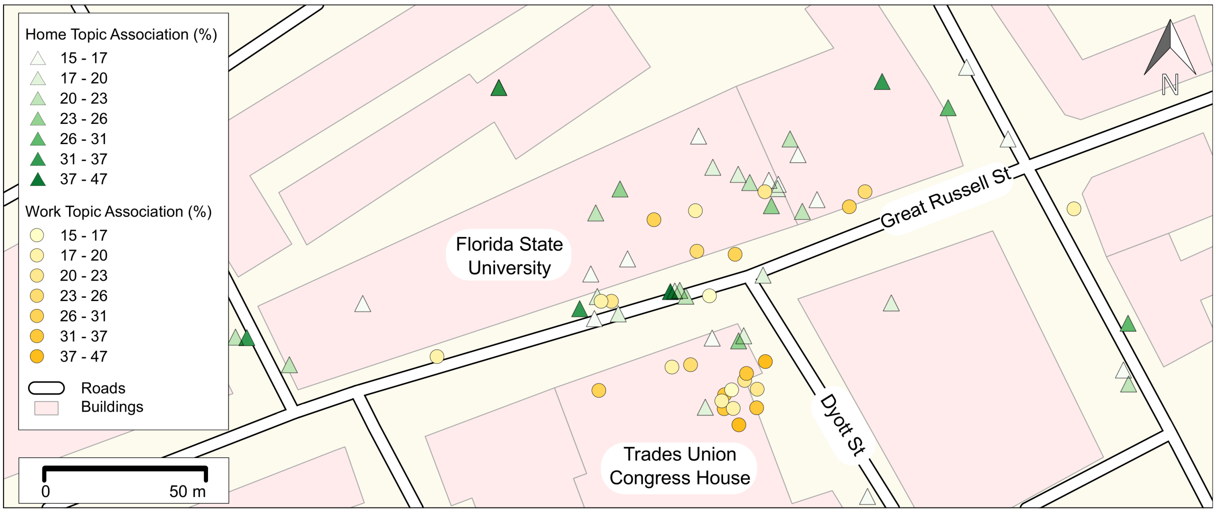


Figure II.2.1: Map showing overlapping tweets in central London. The yellowish tweets represent a semantic "work" topic described in the following section. The greenish tweets, in contrast, were assigned a "home" topic (*cf.* Steiger et al. (2015b) for details on these topics). The background map is based on OpenStreetMap data.

explanation, (Steiger et al. 2015b)). We analysed the spatial pattern of work-related tweets by comparing them against the census workday population. For that purpose we carefully extracted latent topics. The colours in Fig 1 represent a semantic topic: either "home" (green) or "work" topics (yellow). Considering the work-related tweets, note the spatially overlapping variability within the yellow colour-code. We can see that the spatial scales (*i. e.*, the point-spacing) as well as the intensities of the topic assignments (*i. e.*, the attribute values) fluctuate within small areas. This is an indicator for different phenomena or processes being reflected within tweets. It is likely that staff, as well as visitors and the general public, report about different work-related topics in this given area. Besides, the green home topic spatially coincides with the work topic in the northern parts of the observed region. This area is close to a university campus. Intuitively, students can be expected to tweet from this location. Some of their tweets deal with topics being classified as work-related phenomena (*e. g.*, study-related topics), while some others are instead related to leisure activity (home-topic). This shows that both phenomena (home and work) as well as sub-processes of these (within-colour variations) appear in a spatially overlapping manner. Thus, it can be concluded that social media datasets are of multi-categorical nature and spatially intertwined. This indicates an abundance of topological outliers as, unlike with non-overlapping patterns, their topology is highly diverse. They possess geometric characteristics from at least two different processes. Further, the overlap itself creates additional geometric characteristics. These outliers can be expected to influence the outcomes of spatial analyses and are the starting point of this paper.

II.2.1.2 Spatial Analysis and Spatial Heterogeneity

We should first briefly articulate our problem statement in terms of traditional concepts of the field of spatial analysis. The overlap of phenomena, to which we are referring, manifests itself as a specific type of spatial heterogeneity. Spatial heterogeneity traditionally refers to a variable's response to extrinsic spatially varying environmental or socio-economic conditions. This typically leads to varying intensities, which in turn designates spatial heterogeneity as a first-order effect (in contrast to the second-order effect of spatial dependence) (Sui 2004). Common forms of spatial heterogeneity are 'spatial regimes' (patchy areas of varying intensity, abrupt changes) and 'trends' (smooth transitions between means) (Legendre 1993). Regimes are common in urban areas and typically resemble the local-scale variability of such regions (Páez and Scott 2005), while trends are more important to the physical sciences (Atkinson 2001). Dealing with these kinds of spatial heterogeneity is a widely discussed topic in spatial research. It is methodologically reflected by a range of methods such as local measures of spatial dependencies (Anselin 1995; Getis and Ord 1992; Ord and Getis 1995), separate treatments of different regimes (Ord and Getis 2001; Rogerson and Kedron 2012; Rogerson 2015), approaches for determining the local scales of patches (Aldstadt and Getis 2006; Getis and Aldstadt 2004) and local regression models like 'geographically-weighted regression' (Brunsdon et al. 1996; Fotheringham et al. 2002), 'spatial expansion' (Casetti 1972; Casetti 1997) and a localised version of 'spatial eigenvector filtering' (Griffith 2008).

All approaches mentioned above assume that spatially exclusive forms of heterogeneity are observed. That is, they refer to one of the traditional types of spatial data: geostatistical data (spatially continuous phenomena), lattice data (spatially discrete phenomena) or event data (stochastic geometries) (Cressie 1993). Event data incorporates superposition of phenomena to a certain degree, but, however, falls back to a lattice when analysing attributes (different types of the 'mark correlation function', *cf.* Shimatani (2002)). These spatial data types are reasonable with many kinds of spatial data such as census or housing data. The outlined Twitter example from the previous section, however, shows that social media data typically violate the assumption of spatial exclusiveness and cannot be straightforwardly assigned to one of the data types mentioned above. That is, with respect to spatial heterogeneity, social media data show a novel kind of that characteristic. Spatial heterogeneity here is caused by the unstructured data acquisition process (*i. e.*, an extrinsic source) and is characterized by the superposition of phenomena. It expresses itself by the formation of specific (artificial) regimes within the 'zones of overlap'. These zones appear where different phenomena or processes coincide within data and show abnormal behaviour with respect to statistical and topological characteristics. Just as with traditional forms of spatial heterogeneity, this effect is likely to influence the outcomes of spatial analysis, eventually leading to spurious results. These zones of overlap are what we target by our research.

II.2.2 Materials and Methods

II.2.2.1 Ethics Statement

Some of the data used within this study was crawled from the microblogging service Twitter. We have eliminated all references to actual Twitter users. Therefore, the dataset is anonymised and does not violate the privacy of actual persons.

II.2.2.2 Datasets

We use two different datasets for our analyses. One of them is a Twitter dataset consisting of georeferenced tweets. It has been crawled through the publicly available Streaming API during a period of approximately one year. The sample used here is an excerpt of a much larger dataset consisting of 20 Million tweets covering Greater London, which was used in one of our previous studies (Steiger et al. 2015b). We only leveraged explicit coordinates offered in the form of latitude-longitude tuples. This may include GPS-derived locations as well positions determined by WiFi-positioning techniques and check-ins (see Section 'Indications from the Twitter dataset' for further discussion of this point). That is, we did not include location tags like “London, UK”. The latter would blur up the analysis scale as these do not refer to points but to much larger polygons instead. Our pre-processing includes several natural language processing steps such as tokenisation, stop word removal and stemming. Through these steps we remove a great deal of potentially unnecessary noise that might disturb the analysis if not being eliminated. What we did not remove is artefacts such as tweets contributed by bots. Removing these is still an issue of ongoing research (*e. g.*, (Cresci et al. 2015; Gilani et al. 2016; Mukherjee et al. 2016)). Further, we do believe that, through our semantic treatment (see below), a lot of these artificial tweets have been removed implicitly. The semantic modelling was done by means of Latent Dirichlet Allocation (LDA) (Blei et al. 2003), a probabilistic bag-of-words model for extracting latent topics from text corpuses. Please refer to the paper mentioned above for a more detailed explanation of all our conducted pre-processing as well as semantic processing steps. After preprocessing and narrowing down the scope to one latent topic (“work”), approximately 23,000 tweets remain. The attribute used here is a percentage expressing the degree of tweet-topic association. The chosen topic “work” represents a range of business activities and personal reports about individual daily commute, day-to-day work experiences and similar phenomena.

The second dataset used in this paper is a simulated point pattern. It resembles an overlap of two different spatial processes reflected within social media. Attributes attached to the points were drawn from Gaussians. In an initial configuration, these centre on different levels of intensity ($\mu_1 = 250, \mu_2 = 750$) while possessing a similar variance ($\sigma_1^2 = \sigma_2^2 = 22,500$). Each of the involved sub-patterns shows spatial autocorrelation of 0.81 (Moran’s $\mathcal{I}$, IDW-based spatial weights). They do therefore mimic positive spatial autocorrelation and first-order spatial heterogeneity as it is described in Section ’Spatial Analysis and Spatial Heterogeneity’. Both involved sub-patterns operate at different spatial scales, whereby scale is defined in terms of point spacing within this study. The smaller-scale process operates at an interval of [40 m, 50 m], whereas the larger-scale process interacts at distances on [70 m, 80 m]. The geometries of the patterns were generated by a random walk approach. An initial point was placed arbitrarily. Then, starting from that point, 500 points were successively placed by choosing a random angle and distance at each step, both of which are following a uniform distribution constrained by the abovementioned distance intervals. In total, 1,000 points were placed. Now, by overlaying these two patterns, we simulate an overlap as observed within the motivating example above. The degree of overlap has been chosen such that 23.8 % of the points from the large-scale pattern interact with at least one point from the small-scale counterpart. This idealised dataset allows us to vary the scales of the involved sub-patterns in an archetypical way and allows controlling the attached attribute values. It thus allows isolating and investigating different topological effects of sampling-induced spatial heterogeneity on outcomes of spatial analysis while avoiding any cultural, socio-demographic, topographic and other kinds of extrinsic influences (*e. g.*, the bots mentioned earlier) that might bias the analysis outcomes. This guarantees a high level of generalisability of the achieved results. Fig 2 provides an overview of both datasets.

II.2.2.3 Methods

Heat Map of Autocovariance Terms and Variographic Analysis

In a first step, we highlight the problem statement by means of the Twitter dataset. To achieve that we apply two different statistical measures to our Twitter data: sample autocovariance and semivariogram estimation. Sample autocovariance describes the degree of conformity over the mean across the realized

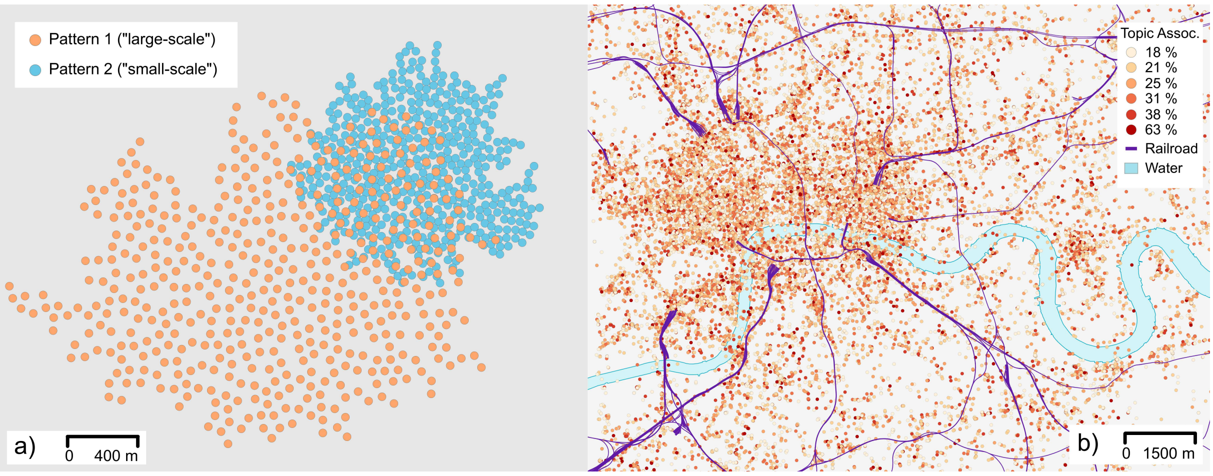


Figure II.2.2: Overview of the two employed datasets. a) Simulated pattern, colors indicate the two primal sub-pattern. b) Twitter data from London. The background map of b) is based on OpenStreetMap data.

tuples of topic associations. In its spatially unweighted form, the pairwise autocovariance matrix is defined as

$$S_{XX} = \frac{1}{n} \cdot \begin{bmatrix} (x_1 - \bar{x})^2 & \dots & (x_1 - \bar{x})(x_n - \bar{x}) \\ \vdots & \ddots & \vdots \\ (x_n - \bar{x})(x_1 - \bar{x}) & \dots & (x_n - \bar{x})^2 \end{bmatrix} \tag{II.2.1}$$

where x_i and x_j, in our case, denote two topic associations indexed over tweets i and j, and $\bar{x}$ is their corresponding mean. We investigate the off-diagonal elements from Equation II.2.1 by relating them to their geographic distances measured between i and j. The result is a heat map of autocovariance mapped against distance. This heat map allows disaggregating the overall autocovariance into its constituting parts. The benefit of this approach is that, other than with a covariogram or a correlogram, we are neither aggregating by distance bands nor by random variables. We thus get a detailed picture of all available pairs of observations within their geographic context. Therefore, these pairwise comparisons reveal local information through geographic space.

We complement the abovementioned local viewpoint by a global summarization of spatial relations. This is done through constructing an empirical semivariogram. Let $\mathcal{P}_i \in \mathbb{R}$ be geometric points (*i. e.*, tweets) over which the topic associations x_i are spatially indexed. The empirical semivariogram is then estimated by (see Bachmaier and Backes (2008))

$$\gamma(h) = \frac{1}{2N(h)} \cdot \sum_{i,j}^{N(h)} (x_i - x_j)^2, \ \ \forall (x_i, x_j) : \|\mathcal{P}_i - \mathcal{P}_j\| \in (h_{\min}, h_{\max}) \tag{II.2.2}$$

where $h_{\min}$ and $h_{\max}$ span non-overlapping distance classes h, $N(h)$ describes the numbers of pairs of points falling into these classes and $\|\cdot\|$ denotes the Euclidean distance measure. A semivariogram thus describes the variance within distance classes h. Our employed distance classes have a width of 25 m. This ensures a fine granularity and acknowledges the large numbers of tweets in dense packing.

Both these measures, autocovariance and semivariogram, are helpful devices for demonstrating the problem statement mentioned in the introduction. We use them to reinforce the issues underlying our research and to show indications for spatial overlaps within Twitter. While the semivariogram comes up with a well-understood interpretation allowing to demonstrate the unexpected behaviour caused by overlaps, the heat map allows for explaining this behaviour in greater detail by uncovering the types of interactions across space.

Moran's $\mathcal{I}$ and Moran Scatterplot

We are interested in analysing general behaviour beyond any specific spatial methods. The universal force underlying spatial methods is called spatial autocorrelation, which quantifies how strongly observations relate with each other in space and how this drives patterns (Fischer and Getis 2010b). Roughly put, spatial autocorrelation refers to "the coincidence of value similarity with locational similarity" (Anselin and Bera 1998, p. 241). Our simulation experiments are therefore based on Moran's $\mathcal{I}$, the quasi-standard measure of spatial autocorrelation. Moran's $\mathcal{I}$ can be roughly characterised as a spatialised version of Pearson r, restricted to observations of a single random variable. Its equation is given by (Cliff and Ord 1973)

$$\mathcal{I} = \frac{n}{\sum_{i,j}^{n} w_{ij}} \cdot \frac{\sum_{i,j}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i}^{n}(x_i - \bar{x})^2} \tag{II.2.3}$$

where n denotes the overall number of observations. The factors w_{ij} denote elements of a spatial weight matrix. This matrix captures the geographic layout of the study area and defines neighbourhood relations. It describes how much resistance the geographic topology bears upon the covariation within the associated attribute. We use an inverse-distance notion for our investigations, because our simulated data was created by underlying distance theory (geometric interaction ranges, see previous section). Clearly, in any empirical studies, the choice of weights is a crucial one and should be undertaken with care and expert knowledge of the underlying phenomenon. Other common weight choices are summarized by (Getis 2009). Investigating how topological social media characteristics influence Moran's $\mathcal{I}$ will allow us to make more general statements about its influence on spatial methods in a broader sense.

The Moran scatterplot is a graphical device complementing Moran's $\mathcal{I}$. It was introduced by Luc Anselin (Anselin 1996) and provides a means to disaggregate spatial autocorrelation into its distinctive parts. Thereby, the regression line through this scatterplot is coincidental with the non-normalized Moran's $\mathcal{I}$ measure. Note that normalising over spatial weights is not necessary here, since we do not vary the spatial layout during our study. Thus, analysing the regression line in the Moran scatterplot is tantamount to analysing Moran's $\mathcal{I}$. For that reason, and because the graphical interpretation allows determining sub-components of spatial autocorrelation in greater detail, we use the Moran scatterplot for analysing systematic disturbances to the spatial pattern caused by topological outliers.

Local Eigenvalues

Analysing the topological configuration requires a measure of overlap and topological heterogeneity. For this purpose, we divide the spatial weight matrix from Moran's $\mathcal{I}$ into local submatrices and calculate their principal eigenvalues. The principal eigenvalues of these localised submatrices represent the local interaction potential attached to each single observation. The higher the eigenvalue, the higher is the variability within the geographic connectivity and thus the contribution of a single spatial unit to the entire region. In turn, high variability within pairwise connectivity relations means that a homogeneous pattern is disrupted by an overlap with another, eventually differently scaled, pattern. The eigenvalues thus summarize the overall degree of overlap of observations within their local geographic context as well as the strength of their influence contributed to spatial pattern assessments. We can calculate the spectra of local eigenvalues for the local matrices as (Tiefelsdorf et al. 1999)

$$\left\{-\sqrt{\sum_{j=1}^{n} w_{ij}^2}, \ldots, 0, \ldots, \sqrt{\sum_{j=1}^{n} w_{ij}^2}\right\} \tag{II.2.4}$$

Only the principal eigenvalues are non-zero. The importance of these eigenvalues for our investigations is that we can use them for summarising local topological effects. They hence allow us to measure the intensity of geometric overlap of sub-patterns because an overlap is expected to produce outlier eigenvalues. We use this measure for analysing the topological influences of overlap on the outcomes of spatial analyses.

Similarly, we also calculate the eigenvalues of the overall global spatial weight matrix. This matrix comes up with n non-zero eigenvalues. Again, the principal eigenvalues are of importance, because they determine the feasible range and the shape of the distribution of Moran's $\mathcal{I}$ values (Tiefelsdorf and Boots 1997; Tiefelsdorf et al. 1999). The eigenvalues do hence predetermine the efficiency and power of Moran's $\mathcal{I}$ as a test statistic. This reinforces how crucial topological outliers are towards spatial pattern assessment and demonstrates why we use them as a useful proxy for topological heterogeneity.

Serial Correlation and Correlogram

The eigenvalues explained above capture the geometric and topological influences that single spatial units exert onto the entire region. Combining these with Moran's $\mathcal{I}$ allows analysing how these (and especially topological outliers) affect the detection of pattern within attributes. This involves investigating how different parts of an overlapping pattern behave with respect to increasing scale differences of the involved overlapping pattern. We are interested in the coherence of these effects. Only when the behaviour is somehow tractable, analysts can try to deal these issues. Chaotic behaviour, in turn, would be hardly treatable. Thus, we estimate the serial correlation of the slope of disturbing components from the Moran scatterplot by means of the one-dimensional sample autocorrelation coefficient:

$$r(\tau) = \frac{\sum_{i}^{n-\tau}(x_i - \bar{x})(x_{i+\tau} - \bar{x})}{\sum_{i}^{n}(x_i - \bar{x})^2} \tag{II.2.5}$$

where τ is the lag (here: the lag of scale differences in meters). We plot these estimates against the lag, which is then called a correlogram. This allows investigating the behaviour of the overlap of patterns through scale differences.

II.2.3 Results

II.2.3.1 Indications from the Twitter Dataset

Before conducting simulation experiments, we should turn our attention toward the Twitter dataset to highlight indications for geometric overlap of different phenomena. Figure II.2.3 visualises two kinds of information: a heat map of all pair-wise entries from the covariance matrix plotted against geographic distance (underlying colour-coded bins) and a semivariogram (dashed line atop).

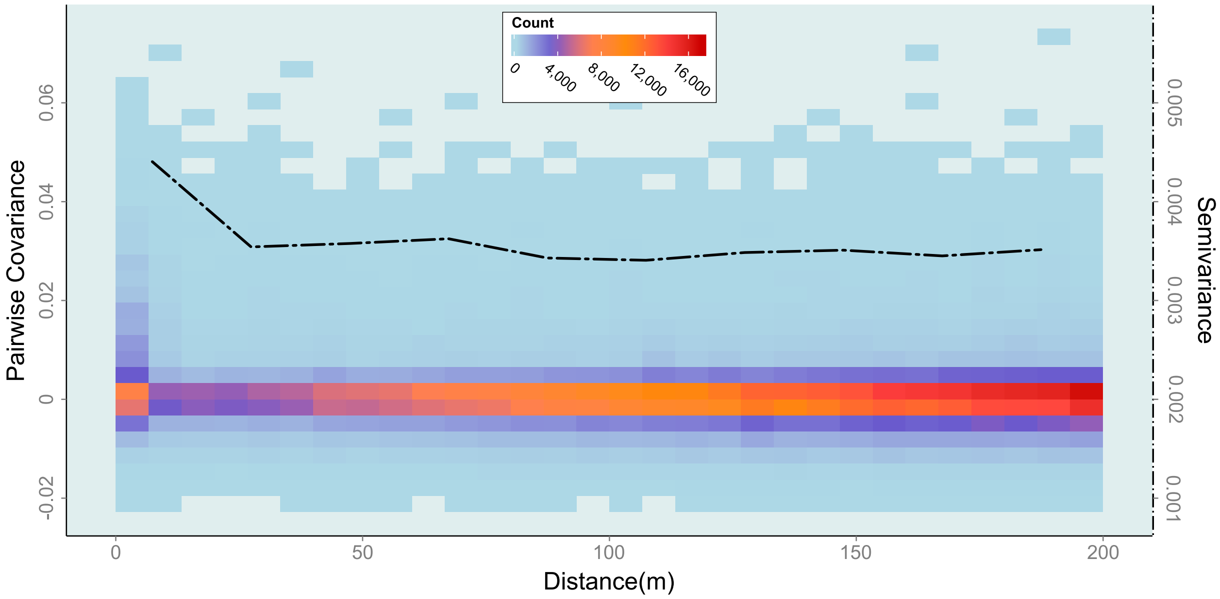


Figure II.2.3: Heat map of pairwise covariance terms and semivariogram of topic associations. The dashed semivariogram refers to the right-hand y-axis (same line-style). The left-hand y-axis refers to the colour-coded bins. Figure bases on the entire Twitter dataset from London, see Section ’Datasets’.

Observe the unusual course of the semivariogram. A typical semivariogram for mono-categorical datasets is of inclining nature when spatial effects are present. The attribute values are typically expected to be most similar in close geographic proximity. Thus, with increasing distance, the variability increases until the so-called ’sill’ is reached. As of that point (called ’range’) the variability levels off to the overall variance and is no longer assumed to be affected by geographic effects. The semivariogram shown in Figure II.2.3, however, indicates a different behaviour. It starts out at a high level of variation, and then progresses towards a constant level. That means that values are dissimilar when they are close to each other. At a first glance this indicates global negative spatial autocorrelation at short-ranges. The topic associations thus seem to possess some kind of repulsion behaviour at a local scale. This kind of association, however, is rarely observed in real-world datasets [51]. We should thus further examine this distinctive behaviour.

The underlying heat map within Figure II.2.3 offers further insight to the notable behaviour of the semivariogram. A significant accumulation of orange bins is observable at distances close to zero. These indicate a large fraction of mutually unrelated tweets at very local scales. Mutually unrelated tweets, however, do not hint on repulsion. They rather show that a great number of neighboured tweets are not related to each other at all, being neither systematically similar nor dissimilar. Apart from that, we can also find some indications for repulsion. Notice the small peak of blueish bins reaching downwards into the negatives. The latter partially supports the observed hint suggested by the semivariogram: we do indeed observe some opposed tweets in close vicinity. Besides these two findings, another important observation from Figure II.2.3 is the high-reaching peak towards higher positive covariation. This peak indicates tweets that are similar to each other, and thus indicate clustering behaviour caused by systematic spatial phenomena. These latter tweets are the ones we are typically interested in when searching for spatial pattern within social phenomena and processes. They hint on common behaviour and thus (in the present case) semantically coherent spatial regions. It is further important to note that the heat map disaggregates the semivariogram. The semivariogram is based on combining multiple tweets, regardless of their underlying phenomena. This is exactly the problem with spatial methodology which we outlined in the introduction. The heat map, in contrast, reveals the underlying causes that ultimately lead to this issue.

Note that Figure II.2.3 is based on all tweets that are found to be semantically associated with the topic “work”. This includes some spatially coincident tweets, which might be due to the Twitter data collection process (*e. g.*, WiFi positioning techniques, Foursquare check-in data). For some kinds of investigations one might want to keep these duplicates (*e. g.*, when characterising places), but in some other cases one might wish to exclude them previously. Figure II.2.20 in the appendix shows that, however, removing these spatial duplicates does not affect the general argumentation outlined above. It only diminishes the magnitude of spatial effects in short distance ranges. The latter is as expected, because removing spatial duplicates is essentially a modification of extremely local tweets. Other than Figure II.2.3, Figure II.2.20 is designed in a relative fashion for the sake of comparability (the numbers of tweets are changed after the removal process, absolute numbers are thus less effective).

In the remainder of this article we will use the controlled simulated dataset to further investigate the consequences of such overlap on different topological aspects of tweets. This allows isolating and controlling precisely the effects we are looking for.

II.2.3.2 Increased Topological Variability

Whenever a single pattern is observed, all points interact at only one or rather few specific scales. In the best case the observed scales match those of the underlying phenomenon. Be it one or multiple scales, the crucial point is that these reflect the underlying causative phenomenon. In such cases, the variability within the relative spatial arrangement of points is relatively low and homogeneous. Most points interact at similar or at least meaningful distances. When patterns overlap, however, points from different patterns are positioned in close proximity to each other. These different patterns might possess different kinds of point spacing characteristics. Thus, the topological diversity is higher and the number of cross-pattern interactions between actually unrelated points increases. This topological diversity is expressed by increased local eigenvalues of the spatial weight matrix as shown by Figure II.2.4 (bottom). We should thus analyse how overlap of patterns influences these local eigenvalues.

The top row of Figure II.2.4 shows eigenvalues for a single *non-overlapping pattern*. The left-most map thereby provides eigenvalues for the respective plain pattern. The two maps at the right-hand side demonstrate the effect of two different kinds of spatial weights normalisations (C and W-coding). These normalizations are often applied for making different spatial weight configurations comparable among each other (see (Bavaud 2014) for an overview). It is well-known that, with non-overlapping patterns, such normalisations lead to topological outliers (Tiefelsdorf et al. 1999). Indeed, we can see that W-coding emphasises the boundaries of the pattern, while C-coding exaggerates its interior. The corresponding outlier observations show a strongly increased variability. Given that the normalization procedures are researcher-induced artefacts, such geographic layouts allow the corresponding outlier units too much interaction with their neighbours. This disrupts subsequent spatial analyses. The plain pattern, however, appears homogeneous. These results confirm previous research (Tiefelsdorf et al. 1999). Moreover, note the generally low intensity of the eigenvalues in case of the plain pattern. None of the values exceeds 0.1. This shows that only one coherent underlying phenomenon is represented by the pattern as the topological heterogeneity is kept fairly low.

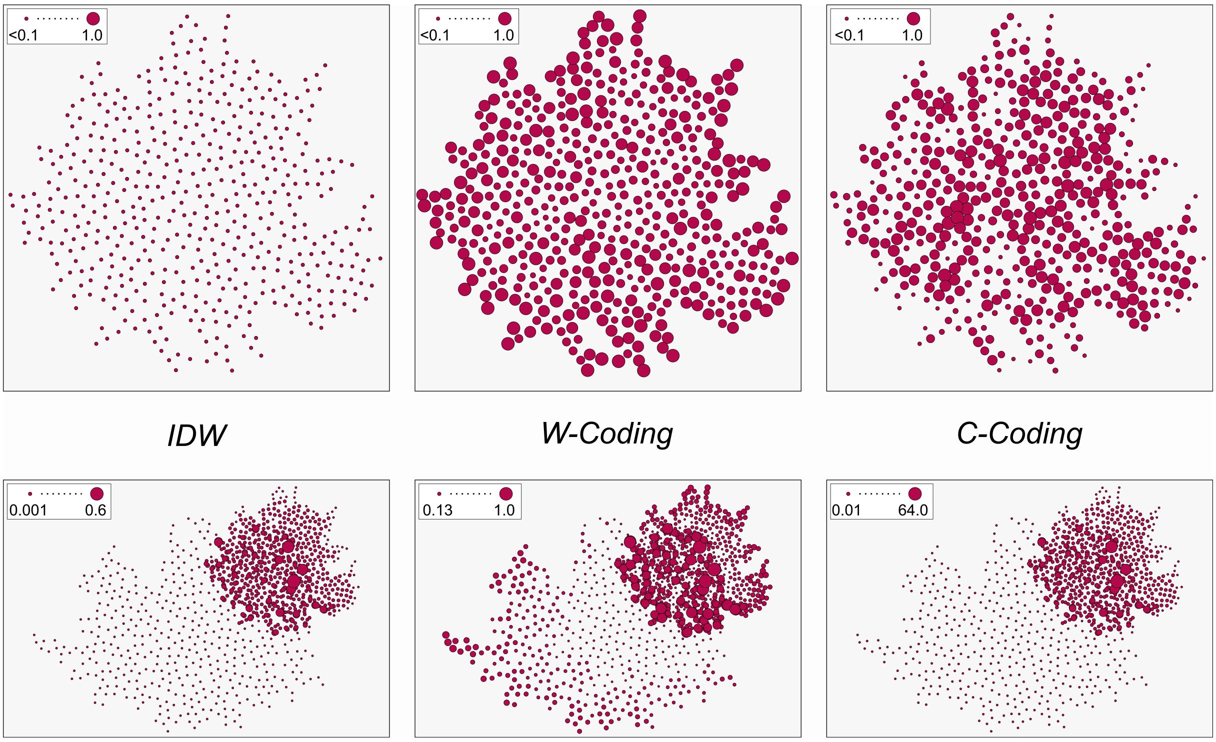


Figure II.2.4: Local eigenvalues of a single pattern (top) and a combined pattern (bottom). Please note the differing value ranges, which are tributes to different distributions of eigenvalues across the maps. Size classification is Jenks natural breaks.

The bottom row of Figure II.2.4 shows eigenvalues for an *overlapping pattern*. Again, the normalised patterns show similar tendencies as their non-overlapping counterparts. However, some differences are noticeable: First, the ranges of the eigenvalues reach up to a way higher intensity, especially with the C-coded pattern where the upper bound reaches up to a value of 64. This demonstrates that geometric overlap does not just produce outliers, but also seems to interact differently with different kinds of normalisation techniques. Second, we can observe that, in contrast to the non-overlapping pattern, even the plain pattern now shows a large number of outliers. The topological variability is thus already increased by the mere fact that different patterns overlap and without having applied any normalisation. The spatial neighbourhoods of such points are composed of different scales simultaneously, making it difficult to adjust any proper analysis scale. The implication of these results is that, when analysing social media data, it is highly likely to observe numerous such outliers. Moreover, overlap may place some of the points in very close proximity at distances shorter than one distance unit. This becomes a severe problem whenever geographic relationships are modelled by means of distance decay functions. Distance decay possesses abnormal behaviour when distances are below one distance unit. Densely covered zones of overlap may thus yield extreme outliers in case of pattern overlap when using distance-based specifications of spatial interactions. This is reflected by Table II.2.2, which provides Moran's $\mathcal{I}$ values for both kinds of patterns from above.

Table II.2.2: Moran's $\mathfrak{I}$ values under different weight specifications. W and C refer to row- and global normalisation.

	IDW	Binary	IDW (W)	IDW (C)	Binary (W)	Binary (C)
Single Pattern	0.81	0.81	0.85	0.81	0.85	0.81
Combined Pattern	-1.07	0.42	0.42	-1.07	0.55	0.42

In order to compare the distance-based weights that were used above toward non-distance weights, we additionally included a binary weighting scheme. Thereby, the upper bounds of the respective point interaction ranges were used as cut-off distances. The attached attribute values are Gaussian as described in Section 'Datasets'. Again, different normalisations were applied *i. e.*, W and C). We see that the variation across different spatial weighting schemes is relatively low for the non-overlapping pattern (top row). As the points used here are placed relatively regular, this is in accordance to results obtained by (Shortridge 2007), who investigated the impact of different weight configurations with regular raster-like patterns.

In contrast, the bottom row outlines results for the combined overlapping pattern. These show marked differences between the weight configurations. Both employed binary schemes behave relatively similar. Contrary, the distance-based weights indicate negative spatial autocorrelation. Recall that both involved patterns are actually positively autocorrelated through space. Thus, the extreme exaggeration of very close but very different points possesses a huge influence on the overall result, ultimately leading to a wrong conclusion about the spatial effects within the pattern. These results show how sensitive the topological outliers react on the type of weights in case of overlapping patterns. Table II.2.2 underpins the importance of a careful choice of weights when analysing geometrically overlapping social media data.

The results from above are also of importance to inferential statistics. Many global spatial statistics like Moran's $\mathfrak{I}$ are defined in an averaging notion. They are defined as a weighted average of local counterparts (in this case: local Moran's $\mathfrak{I}$). This characteristic holds for all statistics and measures of the so-called LISA type (Anselin 1995). With these methods, single outliers control global statistics and influence their distributions. As we have seen above, these outliers are abundant within overlapping patterns. Thus, due to their increased abundance, these cause the probability of extremely high or low degrees of spatial association to increase artificially (Tiefelsdorf and Boots 1997). Whether high or low values are affected thereby depends on the type of outliers observed. The latter point flaws significance procedures and leads to wrong conclusions about spatial effects.

The investigations above are based on *local eigenvalues*, which were in turn calculated from local submatrices. Combining these reveals the overall global spatial weight matrix, and, consequently, the respective *global eigenvalues*. These are of importance for the detection of spatial effects, because they reveal the shape as well as the range of the corresponding reference distribution of analysis outcomes (Jong et al. 1984). In other words: the geographic layout determines the bounds of the strength of detectable effects. Figure II.2.5 visualizes violin plots of these global eigenvalues for the two patterns analysed above. The non-overlapping pattern shows a compact distribution. Half of the values including the median accumulate around the expectation of Moran's $\mathfrak{I}$, which is -0.001 in this particular case. Only few values extend to the extremes. These further span a distinctively narrow overall range. This means that any spatial test statistic which is evaluated on this geographic layout possesses favourable power and efficiency characteristics, as the range of possible outcomes is kept reasonably small. Thus, the measured strength of

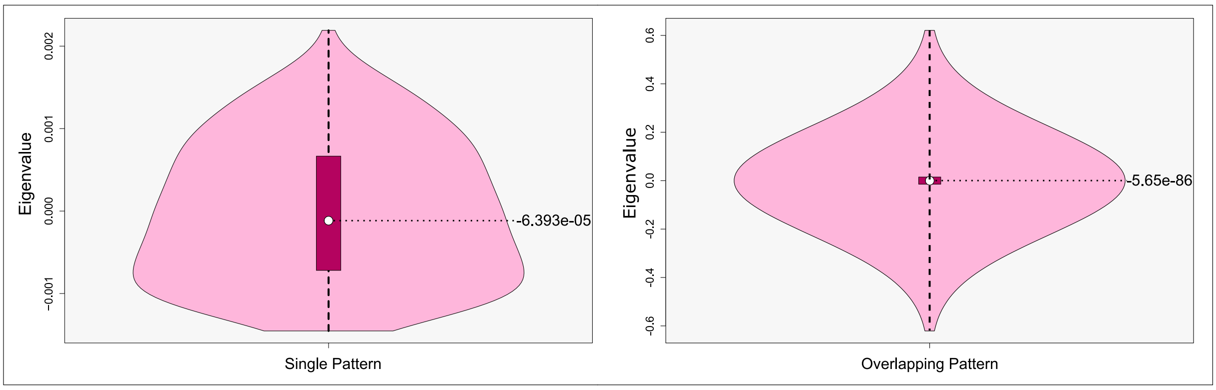


Figure II.2.5: Violin plots (*cf.* Hintze and Nelson (1998)) for a single pattern (left) and an overlapping pattern (right). The central box illustrates data between first and third quartile. The white dot refers to the median.

spatial effects has little room for fluctuating toward unrealistic erroneous choices. In case of the combined pattern we do also observe most of the values around the expectation of Moran's $\mathcal{I}$. This time, however, the markedly broadened range along the y-axis shows that the range has been stretched towards a multiple of the previous one. This demonstrates how strong the outliers caused by the overlap worsen the power as well as the efficiency of spatial test statistics obtained from overlapping patterns.

II.2.3.3 Influences on Spatial Autocorrelation

A naturally arising question now is to ask for the specific consequences of the findings from above on spatial analysis. The analysis of topological variability conducted above is concerned with the geographic layout. However, attribute values were not yet included. A simple yet powerful tool to inspect the strength and type of spatial associations within attributes is the so called Moran scatterplot (Anselin 1996), which enables us to investigate how topological variability influences spatial analyses. Figure II.2.6 showcases a Moran scatterplot for the non-overlapping pattern that was used in the previous section. A supplementary map of the underlying attribute values as well as a corresponding histogram is found in Figure II.2.17 in the appendix. The trend line through these points stretches from the third quadrant into the first one. This is the typical behaviour in case of positive spatial autocorrelation. It indicates that most points are placed in geographic neighbourhoods that consist of similar points. The first quadrant thereby means that high values are spatially surrounded by other high values (HH), while the third quadrant refers to low-low neighbourhoods respectively (LL).

When we construct the same scatterplot for the overlapping pattern we see that a number of additional components appear within the plot (Figure II.2.7, map and histogram in Figure II.2.18). The red points are observations which are unaffected by pattern overlap, and thus do not interact in a cross-pattern manner. The blue points belong to the smaller-scale process but do interact with points from the larger-scale one. In turn, yellow points are part of the larger-scale process but interact with the small-scale pattern. The corresponding lines demonstrate the respective trends for those three point clouds. Observe that the small-scale points from the overlapping area add a positive component which is paralleling the red points. The trend of this component, however, appears flatter than that of the red one. This means that, while still being positive, the blue points weaken the strength of observable spatial effects as they pull down the overall trend. Being even more influential, the yellow trend line shows negative behaviour. The underlying

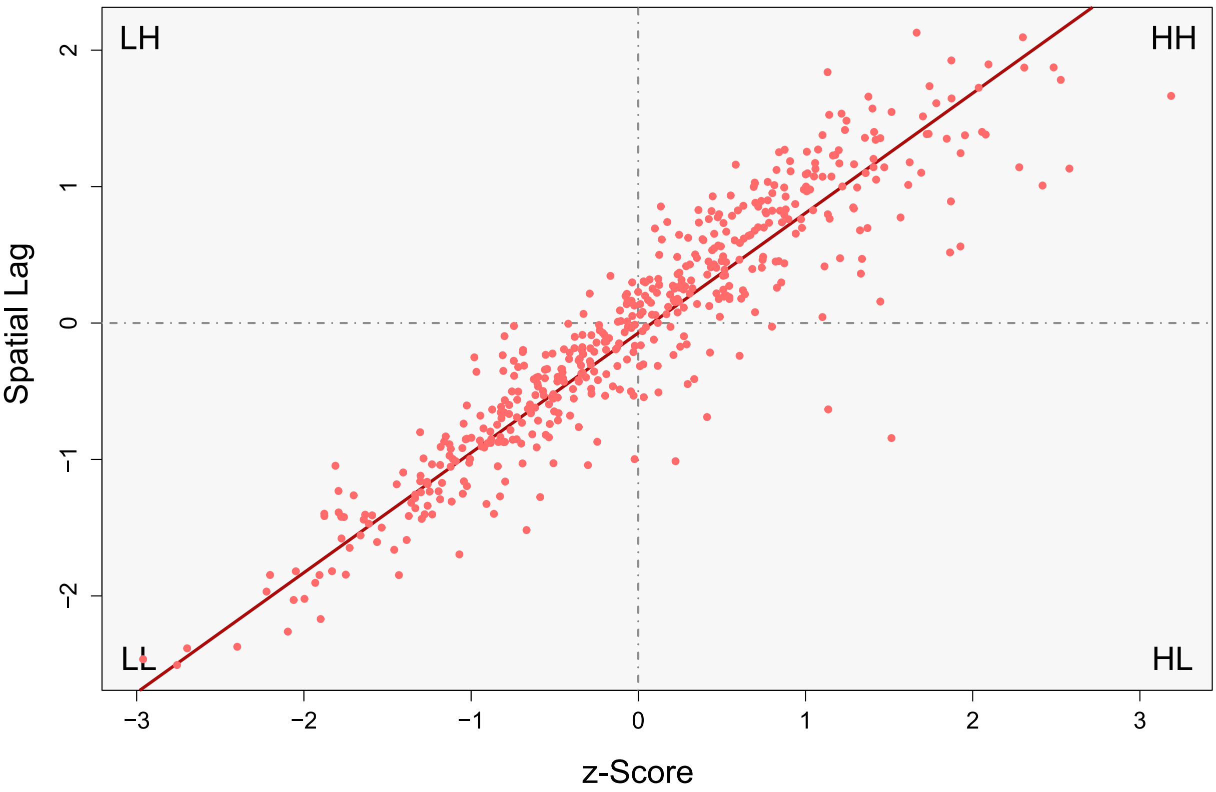


Figure II.2.6: Typical Moran scatterplot for positively spatially autocorrelated data. Blue line shows the trend. HL: High-Low, LH: Low-High, LL: Low-Low and HH: High-High interaction.

points must therefore be negatively correlated with their spatial surrounding. Both these components together obscure the real pattern which is encompassed within the data. The actually searched pattern is inflated with numerous artificial interactions.

Why does the blue component add a positive trend, while the yellow component contributes a negative relationship? Figure II.2.8 partly answers this question for the situation from Figure II.2.7 by showing a magnified detail view from within the zone of overlap. We see that, in case of the small-scale points (II.2.8a), the number of interactions with similar points (*i. e.*, other small-scale points) is still high. That is, although some yellow points are included, the majority of interactions still take place with other blue points. The yellow points are less frequent, because their scale, and therefore their point spacing, is lower. The few cross-pattern interactions between blue and yellow, however, are not without effect. They cause the blue component to be flatter than the red one. In contrast, Figure II.2.8b shows the same situation from a yellow component perspective. Yellow points do interact frequently with blue ones within the zone of overlap. Since the latter operate at a different attribute value intensity (*i. e.*, their attribute mean is higher), these interactions in close proximity appear as repulsion behaviour. Repulsion, in turn, is indicated by negative spatial autocorrelation. This explains why the yellow component runs downwards yet forming a negative relationship.

Figure II.2.7 has shown two different disturbing components and Figure II.2.8 shows that these are caused by different underlying topological constellations. In order to relate these components to topological variation, we relate them to their associated local eigenvalues. Figure II.2.9 shows corresponding 3D plots relating the Moran scatterplots from above to their associated local eigenvalues. In case of a non-overlapping pattern (Figure II.2.9a), the 95 % ellipse appears slightly negatively correlated

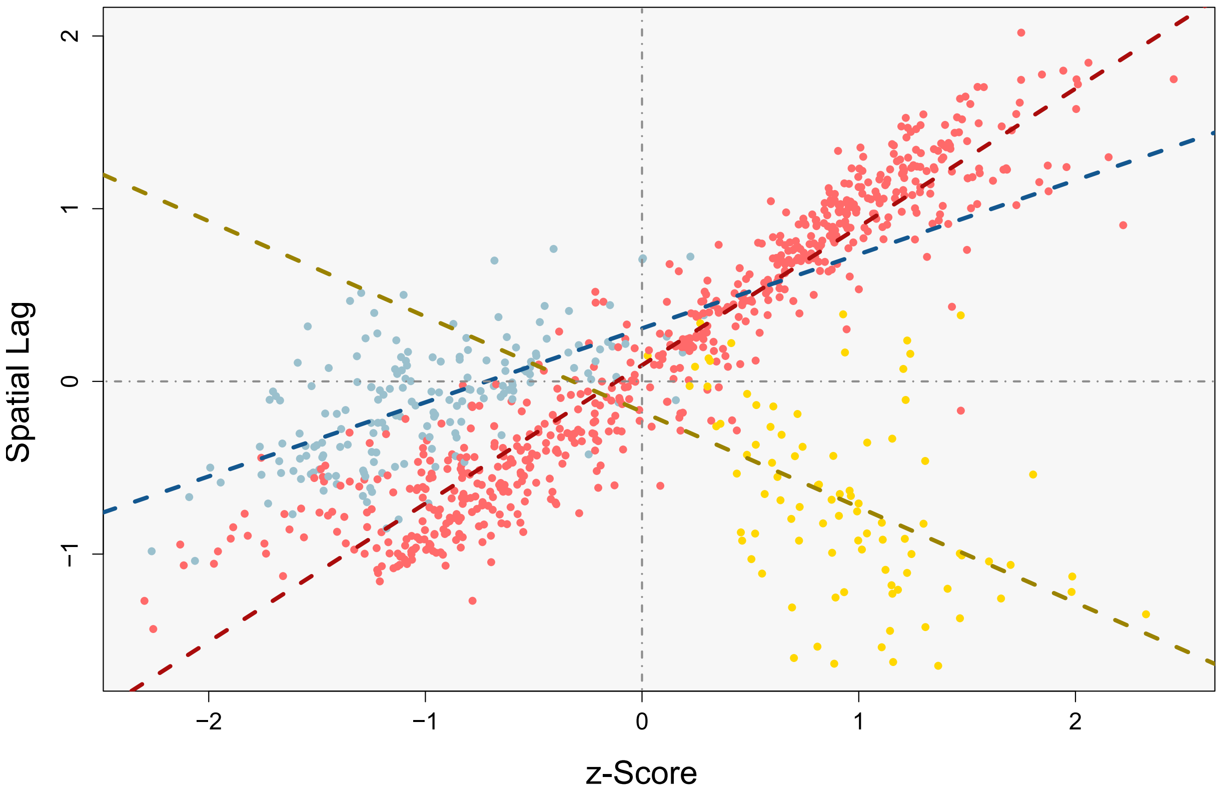


Figure II.2.7: Moran scatterplot for the combined pattern. Dashed lines show the trends of the similarly coloured points.

with the eigenvalues. Thus, as demonstrated above, the behaviour is homogeneous. Figure II.2.9b reveals that the disturbing components react differently on the degree of overlap. While the yellow component possesses a negative relationship with the eigenvalues, the blue component tends towards a positive connection with increasing degrees of overlap. This supports our conclusions drawn from Figure II.2.8, because both these effects seem to become stronger with increasing local eigenvalues. Overall, the plot shows a distinctive shape. It reveals that different parts of the overlapping pattern react in different ways on the topological implications that come along with the overlap.

II.2.3.4 The Role of Scale Differences

The results from above unveil that geometric overlap influences the quantification of patterns. We now turn our attention to effects that govern these influences, namely the effects caused by scale differences between the involved patterns. We investigate this by means of testing a range of scale differences between the involved overlapping patterns.

Influence of Scale Differences on the Numbers of Interactions

All previously stated results are based on analysing a single combined pattern. Yet, we don’t know how differing scales of the involved sub-patterns become effective. When social media patterns overlap, they can interact in two different general ways. One of these is a true geometric overlap. That is, a part of one pattern might be physically overlaying a fraction of another pattern. This kind of interaction manifests itself by an increased number of topological outliers and has been in focus within all previous parts of this

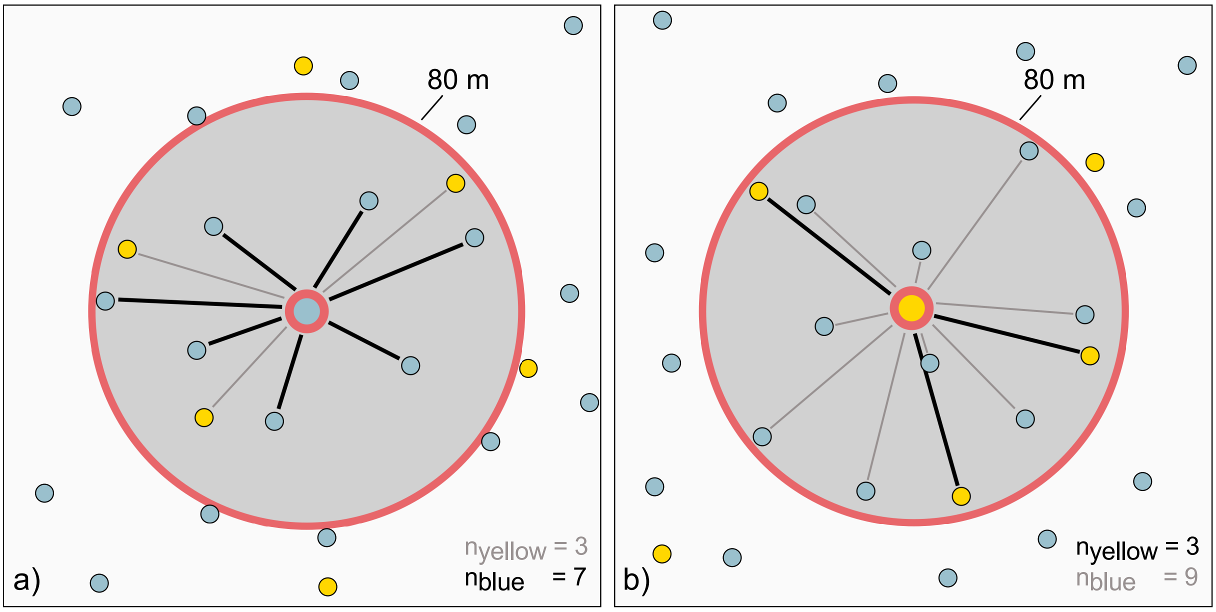


Figure II.2.8: Interrelationships between points within the zone of overlap. a) from a small-scale perspective and b) from a large-scale perspective.

paper. The second possible way of cross-pattern interaction is a cross-wise mutual consideration of points without physical overlap. This refers to the consideration of observations from one pattern, while having adjusted the focus of an analysis to that of another involved pattern. This type of misleading interaction becomes important when the two sub-patterns possess different statistical characteristics (*e. g.*, mean and variance). In such cases, geometric overlap leads to unrealistic mixture distributions not just within the zones of overlap but also when two patterns are closely neighboured.

We investigate these two situations by proceeding in the following way: We first fix a point pattern at a scale range of [1 m,10 m]. Repetitively, we translate the scale range by one meter and, for each range, create new random patterns. We do not alter the span of the scale ranges because we don't intend to introduce additional uncontrolled effects. These random patterns are then moved across the surface until an overlapping degree of 23.8 % is reached. The term "overlapping degree" thereby refers to two perspectives: We either require 23.8 % of the large-scale points to interact with at least one point from their small-scale counterpart ("large-scale perspective"); or adjust the target the other way round ("small-scale perspective"). The value of 23.8 % was thereby chosen to stay in accordance with our previous investigations above. The case of true geometric overlap is simulated by moving the patterns until 23.8 % of points show increased local eigenvalues. Analogously, mutual consideration is achieved by optimising the counts of interactions regardless of the eigenvalues. With increasing scale differences, however, the target is not always reachable. In such cases we rather search for the closest solution. Overall, we generated 9,000 patterns of this kind, 100 per scale difference. All results below are based on averaging over these.

Figure II.2.10a describes the numbers of interactions for the case of a true geometric overlap. Both, small as well as large-scale points do mutually interact in a similar way across scale differences. The only notable difference between them is a differing intensity and can be explained by the generally higher number of points per area for the small-scale process. That is, points from large-scale patterns (the vertical bars in the background) have more small-scale neighbours in their vicinity than vice versa. This

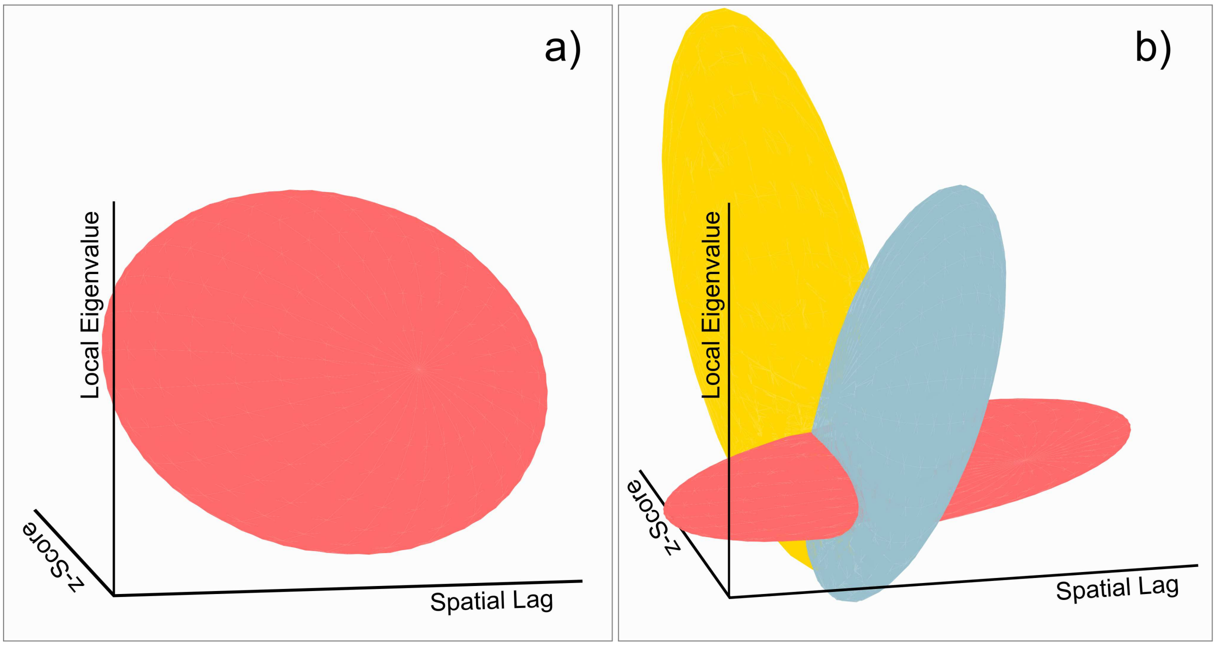


Figure II.2.9: Comparison between the Moran scatterplot and associated local eigenvalues. Colours are in accordance to Figure II.2.7. Shown ellipses mark the respective 95 % confidence ellipses. a) Ellipse for a non-overlapping pattern. b) Ellipse for an overlapping pattern. Note that the magnitudes of the axes differ. Similar sizes were chosen for visualization purposes.

increases the general count level and leads to the observed higher count intensity. The functional decay over scale differences cannot be described by a single function. However, we are able to identify three different regimes. When the patterns' scales are relatively similar, the decay follows a steep exponential curve. Around 15 distance units of scale difference, this relationship is replaced by another, yet flatter, exponential relationship. This function holds up to roughly 61 distance units, where it slowly vanishes into an almost constant level. The latter transition is not an abrupt one, but rather a slow passing over between the functional relations. The thresholds (*i. e.*, 15 and 61 distance units) were assessed by means of cumulatively fitting the different mentioned functions. The supplementary Figure II.2.19 provides the result of this fitting procedure, which in turn reveals the abovementioned thresholds.

Figure II.2.10b illustrates the numbers of interactions for the case of mutual consideration. Other than with the case of geometric overlap, we observe clear differences between large and small-scale patterns. While the large-scale patterns show similar behaviour as within Figure II.2.10a, the small-scale patterns show an almost constant level of interaction counts across scale differences. The few observable fluctuations are merely attributable to the inherent randomness in the pattern generation procedure and the general study design. The constant level is explained as follows: After a certain point (which is reached quickly) only one point from each of the large-scale patterns is left for interaction with a fraction of the neighboured small-scale pattern. The point spacing of the large-scale patterns simply becomes too wide-spread to allow any further interaction. That is, the small-scale process falls completely into one of the gaps between two points of the larger-scale process. Thus, the constancy is resultant to the interaction of one large-scale point with a certain fraction of the small-scale points. This finding is highly relevant, because it demonstrates that, when the patterns' scales are too different from each other, a single point might govern the entire assessment of spatial structure.

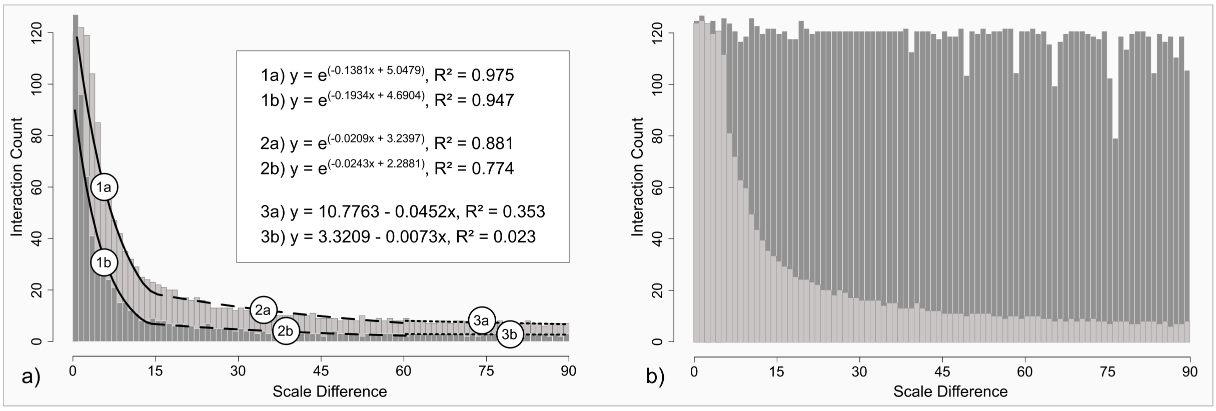


Figure II.2.10: Numbers of interactions between two overlapping patterns across a range of scale differences. a) Overlapping patterns; b) Mutual involvement. Dark gray: small-scale perspective; light gray: large-scale perspective. (1a/b) to (3a/b): fitted decay functions for sub-ranges.

Influence of Scale Differences on the Disturbing Moran Scatterplot Components

Intuitively, one might argue that the more cross-pattern interactions are observed, the more influences can be expected when performing spatial analysis on these. The number of cross-pattern interactions at least depends heavily on scale differences (as shown above). We should therefore investigate how the three different components from the Moran scatterplot (red, yellow and blue) behave across increasing scale differences between the involved patterns. For that purpose we again use the same 9,000 random patterns as in the previous section. However, this time we additionally assign them Gaussian attributes. These attribute values are drawn from the two Gaussians described in Section ’Datasets’. Finally, for each pattern, we calculate the trend lines for the three components and observe their slope over increasingly different scales of the involved overlapping patterns. Thereby, we also vary the direction of the attribute patterns (*i. e.*, increasing from inside to outside vs. decreasing from inside to outside).

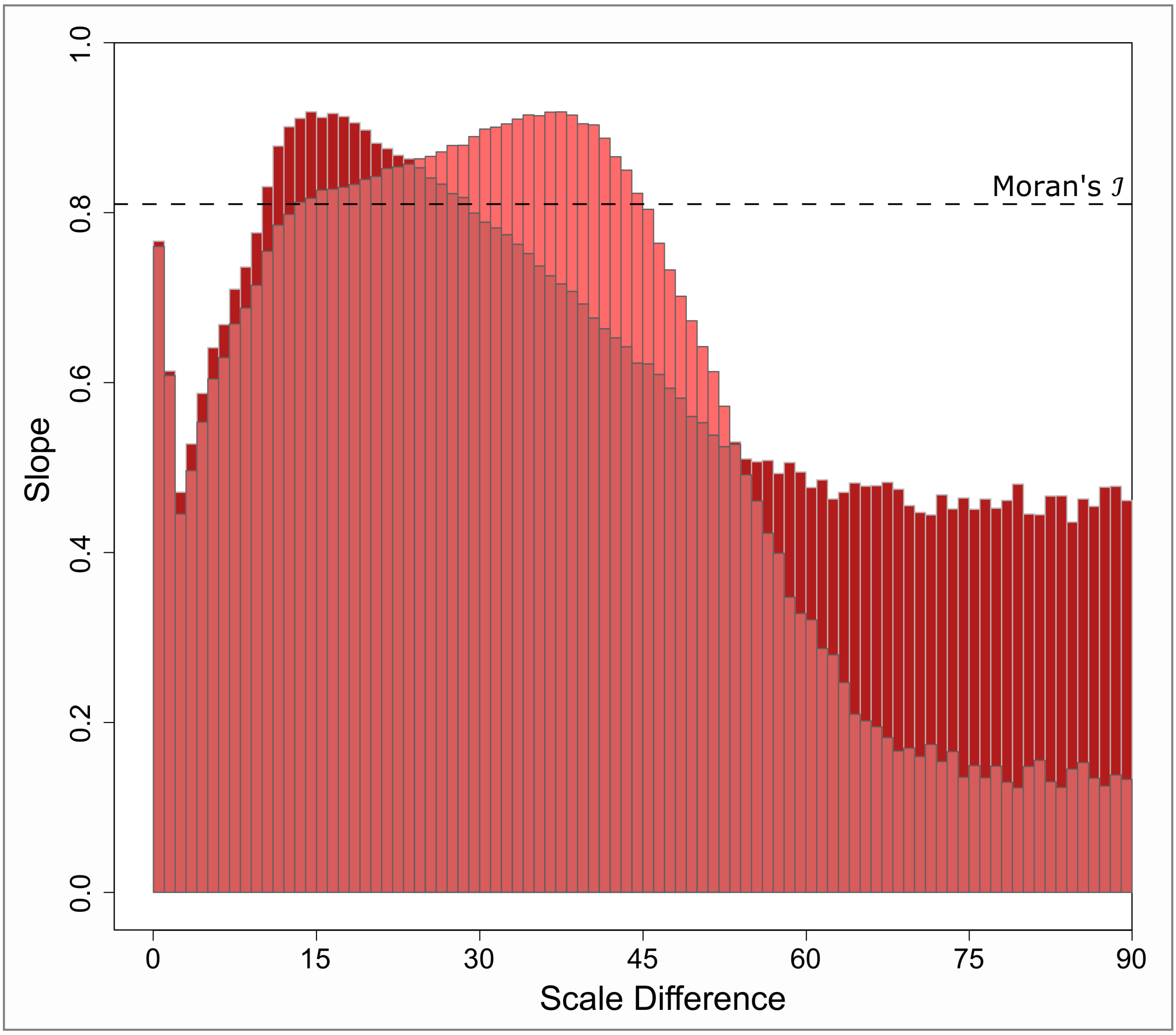


Figure II.2.11: Course of the slope of the red component from the Moran scatterplot. Dark-red: increasing attribute values from pattern centre toward the boundary. Light-red: reversed attribute dispersal. The dashed line indicates the true Moran's $\mathcal{I}$ value of 0.81.

Figure II.2.11 shows two characteristic plots obtained for the red component. Recall that this component reflects the non-overlapping parts of the involved patterns from outside the zone of overlap. The dark-red diagram is based on attribute values that increase from centre to boundary. Contrary, the light-red diagram reflects reversed attribute dispersal. Two differences are notable: The dark-red plot shows a narrow principal peak, and a slow decay. In contrast, the light-red plot possesses a broader saddle, and then decreases more steeply. The small increase in the very beginning, however, is a commonality shared by both plots.

Now, in order to evaluate meaning and significance of these effects, keep in mind that the actual slope of the red component is 0.81 in case of no overlap. Thus, according to Figure II.2.11, we can figure out two types of configurations under which the slope is quite close to that target. One of these is located at the short-range scale differences where the two patterns are almost similarly scaled (the small peak). Here, the patterns' interaction is marginal and cross-pattern effects are mostly caused by small fractions of the two boundary regions overlapping each other. However, there are still many points left without any

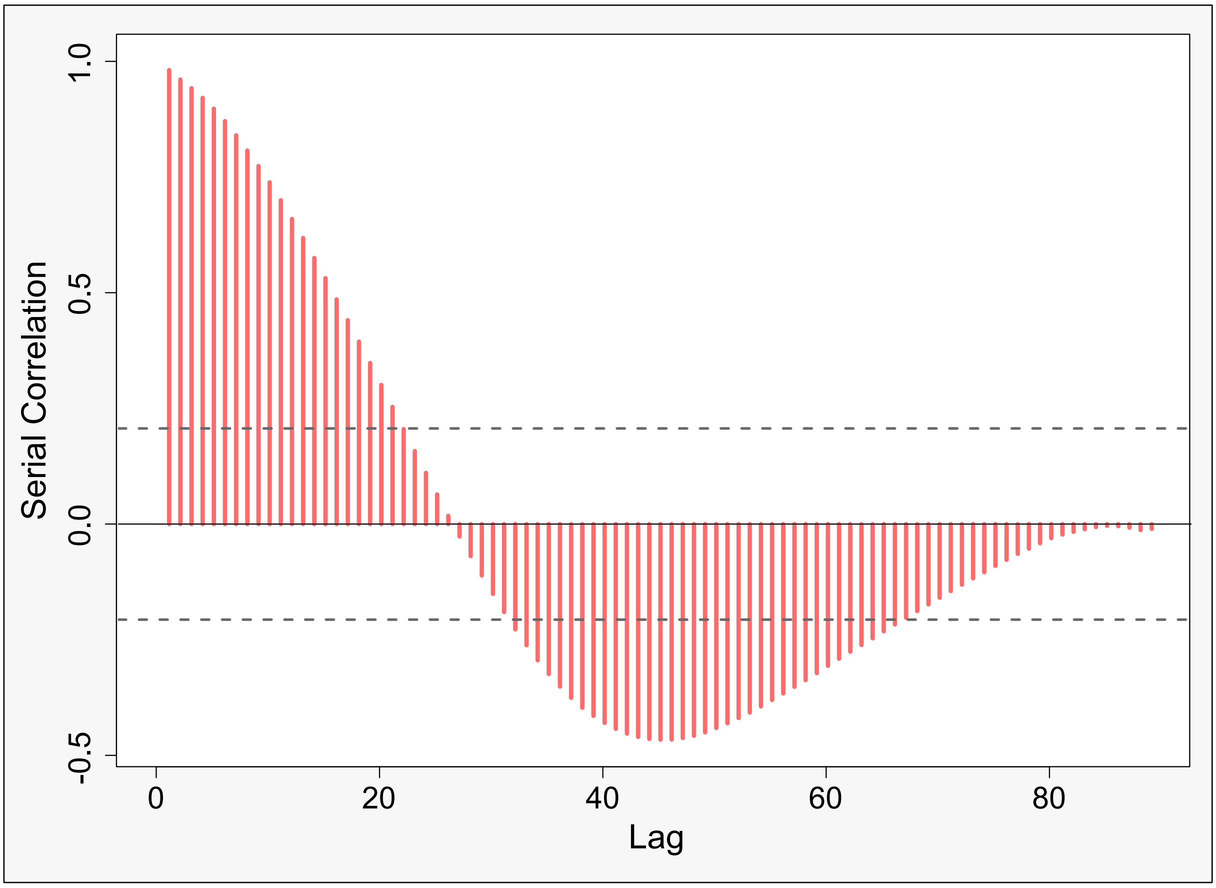


Figure II.2.12: Correlogram of the serial correlation at different lags for the slopes of the red component. Dashed line indicates the 95 % confidence interval.

cross-pattern interaction. This preserves the characteristic spatial pattern of the attribute to a large extent, and leads to an almost uninterrupted red component.

A second favourable configuration is observed when the small-scale pattern covers a fraction of the large-scale pattern in a way such that the overall characteristic distribution of attribute values is preserved. In a radial pattern like it is used here, this is the case whenever the small-scale pattern cuts through the large-scale counterpart in a cross-sectional way. However, when the attribute values are dispersed in different ways, the optimal cut-through might appear in a different fashion. Anyway, the consequence of such overlaps is that the red component is not significantly changed in nature. The left-over non-overlapping points do still possess the characteristic distribution of values and, to a large extent, are able to generate a stable red component. Within Figure II.2.11, this configuration is reflected by the two saddles at medium scale differences. The slight differences between light and dark-red within Figure II.2.11 fall back to the type of pattern possessed within the attributes. Thereby, the way of attribute dispersal within the small-scale pattern governs the width of the saddle at the medium scale differences. In contrast, the attribute dispersal of the large-scale pattern is responsible for the steepness and ultimate level of the decay at larger scale differences.

Investigating the serial correlation within the Moran-scatterplot-related slopes of the red component across the scale differences reveals very systematic behaviour. The estimated correlogram within Figure II.2.12 shows all possible lags across the whole range of scale differences. It appears to be distinctively

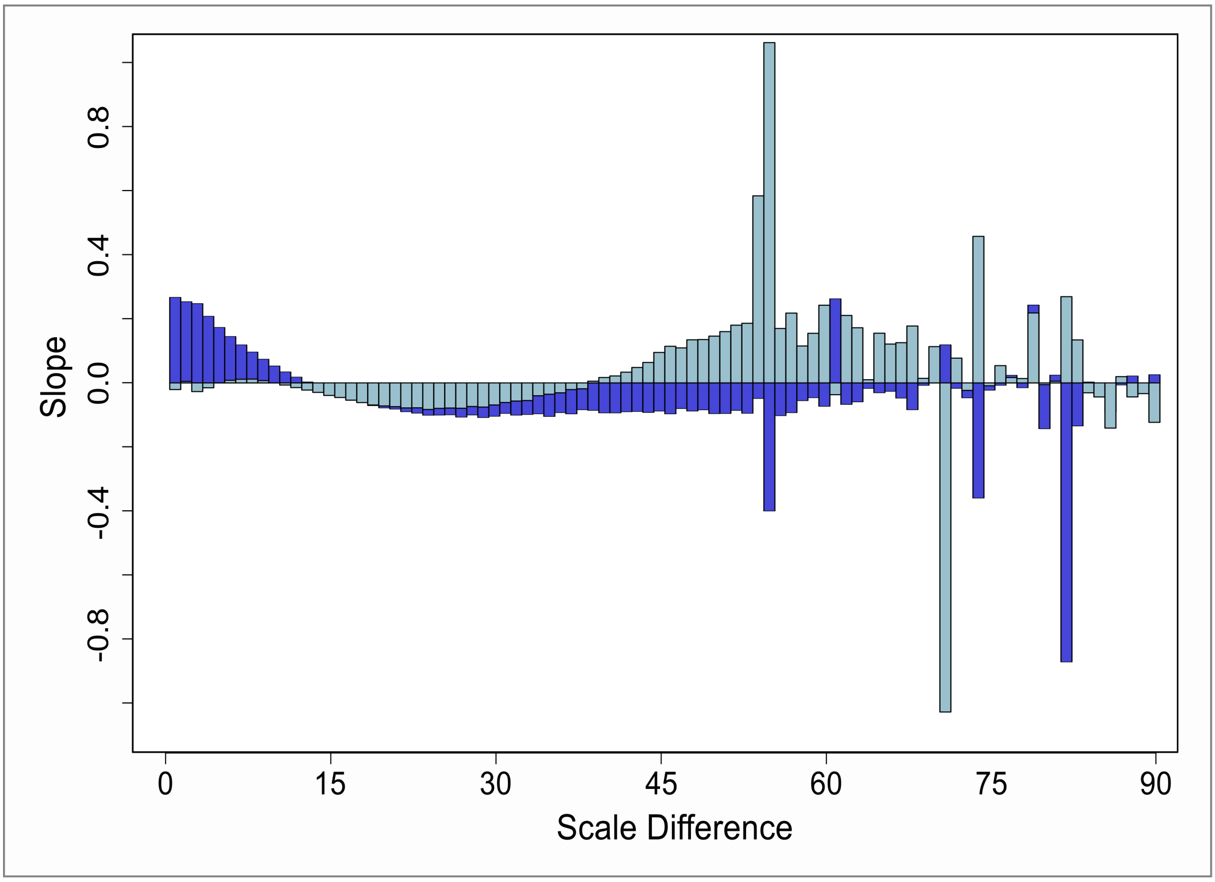


Figure II.2.13: Course of the slope of the blue component from the Moran scatterplot. Dark-blue: increasing attribute values from pattern centre toward the boundary. Light-blue: reversed attribute dispersal.

smooth, whereas bumps and high frequency fluctuations are not observed. It further indicates two regions in which the autocorrelation between nearby scale differences is significant at the 95 % significance level: small lags and medium lags. Thus, as a conclusion, we observe a smooth transition and a slightly sinusoidal seasonality over the sale differences.

The blue component reflects disturbances which are added by overlapping points originating from the small-scale process. Consequently, Figure II.2.13 shows that the small-scale process itself is the main driver of the shape of the slopes across the scale differences. When the pattern of the attribute values increases from centre towards boundary (dark-blue), the component appears slightly positive as long as the involved scales are relatively similar. As the scale differences grow larger, the component transitions into a moderate negative behaviour. At a certain point, the pattern becomes chaotic and less predictable. When the direction of the attribute pattern is reversed (light-blue), the course described above is also reversed at larger scale-differences. However, when the scale ranges are more similar, the component tends towards being negative.

The chaotic behaviour at larger differences is caused by interaction between few points. The small-scale pattern interacts with only one or two points from the large-scale pattern at these scale differences. Moreover, these few points do in turn interact with large parts of the small-scale pattern. Thus, if the attribute values of these few points are somewhat extreme, a large number of either highly positively

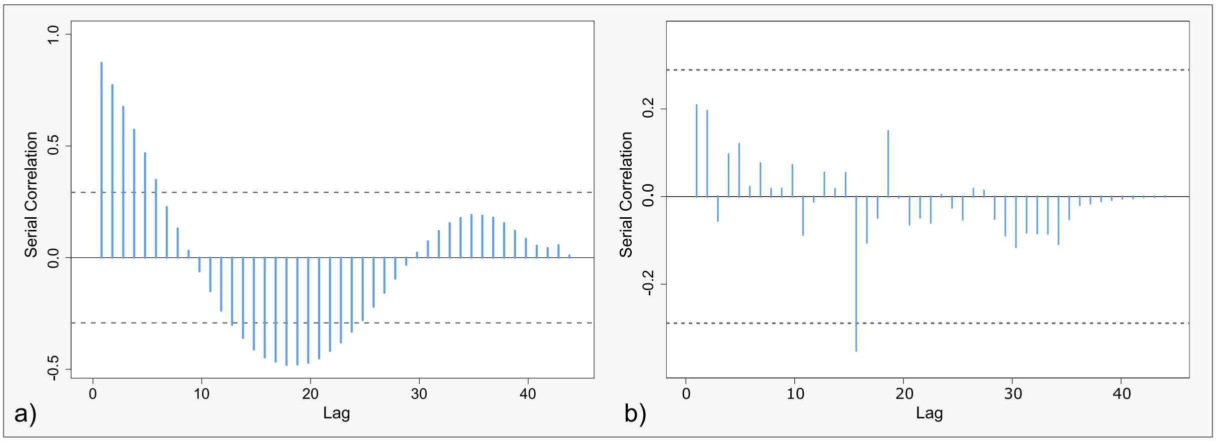


Figure II.2.14: Correlograms of the serial correlation at different lags within the slopes of the blue component. a) Scale differences up to 45 m. b) Scale differences between 45 m and 90 m. Dashed line indicates the 95 % confidence interval.

or negatively correlated comparisons are included. The quintessence is that, as off a scale distance of approximately 45 m, the way how the patterns interact is no longer predictable with respect to the blue component. The explained chaotic behaviour is also well reflected by the serial correlation given by Figure II.2.14. Thereby, unlike with the red component above, we separated the correlogram into two parts. The first of these demonstrates a coherent behaviour up to scale differences of 45 m. Here, the autocorrelation progresses smoothly. The second one reflects the chaotic behaviour at larger differences. Clearly, the high level of fluctuation barely allows any indication for systematic behaviour. However, Figure II.2.13 indicates that there is a slight tendency towards either positive or negative slopes for each of the two investigated attribute patterns. That is, this tendency seems to flip when the pattern gets reversed. Further, note the similarity of the serial correlation at small scale differences and that of the red component. This indicates that the blue component has a strong influence on small scale differences.

In analogy to the small-scale pattern with the blue component, the large-scale pattern is the main driver of the yellow component (Figure II.2.15). When the attribute pattern increases from centre to boundary, the yellow component is positively dominant at small scale differences (light-yellow). When the pattern is reversed, however, this relationship is flipped and the yellow component becomes dominant at larger differences (dark-yellow). Interestingly, the role of the small-scale process here is to control the direction of the component. When the small-scale attribute pattern runs opposite to the larger-scale one, the yellow component is turned to negative either at small or large scale differences.

The serial correlation of the yellow component reveals a similar smoothness as with the blue component. However, as none of the serial correlations is significant, this component is more volatile and less coherent than the blue counterpart. Notice the isolated spike at small lags in Figure II.2.16a. This isolated spike indicates that the pattern does not possess abrupt bumps, because neighbouring values are to a certain extent similar. However, this similarity decreases quickly. Again, the clutter increases strongly for the larger scale differences. Just like with the blue component, this unveils a two-pattern regime (Figure II.2.16a vs. Figure II.2.16b).

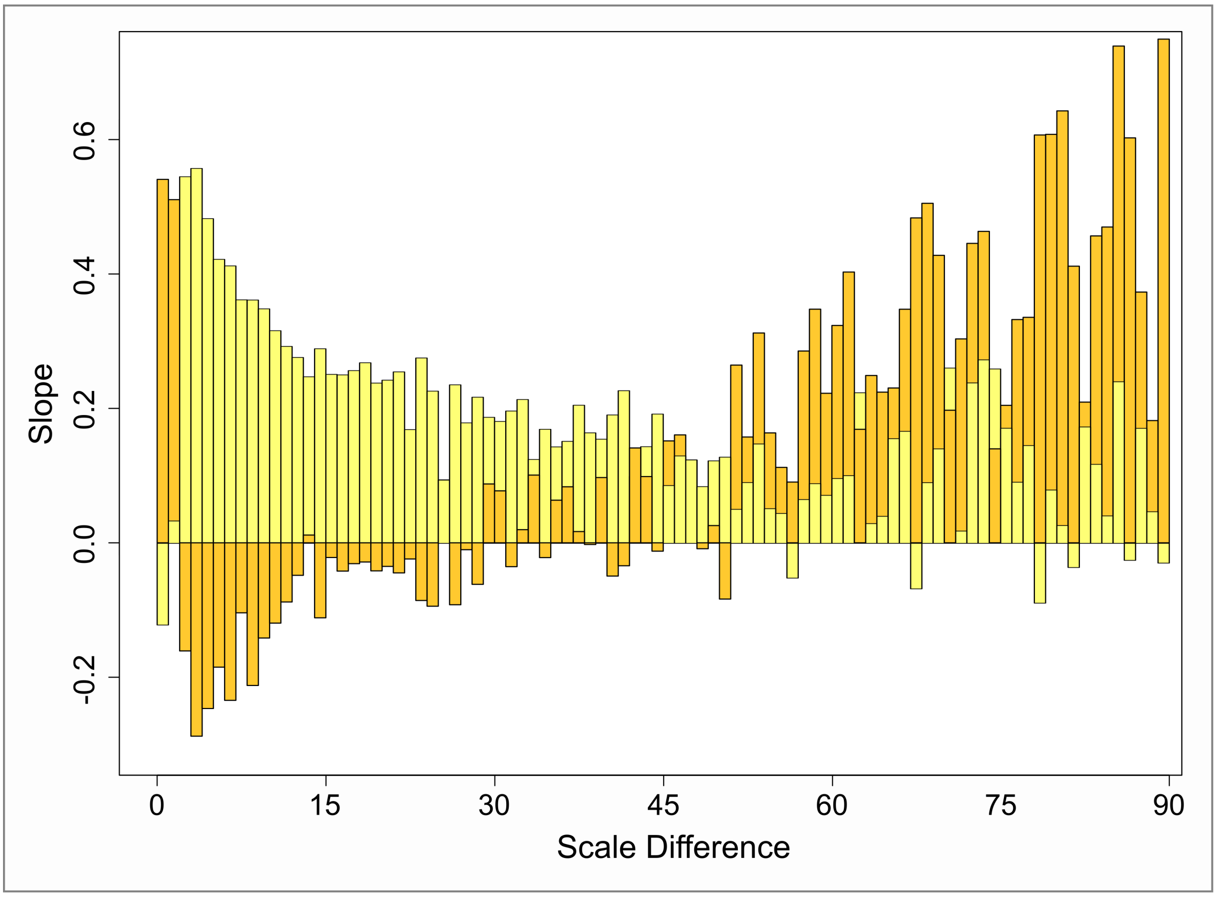


Figure II.2.15: Course of the slope of the blue component from the Moran scatterplot. Light-yellow: increasing attribute values from pattern centre toward the boundary. Dark-yellow: reversed attribute dispersal.

II.2.4 Discussion

The tweets from London used in our study have shown clear indications of geometrically overlapping phenomena and processes. The derived semivariogram shows unusual behaviour and hints on repulsion, and thus negative spatial autocorrelation at local scales. The local-scale activity is not surprising given that urban areas are typically patchy and dense. What is surprising though is the negative (repulsion) behaviour. Besides, the peak in the semivariogram is rather low, which indicates merely negligible spatial behaviour in the variable (which is not plausible in an urban environment). A closer look at the pairwise autocovariance terms and their associated geographic distances reveals that both clustering and repulsion take place in close vicinity to each other, besides a large amount of unrelated tweets. These results demonstrate that the spatial associations of interest (mostly those of clustering nature) may remain hidden, and spurious spatial relationships might instead be detected. These results strongly support our initial hypothesis of overlapping phenomena being reflected within Twitter datasets. This is, however, ultimately leading to a violation of the requirement of second-order stationarity, whereupon many spatial-statistical techniques are based (and so is Moran's $\mathcal{I}$, (Gaetan and Guyon 2010, p. 166)).

We analysed the artificial spatial regime that forms within zones of overlap by means of the eigenvalues of local as well as global spatial weights. This regime is characterized by a large number of topological

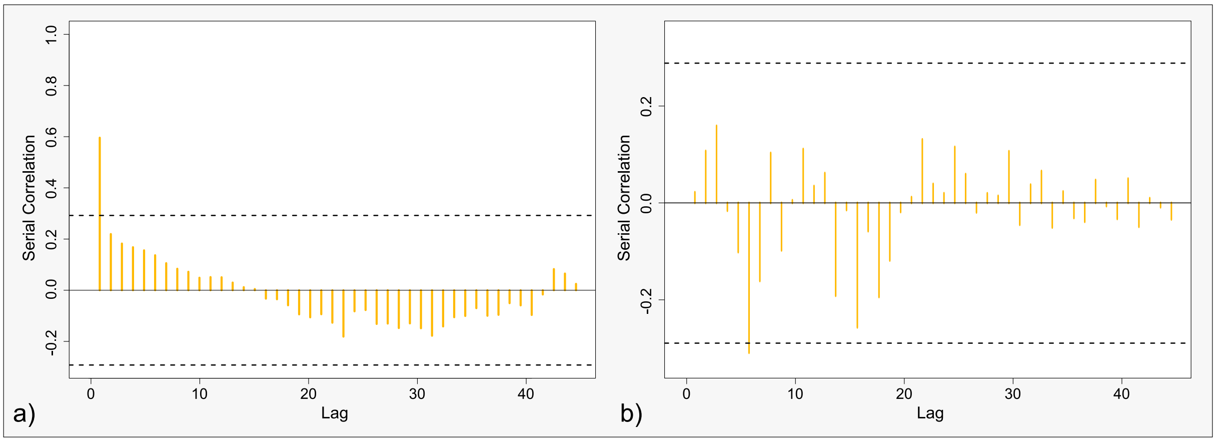


Figure II.2.16: Correlograms of the serial correlation at different lags within the slopes of the yellow component. a) Scale differences up to 45 m. b) Scale differences between 45 m and 90 m. Dashed line indicates the 95 % confidence interval.

outliers. These are known from traditional datasets where they occur after applying some kind of normalization procedure (Tiefelsdorf and Boots 1997; Tiefelsdorf et al. 1999). With social media data, however, these outliers also occur without any further modification of the data as a result of overlap. They increase the topological variability, *i. e.*, the overall chances for detecting spurious spatial interaction and patterns. Further, they decrease the power as well as the efficiency of spatial test statistics, which in turn leads to a higher risk for drawing wrong conclusions (*i. e.*, type I and II errors). We only tested overlapping patterns of roundish shape, which faintly limits the results to these. However, other kinds of patterns should, by principle, behave in a similar way.

These topological outliers have impact on the detection of spatial structure by adding different kinds of disturbances. These manifest themselves in two different ways: One disturbing component is related to overlapping observations which belong to the smaller-scale pattern of the two investigated ones. A second component is related to the points of the large-scale pattern respectively. Both show different behaviours, but, however, are inherently linked. Their mutual relationship is demonstrated by their causal mechanisms. These are both driven by interactions between the two involved patterns. Further, both kinds of disturbances correlate with the degree of topological variation, though in different ways. One component might correlate positively, while the other one associates in a negative way. As a result, these components do in fact lower the strength of detectable spatial effects and might lead to misleading interpretations of spatial patterns. Another interpretation of these nuisances from social media characteristics is to see them as distinct spatial processes. The disturbing components come up with their own spatial interaction behaviour and disturb the actual pattern of interest. The latter is true because they are not caused by real-world social phenomena, which in turn get obscured by them.

Observing these disturbing components over a range of scale differences between both involved patterns unveils several kinds of effects. First of all, the degree of mutual interaction seems to follow several forms of exponential decay functions as scale differences increase. This decay starts out steep and then transitions into a flatter exponential function before slowly vanishing towards an almost constant level at very large differences. When differently scaled patterns are neighboured instead of overlapping geometrically, small and large-scale patterns react differently. Small-scale patterns tend to interact with few points from the large-scale process. Thus, few points govern the results of a spatial analysis in such cases. When these considered large-scale points are extreme with respect to their attribute, any result might be strongly biased. In contrast, the large-scale pattern, again, shows an exponential decay like described above.

In terms of the direction of the components (*i. e.*, adding negative or positive spatial autocorrelation to the Moran scatterplot), the achieved results provide a diverse picture. The red component (consisting of non-overlapping observations) should either overlap in a way such that only smaller parts of the boundaries of the patterns interact with each other. Another low risk option is an overlap that cuts through the attribute values of the larger-scale pattern so that all characteristic parts of the disturbed pattern are retained in accordance to their proportion within that pattern. In all other cases, however, the characteristics of large-scale patterns get disturbed significantly, and results become increasingly unrealistic. However, since we analyse positively autocorrelated data, the red component remains positive across all tested scale-differences.

The blue and yellow disturbances (*i. e.*, those caused by either the smaller or the larger-scale process) behave in more complex ways. These processes are strongly dependent of the actual pattern of the attribute value dispersal. However, as a summarising result, these components do typically provide ranges in the scale differences at which they add negative influences. Similarly, at some other sub-ranges, these relationships turn toward positive respectively. Further, as the scale differences between overlapping patterns become larger, these components show increasingly chaotic, and thus unpredictable behaviour. The latter effect is caused by interactions between only few points from a larger-scale pattern with many of those from a smaller-scale opponent.

Our results reveal some limitations of our research. First of all, we did not remove artefacts like bot-produced tweets from the data. These might contribute content of little explanatory power with respect to real-world social phenomena (see Haustein et al. (2016) for their impact on altmetrics). Therefore, it remains unknown to what extent these tweets play a role in spatial patterning. Further, we investigated a limited number of types of spatial attribute configurations (radial, increasing attributes from inside towards the borders and vice versa) and narrowed down the scope to an overlap of only two patterns. Apart from topological considerations, we also held statistical properties like the means and variances of the attribute patterns constant across our investigations. The reason for both these choices was to keep the analyses tractable and to facilitate their interpretation, but they might play a role in the results. Moreover, we exclusively focused on positive spatial autocorrelation given its higher practical relevance. Nevertheless, findings about negatively correlated patterns under heterogeneous conditions caused by overlaps would be of interest for the study of spatial outliers. From a technological perspective, we restricted our analysis to explicit coordinates by leaving out coordinates obtained through geocoding.

II.2.5 Conclusions

Social media data reflect an ample amount of social phenomena and processes. These are likely to appear overlapping in space and time and are prone to varying interpretations among the contributing users. In this paper we investigate how topological effects caused by these overlaps influence outcomes of spatial analyses. For that purpose, we first analysed the spatial behaviour of LDA-derived semantic topic associations within a Twitter sample from London. Afterwards, we conducted a number of simulation experiments to investigate different aspects related to the topology of overlapping point patterns. We enriched these simulated patterns by Gaussian attribute values at different means but with similar variance. They thus resemble a special case of spatial heterogeneity in which different regimes are not just appearing close to each other (the traditional notion), but form an artificial regime in-between through geometric overlap.

To summarize our results we list the key findings in the following enumeration. These points are also meant to raise scholars' awareness of carefully undertaking spatial analyses of social media data:

- Increased numbers of topological outliers are found and these increase the risk of false positives and negatives in spatial analyses on social media data. Thus, misleading indications regarding spatial relationships within the data must be expected when using established spatial analysis methods.
- The way how spatial proximity is modelled through spatial weight matrices is crucially important in general, but even more so with overlapping patterns. The tested configurations have shown a large variety and thus sensitivity to this issue. Distance-based weights are extremely problematic on that regard, since they possess extreme behaviour at short distances. The latter happens frequently when patterns overlap.
- When differently scaled patterns overlap and when the scale differences are large, single extremal points from the larger-scaled of the involved patterns might control the results significantly.
- When social media patterns are geometrically overlapping, the number of interactions, and thus the chance for detecting spurious effects, decreases exponentially with increasing scale differences. In contrast, when differently scaled patterns are just neighboured in close vicinity, the adjusted analysis scale becomes important for the risk of including wrong observations.
- Besides scales of the involved patterns, the shape of how the attribute values are dispersed possesses great influence on the type of interferences. These might either be expressed in terms of an additional positive or a negative component respectively. The latter act like additional spatial processes that interfere with the actual pattern of interest.

Future research should focus on a range of different aspects that could not be investigated exhaustively within this paper. One of these is the pattern of the attribute values. We used two different kinds of radially dispersed trends within each of our point patterns. However, different kinds of attribute value arrangements might lead to different results. Our tests have shown respective indications for a tremendous sensitivity to this issue. Further, our results indicate interaction with the kind of neighbourhood definition. We used distance-based spatial weights and roughly tested some binary configurations. These experiments, however, unveiled a high variation among different types of weights. This is an issue of high practical relevance and thus deserves particular attention. Further, given their practical relevance and widespread use, the severe behaviour of distance-based weights at short distances should be further examined with respect to social media data as they are commonly used. This should incorporate a critical discussion of results achieved through already conducted spatial studies of social media. Other future prospects might include variations of statistical properties such as means and variances as well as including coordinates from geocoded text-based information from the posted messages.

We close the article with some recommendations to researchers conducting spatial analysis with georeferenced social media feeds. In the first place, one should check the data for signs of heterogeneities. This should incorporate the geometric dimension (*e. g.*, through techniques like Ripley's K function, see Dixon (2013)) as well as the respective attribute (techniques like local spatial heteroscedasticity (LOSH) might be helpful (Ord and Getis 2012; Xu et al. 2014a)). As we have seen with the semivariogram in our analysis, one difficulty is that the heterogeneity might remain hidden because of the noisy nature of the data. Therefore, whenever possible, the target set of observations shall be isolated as far as possible from the rest. Clearly, this is hampered by the oftentimes exploratory character of spatial analysis. Spatial patterns are often part of an early investigation in the hypothesis building phase when the dataset is not well-understood. This requires the acquisition of extensive expert knowledge about the spatial aspects of the target subject of investigation. This leads to the next recommendation which is putting vast effort in properly designing the spatial weights. This is an essential part of each spatial analysis. However, our eigenvalues analysis has shown that it is even more important when it comes to social media. The spatial weights matrix needs to be constrained to the particular research needs as conservatively as possible. In the aftermath when it comes to drawn inference, a double-check needs to be performed whether the reference distribution under the null hypothesis is really appropriate. Again, the outliers might lead to an unexpected shape of this distribution, which would ultimately lead the analyst to wrong conclusions.

Acknowledgements

We sincerely thank the developers of the R packages *foreach*, *doParallel*, *spatstat*, *car*, *scatterplot3d* and *RColorBrewer* for their great software development efforts. Further, we thank *Twitter* for providing public access to at least a certain fraction of their tweets.

References (Chapter II.2)

Albuquerque, J de, B Herfort, A Brenning and A Zipf (2015). 'A Geographic Approach for Combining Social Media and Authoritative Data Towards Identifying Useful Information for Disaster Management'. *International Journal of Geographical Information Science* 29 (4), pp. 667–689. DOI: 10.1080/13658816.2014.996567.

Aldstadt, J and A Getis (2006). 'Using AMOEBA to Create a Spatial Weights Matrix and Identify Spatial Clusters'. *Geographical Analysis* 38 (4), pp. 327–343. DOI: 10.1111/j.1538-4632.2006.00689.x.

Anselin, L (1995). 'Local Indicators of Spatial Association - LISA'. *Geographical Analysis* 27 (2), pp. 93–115. DOI: 10.1111/j.1538-4632.1995.tb00338.x.

— (1996). 'The Moran Scatterplot as an ESDA Tool to Assess Local Instability in Spatial Association'. In: *Spatial Analytical Perspectives on GIS in Environmental and Socio-Economic Sciences*. Ed. by M Fischer, H Scholten and D Unwin. London, UK: Taylor & Francis, pp. 111–125.

Anselin, L and A Bera (1998). 'Spatial Dependence in Linear Regression Models with an Introduction to Spatial Econometrics'. In: *Handbook of Applied Economic Statistics*. Ed. by A Ullah. London, UK: CRC Press, pp. 237–289.

Atkinson, P (2001). 'Geographical Information Science: GeoComputation and Nonstationarity'. *Progress in Physical Geography* 25 (1), pp. 111–122. DOI: 10.1177/030913330102500106.

Bachmaier, M and M Backes (2008). 'Variogram or Semivariogram? Understanding the Variances in a Variogram'. *Precision Agriculture* 9, pp. 173–175. DOI: 10.1007/s11119-008-9056-2.

Bavaud, F (2014). 'Spatial Weights: Constructing Weight-Compatible Exchange Matrices from Proximity Matrices'. In: *LNCS: Geographic Information Science*. Ed. by M Duckham, E Pebesma, K Stewart and A Frank. Heidelberg: Springer, pp. 81–96. DOI: 10.1007/978-3-319-11593-1_6.

Blei, D, A Ng and M Jordan (2003). 'Latent Dirichlet Allocation'. *The Journal of Machine Learning Research* 3, pp. 993–1022.

Brunsdon, C, A Fotheringham and M Charlton (1996). 'Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity'. *Geographical Analysis* 28 (4), pp. 281–298. DOI: 10.1111/j.1538-4632.1996.tb00936.x.

Casetti, E (1972). 'Generating Models by the Expansion Method: Applications to Geographical Research'. *Geographical Analysis* 4 (1), pp. 81–91. DOI: 10.1111/j.1538-4632.1972.tb00458.x.

— (1997). 'The Expansion Method, Mathematical Modeling, and Spatial Econometrics'. *International Regional Science Review* 20 (1-2), pp. 9–33. DOI: 10.1177/016001769702000102.

Cliff, A and J Ord (1973). *Spatial Autocorrelation*. London, UK: Pion.

Cranshaw, J, R Schwartz, J Hong and N Sadeh (2012). 'The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City'. In: *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media*. Dublin.

Cresci, S, R Di Pietro, M Petrocchi, A Spognardi and M Tesconi (2015). 'Fame for Sale: Efficient Detection of Fake Twitter Followers'. *Decision Support Systems* 80, pp. 56–71. DOI: 10.1016/j.dss.2015.09.003.

Cressie, N (1993). *Statistics for Spatial Data*. Wiley Series in Probability and Statistics. Hoboken, NJ: John Wiley & Sons. DOI: 10.1002/9781119115151.

Crooks, A, A Croitoru, A Stefanidis and J Radzikowski (2013). '#Earthquake: Twitter as a Distributed Sensor System'. *Transactions in GIS* 17 (1), pp. 124–147. DOI: 10.1111/j.1467-9671.2012.01359.x.

Dixon, P (2013). 'Ripley's K Function'. In: *Encyclopedia of Environmetrics*. Chichester, UK: John Wiley & Sons. DOI: 10.1002/9780470057339.var046.pub2.

Ester, M, H Kriegel, J Sander and X Xu (1996). 'A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise'. In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. Ed. by E Simoudis, H Jiawei and F Usama. Portland, OR: AAAI Press, pp. 226–231.

Fischer, M and A Getis (2010b). 'Introduction'. In: *Handbook of Applied Spatial Analysis*. Ed. by M Fischer and A Getis. Heidelberg: Springer, pp. 1–24. DOI: 10.1007/978-3-642-03647-7_1.

Fotheringham, A, C Brunsdon and M Charlton (2002). *Geographically Weighted Regression : The Analysis of Spatially Varying Relationships*. Chichester, UK: Wiley.

Gaetan, C and X Guyon (2010). *Spatial Statistics and Modeling*. Springer Series in Statistics. New York, NY: Springer. DOI: 10.1007/978-0-387-92257-7.

Gayo-Avello, D (2012). 'No, You Cannot Predict Elections with Twitter'. *IEEE Internet Computing* 16 (6), pp. 91–94. DOI: 10.1109/MIC.2012.137.

Getis, A (2009). 'Spatial Weights Matrices'. *Geographical Analysis* 41 (4), pp. 404–410. DOI: 10.1111/j.1538-4632.2009.00768.x.

Getis, A and J Aldstadt (2004). 'Constructing the Spatial Weights Matrix Using a Local Statistic'. *Geographical Analysis* 34 (2), pp. 130–140. DOI: 10.1353/geo.2004.0002.

Getis, A and J Ord (1992). 'The Analysis of Spatial Association by Use of Distance Statistics'. *Geographical Analysis* 24 (3), pp. 189–206. DOI: 10.1111/j.1538-4632.1992.tb00261.x.

Gilani, Z, L Wang, J Crowcroft, M Almeida and R Farahbakhsh (2016). 'Stweeler: A Framework for Twitter Bot Analysis'. In: *Proceedings of the 25th International Conference Companion on World Wide Web*. Ed. by J Bourdeau, J Hendler and R Nkambou. Montréal: ACM Press, pp. 37–38. DOI: 10.1145/2872518.2889360.

Griffith, D (2006). ‘Hidden Negative Spatial Autocorrelation’. *Journal of Geographical Systems* 8 (4), pp. 335–355. DOI: 10.1007/s10109-006-0034-9.

— (2008). ‘Spatial-Filtering-Based Contributions to a Critique of Geographically Weighted Regression (GWR)’. *Environment and Planning A* 40 (11), pp. 2751–2769. DOI: 10.1068/a38218.

Haustein, S, T Bowman, K Holmberg, A Tsou, C Sugimoto and V Larivi?re (2016). ‘Tweets as Impact Indicators: Examining the Implications of Automated "Bot" Accounts on Twitter’. *Journal of the Association for Information Science and Technology* 67 (1), pp. 232–238. DOI: 10.1002/asi.23456.

Hawelka, B, I Sitko, E Beinat, S Sobolevsky, P Kazakopoulos and C Ratti (2014). ‘Geo-Located Twitter as Proxy for Global Mobility Patterns’. *Cartography and Geographic Information Science* 41 (3), pp. 260–271. DOI: 10.1080/15230406.2014.890072.

Hegarty, M, D Montello, A Richardson, T Ishikawa and K Lovelace (2006). ‘Spatial Abilities at Different Scales: Individual Differences in Aptitude-Test Performance and Spatial-Layout Learning’. *Intelligence* 34 (2), pp. 151–176. DOI: 10.1016/j.intell.2005.09.005.

Hintze, J and R Nelson (1998). ‘Violin Plots: A Box Plot - Density Trace Synergism’. *The American Statistician* 52 (2), pp. 181–184. DOI: 10.1080/00031305.1998.10480559.

Iosa, M, A Fusco, G Morone and S Paolucci (2012). ‘Walking There: Environmental Influence on Walking-Distance Estimation’. *Behavioural Brain Research* 226 (1), pp. 124–132. DOI: 10.1016/j.bbr.2011.09.007.

Jong, P, C Sprenger and F Veen (1984). ‘On Extreme Values of Moran’s I and Geary’s c’. *Geographical Analysis* 16 (1), pp. 17–24. DOI: 10.1111/j.1538-4632.1984.tb00797.x.

Lee, R, S Wakamiya and K Sumiya (2013). ‘Urban area characterization based on crowd behavioral lifelogs over Twitter’. *Personal and Ubiquitous Computing* 17 (4), pp. 605–620. DOI: 10.1007/s00779-012-0510-9.

Legendre, P (1993). ‘Spatial Autocorrelation: Trouble or New Paradigm?’ *Ecology* 74 (6), pp. 1659–1673. DOI: 10.2307/1939924.

Lenormand, M, M Picornell, O Cantu-Ros, A Tugores, T Louail, R Herranz, M Barthelemy, E Frias-Martinez and J Ramasco (2014). ‘Tweets on the Road’. *PLoS ONE* 9, e105407. DOI: 10.1371/journal.pone.0105184.

Longley, P, M Adnan and G Lansley (2015). ‘The Geotemporal Demographics of Twitter Usage’. *Environment and Planning A* 47 (2), pp. 465–484. DOI: 10.1068/a130122p.

Mislove, A, S Lehmann, Y Ahn, J Onnela and J N Rosenquist (2011). ‘Understanding the Demographics of Twitter Users’. In: *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*. Barcelona.

Mukherjee, S, A Sarkar, S Goswami and D Amit Kumar (2016). ‘A Spam Detection Study of Tweets in Indian Healthcare’. *Artificial Intelligent Systems and Machine Learning* 8 (4), pp. 123–127.

Newsome, T, W Walcott and P Smith (1998). ‘Urban Activity Spaces: Illustrations and Application of a Conceptual Model for Integrating the Time and Space Dimensions’. *Transportation* 25 (4), pp. 357–377. DOI: 10.1023/A:1005082827030.

Oliver, M (2010). ‘The Variogram and Kriging’. In: *Handbook of Applied Spatial Analysis*. Ed. by M Fischer and A Getis. Heidelberg: Springer, pp. 319–352. DOI: 10.1007/978-3-642-03647-7_17.

Ord, J and A Getis (1995). 'Local Spatial Autocorrelation Statistics: Distributional Issues and an Application'. *Geographical Analysis* 27 (4), pp. 286–306. DOI: `10.1111/j.1538-4632.1995.tb00912.x`.

— (2001). 'Testing for Local Spatial Autocorrelation in the Presence of Global Autocorrelation'. *Journal of Regional Science* 41 (3), pp. 411–432. DOI: `10.1111/0022-4146.00224`.

— (2012). 'Local Spatial Heteroscedasticity (LOSH)'. *The Annals of Regional Science* 48 (2), pp. 529–539. DOI: `10.1007/s00168-011-0492-y`.

Páez, A and D Scott (2005). 'Spatial Statistics for Urban Analysis: A Review of Techniques with Examples'. *GeoJournal* 61 (1), pp. 53–67. DOI: `10.1007/s10708-005-0877-5`.

Rae, A and A Singleton (2015). 'Putting big data in its place: a Regional Studies and Regional Science perspective'. *Regional Studies, Regional Science* 2 (1), pp. 1–5. DOI: `10.1080/21681376.2014.990678`.

Rai, R., M Balmer, M Rieser, V Vaze, S Schönfelder and K Axhausen (2007). 'Capturing Human Activity Spaces: New Geometries'. *Transportation Research Record* 2021, pp. 70–80. DOI: `10.3141/2021-09`.

Rogerson, P (2015). 'Maximum Getis-Ord Statistic Adjusted for Spatially Autocorrelated Data'. *Geographical Analysis* 47 (1), pp. 20–33. DOI: `10.1111/gean.12055`.

Rogerson, P and P Kedron (2012). 'Optimal Weights for Focused Tests of Clustering Using the Local Moran Statistic'. *Geographical Analysis* 44 (2), pp. 121–133. DOI: `10.1111/j.1538-4632.2012.00840.x`.

Sengstock, C and M Gertz (2012). 'Latent Geographic Feature Extraction from Social Media'. In: *Proceedings of the 20th International Conference on Advances in Geographic Information Systems*. New York, NY: ACM Press, pp. 149–158. DOI: `10.1145/2424321.2424342`.

Shimatani, K (2002). 'Point Processes for Fine-Scale Spatial Genetics and Molecular Ecology'. *Biometrical Journal* 44 (3), pp. 325–352. DOI: `10.1002/1521-4036(200204)44:3<325::AID-BIMJ325>3.0.CO;2-B`.

Shortridge, A (2007). 'Practical Limits of Moran's Autocorrelation Index for Raster Class Maps'. *Computers, Environment and Urban Systems* 31 (3), pp. 362–371. DOI: `10.1016/j.compenvurbsys.2006.07.001`.

Steiger, E, J de Albuquerque and A Zipf (2015a). 'An Advanced Systematic Literature Review on Spatiotemporal Analyses of Twitter Data'. *Transactions in GIS* 19 (6), pp. 809–834. DOI: `10.1111/tgis.12132`.

Steiger, E, R Westerholt, B Resch and A Zipf (2015b). 'Twitter as an Indicator for Whereabouts of People? Correlating Twitter with UK Census Data'. *Computers, Environment and Urban Systems* 54, pp. 255–265. DOI: `10.1016/j.compenvurbsys.2015.09.007`.

Sui, D (2004). 'Tobler's First Law of Geography: A Big Idea for a Small World?' *Annals of the Association of American Geographers* 94 (2), pp. 269–277. DOI: `10.1111/j.1467-8306.2004.09402003.x`.

Tiefelsdorf, M and B Boots (1997). 'A Note on the Extremities of Local Moran's Iis and Their Impact on Global Moran's I'. *Geographical Analysis* 29 (3), pp. 248–257. DOI: `10.1111/j.1538-4632.1997.tb00960.x`.

Tiefelsdorf, M, D Griffith and B Boots (1999). 'A Variance-Stabilizing Coding Scheme for Spatial Link Matrices'. *Environment and Planning A* 31 (1), pp. 165–180. DOI: `10.1068/a310165`.

Xu, M, C Mei and N Yan (2014a). 'A Note on the Null Distribution of the Local Spatial Heteroscedasticity (LOSH) Statistic'. *The Annals of Regional Science* 52 (3), pp. 697–710. DOI: 10.1007/s00168-014-0605-5.

II.2.6 Supporting Information

S1 Dataset. Twitter Sample

Data is available online: https://doi.org/10.1371/journal.pone.0162360.s001.

S2 Dataset. Simulated Data Used in Sections 'Influences on Spatial Autocorrelation' and 'Increased Topological Variability'

Data is available online: https://doi.org/10.1371/journal.pone.0162360.s002.

S3 Dataset. Simulated Data ("Inclusion" and "Large-Scale Perspective") Used in Section 'Influence of Scale Differences on the Numbers of Interactions'

Data is available online: https://doi.org/10.1371/journal.pone.0162360.s003.

S4 Dataset. Simulated Data ("Inclusion" and "Small-Scale Perspective") Used in Section 'Influence of Scale Differences on the Numbers of Interactions'

Data is available online: https://doi.org/10.1371/journal.pone.0162360.s004.

S5 Dataset. Simulated Data ("Overlap" and "Large-Scale Perspective") Used in Section 'Influence of Scale Differences on the Numbers of Interactions'

Data is available online: https://doi.org/10.1371/journal.pone.0162360.s005.

S6 Dataset. Simulated Data ("Overlap" and "Small-Scale Perspective") Used in Section 'Influence of Scale Differences on the Numbers of Interactions'

Data is available online: https://doi.org/10.1371/journal.pone.0162360.s006.

S1 Fig. Spatial Distribution of the Gaussian Attribute Values Across the Single Pattern (Top) and Their Histogram (Bottom)

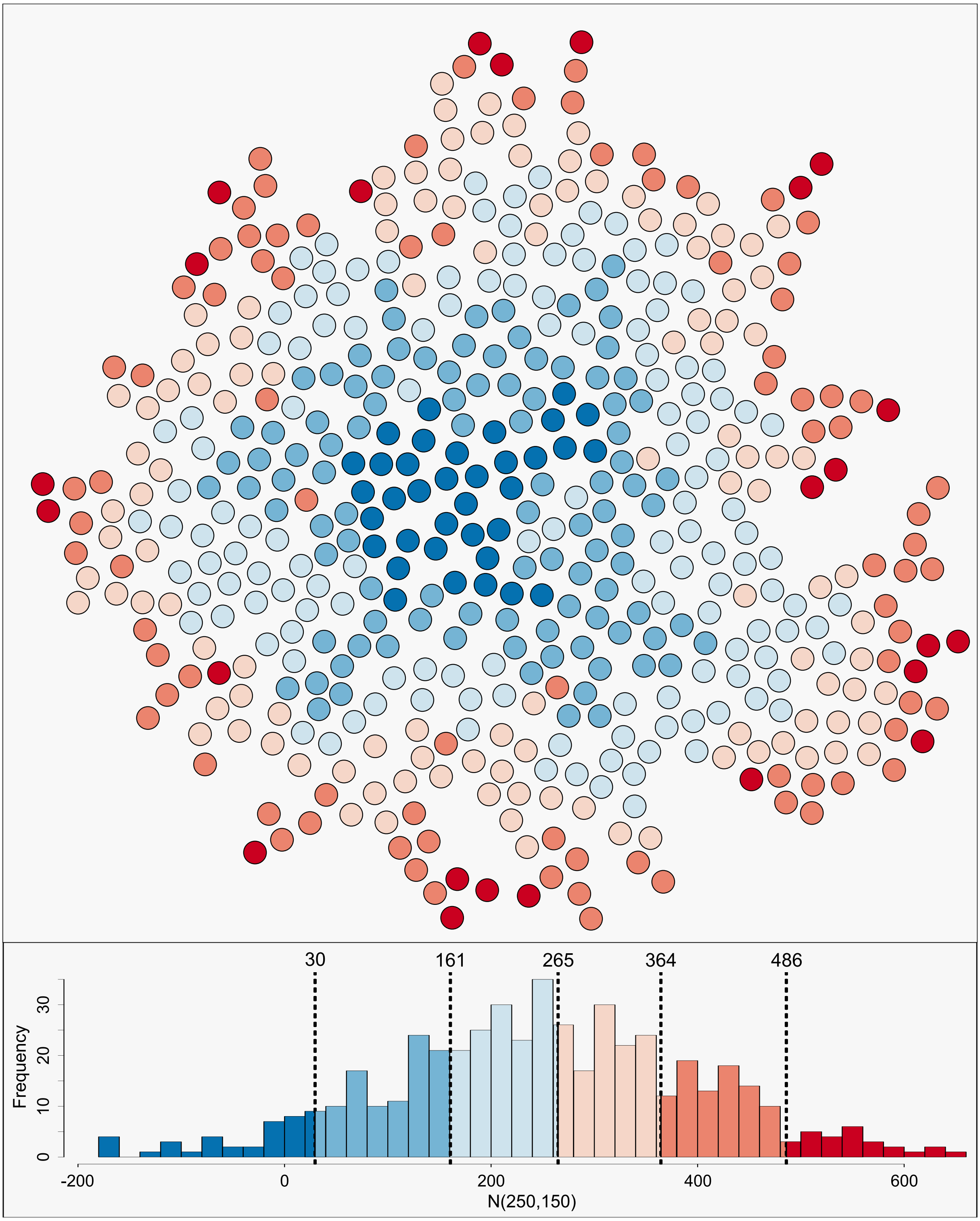


Figure II.2.17: Spatial distribution of the Gaussian attribute values across the single pattern (top) and their histogram (bottom).

S2 Fig. Spatial Distribution of the Gaussian Mixture Across a Combined Pattern (Top) and Their Joint Histogram (Bottom)

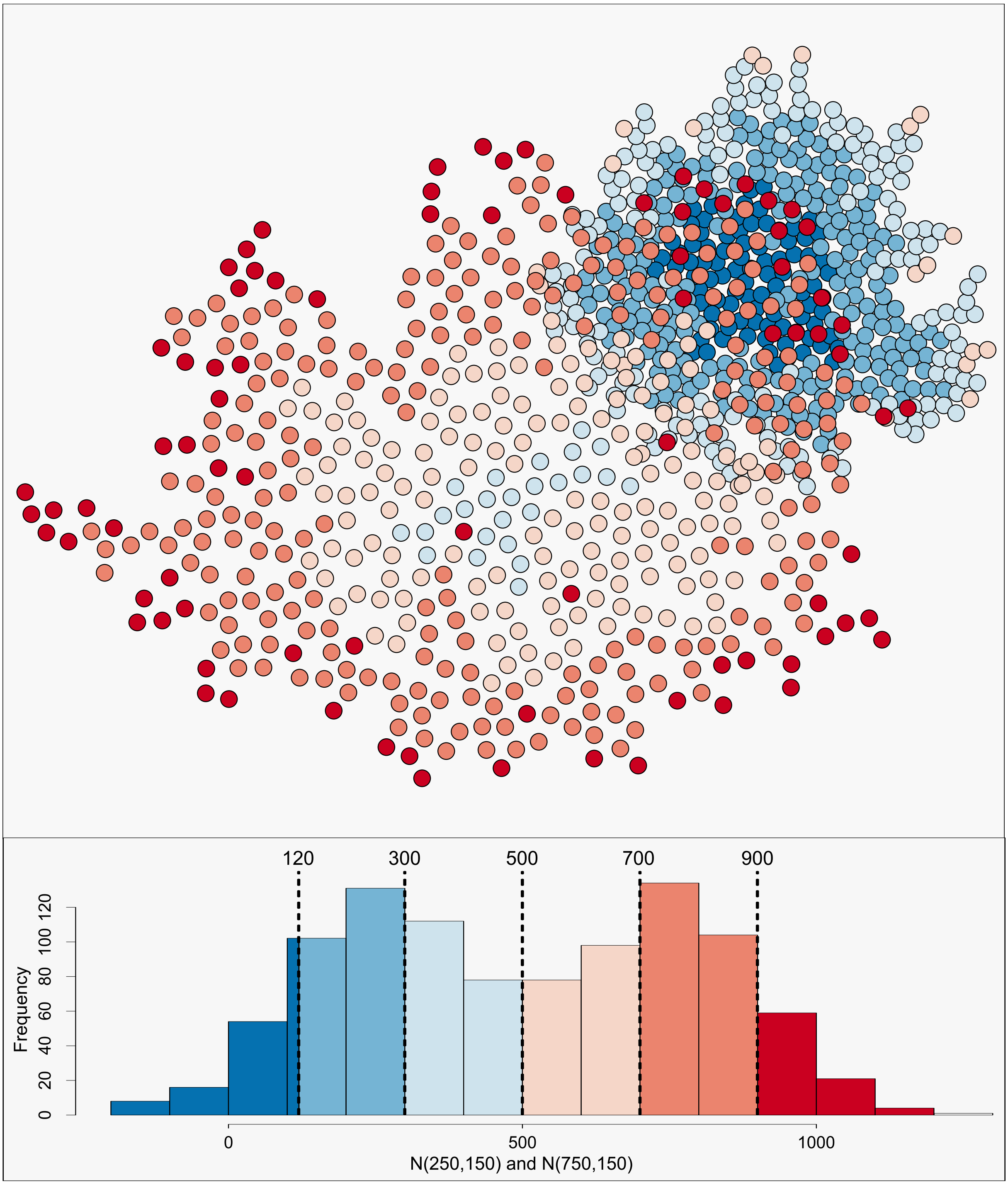


Figure II.2.18: Spatial distribution of the Gaussian mixture across a combined pattern (top) and their joint histogram (bottom).

S3 Fig. Goodness of Fit for the Fitted Functions

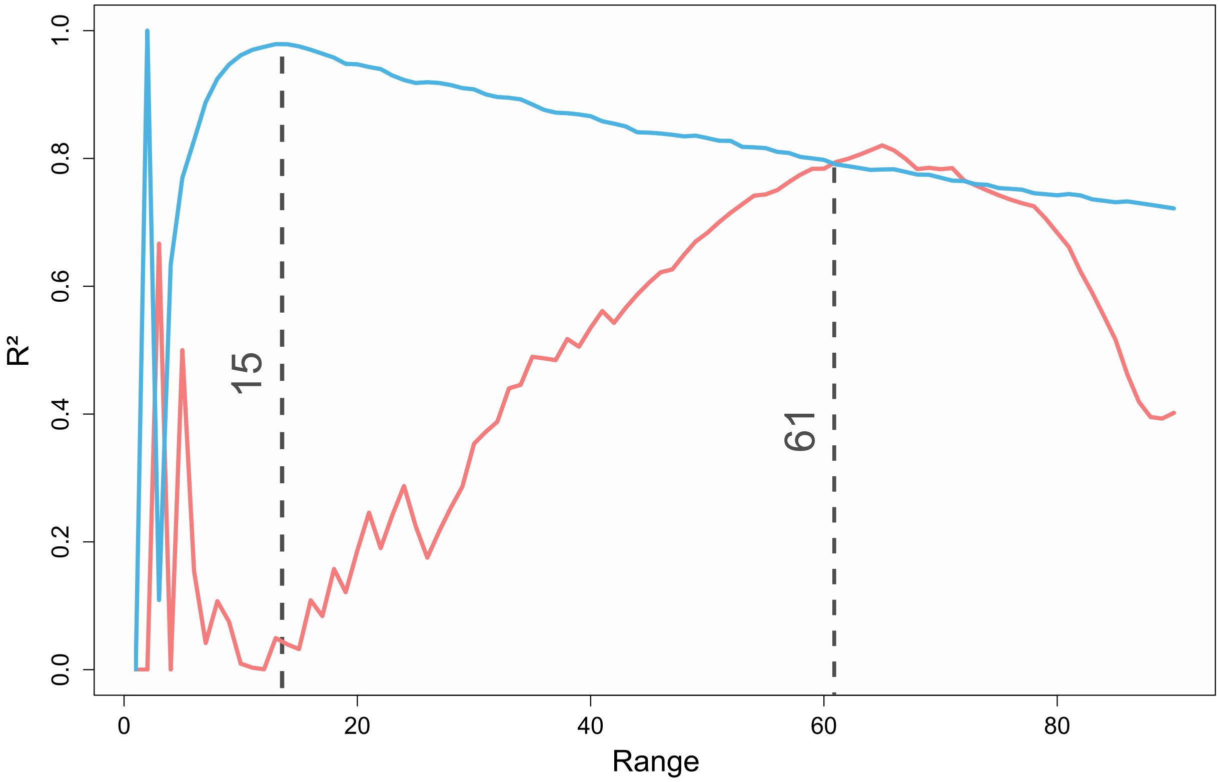


Figure II.2.19: Goodness of fit for the fitted functions. Blue: exponential function; red: linear function. Please read the fits in a cumulative way. The exponential function was evaluated from left to right. That is, the determined optimum at 15 means that the first 15 m of the course follow the respective function. In contrast, the red linear function needs to be read in a reversed order. The trailing scale differences as off 61 m proceed like the fitted function.

S4 Fig. Heat Map of Pairwise Covariance Terms and Semivariogram of Topic Associations

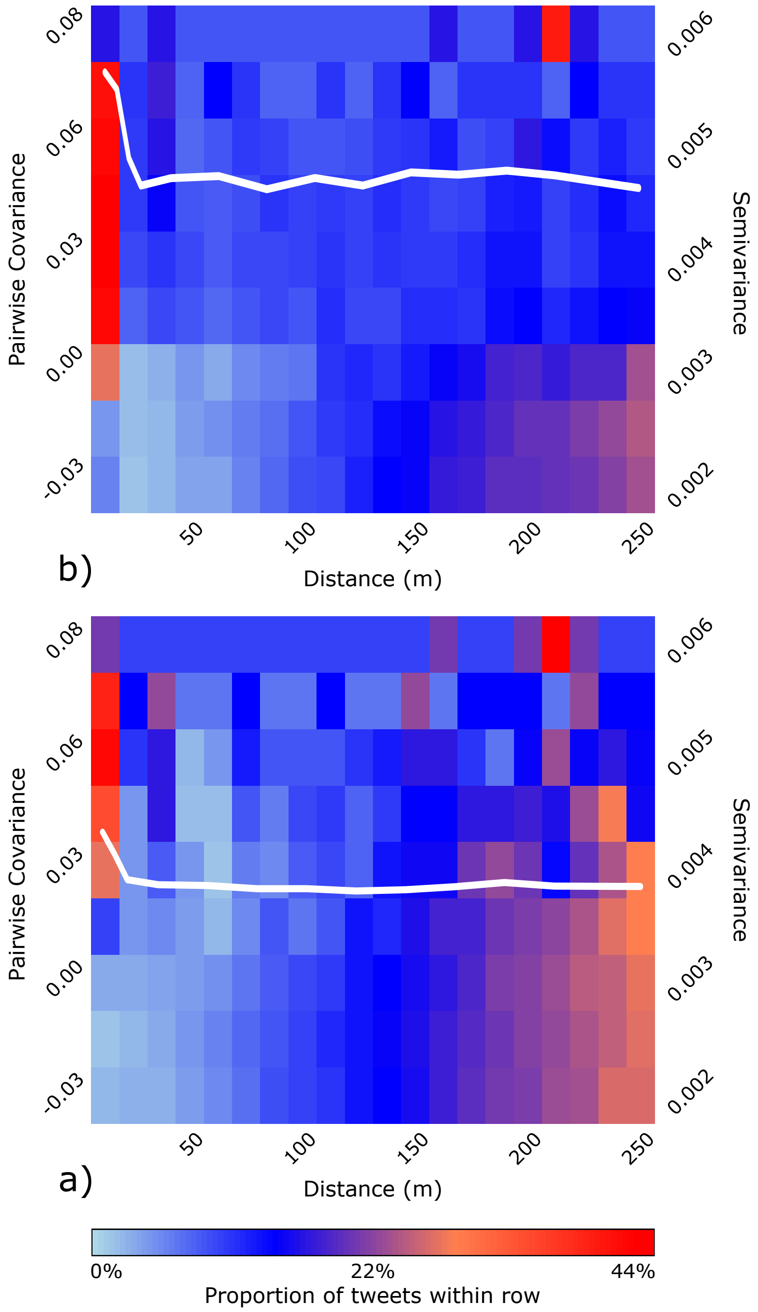


Figure II.2.20: Heat map of pairwise covariance terms and semivariogram of topic associations. The white semivariogram plotted atop of the heat map refers to the right-hand y-axis. The left-hand y-axis is associated with the underlying colour-coded bins of the heat map. This figure is similar to Fig II.2.3, but shows relative heat map values for reasons of comparison (*i. e.*, heat map values are normalised by rows). Part a) contains all tweets while part b) is adjusted for spatial coincident tweets.

II.3 A Statistical Test on the Local Effects of Spatially Structured Variance

Abstract

Spatial variance is an important characteristic of spatial random variables. It describes local deviations from average global conditions and is thus a proxy for spatial heterogeneity. Investigating instability in spatial variance is a useful way of detecting spatial boundaries, analysing the internal structure of spatial clusters and revealing simultaneously acting geographic phenomena. Recently, a corresponding test statistic called 'Local Spatial Heteroscedasticity' (LOSH) has been proposed. This test allows locally heterogeneous regions to be mapped and investigated by comparing them with the global average mean deviation in a dataset. While this test is useful in stationary conditions, its value is limited in a global heterogeneous state. There is a risk that local structures might be overlooked and wrong inferences drawn. In this article, we introduce a test that takes account of global spatial heterogeneity in assessing local spatial effects. The proposed measure, which we call 'Local Spatial Dispersion' (LSD), adapts LOSH to local conditions by omitting global information beyond the range of the local neighbourhood and by keeping the related inferential procedure at a local level. Thereby, the local neighbourhoods might be small and cause small-sample issues. In the view of this, we recommend an empirical Bayesian technique to increase the data that is available for resampling by employing empirical prior knowledge. The usefulness of this approach is demonstrated by applying it to a LiDAR-derived dataset with height differences and by making a comparison with LOSH. Our results show that LSD is uncorrelated with non-spatial variance as well as local spatial autocorrelation. It thus discloses patterns that would be missed by LOSH or indicators of spatial autocorrelation. Furthermore, the empirical outcomes suggest that interpreting LOSH and LSD together, is of greater value than interpreting each of the measures individually. In the given example, local interactions can be statistically detected between variance and spatial patterns in the presence of global structuring, and thus reveal details that might otherwise be overlooked.

Keywords: Spatial analysis, Spatial heterogeneity, Spatial hypothesis testing, Spatial non-stationarity

II.3.1 Introduction

Geographic instability in statistical parameters (called 'spatial heterogeneity', s. Dutilleul and Legendre 1993) has long been of scientific interest. Alexander von Humboldt noted a distinctive geographic patchiness in the 19th century(Sparrow 1999), and Darwin's theory of evolution was largely driven by his recognition of a geographic distribution of phenotypic variants (Jacquez 2010). Currently, scholars from empirical research subjects such as ecology, epidemiology or sociology, leverage knowledge out of spatial heterogeneity to either detect and specify zones of transition (*e. g.*, transitions from terrestrial to aquatic habitats (Turner 1989)), to link local to global processes (Berkes et al. 2006), or to acquire a better understanding of the ecological complexity of urban areas (Cadenasso et al. 2007).

Despite its useful properties, spatial heterogeneity plays a relatively minor role in spatial analysis techniques, which are mostly designed for clustering. Measures of spatial autocorrelation and hot spot techniques are prevalent, and these are used to assess associations within spatial random variables (Getis 2010). In contrast, spatial heterogeneity is often deemed to be a technical nuisance and seldom regarded as a source of valuable information. It either requires a methodological approach (Anselin 1988a; Páez and Scott 2005; Graif and Sampson 2009), or is considered to be reminiscent of large-scale structures that influence local patterns (*e. g.*, Ord and Getis 2001). Spatial heterogeneity indeed undermines the stationarity assumptions that form the basis of many spatial techniques (Gaetan and Guyon 2010, 166 ff.). Thus, regarding it from the standpoint of a nuisance is partly justified.

Nonetheless, spatial heterogeneity often contains useful information. A good illustration of this is the recent investigation of the domiciles of newly arrived migrants from rural areas to Accra, Ghana (Getis 2015). The use of spatial variance as a proxy for spatial heterogeneity allows transitional zones to be detected between the underdeveloped and wealthy districts of the city. Incoming migrants from rural parts of Ghana first settle in these transitional areas after they first arrive in the city. This recent example shows that spatial heterogeneity can supply important information to investigations of complex geographic situations, and lead to useful conclusions of both a theoretical and practical value.

Spatial heterogeneity is also important for the analysis of intrinsically heterogeneous and novel data sources. Social media data, for instance, are sometimes called the ‘big noise’ (Lovelace et al. 2016), because they are characterised by unstable ‘wild variance’ (Jiang 2015). The latter is characterised by an interaction between spatial patterns and variance, which influences analysis results. Westerholt et al. (2015) and Westerholt et al. (2016) recently found that spatial heterogeneity causes type I errors, topological outliers and some further problems that are relevant to the spatial analysis of Twitter data. As a result, many researchers are now investigating social media data in an attempt to mitigate its noisy features (*e. g.*, Sengstock et al. 2013; Lovelace et al. 2016; Steiger et al. 2016b). The investigation of heterogeneity, however, might provide a clue about the spatial perceptions of people and help to characterise the users’ everyday behaviours more accurately. Similar arguments hold true for the data obtained from multi-temporal analysis. The differences between multi-temporal data acquired by ‘Light Detection and Ranging’ (LiDAR) (Fang and Huang 2004; Tian et al. 2014), for example, are a means of detecting heterogeneous changes in surface phenomena. When investigating the LiDAR recordings of landslides (Jaboyedoff et al. 2012), it was found that spatial heterogeneity can provide a wealth of information about significant morphological features like differently shaped earth deposits (Hungr et al. 2014). These two rather different examples demonstrate the potential value of investigating spatial heterogeneity in a number of application scenarios.

Recently, a statistical measure of spatial variance called ‘Local Spatial Heteroscedasticity’ (LOSH; Ord and Getis 2012) was put forward as a means of investigating spatial heterogeneity. LOSH assesses the effects of spatial patterns on the variance of an attribute. It identifies regions in which the local spatial variance deviates from the global average variability. The measure thus reveals and maps structures of the variance that are at least partially global in nature, whereas the weaker structures that are entirely local remain hidden. The latter only feature prominently in local circumstances but remain undetected by the global reference framework of LOSH.

We set out a technique that extends LOSH by making it a measure for the influence of local spatial patterns on local variance. The test, which we call ‘Local Spatial Dispersion’ (LSD), makes it possible to detect whether the local geographic arrangement of random variables increases or reduces the variance. This is carried out in an entirely local manner and takes no account of global characteristics. In addition, we propose an entirely local bootstrapping approach for drawing inferences. Drawing inferences that are only local, however, entails limited amounts of data from small local neighbourhoods. As a means of circumventing the problem of small-size samples within these local subsets, (which particularly arises when adjusting small analytical scales), the inference technique includes an empirical Bayesian prediction of additional synthetic local data. The usefulness of the proposed technique can be demonstrated by applying it to a high-resolution 3D change detection dataset. The data is derived from a long-term 'automatic terrestrial laser scanning station' (ATLS) that covers a slow-moving landslide in Gresten, Austria and provides a useful scenario because it contains both a distinct global structure and additional local patterns.

The paper starts with a detailed review of spatial heterogeneity and spatial heteroscedasticity, and includes a brief discussion of related statistical methodologies (Section II.3.2). Following this outline, LOSH and our proposed measure are introduced (Sections II.3.3 and II.3.4). Then, there is a Bayesian prediction of residuals as well as the bootstrap method for developing predictive models (Section II.3.5), before the empirical results are discussed (Session II.3.6) and the final conclusions are drawn (Section II.3.7).

II.3.2 Related Work: Spatial Heterogeneity and Spatial Heteroscedasticity

Spatial heterogeneity refers to non-uniformity and instability in geographic random variables (Dutilleul and Legendre 1993). Their corresponding zones "where variables change rapidly" (Jacquez 2010, p. 210) are of scientific and practical interest. They can i) represent regions of habitat use and ecological interactions (Fagan et al. 1999; Lohrer et al. 2013), ii) assist in testing ethno-racial diversity (Abascal and Baldassarri 2015; Legewie and Schaeffer 2016), or iii) touch on the question of disease transmission (Grillet et al. 2010; Perkins et al. 2013). Spatial heterogeneity is also important for urban studies. Metaphorically speaking, just as prices in economic markets do not 'glide' but often 'leap' (Mandelbrot and Hudson 2004), urban regions tend to be heterogeneous and disruptive in nature (Cadenasso et al. 2007). Analysing heterogeneity is thus of crucial importance for understanding urban social processes, while an analysis of boundaries can assist in distinguishing subpopulations. Furthermore, spatial heterogeneity provides guidance in testing assumptions and theories about the relationships between variables (Jacquez 2010), as well as assisting in data aggregation and dynamic modelling (Anselin 1990).

Different structural types of spatial heterogeneity are distinguishable. These are characterised by their causal origins, maintenance mechanisms, spatial structures, and functional and temporal dynamics (Strayer et al. 2003). Other more technical distinguishing factors include the types of investigated variables (Wagner and Fortin 2005), the underlying spatial indexes (spatially discrete vs continuous; Anselin 2010) and even the methodological perspectives that researchers adopt (dynamic modelling vs hypothesis testing; Fagan et al. 2003). In structural terms, heterogeneous zones sometimes condense to thin and crisp boundaries, while they can also appear fuzzy (Jacquez et al. 2000).

For functional purposes, heterogeneous zones can act as semipermeable filters or conduits and as devices from which spatial processes either originate or where they are impeded (Forman 1995). Steep gradients or threshold conditions, at which variable states change suddenly, can also be found in heterogeneous areas (Fagan et al. 2003). These characteristics allow the spatial heterogeneity to exert a short or long-range influence on dynamic processes (Fagan et al. 2003). Sometimes these influences get strengthened by the interrelations between the effects mentioned earlier, especially by the interplay between the structural and functional characteristics (Laurance et al. 2001). Hence, the various features, together with the number of functional influences, show the importance of investigating spatially heterogeneous zones.

Techniques to detect heterogeneous zones (especially crisp boundaries), first appeared in image processing. Some corresponding methods have been designed for segmenting synthetic images, although they are not capable of depicting dynamic real-world systems in their entirety (Goovaerts 2010). A range of more suitable methods has thus evolved, including techniques based on moving split-windows (Fortin 1994; Fortin 1999; Kent et al. 1997; Kent et al. 2006), first-order derivatives ('Wombling' Womble 1951; Barbujani et al. 1989; Gelfand and Banerjee 2015), second-order derivatives (Fagan et al. 2003; Lillesand et al. 2015), spatially constrained clustering (Jacquez et al. 2000; Patil et al. 2006; Bravo and Weber 2011), fuzzy set modelling (Arnot and Fisher 2007; Fisher and Robinson 2014), wavelets (Csillag and Sándor 2002; Keitt and Urban 2005; Ye et al. 2015) and several further parametric as well as non-parametric techniques (Jacquez et al. 2008; Wang et al. 2016a). Another closely related research field is concerned with integrating spatial heterogeneity with quantitative models. The respective approaches include the following: hierarchical and Bayesian concepts (Lee and Mitchell 2012; Anderson et al. 2014; Hanson et al. 2015), geostatistical techniques (Garrigues et al. 2006; Goovaerts 2008; Hu et al. 2015), extensions to global spatial regression methods (Anselin 2001), and the local geographically-weighted regression approach (Fotheringham et al. 1996; Fotheringham et al. 2002; Brunsdon et al. 1998).

In statistics, heterogeneity either refers to single parameters (*e. g.*, mean or variance) or to complete distributions (Kolasa and Rollo 1991; Dutilleul and Legendre 1993). Spatial heterogeneity can be decomposed into a deterministic, random and chaotic parts (Dutilleul 2011). The deterministic part reflects the varying average component ('large-scale trend'), while the latter two together reflect variations caused by variance instability ('unstable mean deviations') and spatial autocorrelation ('variation through interaction'). It is necessary to differentiate between heterogeneities in different parameters and also between the three parts outlined above, to achieve a thorough understanding of the behaviour of random variables and related phenomena.

In spatial analysis, varying means are analysed by hot spot techniques like the G and the O-statistic (Getis and Ord 1992; Ord and Getis 1995; Ord and Getis 2001). By analogy, variations caused by autocorrelation are analysed through local measures of spatial autocorrelation like the 'Local Indicators of Spatial Association' (LISA, Anselin 1995). While these cases have been widely investigated, there has been comparatively little research on variability in the variance (called 'spatial heteroscedasticity'; Dutilleul and Legendre 1993). Roughly speaking, spatial heteroscedasticity refers to 'wild variance' (Jiang 2015). Ord and Getis (2012) recently put forward a local measure called 'Local Spatial Heteroscedasticity' (LOSH), which assesses spatial structure in variance and is akin to a spatial χ^2 test. Xu et al. (2014a) investigated the distributional properties of LOSH and found that the χ^2 approximation proposed by Ord and Getis (2012) is not always suitable, and that a Monte Carlo bootstrap should be used instead.

LOSH is ideally suited to detecting boundary-like sub-regions lying *between* homogeneous regimes. However, it cannot describe in detail how local spatial arrangements of random variables in place affect the heterogeneity *within* the individual sub-regions. This is where our study is able to make a contribution to the field because it supplements LOSH by conducting a test involving the local spatial microstructure of the variance of georeferenced random variables.

II.3.3 Local Spatial Heteroscedasticity (LOSH)

The LOSH measure (Ord and Getis 2012) calculates local deviations from the global average variance. It is derived from the hot spot technique called ‘G-statistic’ (Getis and Ord 1992; Ord and Getis 1995) and allows boundaries and hot spots of high variability to be detected. LOSH tests the following hypotheses:

> H_0^{LOSH}: *The variance in a region does not deviate markedly from its global average.*
>
> H_1^{LOSH}: *The variance in a region deviates from overall variance homogeneity.*

LOSH proceeds as follows: In the first stage, residuals that describe the difference between an attribute value and its local spatially weighted mean value are estimated. In each location, the spatially weighted averages of these residuals are then compared with their global counterpart. The latter is estimated with data from all locations by randomising the spatial pattern at the same time. The calculated ratio of these two averages then forms a test statistic from which inferences can be drawn. Let X be a set of n real-valued random variables X_i referenced in an index set $\mathcal{N} = \{1, \ldots, n\}$ that indicates discrete spatial units. By analogy, let $\mathcal{N}_i = \{j \in \mathcal{N} \mid \exists\, i \in \mathcal{N}\colon w_{ij} \neq 0\}$ be the local neighbourhood of spatial unit i that can be defined by suitable spatial weights, whereby the choice of the latter depends on the application scenario. These weights, which are given by W, a symmetric matrix of elements w_{ij} that map pairs of spatial units to positive real weights, are a mathematical representation of the geographical layout of the investigated region (Dray 2011). The weight matrix thereby limits the entire geographic layout to those geographic features that are relevant to a particular phenomenon under study. These weights can be of an arbitrary shape (s. Bavaud 2014, for an overview) and no specific form is required for the remainder. LOSH (H_i is the notation for LOSH chosen by (Ord and Getis 2012)) then reads as

$$\begin{aligned} H_i &= \frac{\sum_{j\in\mathcal{N}} w_{ij} \cdot |e_j|^a}{h_1 \cdot \sum_{j\in\mathcal{N}} w_{ij}}, & e_j &= x_j - \bar{x}_j, \\ \bar{x}_j &= \frac{\sum_{k\in\mathcal{N}_j} w_{jk} \cdot x_k}{\sum_{k\in\mathcal{N}_j} w_{jk}}, & h_1 &= \frac{\sum_{j\in\mathcal{N}} |e_j|^a}{n} \end{aligned} \tag{II.3.1}$$

where e_j is a residual about a local spatially weighted mean $\bar{x}_j$ and h_1 is the overall average residual estimated from all the spatial units in the region. Note that $\mathcal{N}_j$ is the neighbourhood around unit j, that is defined analogously to $\mathcal{N}_i$. Exponent a allows different types of mean deviations to be investigated. For the remainder of this paper, we adjust $a = 2$ and confine the discussion to a measure of variance.

An inference about LOSH assumes random permutations of the residuals. When an average residual h_1 is employed, it thus makes clear that LOSH assumes weak stationarity in the null hypothesis. The successful detection of a local pattern thus depends on the global reasonability of h_1. Through a random permutation of the residuals, the statistic obtains an expected value of $E[H_i] = 1$ and has a variance of

$$\begin{aligned} V_i[H_i] = \frac{1}{n-1} \cdot \left(\frac{1}{h_1 \sum_{j\in\mathcal{N}} w_{ij}}\right)^2 \cdot \left[\frac{1}{n}\left(\sum_{j\in\mathcal{N}} |e_j|^{2a} - \left[\sum_{j\in\mathcal{N}} |e_j|^a\right]^2\right)\right] \\ \cdot \left(n \sum_{j\in\mathcal{N}} w_{ij}^2 - \left(\sum_{j\in\mathcal{N}} w_{ij}\right)^2\right). \end{aligned} \tag{II.3.2}$$

Ord and Getis (2012) propose an adjusted χ^2 approximation to the null distribution as a parametric solution to statistical inference. The χ^2 distribution stems from the design of the statistic as a spatialized variant of the classic χ^2 test for testing deviations from a hypothesised variance. This is seen by writing out the individual terms of the sum from Equation II.3.1:

$$H_i = \frac{w_{i1}}{\sum_{j \in \mathbb{N}} w_{ij}} \frac{|e_1|^2}{h_1} + \cdots + \frac{w_{ij}}{\sum_{j \in \mathbb{N}} w_{ij}} \frac{|e_j|^2}{h_1} + \cdots + \frac{w_{in}}{\sum_{j \in \mathbb{N}} w_{ij}} \frac{|e_n|^2}{h_1}. \tag{II.3.3}$$

Variable h_1 is the hypothesised variance and the summands are (spatially weighted) squared standardised residuals. Under normality constraints, these are χ^2 with one degree of freedom. Their sum is then χ^2 with additive degrees of freedom (Cochran 1934). On the basis of the findings from Box (1953), Ord and Getis (2012) adjust LOSH to take better account of non-normality by including the empirical variance V_i. This matches the χ^2 approximation to the observed outcomes and controls the shape of the reference distribution. The skew and the excess kurtosis of the reference distribution are given by $\gamma_1 = 2\sqrt{V_i}$ and $\gamma_2 = 6V_i$, and the test statistic is $Z_i = 2H_i/V_i$ with $2/V_i$ degrees of freedom. However, Xu et al. (2014a) found deviations between empirical distributions obtained from data and the adjusted approximation outlined above. These even occur with normal variables, which is why Xu et al. (2014a) suggest adopting a nonparametric bootstrap procedure instead.

II.3.4 Local Spatial Dispersion (LSD)

Instead of comparing local regions with a global average like LOSH, the proposed measure LSD is concerned with the effect of the local spatial pattern on local variances. The underlying assumption is that the way random variables are arranged geographically increases or reduces the variance, or else is unrelated to its characterisation. The measure is only defined in a *local* context and does not take account of global information. The same principle also applies for the related inference procedure, which is conducted locally.

The proposed LSD is useful when a dataset comprises statistically differing sub-regions or when spatially coexisting phenomena are observed. However, global information such as the average residual h_1 is not meaningful in these circumstances. This means the LOSH approach causes problems because it is unrealistic to assume there is weak stationarity in these cases. Instead, the variance patterns might be strongly interacting with the geographic layout locally, although they might not be recognised when a global comparison is made with sub-regions that show a stronger dispersal. Thus LOSH cannot be employed to assess entirely local effects and an entirely local measure of spatial variance, such as LSD, can prove to be useful.

II.3.4.1 Hypotheses

The proposed test determines whether the local spatial arrangement of random variables increases or reduces the local variance. The following two hypotheses for LSD are formulated:

H_0^{LSD}: *The local geographic layout has no systematic effect on the variance.*

H_1^{LSD}: *The local geographic layout causes local over- or underdispersion.*

The null model assumes that the local variance is unrelated to the geographic arrangement. If the null is accepted, it means that the investigated data gives no indication that geographical factors are responsible for the variance effects. Note that variability can still be related to its particular location. The average level of variability is still treated as a function of location. This is achieved through a local average residual h_i (see Equation II.3.5). However, LSD tests the local spatial influence on the dispersal behaviour above the general local variability level. In conceptual terms, the hypothesis testing scheme of H_0^{LSD} and H_1^{LSD} derives from a linear autoregressive framework. Let $E_i = (|e_j|^a)_j$ with $j \in \mathcal{N}_i$ be a vector of exponentiated residuals from a local neighbourhood i with e_j as defined in Equation II.3.1. Let $\alpha_i = E\left[|e_j|^a\right]$ be the expected (non-geographic) exponentiated residual within a local neighbourhood i. The two presented hypotheses can be derived from a linear regression model:

$$\mathfrak{e}_i = \alpha_i + \rho_i W_i^T E_i + \varepsilon_i, \tag{II.3.4}$$

where $\mathfrak{e}_i$ denotes the mean deviation influenced by geographical factors, ε_i captures the regression residuals and W_i is the vector of spatial weights for spatial unit i. The null model occurs when the coefficient ρ_i is close to zero. Hence, LSD tests to what degree this coefficient deviates from zero. If a left-side test is conducted, the alternative model represents a significantly negative ρ_i. Its acceptance thus means that the geographic arrangement, as defined through W, reduces the variance more than it would be the case when geographical factors have no effect. By analogy, acceptance of the alternative in a test on the right-side indicates a significantly positive magnitude of ρ_i, which means that the local geography increases the variability within the random variables. The hypotheses outlined here are thus useful devices to test the role of geographic layout in the local dispersal behaviour of the spatial random variables.

II.3.4.2 Mathematical Definition

The LSD measure is formulated mathematically as a ratio of the spatially weighted local residuals and their own spatially randomised local average. Therefore, LSD is given by

$$LSD_i = \frac{\sum_{j \in \mathcal{N}} w_{ij} \cdot |e_j|^a}{h_i \cdot \sum_{j \in \mathcal{N}} w_{ij}}, \quad h_i = \frac{\sum_{j \in \mathcal{N}_i} |e_j|^a}{n_i} \tag{II.3.5}$$

where n_i denotes the cardinality of $\mathcal{N}_i$ and h_i is the local mean residual. Residuals e_j are as defined in Equation II.3.1. The term h_i is a replacement of h_1 and allows a strictly local analysis to be conducted. The datasets can thus be heterogeneous with regard to mean and variance. This important difference from LOSH is further illustrated through the relationship between LOSH and LSD (Appendix II.3.9.1):

$$LSD_i = \frac{H_i \cdot h_1}{h_i} \tag{II.3.6}$$

Equation II.3.6 shows that LSD is a rescaled version of LOSH. Whenever h_i equals h_1, LOSH and LSD are equivalent. This is the case when the local variability equals the global average dispersal behaviour. LSD is particularly valuable when $h_i < h_1$, because LOSH tends to overlook these kinds of weak local structures. In contrast, LSD adapts to specific local conditions and enables truly local variance patterns to be investigated. On the contrary, local deviations detected by LOSH are, at least in part, caused by global instability in the first two moments.

An intrinsically local perspective of LSD is useful in a wide range of situations: i) it can be adopted to describe variegated geographic phenomena occurring at the same time; ii) it allows regions with similar spatial dispersal mechanisms to be revealed beyond the variance magnitudes; iii) it can support in constructing hypotheses regarding the causal mechanisms of phenomena that are spatially coincident; and

iv) it is a diagnostic tool for investigating local non-stationarities. Interpreting LSD and LOSH together should provide a clearer insight into spatial variance patterns: LOSH discloses and maps the overall global variance volatility including distinctive boundaries, whereas LSD is able to discover the local patterning mechanisms that influence heterogeneity in a given place. Section II.3.6 demonstrates some of these possible uses.

II.3.5 Inference Procedure

Two issues complicate the task of making inferences about LSD: potential deviations from normality and the constraint of having to keep the inference local. In the case of normal attributes, LSD can technically be evaluated as a χ^2 test, even though the mean and variance might vary (Walck 2007, p. 38). However, in the light of the results of (Xu et al. 2014a), we do not want to restrict the test to normal populations that seldom occur in real geographic conditions. Furthermore, the intended local nature of an inference approach might cause problems by the small-size local samples. This is particularly the case when the analytical scale is small. In such cases, there is a serious lack of data available for local resampling and bootstrap distributions are unreliable. The χ^2 approach is thus not applicable and a different inferential strategy is required.

A two-step approach is put forward as a means of overcoming this difficulty:

1. A Bayesian prediction of synthetic data to increase the size of the local database through
   (a) determining suitable prior distributions and
   (b) a Bayesian updating for adjusting priors to local conditions.
2. Arranging of local bootstrap distributions using the data from step 1.

The Bayesian approach in the first stage is used to boosting the amount of available data. The purpose of this is to predict additional local mean values, from which auxiliary residuals can be generated. These can then be plugged into LSD during the Monte Carlo iterations in the bootstrap. The second stage describes the final estimation of a reference distribution that is used for inference purposes. The following sub-sections outline these two stages in more detail.

II.3.5.1 Bayesian Mean Prediction

The first part in the inferential approach is to supply the available local subsets with additional information. This is carried out by predicting the synthetic mean values that are used for drawing additional local residuals. The mean estimation is subject to the central limit theorem. This allows us to exploit the advantage of well-known a priori knowledge about the underlying distributional characteristics of mean estimations. Arithmetic means converge to normal distributions. Predicting the means is thus conceptually simpler than drawing the residuals, and for this reason, we have chosen to follow this path rather than predicting residuals directly.

The synthetic means are constructed through a semi-global empirical Bayesian procedure that takes advantage of two sources of information: global information from the overall dataset and local information from the neighbourhoods under consideration. Our proposed approach utilises the observed sampling variability of all the observed mean estimations as prior belief. This global prior reflects strongly averaged information. Hence, the prior belief is further adapted to local conditions by taking account of the local features. The latter step mitigates the global averaging and fits the distribution better to the local conditions in a particular location. In other words, the outlined two-step approach reduces the risk of adapting to local situations too far by taking into account the global setting (note that observed data might represent outlier situations). At the same time, the approach does not entirely rely on global average information.

The partial inclusion of global information contradicts the stated objectives of LSD. However, the use of global data in the auxiliary Bayesian stage, which precedes the arranging of bootstrap distributions, is a pragmatic compromise and its influence should be kept to a minimum. Apart from predicting means, the global information is not transferred to other parts of the inference procedure such as the bootstrap. The alternative to using global information would be an objective Bayesian approach with an uninformed prior. However, this could result in an excessively overfitted predictive posterior distribution as such approach implies only using local information. In other words, the problems of uninformed objective priors parallel those of local bootstrapping without generating any additional information. An objective Bayesian approach would thus not address the two major issues outlined earlier. The following two sub-sections describe the design of the prior distribution and of the updating step.

II.3.5.2 An Informed Prior

The first stage of the Bayesian predictive approach is to construct a prior that models previous knowledge about the sampling variability of local spatial mean values. The prior must maintain realism, but, at the same time, it should not interfere with the likelihood of the local data that is used in the posterior. That latter likelihood will be obtained from information from the neighbourhood of interest, which must thus then be kept for the updating step. The dual use of data might otherwise lead to a dominant prior that drives the posterior too far, especially with small datasets (Berger 2006; Darnieder 2011; Gelman et al. 2013). The dataset is therefore subsetted. In addition to $\mathcal{N}$ and $\mathcal{N}_i$, we define

$$\mathcal{N}_{i+} = \{k \in \mathcal{N} \mid \exists j \in \mathcal{N}_i \colon w_{jk} \neq 0\}, \quad \mathcal{N}_i \subseteq \mathcal{N}_{i+} \subseteq \mathcal{N}, \tag{II.3.7a}$$

$$\mathcal{A}_i = \mathcal{N} \setminus \mathcal{N}_{i+}. \tag{II.3.7b}$$

Subset $\mathcal{N}_{i+}$ (Equation II.3.7a) includes the neighbours of the neighbours of unit i. Set $\mathcal{A}_i$ (Equation II.3.7b) contains all the units outside the extended neighbourhood $\mathcal{N}_{i+}$. Figure II.3.1 illustrates these subsets.

Constructing an informed prior requires *a priori* distributional knowledge. While making the allowance for global non-stationarity, it is not guaranteed that the underlying random variables X_j will be distributed in an identical manner. However, through the central limit theorem and assuming the sample size to be reasonably large, it can be assumed that the spatially weighted mean values $Y_j = \sum_{k \in \mathcal{N}_j} w_{jk} x_k / \sum_{k \in \mathcal{N}_j} w_{jk}$ are approximately normal. We thus have $Y_j \sim N(\mu_{X_j}, a_j \sigma^2_{X_j})$, where μ_{X_j} and $\sigma^2_{X_j}$ are the unknown expectation and variance of the variates $X_{\mathcal{N}_j}$ (*i. e.*, the variates from $\mathcal{N}_j$). The factor $a_j = \sum_{k \in \mathcal{N}_j} w^2_{jk} / W^2_j$ (see Appendix II.3.9.2) reflects the geographic constraints from the spatial weights matrix W. We can ignore this latter constant for the moment but will need it later in the bootstrap.

We seek to predict the parameters μ_{X_j} and $\sigma^2_{X_j}$. It must be remembered that the prior should be backed up by a sufficient amount of data. Instead of estimating the parameters multiple times from small neighbourhoods, our aim is to combine all the information from $\mathcal{A}_i$. Since $\mathcal{A}_i$ varies across locations,

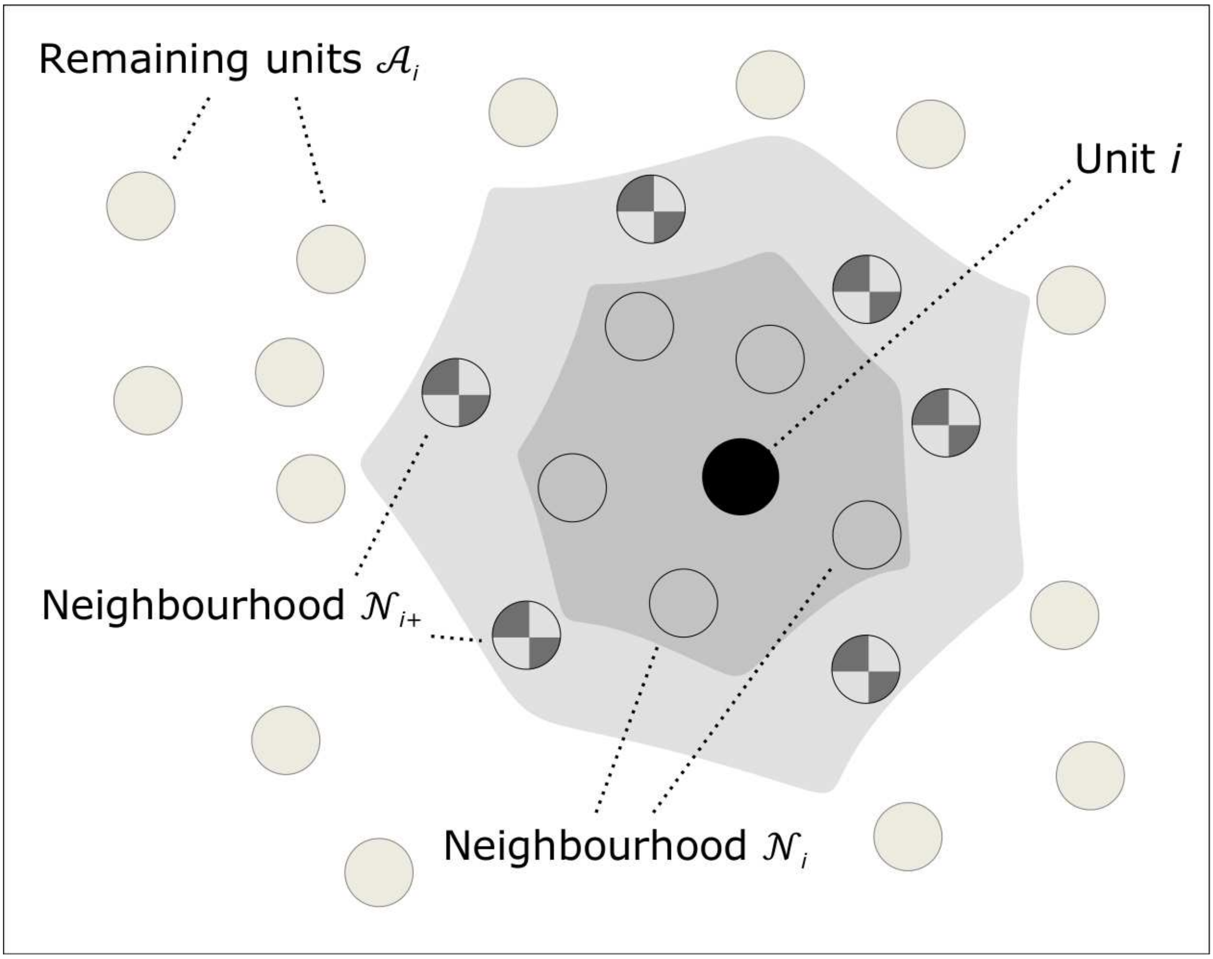

Remaining units $\mathcal{A}_i$
Unit i
Neighbourhood $\mathcal{N}_{i+}$
Neighbourhood $\mathcal{N}_i$

Figure II.3.1: Schematic illustration of region $\mathcal{N}$ separated into $\mathcal{N}_i$, $\mathcal{N}_{i+}$ and $\mathcal{A}_i$.

individual priors must be obtained for each neighbourhood. The combined mean and variance estimators are given by (see Appendix II.3.9.3):

$$\bar{x}_c = \frac{\sum_{j\in\mathcal{A}_i} n_j \bar{x}_j}{\sum_{j\in\mathcal{A}_i} n_j} \quad \text{and} \quad s_c^2 = \frac{\sum_{j\in\mathcal{A}_i} (n_j - 1)\left(s_j^2 + \bar{x}_j^2\right)}{\left(\sum_{j\in\mathcal{A}_i} n_j\right) - n_{\mathcal{A}_i}} - \bar{x}_c^2 \tag{II.3.8}$$

These estimators account for mutually overlapping spatial neighbourhoods. Variable $n_{\mathcal{A}_i}$ is the cardinality of $\mathcal{A}_i$ and subscript c illustrates the combinatorial nature of the proposed estimators from Equation II.3.8.

The prior is the product of the two marginal densities of mean and variance outlined above. The mean of Gaussian random variables Y_j is itself a normal random variable centred on $\mu_0 = \bar{x}_c$ and depends on knowledge of the variance:

$$\mu_{X_j} \mid \sigma_{X_j}^2 \sim N\left(\mu_0, \sigma_0 = \frac{\sigma_{X_j}^2}{n_i}\right) \tag{II.3.9}$$

Technical, but non-substantive parameters (*i. e.*, hyperparameters) are indicated by subscript 0. Variable n_i gives the measurement scale of the neighbourhood i of interest. We use n_i rather than the scale that is actually associated with $\bar{x}_c$ to increase the realism of the prior. The much larger cardinality of $\mathcal{A}_i$ would otherwise cause the prior to be underdispersed. The influence of the prior on predictions could

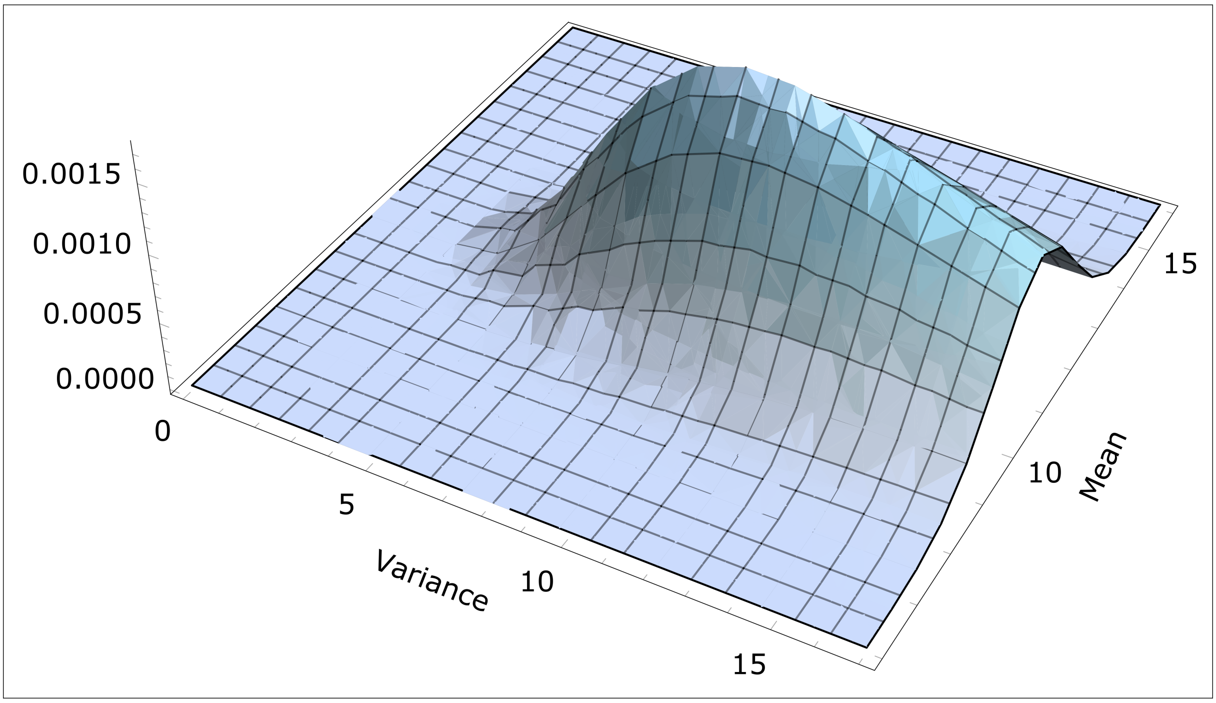


Figure II.3.2: Illustration of the prior density for $n_i = 10$, $\mu_0 = 11$, $\upsilon_0 = 5$ and $\tau_0^2 = 16$.

then become overly dominant. Employing n_i instead, is a means of matching the prior scale to that of the neighbourhood of interest and is thus more appropriate.

The variance $\sigma^2_{X_j}$ follows a normal scaled inverse-chi-squared distribution (Gelman et al. 2013, 67 f.). This results from the χ^2-distributed scaled ratio of the sample variance to the variance of the population:

$$\frac{(n_i - 1) \cdot s_c^2}{\sigma^2_{X_j}} \sim \chi^2_{n_i - 1} \quad \Longrightarrow \quad \sigma^2_{X_j} \sim \chi^{-2}_{\text{scaled}}\left(\upsilon_0, \tau_0^2\right) \tag{II.3.10}$$

As in the case of the mean, the degrees of freedom υ_0 is adjusted to $n_i - 1$ instead of $\left(\sum_{i=1}^{n} n_i\right) - n$, as it is necessary for the prior to be informative about predicting data for $\mathcal{N}_i$ rather than $\mathcal{A}_i$. The scale parameter τ_0^2 is equal to the variance estimate s_c^2.

A combination of the two marginal densities from Equations II.3.8 and II.3.9 yields the prior (see Appendix II.3.9.4)

$$\pi\left(\mu, \sigma^2\right) \propto \frac{1}{\sigma^{3+\upsilon_0}} \cdot \exp\left(-\frac{n_i\left(\mu - \mu_0\right)^2 - \upsilon_0 \tau_0^2}{2\sigma^2}\right). \tag{II.3.11}$$

This prior represents the non-spatial global *a priori* belief about mean values estimated from samples of size n_i. It thus represents information about the variability of mean estimations from across the entire study area beyond location i of interest. Figure II.3.2 provides a parameterised illustration of the constructed prior density.

II.3.5.3 Posterior Distribution

The posterior combines the prior with the likelihood of the observed local spatial mean value $Y_i \sim N\left(\mu_{X_i}, a_i \sigma^2_{X_i}\right)$. Our aim is to predict suitable values for μ_{X_i} and $\sigma^2_{X_i}$. These parameters specify the final Gaussian from which the additional means are drawn. Constant a_i is, again, fixed because it is a non-random property of the neighbourhood of interest. The respective posterior follows a normal scaled inverse-chi-squared distribution and yields (see Appendix II.3.9.5)

$$f\left(\mu_{X_i}, \sigma^2_{X_i}\right) \propto \frac{1}{\sigma^{4+\upsilon_0}_{X_i}} \cdot \exp\left(-\frac{n_i\left(\mu_{X_i}-\mu_0\right)^2 + \left(Y_i - \mu_{X_i}\right)^2 + \upsilon_0\tau_0^2}{2\sigma^2_{X_i}}\right). \tag{II.3.12}$$

Drawing values for μ_{X_i} and $\sigma^2_{X_i}$ requires deriving the conditional posterior $\mu_{X_i} \mid \sigma^2_{X_i}, Y_i$ and, since this in turn requires a known $\sigma^2_{X_i}$, the corresponding marginal posterior $\sigma^2_{X_i} \mid Y_i$. By building on the results obtained from Gelman et al. (2013), we derive

$$\mu_{X_i} \mid \sigma^2_{X_i}, Y_i \sim N\left(\frac{\mu_0 + Y_i}{2}, \frac{\sigma^2_{X_i}}{2n_i}\right), \tag{II.3.13a}$$

$$\sigma^2_{X_i} \mid Y_i \sim \chi^{-2}_{\text{scaled}}\left(\upsilon_0 + n_i, \tilde{\tau}^2\right) \quad \text{and} \quad \tilde{\tau}^2 = \frac{\upsilon_0\tau_0^2 + \left(n_i - 1\right)s^2 + \frac{n_i}{2}\left(Y_i - \mu_0\right)^2}{\upsilon_0 + n_i}, \tag{II.3.13b}$$

where s^2 is the sample variance from neighbourhood $\mathcal{N}_i$. The conditional mean posterior in Equation II.3.13a is a trade-off between prior belief and observed local information. Its mean averages the combined means, while its scale shows that the posterior is supported by twice the amount of information, as it is based on two separate mean estimations. The marginal variance posterior in Equation II.3.13b has additive degrees of freedom, whereas the updated scale parameter $\tilde{\tau}^2$ combines the prior and observed sum of squares. The latter are dilated by extra uncertainty from the deviation between the combined means. While the two individual sums of squares in the numerator represent the variability within the individual distributions, the additional uncertainty stems from the likelihood of both occurring together.

Equations II.3.13a and II.3.13b demonstrate that the prior and the local information each supply half of the posterior information. The benefit of this is that the posterior is robust against inflation, which might be caused by local boundary conditions or by extreme global imbalance.

II.3.5.4 Bootstrapping

The final methodological stage is to generate a bootstrap distribution for LSD that involves the Bayesian procedure outlined above. In each bootstrap iteration, the following steps must be repeated:

1. random resampling with replacement within the local neighbourhoods,
2. drawing of new synthetic means and recalculation of the residuals,
3. recalculation of LSD for each drawn pseudo-sample with substituted means,
4. estimation of an empirical distribution of LSD and assessment of pseudo p-values p^*.

These four stages resemble the Monte Carlo approach outlined in (Hope 1968). A concise description of the stages that are usually involved in this kind of approach, is also found in (Dray 2011, 129 f.). What differentiates our approach from these two studies is that the proposed bootstrap is locally constrained. The drawing of additional means in Stage 2 involves the Bayesian approach from Sections II.3.5.2 and II.3.5.3 and is achieved in three phases:

1. drawing of a posterior variance $\sigma^2_{X_i}$ from Equation II.3.13b,
2. substitution of $\sigma^2_{X_i}$ into Equation II.3.13a and drawing of a posterior mean μ_{X_i},
3. drawing of new mean values from $N\left(\mu_{X_i}, a_i\sigma^2_{X_i}\right)$.

The pseudo p-values p^* that are needed for inference can then be calculated in different ways, depending on the desired hypothesis testing scheme (Table II.3.1).

Table II.3.1: Overview of estimators of pseudo p-values p^* for different types of hypotheses. LSD_{i_0} denotes an observed LSD value, LSD_{i_k} is the LSD value obtained from the k-th bootstrap, m is the overall number of iterations and α is the adjusted significance level. We use $\#$ to denote the cardinality of a set to avoid notational ambiguity.

Testing scheme	**Pseudo p-value estimator**	**Interpretation of** $p^* < \alpha$
Right-tailed	$p^* = \frac{1}{m}\# \{k \mid LSD_{i_k} > LSD_{i_0}\}$	Geographic arrangement increases the variance
Left-tailed	$p^* = \frac{1}{m}\# \{k \mid LSD_{i_k} < LSD_{i_0}\}$	Geographic arrangement reduces the variance
Two-tailed	$p^* = \frac{1}{m}\# \left\{k \mid LSD^*_{i_k} > LSD_{i_0}\right\}$ where $LSD^*_{i_k} = \lvert LSD_{i_k} - L\bar{S}D_{i_k}\rvert$	Geographic arrangement affects the variance

II.3.6 Empirical Results from a LiDAR-Derived Dataset

Both LSD and LOSH are applied to a subset of 4,436 height differences calculated from two co-registered and filtered LiDAR datasets between 20th and 24th August 2016. These are taken from an 'automatic terrestrial laser scanning station' (ATLS) monitoring project of daily scans, which involves surveying a slow-moving landslide in Gresten, Austria (Figure II.3.3, *cf.* (Canli et al. 2015; Höfle et al. 2016)). The height differences were obtained from the 'Multiscale Model to Model Cloud Comparison' (Barnhart and Crosby 2013; Lague et al. 2013), a point-based comparison method that recognises the existence of sampling variability and measurement error. The eastern part of the scanned area got mown in between the two dates (a figure showing the study area before and after the mowing is provided in the online supplementary material II.3.11). The dataset thus comprises a distinctive global structure (mown vs unmown; diagonal dividing line) and, in addition, weaker local structures within the sub-regions. This two-stage structure makes the data a suitable test case for LSD and LOSH. These techniques are applied with inverse-distance weighting and a cut-off at a distance of one meter. This scheme is useful because the observed process has a positive spatial autocorrelation and does not show abrupt changes within the regimes. Note that the obtained results should not be understood as outcomes of an empirical investigation, but rather as a scenario for demonstrating differences between LSD and LOSH.

II.3.6.1 Interpretation of LSD and LOSH

The results from LSD and LOSH reveal different features of variance patterns. Thus, when they are interpreted together, it is easier to make a direct comparison. Figure II.3.4 maps statistically significant LOSH and LSD outcomes. We randomise locally, but omit the Bayesian approach for the moment, as all the involved neighbourhoods are sufficiently large. The smallest available neighbourhood size is $n_i = 8$, which allows $8! = 40,320$ permutations. The average of $n_i = 54$, however, allows ca. $2.31 \times 1,071$ permutations, which is enough for virtually all the application scenarios. Despite this, $n_i = 8$ is still a small number of observations and hence contains little information. The Bayesian technique thus proves to be useful, as will be discussed in Sub-section II.3.6.5.

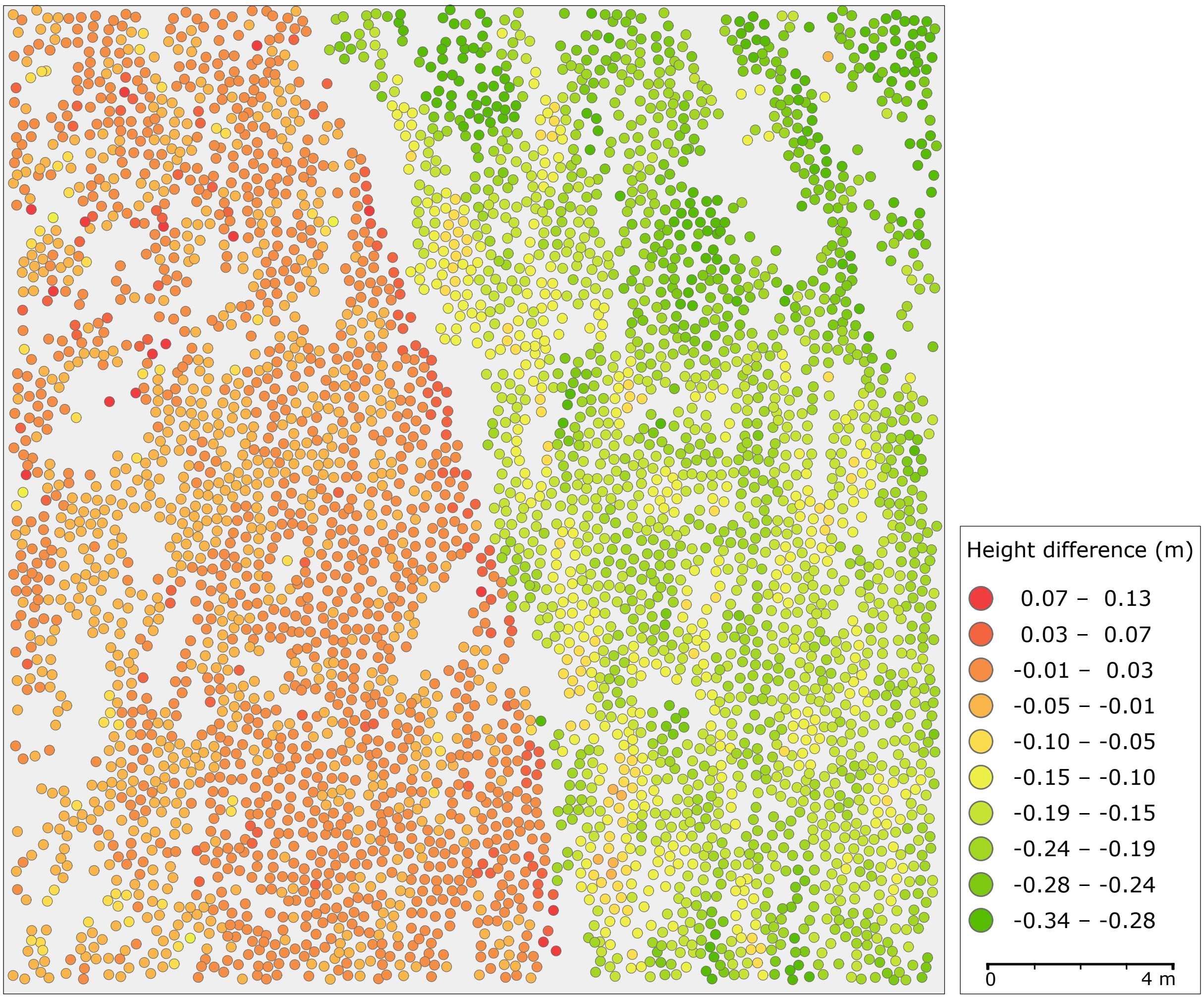


Figure II.3.3: Height differences between two ATLS datasets.

The global dividing line cutting across the centre of the region, is a feature where significant LOSH values from the right tail of the reference distribution accumulate (Figure II.3.4a). Thereby, the southern part dominates, while the northern part of the line is influenced by a spatial gap (a gentle slope in the terrain) which is an obstacle to high LOSH scores. Further high values are found in the mown regime, in particular in the northernmost part (disturbances from artefacts) and in the South (these vanish when the false discovery rate is controlled by following Benjamini and Hochberg (1995)). In contrast, the western unmown part is dominated by significantly low LOSH values. These are caused by the global resampling scheme of LOSH, which shifts statistically differing values from the mown part into the unmown region. This biases the p-values towards the left tail of the bootstrap distribution and makes it impossible to disclose local variance patterns. The eastern mown regime is not as homogeneous as expected. Grass cuttings produced from the lawn mower were being left on the meadow. This increases the global average residual h_1 and in turn leads to a more homogeneous appearance of the unmown regime, as explained earlier. Nevertheless, LOSH reveals and maps the global variance structure in the locality by identifying the most (the dividing line) and least dispersed areas (the unmown part).

The LSD values (Figure II.3.4b) are more evenly distributed than the LOSH values. Unexpectedly and in stark contrast with LOSH, the dividing line no longer appears on the map except for a small part in

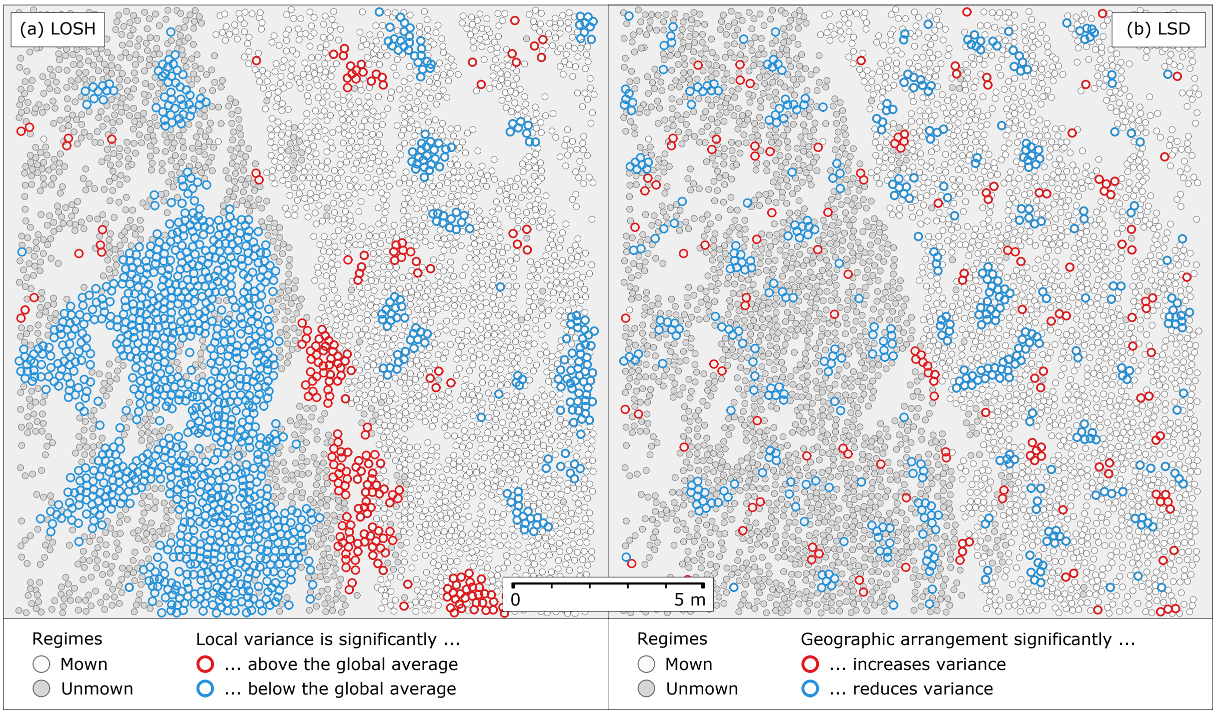

(a) LOSH
(b) LSD
0
5 m
Regimes
○ Mown
● Unmown
Local variance is significantly ...
○ ... above the global average
○ ... below the global average
Regimes
○ Mown
● Unmown
Geographic arrangement significantly ...
○ ... increases variance
○ ... reduces variance

Figure II.3.4: Significant scores from (a) LOSH and (b) LSD (two-sided test; $\alpha = 0.05$; 1,000 iterations).

the centre. The local spatial arrangement is thus not leading to the variability of the features and it can be concluded that the dividing line is a truly global feature that is only caused by the existence of two different regimes. Apart from this, the western part is no longer as homogeneous as it appeared with LOSH. LSD reveals certain significant local features that are interspersed and like small spots of high variability within the unmown part. The overall distribution of the LSD values is, however, rather homogeneous across the two regimes (Table II.3.2). The structures in the mown and unmown parts therefore do not seem to differ noticeably and behave in a relatively similar way. A different significance evaluation will be seen when the Bayesian mean prediction is incorporated in Sub-section II.3.6.5.

Table II.3.2: Descriptive statistics for LSD scores within the mown and unmown regimes.

Regime	**Min**	**Max**	**Mean**	**Median**	**Standard deviation**	**Interquartile range**
Mown	0.146	3.185	0.907	0.823	0.363	0.446
Unmown	0.249	4.426	0.904	0.782	0.453	0.501

In summary it can be stated that an evaluation of LOSH and LSD scores in combination reveals both global and local variance patterns. The observed LSD values further confirm that, at least for the adjusted analytical scale, the dividing line is a global feature. It is also clear that local structures that remain hidden with LOSH are present in the map when LSD is considered. The LSD scores thus provide an additional insight into the dataset.

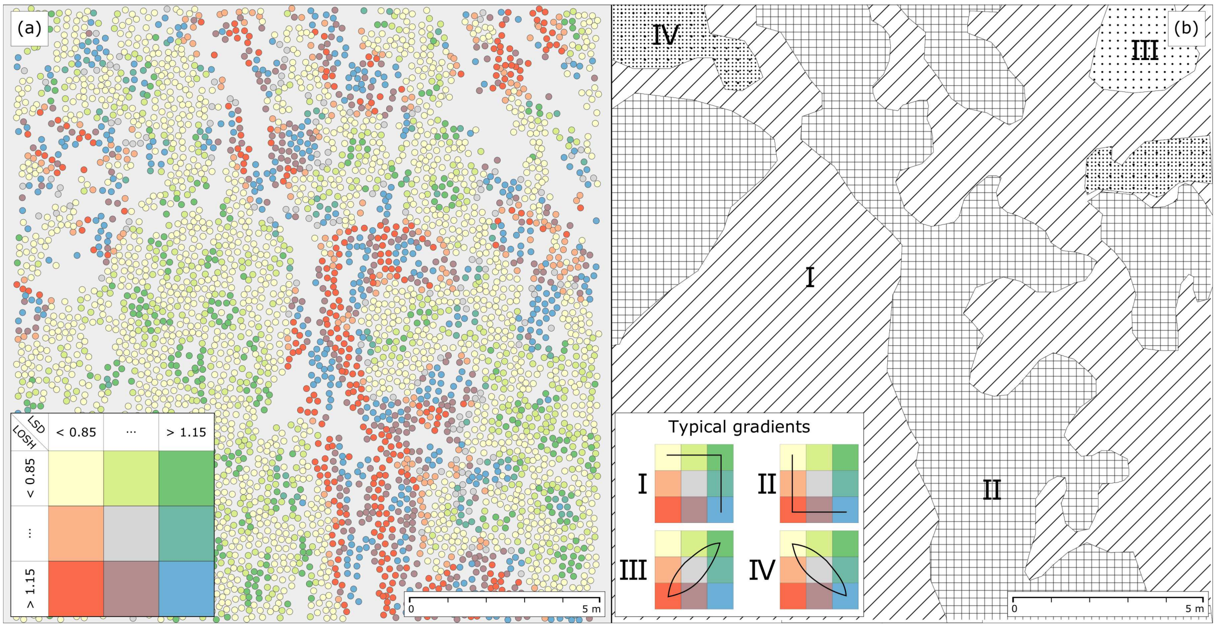


Figure II.3.5: Variance patterns within LiDAR-derived height differences. (a) A detailed characterisation of local and global effects: the variance can be above the global mean and increased at a local level by the geographic pattern (blue) or below the global mean and reduced further at a local level at the same time (yellow). The locations can also be homogeneous from a global standpoint while the local pattern increases the variance (dark green) or vice versa (red), with all sorts of possible transitional effects (intermediate colours); (b) A schematic sketch of LOSH-LSD configurations: the prevailingly local structures (I), the prevailingly global structures (II), locally homogeneous, and globally dispersed (III), global and local variance fluctuations (IV).

II.3.6.2 A Map of Global and Local Spatial Variance Patterns

It is worth flagging the significance of the LSD and LOSH values, but they are not exhaustive in terms of their interpretation. The maps in Figure II.3.5 thus provide a classification scheme for LOSH-LSD tuples. Four standard gradients can be derived from these, each characterising different sub-regions in the map.

Figure II.3.5 shows a way to classify LOSH and LSD outcomes together. A prominent feature in Figure II.3.5a is, again, the dividing line (see also type II gradient in Figure II.3.5b). The figure, however, shows the line in more detail: The centre of the line appears to be narrow and elongated and reflects the thin crisp edge of the boundary where the two regimes meet. The spatial pattern strongly increases the variance in this from both a global and local standpoint. Adjoining this is a fuzzy region where local spatial effects are negligible, while the spatial variance is generally high in a global comparison. In other words, while the global variance gradient features prominently, the local spatial variance pattern is closer to randomness and regularity. The local geographic arrangement is thus not related to the increased variance in these regions.

When one moves farther away from the dividing line, the effects prevail at a local level. The variance structures turn into insular regions of small areas where the local pattern increases the variance, which are surrounded by a homogenising geographic arrangement (type I). The northern part, which is affected by artefacts, is further characterised by two volatile variance patterns (types III and IV). The type III pattern, which is featured in the north-eastern part, is caused by a larger haystack. This appears to be regular in local terms (its internal structure), but is disruptive globally as it is a prominent feature (above the global mean variance). In contrast, the type IV pattern reflects taller bunch grass that is characterised by abrupt fluctuations between regular and heterogeneous conditions caused by the related clumps of culms. An interpretation that combined LOSH and LSD made it possible to distinguish these rather different features in the data.

The detailed interpretations given above, demonstrate the additional value that LSD provides. Global structures are detected and mapped locally by LOSH where these dominate, but local details are missed out. In contrast, LSD assesses local structures and describes in greater detail the internal structure of global features (*e. g.*, the nature of the central boundary or of the homogeneous sub-regions). The measure thus not only assesses different structures, but also reveals additional information about features obtained from LOSH.

II.3.6.3 Interplay with Variance

Since both LSD and LOSH are measures of variance, it is worth investigating how they relate to the magnitude of the non-spatial local variance. This illustrates the ability of LOSH and LSD to separate effects of spatial patterning from other influences of general variability.

The design of LOSH implies there is a strong dependence on general local variability through its constant denominator. When a sub-region is generally diverse, the prospect of assessing high LOSH scores is also high, regardless of the local spatial patterning. Figure 6a illustrates this link, and Kendall's Tau-b, an ordinal correlation measure that accounts for non-normality and ties, gives further support through a strongly significant test score of $\tau = 0.679\,(p < 0.001)$. However, this relationship is not uniform since LOSH is more dispersed when the local variability is stronger. Regressing LOSH on variance and conducting a non-parametric Koenker-Bassett test (Koenker and Bassett 1982; Godfrey 1996) on the residuals, confirms the heteroscedasticity that is visible in Figure II.3.6a. The two diverging quartile trend lines in the biquantile regressogram (Figure II.3.6b) underpin this outcome, while the median trace shows that variance is a good predictor of LOSH. The measure is thus dominated by non-spatial variability. This result is in accordance with the intended purpose of LOSH to detect both the most and least dispersed regions in geographic data. However, it also shows that LOSHs power to detect solely spatial effects in local circumstances is limited.

In contrast, Figure II.3.7a shows that LSD is only weakly related to variance ($\tau = 0.023, p < 0.001$), and the median trend line in the regressogram (Figure II.3.7b) represents a relationship that varies with the strength of LSD. Variance is a sufficient predictor of LSD when it is strong and when, at the same time, the influence of the spatial patterning is weak (*i. e.*, the right part of the scatter plot in Figure II.3.7a). However, the ascending slopes of the median, as well as the quartile trend lines (Figure II.3.7b) show that variance systematically overestimates high LSD scores. The spatial pattern thus dominates the (more interesting) high LSD values. These characteristics are desirable properties: LSD only has a negligible link with local variance, while extremal outcomes are controlled by the spatial effects that they are supposed to quantify.

II.3.6.4 Relationship with Spatial Autocorrelation

The two measures quantify different aspects of ´dissimilarity' within random variables. As mentioned in Section II.3.2, spatial autocorrelation represents an additional, covariance-based dimension of heterogeneity. Local estimators like local Moran's *I* (Anselin 1995) can be used to quantify spatial autocorrelation, and Figure II.3.8a shows its relation to LOSH. The Moran interval $[0.0, 1.4]$ shows a significant negative

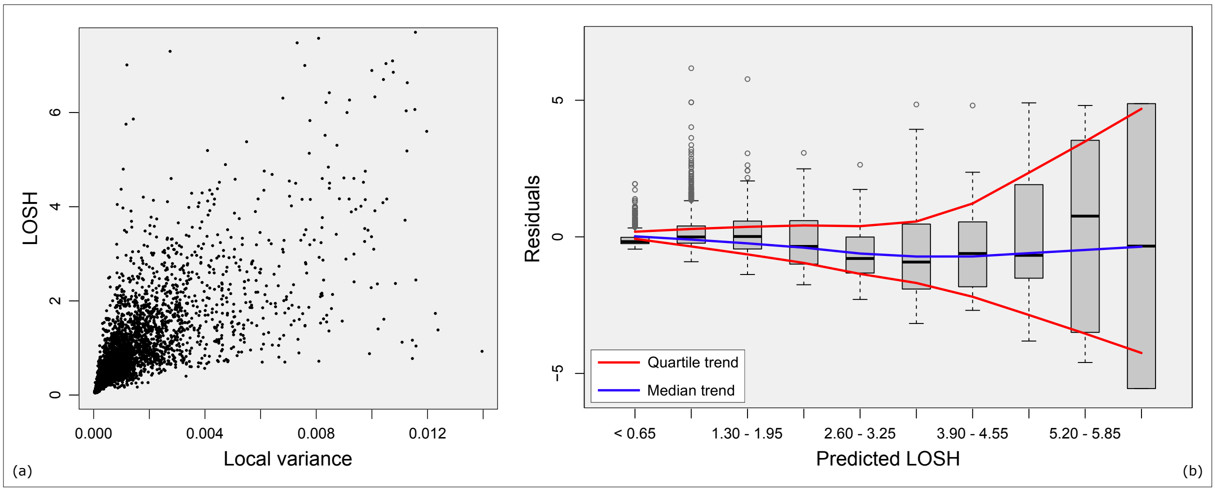


Figure II.3.6: Relationship between LOSH and local variance. (a) A scatter plot of variance and LOSH; and (b) A biquantile regressogram (Tukey 1977) illustrating heteroscedasticity in LOSH.

relationship ($\tau = 0.27$). LOSH is high when the association between neighbours is random and low when observations occur in a clustered form. Both measures therefore, to some extent, highlight similar structures from different perspectives (variance vs. covariance). Observations showing autocorrelations higher than 1.4 belong to the northern artefacts and thus can reasonably be regarded as outliers, that do not conform to the general observations made above. Overall, LOSH reveals roughly similar structures to those of Moran's I, as is evident from their antipodal behavioural pattern.

In contrast, LSD is almost unrelated to Moran's I, when the latter is on the interval $[0.0, 1.4]$. Most of the data points accumulate on the left side of the scatter plot in Figure II.3.8b without showing any notable trend ($\tau = -0.09$). This strengthens the likelihood indicated above that LSD is able to reveal patterns that cannot be detected by LOSH and Moran's I. These detected patterns are not linked to the clustering tendency of the attribute values. Rather, they are features in their own right, which makes them of value for empirical investigations since they might supply important details about the disclosure of the mechanisms in spatial random variables.

II.3.6.5 Influence of the Bayesian Prediction of Mean Values

The Bayesian procedure from Section II.3.5 extends local resampling by the use of synthetic data generated from empirical prior knowledge combined with local information. This approach differs from conventional bootstrapping that only relies on observed information. There is a need to investigate how the Bayesian approach influences drawn inferences.

Figure II.3.9 shows a sigmoidal relationship between conventional p-values (*i. e.*, those that were used in the previous paragraphs) and those involving synthetic means. They show a strong monotonic association of $\tau = 0.77$ at medium ranges. There is a significant fall in this association in both tails ($\tau = 0.23$), which is an important observation as the tails possess values which are important for drawing inferences. In the Bayesian approach, the p-values tend to concentrate around the extremes of 0 and 1. In contrast, conventional p-values show a higher level of dispersion in the tails. This is caused by the number of available observations, which have limited explanatory power because they only represent a small fraction of all possible values. In contrast, the Bayesian approach extends this spectrum, which increases its ability to detect spatial effects because the comparative values are not biased towards a certain range.

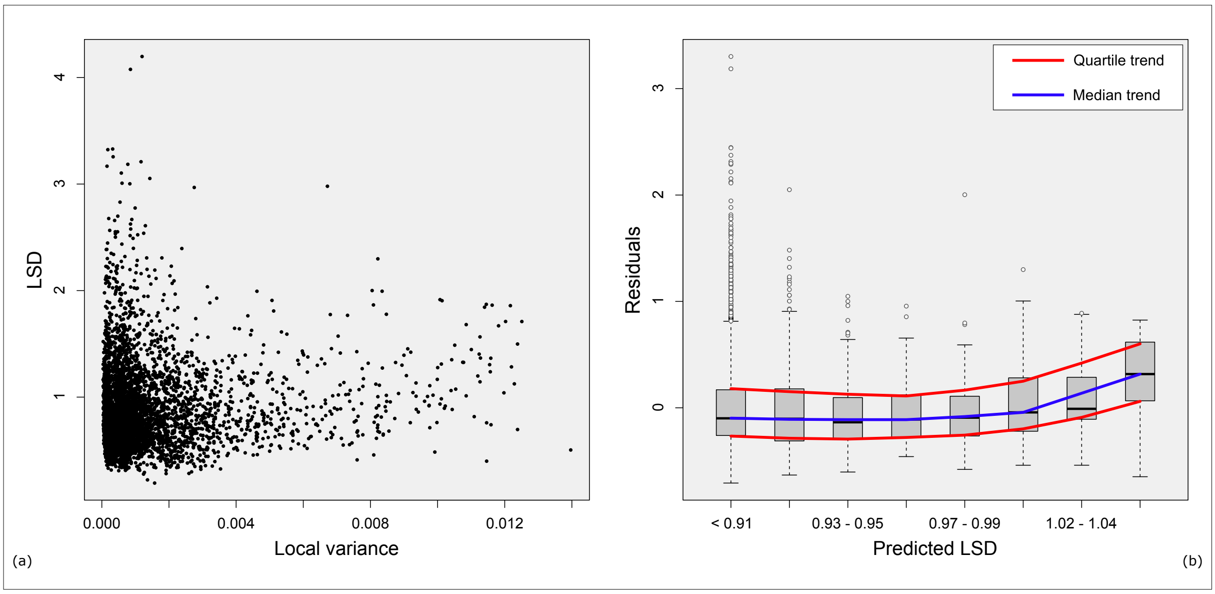


Figure II.3.7: Relationship between LSD and local variance. (a) A scatter plot of variance with regard to LSD; and (b) a biquantile regressogram (Tukey 1977) illustrating heteroscedasticity within LSD.

The increased ability to detect effects with the Bayesian approach is further evident after the p-values have been corrected for multiple hypothesis testing. Note that LSD tests n hypotheses with one dataset. This repeated use of the data leads to an increase in the type I error rate and requires correction. When the false-discovery rate is controlled at $\alpha = 0.05$ (FDR; Benjamini and Hochberg 1995) and the p-values are corrected accordingly, it is seen that the non-Bayesian approach is very conservative. Only 0.5% of all the null hypotheses are rejected, which is way below the significance level that was envisaged. In other words, many actual effects might be missed out. In contrast, when the Bayesian-generated p-values are adjusted, they yield a ratio of 5.2%, which is close to the desired α level.

Figure II.3.10 illustrates the FDR-corrected p-values of significant observations by incorporating the Bayesian-generated means. The significant features in the eastern part (the 'mown') show a general North-South bearing (Figure II.3.10a) and resemble the direction of the mowing process, which is illustrated in the background of II.3.10b through a hill-shading raster. The blue features, where the geographic layout reduces the variance, either accumulate alongside the small piles of hay that were left on the meadow or in the furrows in-between. In contrast, the western part (the 'unmown') is not affected by the after-mowing topography. The patterns of significant features observed in this part are mostly unrelated to the hill-shading. This makes sense given that the height differences that were analysed are affected by physical, biological and other factors that do not necessarily correspond to the topography shown in Figure II.3.10b, especially in the unmown area.

A comparison of Figure II.3.10 with Figure II.3.4b shows that the conventional p-values generated from the local bootstrapping, do not show the features described above. In fact, there is no noticeable difference between the mown and unmown parts in this case. The p-values generated by the inclusion of predicted means are closer to the phenomenon (especially in the mown part) and can thus be considered to be of greater value. Hence, this comparison implies that the proposed Bayesian approach is a reasonable alternative to more conventional forms of pseudo p-value estimation.

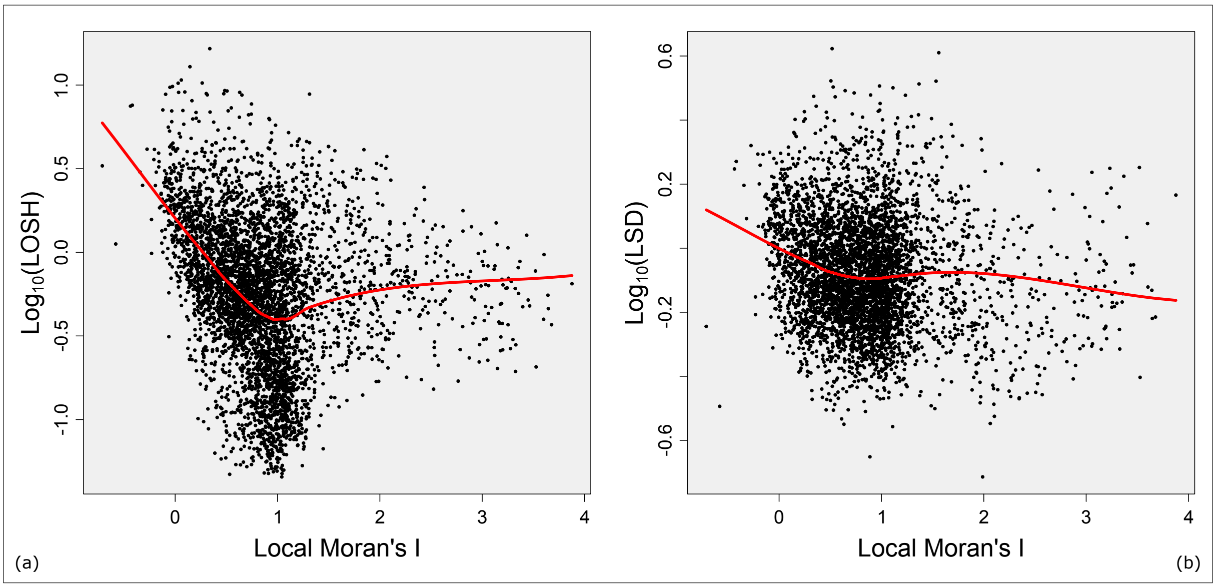


Figure II.3.8: The relationship of LOSH and LSD with local Moran's I. The logarithms of the two measures were chosen to improve interpretability. The red line represents a first-order LOESS trend. (a) LOSH; and (b) LSD.

II.3.7 Discussion and Conclusions

This paper introduces a test called 'Local Spatial Dispersion' (LSD), which is able to determine the local influence of geographic arrangements on variance. It does not incorporate global information and allows local patterns to be detected in the presence of a global structure. The strictly local nature of the test, however, increases the risk of problems arising from small-size samples within local neighbourhoods. To mitigate this risk, a stratified bootstrapping procedure is introduced that combines traditional resampling with a Bayesian prediction of synthetic data. The proposed LSD supplements LOSH, which is a recently devised technique to map global variance structures locally. The measure adapts LOSH to strictly local circumstances. Conceptually, LSD forms a part of a series of localised techniques like the hot spot method called 'O-statistic' (Ord and Getis 2001), or locally adaptive geometric clustering techniques such as the inhomogeneous marked and unmarked K-functions (Cuzick and Edwards 1990; Baddeley et al. 2000).

Its application to a dataset for height differences derived from LiDAR data demonstrates the ability of LSD to detect local patterns within a distinct global structure. An interpretation combined with LOSH reveals further characteristic variance patterns, which would not have been detected by using either measure alone. Furthermore, the obtained results show that LOSH is closely correlated with general non-spatial variability, which hampers the separation of genuinely spatial from other effects. In contrast, LSD is uncorrelated with non-spatial variation and is capable of exposing entirely spatial variance effects. Notably, LSD is also unrelated to positive spatial autocorrelation. This allows the measure to assess other complex patterns apart from general attribute clustering, such as the internal structures of clusters and the detailed contours of geographic boundaries. The proposed inference mechanism further facilitates the detection of local structures. While conventional stratified bootstrapping turns out to be overly conservative, the synthetic expansion of the available local data keeps the α-rate in compliance with the adjusted significance level, which increases its ability to detect meaningful patterns. Overall, LSD has been shown to be a useful extension to the spatial analysis toolbox. In the given example, it is possible, in

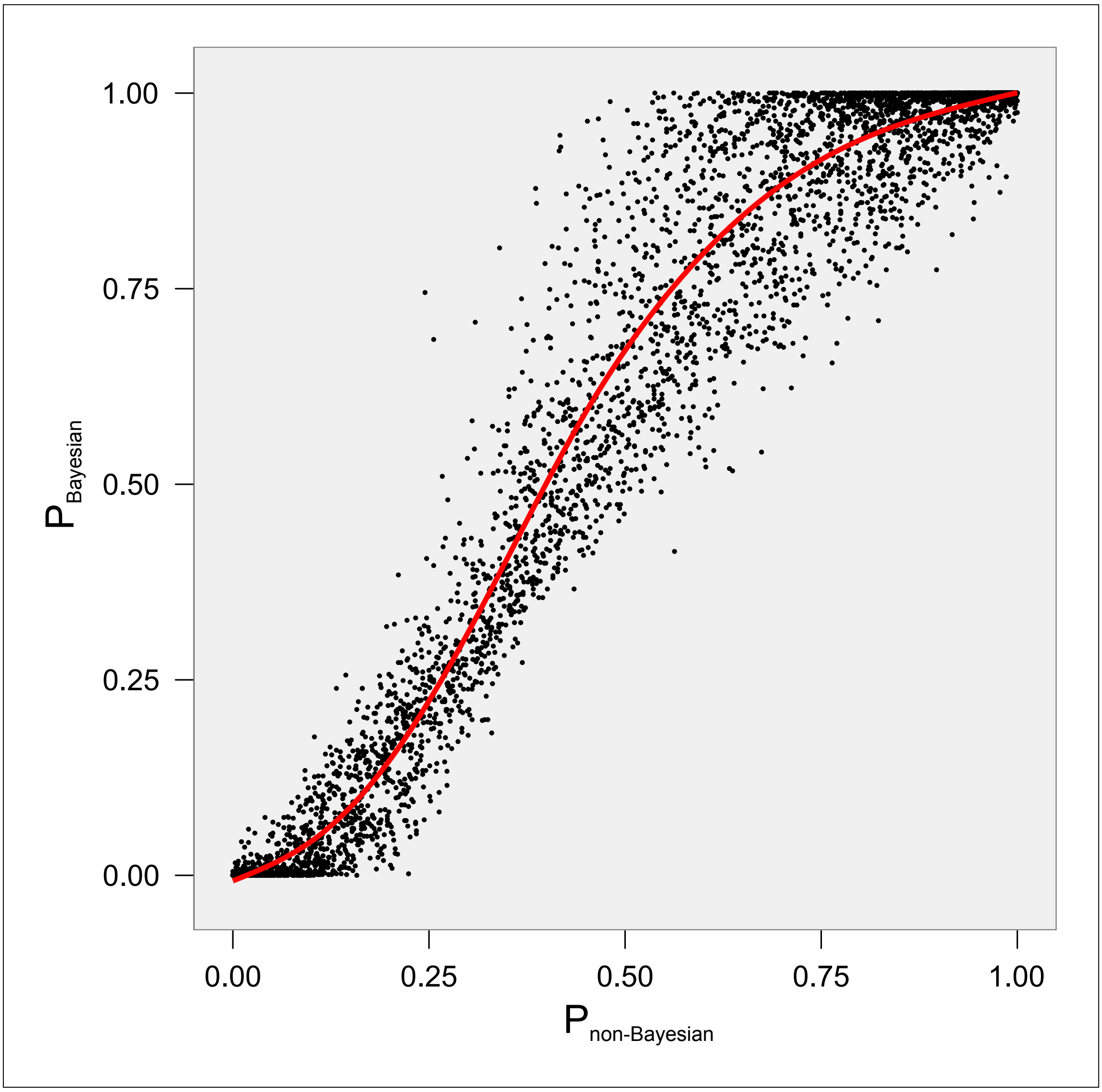


Figure II.3.9: The relationship between the Bayesian and the non-Bayesian p-values.

statistical terms, to detect local interaction between variance and spatial patterns within global structures and thus to disclose details that would otherwise have been overlooked.

The anonymous reviewers pointed out that there was a relationship between the proposed LSD technique and local variograms. Variograms quantify the variance of the spatial increment between two locations separated by a certain distance (Bachmaier and Backes 2008; Cuba et al. 2012). Both, LSD and variograms are thus concerned with variance estimation. What differentiates them is that LSD is a) a hypothesis test designed to determine the influence of a specific spatial arrangement on variance and b) that it is concerned with in-place variance rather than with the variance of the incremental process. In contrast, variograms estimate variance within certain distance bands by relying on the validity of the employed spatial weights (instead of testing their influence). The estimates of the variograms are then used for modelling (*e. g.*, in Kriging), which means that our proposed test can be used as a diagnostic tool for geostatistics. For instance, LSD can be used to fully investigate the possible sources of local

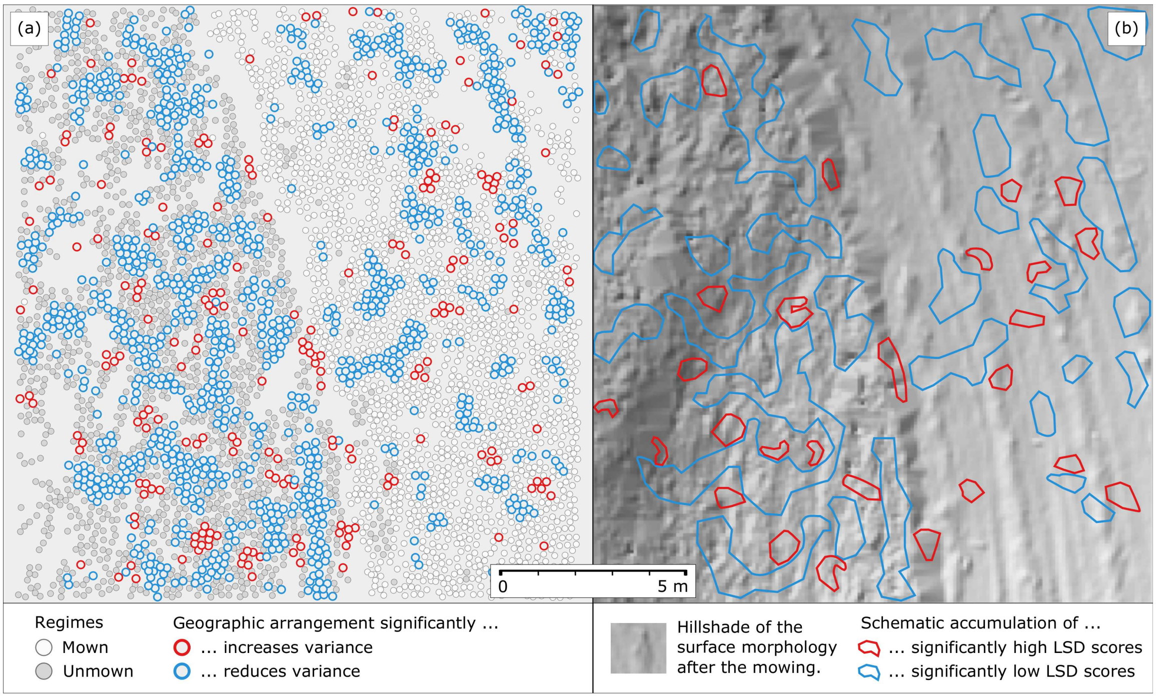

(a)
(b)
0
5 m
Regimes
Mown
Unmown
Geographic arrangement significantly ...
... increases variance
... reduces variance
Hillshade of the surface morphology after the mowing.
Schematic accumulation of ...
... significantly high LSD scores
... significantly low LSD scores

Figure II.3.10: Significant LSD scores involving Bayesian-predicted means (two-sided test; $\alpha = 0.05$; $1,000$ iterations). (a) Map of significant features. (b) Schematic sketch of significant accumulated features, against the background of the hill-shading of the surface after the mowing.

non-stationarities, which might lead to a lack of stationarity in the difference processes between locations. Thus, LSD might also be a useful device in the area of geostatistics.

However, there are some shortcomings in this paper that could not be addressed. One of these is that our data only have a positive spatial autocorrelation. A negative spatial autocorrelation is different in nature, since it involves a certain degree of heterogeneity, which in turn is related to variance. An interesting relation between LSD and negative spatial autocorrelation might thus exist, which would be worth exploring in a systematic way in a future research project. In terms of inference, the forms adopted for the prior and likelihood are strongly supported by the central limit theorem and leave little room for variation. However, the way that the prior and likelihood enter the posterior distribution, need to be analysed with regard to suitable combinations other than the applied 'half-and-half scheme'. For instance, an adaptive solution could be useful, in which the likelihood is given more weight in larger neighbourhoods that are backed up by a more solid database. In terms of LOSH, our empirical results show a strong heteroscedasticity with regard to local variance. Future research should therefore seek to achieve a variance stabilisation in order to make the outcomes of LOSH more robust for inhomogeneous populations and assist its interpretation. From a technological standpoint, the proposed solution is computationally expensive as it includes bootstrapping. The application of LSD to large datasets would hence clearly benefit from an efficient implementation strategy. All in all, LSD provides the means of obtaining a valuable and detailed insight into variance mechanisms of geographic random variables and offers the prospect of achieving significant new empirical results in various fields.

Acknowledgements

This work was supported by the *German Research Foundation* (*DFG*) through the priority programme 'Volunteered Geographic Information: Interpretation, Visualisation and Social Computing' (SPP 1894). Special thanks go to *Prof Thomas Glade* (*University of Vienna*) for the provision of LiDAR data from the '4DEMON' project. We would also like to thank the Lower Austrian government for their financial support for the LiDAR data collection within the framework of the 'NoeSLIDE' project.

References (Chapter II.3)

Abascal, M and D Baldassarri (2015). 'Love Thy Neighbor? Ethnoracial Diversity and Trust Reexamined'. *American Journal of Sociology* 121 (3), pp. 722–782. DOI: 10.1086/683144.

Anderson, C, D Lee and N Dean (2014). 'Identifying Clusters in Bayesian Disease Mapping'. *Biostatistics* 15 (3), pp. 457–469. DOI: 10.1093/biostatistics/kxu005.

Anselin, L (1988a). 'Lagrange Multiplier Test Diagnostics for Spatial Dependence and Spatial Heterogeneity'. *Geographical Analysis* 20 (1), pp. 1–17. DOI: 10.1111/j.1538-4632.1988.tb00159.x.

— (1990). 'Spatial Dependence and Spatial Structural Instability in Applied Regression Analysis'. *Journal of Regional Science* 30 (2), pp. 185–207.

— (1995). 'Local Indicators of Spatial Association - LISA'. *Geographical Analysis* 27 (2), pp. 93–115. DOI: 10.1111/j.1538-4632.1995.tb00338.x.

— (2001). 'Spatial Econometrics'. In: *A Companion to Theoretical Econometrics*. Ed. by B Baltagi. Hoboken, NJ: Wiley-Blackwell, pp. 310–330. DOI: 10.1002/9780470996249.

— (2010). 'Thirty Years of Spatial Econometrics'. *Papers in Regional Science* 89 (1), pp. 3–25. DOI: 10.1111/j.1435-5957.2010.00279.x.

Arnot, C and P Fisher (2007). 'Mapping the Ecotone with Fuzzy Sets'. In: *Geographic Uncertainty in Environmental Security*. Ed. by A Morris and S Kokhan. Dordrecht: Springer Netherlands, pp. 19–32. DOI: 10.1007/978-1-4020-6438-8_2.

Bachmaier, M and M Backes (2008). 'Variogram or Semivariogram? Understanding the Variances in a Variogram'. *Precision Agriculture* 9, pp. 173–175. DOI: 10.1007/s11119-008-9056-2.

Baddeley, A, J Moller and R Waagepetersen (2000). 'Non- and Semi-Parametric Estimation of Interaction in Inhomogeneous Point Patterns'. *Statistica Neerlandica* 54 (3), pp. 329–350. DOI: 10.1111/1467-9574.00144.

Barbujani, G, N Oden and R Sokal (1989). 'Detecting Regions of Abrupt Change in Maps of Biological Variables'. *Systematic Zoology* 38 (4), pp. 376–389. DOI: 10.2307/2992403.

Barnhart, T and B Crosby (2013). 'Comparing Two Methods of Surface Change Detection on an Evolving Thermokarst Using High-Temporal-Frequency Terrestrial Laser Scanning, Selawik River, Alaska'. *Remote Sensing* 5 (6), pp. 2813–2837. DOI: 10.3390/rs5062813.

Bavaud, F (2014). 'Spatial Weights: Constructing Weight-Compatible Exchange Matrices from Proximity Matrices'. In: *LNCS: Geographic Information Science*. Ed. by M Duckham, E Pebesma, K Stewart and A Frank. Heidelberg: Springer, pp. 81–96. DOI: 10.1007/978-3-319-11593-1_6.

Benjamini, Y and Y Hochberg (1995). 'Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing'. *Journal of the Royal Statistical Society: Series B (Methodological)* 57 (1), pp. 289–300.

Berger, J (2006). 'The Case for Objective Bayesian Analysis'. *Bayesian Analysis* 1 (3), pp. 385–402. DOI: 10.1214/06-BA115.

Berkes, F, T Hughes, R Steneck, J Wilson, D Bellwood, B Crona, C Folke, L Gunderson, H Leslie, J Norberg, M Nyström, P Olsson, H Österblom, M Scheffer and B Worm (2006). 'Globalization, Roving Bandits, and Marine Resources'. *Science* 311 (5767), pp. 1557–1558. DOI: 10.1126/science.1122804.

Box, G (1953). 'Non-Normality and Tests on Variances'. *Biometrika* 40 (3/4), pp. 318–335. DOI: 10.2307/2333350.

Bravo, C and R Weber (2011). 'Semi-Supervised Constrained Clustering with Cluster Outlier Filtering'. In: *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*. Ed. by C San Martin and S Kim. Heidelberg: Springer, pp. 347–354. DOI: 10.1007/978-3-642-25085-9_41.

Brunsdon, C, A Fotheringham and M Charlton (1996). 'Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity'. *Geographical Analysis* 28 (4), pp. 281–298. DOI: 10.1111/j.1538-4632.1996.tb00936.x.

Brunsdon, C, S Fotheringham and M Charlton (1998). 'Geographically Weighted Regression'. *Journal of the Royal Statistical Society: Series D (The Statistician)* 47 (3), pp. 431–443. DOI: 10.1111/1467-9884.00145.

Cadenasso, M, S Pickett and K Schwarz (2007). 'Spatial Heterogeneity in Urban Ecosystems: Reconceptualizing Land Cover and a Framework for Classification'. *Frontiers in Ecology and the Environment* 5 (2), pp. 80–88. DOI: 10.1890/1540-9295(2007)5[80:SHIUER]2.0.CO;2.

Canli, E, B Höfle, M Hämmerle, B Thiebes and T Glade (2015). 'Permanent 3D Laser Scanning System for an Active Landslide in Gresten (Austria)'. In: *Geophysical Research Abstracts EGU General Assembly*. Vol. 17, p. 2885.

Cochran, W (1934). 'The Distribution of Quadratic Forms in a Normal System, with Applications to the Analysis of Covariance'. *Mathematical Proceedings of the Cambridge Philosophical Society* 30 (2), pp. 178–191. DOI: 10.1017/S0305004100016595.

Csillag, F and K Sándor (2002). 'Wavelets, Boundaries, and the Spatial Analysis of Landscape Pattern'. *Écoscience* 9 (2), pp. 177–190.

Cuba, M, O Leuangthong and J Ortiz (2012). 'Detecting and Quantifying Sources of Non-Stationarity via Experimental Semivariogram Modeling'. *Stochastic Environmental Research and Risk Assessment* 26 (2), pp. 247–260. DOI: 10.1007/s00477-011-0501-9.

Cuzick, J and R Edwards (1990). 'Spatial Clustering for Inhomogeneous Populations'. *Journal of the Royal Statistical Society . Series B (Methodological)* 52 (1), pp. 73–104. DOI: 10.2307/2346101.

Darnieder, W (2011). *Bayesian Methods for Data-Dependent Priors*. Columbus, OH: The Ohio State University.

Dray, S (2011). 'A New Perspective about Moran's Coefficient: Spatial Autocorrelation as a Linear Regression Problem'. *Geographical Analysis* 43 (2), pp. 127–141. DOI: 10.1111/j.1538-4632.2011.00811.x.

Dutilleul, P (2011). *Spatio-Temporal Heterogeneity: Concepts and Analyses*. Cambridge, UK: Cambridge University Press.

Dutilleul, Pierre and Pierre Legendre (1993). 'Spatial Heterogeneity Against Heteroscedasticity: An Ecological Paradigm Versus a Statistical Concept'. *Oikos* 66 (1), pp. 152–171. DOI: 10.2307/3545210.

Fagan, W, R Cantrell and C Cosner (1999). 'How Habitat Edges Change Species Interactions'. *The American Naturalist* 153 (2), pp. 165–182. DOI: 10.1086/303162.

Fagan, W, M Fortin and C Soykan (2003). 'Integrating Edge Detection and Dynamic Modeling in Quantitative Analyses of Ecological Boundaries'. *BioScience* 53 (8), p. 730. DOI: 10.1641/0006-3568(2003)053[0730:IEDADM]2.0.CO;2.

Fang, H and D Huang (2004). 'Noise Reduction in LiDAR Signal Based on Discrete Wavelet Transform'. *Optics Communications* 233 (1-3), pp. 67–76. DOI: 10.1016/j.optcom.2004.01.017.

Fisher, P and V Robinson (2014). 'Fuzzy Modelling'. In: *GeoComputation*. Ed. by R Abrahart and L See. 2nd ed. Boca Raton, FL: CRC Press, pp. 283–306.

Forman, R (1995). *Land Mosaics: The Ecology of Landscapes and Regions*. Cambridge, UK: Cambridge University Press.

Fortin, M (1994). 'Edge Detection Algorithms for Two-Dimensional Ecological Data'. *Ecology* 75 (4), pp. 956–965. DOI: 10.2307/1939419.

— (1999). 'Spatial Statistics in Landscape Ecology'. In: *Landscape Ecological Analysis*. Ed. by J Klopatek and R Gardner. New York, NY: Springer, pp. 253–279. DOI: 10.1007/978-1-4612-0529-6_12.

Fotheringham, A, C Brunsdon and M Charlton (2002). *Geographically Weighted Regression : The Analysis of Spatially Varying Relationships*. Chichester, UK: Wiley.

Fotheringham, A, M Charlton and C Brunsdon (1996). 'The Geography of Parameter Space: An Investigation of Spatial Non-Stationarity'. *International Journal of Geographical Information Systems* 10 (5), pp. 605–627. DOI: 10.1080/02693799608902100.

Gaetan, C and X Guyon (2010). *Spatial Statistics and Modeling*. Springer Series in Statistics. New York, NY: Springer. DOI: 10.1007/978-0-387-92257-7.

Garrigues, S, D Allard, F Baret and M Weiss (2006). 'Quantifying Spatial Heterogeneity at the Landscape Scale Using Variogram Models'. *Remote Sensing of Environment* 103 (1), pp. 81–96. DOI: 10.1016/j.rse.2006.03.013.

Gelfand, A and S Banerjee (2015). 'Bayesian Wombling: Finding Rapid Change in Spatial Maps'. *Wiley Interdisciplinary Reviews: Computational Statistics* 7 (October), pp. 307–315. DOI: 10.1002/wics.1360.

Gelman, A, J Carlin, H Stern, D Dunson, A Vehtari and D Rubin (2013). *Bayesian Data Analysis*. 3rd ed. Boca Raton, FL: CRC Press.

Getis, A (2010). 'Spatial Autocorrelation'. In: *Handbook of Applied Spatial Analysis*. Ed. by M Fischer and A Getis. Heidelberg: Springer, pp. 255–278.

— (2015). 'Analytically Derived Neighborhoods in a Rapidly Growing West African City: The Case of Accra, Ghana'. *Habitat International* 45 (Part 2), pp. 126–134. DOI: 10.1016/j.habitatint.2014.06.021.

Getis, A and J Ord (1992). 'The Analysis of Spatial Association by Use of Distance Statistics'. *Geographical Analysis* 24 (3), pp. 189–206. DOI: 10.1111/j.1538-4632.1992.tb00261.x.

Godfrey, L (1996). 'Some Results on the Glejser and Koenker Tests for Heteroskedasticity'. *Journal of Econometrics* 72 (1-2), pp. 275–299. DOI: 10.1016/0304-4076(94)01723-9.

Goovaerts, P (2008). 'Accounting for Rate Instability and Spatial Patterns in the Boundary Aanalysis of Cancer Mortality Maps'. *Environmental and Ecological Statistics* 15 (4), pp. 421–446. DOI: 10.1007/s10651-007-0064-6.

Goovaerts, P (2010). 'How do Multiple Testing Correction and Spatial Autocorrelation Affect Areal Boundary Analysis?' *Spatial and Spatio-temporal Epidemiology* 1 (4), pp. 219–229. DOI: 10.1016/j.sste.2010.09.004.

Graif, C and R Sampson (2009). 'Spatial Heterogeneity in the Effects of Immigration and Diversity on Neighborhood Homicide Rates'. *Homicide Studies* 13 (3), pp. 242–260. DOI: 10.1177/1088767909336728.

Grillet, M, G Jordan and M Fortin (2010). 'State Transition Detection in the Spatio-Temporal Incidence of Malaria'. *Spatial and Spatio-temporal Epidemiology* 1 (4), pp. 251–259. DOI: 10.1016/j.sste.2010.09.007.

Hanson, T, S Banerjee, P Li and A McBean (2015). 'Spatial Boundary Detection for Areal Counts'. In: *Nonparametric Bayesian Inference in Biostatistics*. Ed. by R Mitra and P Müller. New York, NY: Springer, pp. 377–399. DOI: 10.1007/978-3-319-19518-6_19.

Höfle, B, E Canli, E Schmitz, S Crommelinck and D Hoffmeister (2016). '4D Near Real-Time Environmental Monitoring Using Highly Temporal LiDAR'. In: *Geophysical Research Abstracts of the EGU General Assembly*. Vol. 18.

Hope, A (1968). 'A Simplified Monte Carlo Significance Test Procedure'. *Journal of the Royal Statistical Society. Series B (Methodological)* 30 (3), pp. 582–598.

Hu, T, Q Liu, Y Du, H Li and H Huang (2015). 'Analysis of Land Surface Temperature Spatial Heterogeneity Using Variogram Model'. In: *IEEE International Geoscience and Remote Sensing Symposium 2015*. Milan: IEEE, pp. 132–135. DOI: 10.1109/IGARSS.2015.7325716.

Hungr, O, S Leroueil and L Picarelli (2014). 'The Varnes Classification of Landslide Types, an Update'. *Landslides* 11 (2), pp. 167–194. DOI: 10.1007/s10346-013-0436-y.

Jaboyedoff, M, T Oppikofer, A Abellán, M Derron, A Loye, R Metzger and A Pedrazzini (2012). 'Use of LiDAR in Landslide Investigations: A Review'. *Natural Hazards* 61 (1), pp. 5–28. DOI: 10.1007/s11069-010-9634-2.

Jacquez, G (2010). 'Geographic Boundary Analysis in Spatial and Spatio-Temporal Epidemiology: Perspective and Prospects'. *Spatial and Spatio-Temporal Epidemiology* 1 (4), pp. 207–218. DOI: 10.1016/j.sste.2010.09.003.

Jacquez, G, A Kaufmann and P Goovaerts (2008). 'Boundaries, Links and Clusters: A New Paradigm in Spatial Analysis?' *Environmental and Ecological Statistics* 15 (4), pp. 403–419. DOI: 10.1007/s10651-007-0066-4.

Jacquez, G, S Maruca and M Fortin (2000). 'From Fields to Objects: A Review of Geographic Boundary Analysis'. *Journal of Geographical Systems* 2 (3), pp. 221–241. DOI: 10.1007/PL00011456.

Jiang, B (2015). 'Geospatial Analysis Requires a Different Way of Thinking: The Problem of Spatial Heterogeneity'. *GeoJournal* 80 (1), pp. 1–13. DOI: 10.1007/s10708-014-9537-y.

Keitt, T and D Urban (2005). 'Scale-Specific Inference Using Wavelets'. *Ecology* 86 (9), pp. 2497–2504. DOI: 10.1890/04-1016.

Kent, M, W Gill, R Weaver and R Armitage (1997). 'Landscape and Plant Community Boundaries in Biogeography'. *Progress in Physical Geography* 21 (3), pp. 315–353. DOI: 10.1177/030913339702100301.

Kent, M, R Moyeed, C Reid, R Pakeman and R Weaver (2006). 'Geostatistics, Spatial Rate of Change Analysis and Boundary Detection in Plant Ecology and Biogeography'. *Progress in Physical Geography* 30 (2), pp. 201–231. DOI: 10.1191/0309133306pp477ra.

Koenker, R and G Bassett (1982). ‘Robust Tests for Heteroscedasticity Based on Regression Quantiles’. *Econometrica* 50 (1), pp. 43–61. DOI: 10.2307/1912528.

Kolasa, J and C Rollo (1991). ‘The Heterogeneity of Heterogeneity: A Glossary’. In: *Ecological Heterogeneity*. Ed. by J Kolasa and S Pickett. Heidelberg: Springer. Chap. 1, pp. 1–23. DOI: 10.1007/978-1-4612-3062-5_1.

Lague, D, N Brodu and J Leroux (2013). ‘Accurate 3D Comparison of Complex Topography with Terrestrial Laser Scanner: Application to the Rangitikei Canyon (N-Z)’. *ISPRS Journal of Photogrammetry and Remote Sensing* 82, pp. 10–26. DOI: 10.1016/j.isprsjprs.2013.04.009.

Laurance, W, R Didham and M Power (2001). ‘Ecological Boundaries: A Search for Synthesis’. *Trends in Ecology & Evolution* 16 (2), pp. 70–71. DOI: 10.1016/S0169-5347(00)02070-X.

Lee, D and R Mitchell (2012). ‘Boundary Detection in Disease Mapping Studies’. *Biostatistics* 13 (3), pp. 415–426. DOI: 10.1093/biostatistics/kxr036.

Legewie, J and M Schaeffer (2016). ‘Contested Boundaries: Explaining Where Ethno-Racial Diversity Provokes Neighborhood Conflict’. *American Journal of Sociology* 122 (1), pp. 125–161. DOI: 10.1086/686942.

Lillesand, T, R Kiefer and J Chipman (2015). *Remote Sensing and Image Interpretation*. 7th ed. Hoboken, NJ: Wiley & Sons.

Lohrer, A, I Rodil, M Townsend, L Chiaroni, J Hewitt and S Thrush (2013). ‘Biogenic Habitat Transitions Influence Facilitation in a Marine Soft-Sediment Ecosystem’. *Ecology* 94 (1), pp. 136–145. DOI: 10.1890/11-1779.1.

Lovelace, R, M Birkin, P Cross and M Clarke (2016). ‘From Big Noise to Big Data: Toward the Verification of Large Data Sets for Understanding Regional Retail Flows’. *Geographical Analysis* 48 (1), pp. 59–81. DOI: 10.1111/gean.12081.

Mandelbrot, B and R Hudson (2004). *The (Mis)behavior of Markets: A Fractal View of Risk, Ruin, and Reward*. New York, NY: Basic Books, p. 328.

Ord, J and A Getis (1995). ‘Local Spatial Autocorrelation Statistics: Distributional Issues and an Application’. *Geographical Analysis* 27 (4), pp. 286–306. DOI: 10.1111/j.1538-4632.1995.tb00912.x.

— (2001). ‘Testing for Local Spatial Autocorrelation in the Presence of Global Autocorrelation’. *Journal of Regional Science* 41 (3), pp. 411–432. DOI: 10.1111/0022-4146.00224.

— (2012). ‘Local Spatial Heteroscedasticity (LOSH)’. *The Annals of Regional Science* 48 (2), pp. 529–539. DOI: 10.1007/s00168-011-0492-y.

Páez, A and D Scott (2005). ‘Spatial Statistics for Urban Analysis: A Review of Techniques with Examples’. *GeoJournal* 61 (1), pp. 53–67. DOI: 10.1007/s10708-005-0877-5.

Patil, G, R Modarres, W Myers and P Patankar (2006). ‘Spatially Constrained Clustering and Upper Level Set Scan Hotspot Detection in Surveillance Geoinformatics’. *Environmental and Ecological Statistics* 13 (4), pp. 365–377. DOI: 10.1007/s10651-006-0017-5.

Perkins, T, T Scott, A Le Menach and D Smith (2013). ‘Heterogeneity, Mixing, and the Spatial Scales of Mosquito-Borne Pathogen Transmission’. *PLOS Computational Biology* 9 (12), e1003327. DOI: 10.1371/journal.pcbi.1003327.

Sengstock, C, M Gertz, F Flatow and H Abdelhaq (2013). ‘A Probablistic Model for Spatio-Temporal Signal Extraction from Social Media’. In: *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL’13)*. Ed. by Craig Knoblock and Markus Schneider. Orlando, FL: ACM, pp. 264–273. DOI: 10.1145/2525314.2525353.

Sparrow, A (1999). 'A Heterogeneity of Heteogeneities.' *Trends in Ecology & Evolution* 14 (11), pp. 422–423. DOI: 10.1016/S0169-5347(99)01735-8.

Steiger, E, B Resch and A Zipf (2016b). 'Exploration of Spatiotemporal and Semantic Clusters of Twitter Data Using Unsupervised Neural Networks'. *International Journal of Geographical Information Science* 30 (9), pp. 1694–1716. DOI: 10.1080/13658816.2015.1099658.

Strayer, D, M Power, W Fagan, S Pickett and J Belnap (2003). 'A Classification of Ecological Boundaries'. *BioScience* 53 (8), pp. 723–729. DOI: 10.1641/0006-3568(2003)053[0723:ACOEB]2.0.CO;2.

Tian, P, X Cao, J Liang, L Zhang, N Yi, L Wang and X Cheng (2014). 'Improved Empirical Mode Decomposition Based Denoising Method for LiDAR Signals'. *Optics Communications* 325, pp. 54–59. DOI: 10.1016/j.optcom.2014.03.083.

Tukey, J (1977). *Exploratory Data Analysis*. Boston, MA: Addison-Wesley.

Turner, M (1989). 'Landscape Ecology: the Effect of Pattern on Process'. *Annual Review of Ecology and Systematics* 20 (1), pp. 171–197. DOI: 10.1146/annurev.es.20.110189.001131.

Wagner, H and M Fortin (2005). 'Spatial Analysis of Landscapes: Concepts and Statistics'. *Ecology* 86 (8), pp. 1975–1987. DOI: 10.1890/04-0914.

Walck, C (2007). *Hand-Book on Statistical Distributions for Experimentalists*. Stockholm: University of Stockholm.

Wang, J, T Zhang and B Fu (2016a). 'A Measure of Spatial Stratified Heterogeneity'. *Ecological Indicators* 67, pp. 250–256. DOI: 10.1016/j.ecolind.2016.02.052.

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 (5), pp. 868–887. DOI: 10.1080/13658816.2014.1002499.

Westerholt, R, E Steiger, B Resch and A Zipf (2016). 'Abundant Topological Outliers in Social Media Data and Their Effect on Spatial Analysis'. *PLOS ONE* 11 (9), e0162360. DOI: 10.1371/journal.pone.0162360.

Womble, W (1951). 'Differential Systematics'. *Science* 114 (2961), pp. 315–322. DOI: 10.1126/science.114.2961.315.

Xu, M, C Mei and N Yan (2014a). 'A Note on the Null Distribution of the Local Spatial Heteroscedasticity (LOSH) Statistic'. *The Annals of Regional Science* 52 (3), pp. 697–710. DOI: 10.1007/s00168-014-0605-5.

Ye, X, T Wang, A Skidmore, D Fortin, G Bastille-Rousseau and L Parrott (2015). 'A Wavelet-Based Approach to Evaluate the Roles of Structural and Functional Landscape Heterogeneity in Animal Space Use at Multiple Scales'. *Ecography* 38 (7), pp. 740–750. DOI: 10.1111/ecog.00812.

II.3.8 Supporting Information

S1 Fig. Overview of the investigated study area

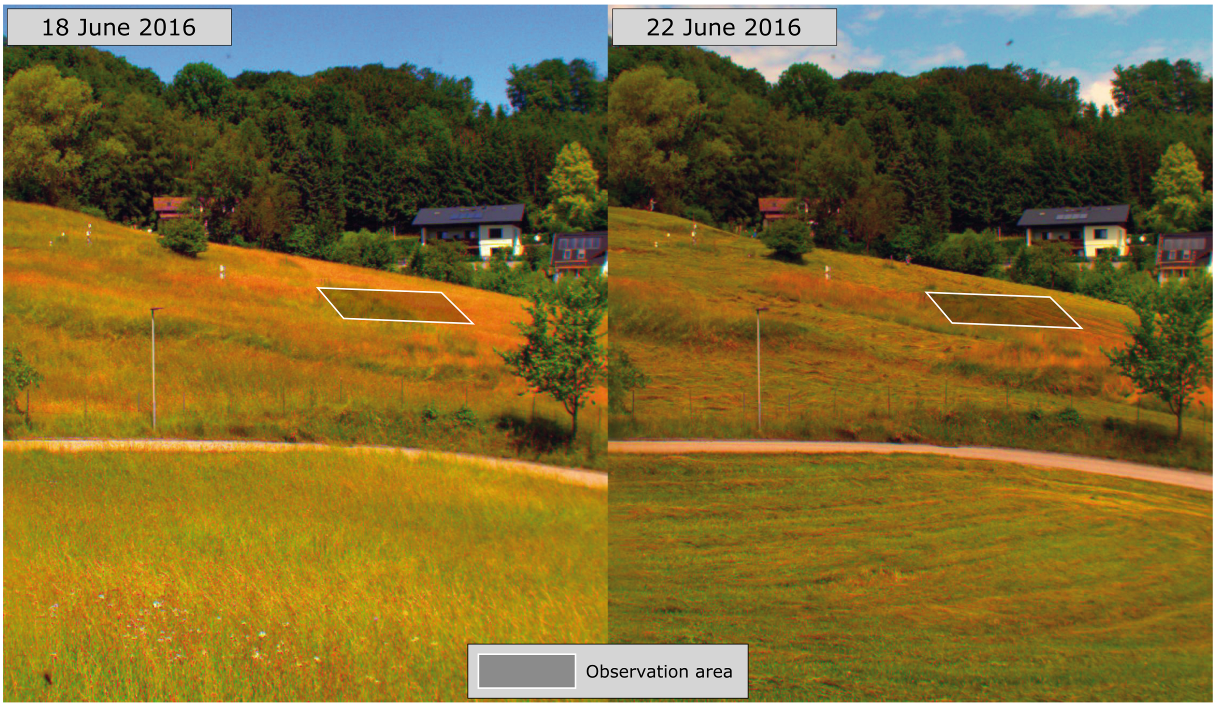


Figure II.3.11: The investigated study area in Gresten, Austria, before and after the mowing.

II.3.9 Appendix

II.3.9.1 A1. Relationship between LOSH and LSD

The ratio between LOSH and LSD is given by

$$\frac{H_i}{LSD_i} = \frac{\frac{1}{n_i}\sum_{j\in\mathcal{N}_i}|e_j|^2 \cdot \sum_{j\in\mathcal{N}} w_{ij}|e_j|^2 \cdot \sum_{j\in\mathcal{N}} w_{ij}}{\frac{1}{n}\sum_{j\in\mathcal{N}}|e_j|^2 \cdot \sum_{j\in\mathcal{N}} w_{ij}|e_j|^2 \cdot \sum_{j\in\mathcal{N}} w_{ij}} = \frac{n\cdot h_i}{\sum_{j\in\mathcal{N}}|e_j|^2} = h_i \cdot h_1^{-1}.$$

From this, LOSH and LSD can be inferred:

$$H_i = \frac{LSD_i \cdot h_i}{h_1} \qquad \text{and} \qquad LSD_i = \frac{H_i \cdot h_1}{h_i}.$$

The ratio above shows that LSD can be turned into LOSH and vice versa, demonstrating that both measures represent a scaled version of the respective other.

II.3.9.2 A2. Expectation and variance of spatially weighted mean estimates

The mean and the variance of the spatially weighted mean estimates Y_j are affected by the spatial weighting structure. Let $\{X_k\}$ be independent real random variables indexed over the neighbourhood set of spatial units $\mathcal{N}_j$. Let further $\{w_{jk}\}$ denote the set of spatial weights upon $\mathcal{N}_j$ that sums up to

$W_j = \sum_{k \in \mathcal{N}_j} w_{jk}$, and $Y_j = (1/W_j) \sum_{k \in \mathcal{N}_j} w_{jk} x_k$ be a local spatial average as defined in Equation II.3.1. Under local randomisation the expectation $E[Y_j]$ is given by

$$E\left[\frac{1}{W_j}\right] \sum_{k \in \mathcal{N}_j} w_{jk} X_k = \frac{1}{W_j} \sum_{k \in \mathcal{N}_j} w_{jk} E[X_k] = \frac{1}{W_j} \sum_{k \in \mathcal{N}_j} w_{jk} \mu_{X_j} = \mu_{X_j}.$$

The location of the mean is thus not affected by the spatial weights. Note that the weighted sample mean is a linear combination $(w_{j1}/W_j) X_1 + \cdots + \left(w_{jn_j}/W_j\right) X_{n_j}$ of independent random variables (*i. e.*, local independent under the randomisation assumption of H_0 of LSD). The variance of Y_j is therefore obtained by applying the rule for the variance of linear combinations of independent random variables, which is given by $Var\left[\sum_k a_k X_k\right] = \sum_k a_k^2 Var[X_k]$, and thus for Y_j yields

$$Var[Y_j] = \sum_{k \in \mathcal{N}_j} \left(\frac{w_{jk}}{W_j}\right)^2 \sigma^2_{\mathcal{N}_j} = \sigma^2_{\mathcal{N}_j} \cdot \sum_{k \in \mathcal{N}_j} \left(\frac{w_{jk}^2}{W_j^2}\right).$$

Unlike the mean value, the variance is scaled by the weighting scheme. The above relationship for the variance is demonstrated for the ordinary unweighted case through substituting 1 for each weight w_{jk}. Their sum then yields n_j and the above equation reduces to the variance of the unweighted sample mean $\sigma^2_{N_j} \sum_{k=1}^{n_j} \left(1^2/n_j^2\right) = \sigma^2_{N_j}/n_j$.

II.3.9.3 A3. Averaging of several local variances

The variance of a random variable X is generally given by the shift rule $Var[X] = E\left[X^2\right] - E[X]^2$. We already determined the estimator of $E[X]^2$ in Equation II.3.8, which is $\bar{x}_c^2$. The estimator of $E\left[X^2\right]$ takes the form $\sum_{i=1}^n x_i^2/n$, though it must take into account the grouping in the data (*i. e.*, the spatially overlapping neighbourhoods). To simplify the following steps, the Bessel correction of the unbiased sample variance s^2_{n-1} is reversed first:

$$\dot{s}^2 = \left(\frac{n-1}{n}\right) s^2_{n-1} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \left(\frac{1}{n} \sum_{i=1}^n x_i^2\right) - \bar{x}^2.$$

From this, the corresponding sum of squares is obtained:

$$n\left(\dot{s}^2 + \bar{x}^2\right) = \sum_{i=1}^n x_i^2.$$

This sum of squares can be split up into a series of partial sums. For the case of partly overlapping spatial neighbourhoods this gives $\sum_{i \in \mathcal{N}_1} x_i^2 + \cdots + \sum_{i \in \mathcal{N}_n} x_i^2$. Each of these summations can be represented through their respective local sample variance and local mean value. We get that

$$\sum_{i \in \mathcal{A}_i} \sum_{j \in \mathcal{N}_i} x_j^2 = \sum_{i \in \mathcal{A}_i} n_i \left(\dot{s}_i^2 + \bar{x}_i^2\right),$$

and the substitution of this back into $Var\left[X\right]=E\left[X^2\right]-E\left[X\right]^2$ yields

$$\dot{s}_c^2=\frac{\sum_{j\in\mathcal{A}_i}n_j\left(\dot{s}_j^2+\bar{x}_j^2\right)}{\sum_{j\in\mathcal{A}_i}n_j}-\bar{x}_c^2.$$

In order to obtain an unbiased result, we reverse the previous elimination of the Bessel correction. The rescaled version of $\dot{s}_c^2$ is

$$s_c^2=\frac{\sum_{j\in\mathcal{A}_i}n_j}{\left(\sum_{j\in\mathcal{A}_i}n_j\right)-n_{\mathcal{A}_i}}\cdot\dot{s}_c^2=\frac{\sum_{j\in\mathcal{A}_i}\left(n_j-1\right)\left(s_j^2+\bar{x}_j^2\right)}{\left(\sum_{j\in\mathcal{A}_i}n_j\right)-n_{\mathcal{A}_i}}-\bar{x}_c^2.$$

II.3.9.4 A4. Derivation of the prior

The prior combines the two marginal densities

$$f\left(\mu_{X_j}\mid\sigma_{X_j}^2;\mu_0,\sigma_0^2=\sigma_{X_j}^2/n_i\right)=\frac{\sqrt{n_i}}{\sqrt{2\pi}\sigma_{X_j}}\exp\left(-\frac{n_i\left(\mu_{X_j}-\mu_0\right)^2}{2\sigma_{X_j}^2}\right)$$

and

$$f\left(\sigma_{X_j}^2;\upsilon_0,\tau_0^2\right)=\frac{\left(\frac{\tau_0^2\upsilon_0}{2}\right)^{\upsilon_0/2}}{\Gamma\left(\frac{\upsilon_0}{2}\right)\sigma_{X_j}^{2+\upsilon_0}}\exp\left(-\frac{\upsilon_0\tau_0^2}{2\sigma_{X_j}^2}\right)$$

into their joint product

$$\pi\left(\mu_{X_j},\sigma_{X_j}^2\right)=\frac{\left(\frac{\tau_0^2\upsilon_0}{2}\right)^{\upsilon_0/2}}{\sqrt{2\pi}\sigma_{X_j}\sqrt{n_i}\cdot\Gamma\left(\frac{\upsilon_0}{2}\right)\sigma_{X_j}^{2+\upsilon_0}}\exp\left(-\frac{n_i\left(\mu_{X_j}-\mu_0\right)^2+\upsilon_0\tau_0^2}{2\sigma_{X_j}^2}\right),$$

where Γ is the gamma function. If all normalising constants are omitted, the prior is obtained as follows:

$$\pi\left(\mu_{X_j},\sigma_{X_j}^2\right)\propto\frac{1}{\sigma_{X_j}^{3+\upsilon_0}}\exp\left(-\frac{n_i\left(\mu_{X_j}-\mu_0\right)^2+\upsilon_0\tau_0^2}{2\sigma_{X_j}^2}\right).$$

II.3.9.5 A5. Derivation of the posterior

The observed data $Y_i\sim N\left(\mu_{X_i},\sigma_{X_i}^2\right)$ is described by the normal likelihood function

$$f\left(Y_i\mid\mu_{X_i},\sigma_{X_i}^2\right)=\frac{1}{\sqrt{2\pi}\sigma_{X_i}}\exp\left(-\frac{\left(Y_i-\mu_{X_i}\right)^2}{2\sigma_{X_i}^2}\right)\propto\sigma^{-1}\exp\left(-\frac{\left(Y_i-\mu_{X_i}\right)^2}{2\sigma_{X_i}^2}\right).$$

The multiplication of this likelihood by the prior from Appendix II.3.9.4 yields the following posterior density:

$$f\left(\mu_{X_i},\sigma_{X_i}^2\mid Y_i\right)\propto\frac{\left(\frac{\tau_0^2\upsilon_0}{2}\right)^{\upsilon_0/2}}{2\pi\sqrt{n_i}\cdot\Gamma\left(\frac{\upsilon_0}{2}\right)\sigma_{X_i}^{4+\upsilon_0}}\exp\left(-\frac{n_i\left(\mu_{X_i}-\mu_0\right)^2+\left(Y_i-\mu_{X_i}\right)^2+\upsilon_0\tau_0^2}{2\sigma_{X_i}^2}\right).$$

After again omitting all normalising constants, we arrive at

$$f\left(\mu_{X_i}, \sigma_{X_i}^2 \mid Y_i\right) \propto \frac{1}{\sigma_{X_i}^{4+\upsilon_0}} \exp\left(-\frac{n_i\left(\mu_{X_i}-\mu_0\right)^2+\left(Y_i-\mu_{X_i}\right)^2+\upsilon_0\tau_0^2}{2\sigma_{X_i}^2}\right),$$

which is the non-normalised posterior density.

II.4 Integrating Geographic Information into Survey Research: Current Applications, Challenges and Future Avenues

Abstract

Geographic information science (GIScience) offers survey researchers a plethora of rapidly evolving research strategies and tools for data acquisition and analysis. However, the potential for incorporating geographic information systems (GIS) tools into traditional survey research has not yet been fully appreciated by survey researchers. In this article, we provide a comprehensive overview of recent advances and challenges in leveraging this potential. First, we present state-of-the-art applications of GIS tools in traditional survey research, drawing mainly on examples from psychological survey research (e. g., socioecological psychology). We also discuss innovative GIS tools (e. g., wearables) and GIScience methods (e. g., citizen sensing) that expand the scope of traditional surveys. Second, we highlight a number of challenges and problems (e. g., choice of spatial scale, statistical issues, privacy concerns) and—where possible—suggest remedies. With increasing awareness of the potential that GIS tools hold for survey research, and intensified dialogue between researchers from both sides, more fruitful collaboration appears within reach.

Keywords: GIScience, GIS Tools, Socioecological Psychology, Wearables, Social Media

II.4.1 Introduction and Overview

II.4.1.1 Relevance of Geospatial Aspects

Recent advances in integrating geographic information science (GIScience) into psychology and survey methodology may be considered *evolutionary* by some researchers and *revolutionary* by others. Some observers view these advances as a paradigmatic shift that justifies the term “spatial turn” (*e. g.*, (Richardson et al. 2013)). At any rate, the sheer number of new research strategies and geographic information system (GIS) tools can be daunting, and it is hard for those outside GIScience to keep pace with recent developments. Perhaps for this reason, psychology has, for the most part, been slow to adopt some of the promising innovations offered by GIScience and related disciplines (see Appendix for a glossary of terms). There is the danger of a widening gap, if not detachment, between scientific communities in terms of concepts, methods, and tools.

To avoid this pitfall, we think it is vital to inform survey researchers in general, and psychological researchers in particular, about the methodological potential that GIS tools (*e. g.*, techniques for acquiring, analysing, and visualising geographic data) hold. Although in-depth discussions of single geospatial techniques abound, an up-to-date integrative overview of these innovations is absent from the literature. The present paper is the first to offer such an overview from the perspective of GIScientists and psychological researchers practically involved in the application of GIS tools in survey research. We restrict our overview to those GIScience techniques that concern the analysis of survey *data*, that is, techniques that augment traditional surveys by incorporating geospatial information, or that complement traditional surveys by providing new forms of data and data analysis. We do not consider how GIScience techniques can be used to improve traditional survey *methodology* (*e. g.*, for monitoring field work and interviewer behaviour, or for designing sampling frames and weighting schemes).

The purpose of our paper is twofold. First, it is intended as a starting point for survey researchers interested in applying innovative GIS tools in their work. Despite the rise of other approaches (*e. g.*, neuroscientific methods, computer-based cognitive tests), questionnaires and population surveys are still among the most widely used tools for collecting data on individuals and social groups, notably in psychology. Second, we hope to foster an informed discussion between two inherently methodological disciplines—GIScience and survey research. We encourage researchers from both sides to critically examine and make use of solutions offered by each field, to develop a common theoretical framework, and to adopt each others' insights, tools, and methods. Such an intensified dialogue, we believe, may challenge our traditional understanding of survey data and methodology, thus paving the way for future innovations in survey research (see Warf and Arias (2009)).

Yet why should survey researchers consider spatial aspects at all? Geospatial information might be useful at various stages of the survey design. Surveys can be supplemented with geographic coordinates, such as the location at which a respondent completed the questionnaire (*i. e.*, the point of origin) or where the respondent predominantly lives (*i. e.*, place of residence). Either respondents' exact geographic locations or their approximated locations (*e. g.*, via regional codes) may be available. These geographic coordinates can then be used to incorporate contextual data into subsequent analyses. For instance, socioeconomic data on households in a neighbourhood, or regional divorce rates as a proxy for individualisation in a society, may be contextual variables and may complement individual-level data analyses (Lechner et al. 2017). Geographic visualisation techniques may also provide additional insights into the spatial distribution of survey results. By analysing and mapping biophysiological data from study participants who wear trackable devices while moving through space (Tröndle et al. 2014), or by plotting Twitter-based information almost in real time to geographic maps (Curini et al. 2015), researchers can follow social processes at unprecedented spatiotemporal resolutions.

II.4.1.2 Survey Concepts

It is vital to note that the understanding of the term "survey" often differs between survey researchers and GIScientists. For survey researchers, the survey questions—augmented with georeferenced data—are typically meant to reveal information *about individuals*. Spatial information (*e. g.*, location-related information) is added to better *understand humans or social systems*. Survey researchers often operate with a survey definition that involves the collection of data from a sample of elements drawn from a well-defined population through the use of a questionnaire (Visser et al. 2000). Hence, "a survey can be seen as a research strategy in which quantitative information is systematically collected from a relatively large sample taken from a population" (Leeuw et al. 2008, p. 3). GIScientists, by contrast, typically focus on extracting information *about places and spaces* (*e. g.*, street corners, cities, regions, or countries) and the phenomena and developments they undergo over time. For them, survey questions add information from a human perspective to better *understand the environment*. Yet GIScientists often endorse a rather minimalistic definition of "survey" as a method of gathering information from any sample of individuals (Scheuren 2004). Their broader term may include gathering data *en passant* from social media users, rather than collecting answers from survey *respondents*.

Section II.4.2 illustrates the breadth of potential applications of GIScience methods and GIS tools in survey research. These applications include the augmentation of classic survey data with georeferenced information, on the one hand, and new techniques to obtain spatiotemporally distributed data from GIScience, on the other. In Section II.4.3, we outline a number of challenges shared by most of these applications. They include, for example, spatial scale usage, the modifiable areal unit problem, the arbitrary nature of maps and visualisation techniques potential pitfalls in the analysis of contextual data, fallacies when dealing with different levels of analysis (individual and aggregate data); issues surrounding user-generated data, and privacy and data protection issues. Where possible, we make methodological recommendations to deal with these challenges. Finally, in Section II.4.3, we also consider how traditional survey data and methodology may be of benefit to GIScience.

II.4.2 Recent Applications of GIS Tools to Augment Survey Data

Before we begin our overview of recent applications, let us follow Agnew and Livingstone (2011) and Tuan (1977) and clarify the distinctions between important GIScience terms that readers may or may not be familiar with: space, location, and place. *Space* is understood as an abstract, non-semantically enriched geographic space spanning planet Earth, in which processes of interest occur. *Location* demarcates a specific point or area in this space, mostly delimited by crisp boundaries, which can be represented in GIS. In contrast, *place* is defined as space infused with human meaning. As this meaning is almost never specified with perfect intersubjectivity, its borders are often fuzzy and ambiguous. Several scientific disciplines deal with these concepts. The methodological companion to geography is geographic information science, or GIScience (Goodchild 1992; Goodchild 2010). GIScience and GIS tools are closely related, and they provide partly overlapping innovations (see Appendix II.4.5.1).

Augmenting classic survey data with georeferenced data represents a first way in which geographic information about individuals and their backgrounds is utilized in the social sciences (Hoffmeyer-Zlotnik 2013; Okner 1972; Schnell 2013b). Survey datasets that are geocoded—that is, datasets that contain one or more variables assigning a geographic location (*e. g.*, an exact location or a more coarse location such as a postal code or an administrative unit) to each response unit—can be merged with geotagged contextual information, thereby greatly enhancing the value of these augmented datasets to investigate new research questions (Meyer and Enzler 2013; Okner 1972; Schnell 2013b). To take the Swiss Environmental Survey as an example, regional statistics on environmental factors (*e. g.*, pollution, emissions) were linked to respondents' subjective impressions of environmental stress to gain a better understanding of the relationship between objective contextual variables and participants' subjective responses (Diekmann and Meyer 2010). By adding contextual information to individual respondent data, cross-level relationships can be explored (Hoffmeyer-Zlotnik 2013; RatSWD 2012). Rich contextual data are now offered by various public institutions (*e. g.*, register, census, and economic data), private organisations and companies (*e. g.*, operational and customer-tracking data), and accumulated sources (*e. g.*, social media, representative surveys; for a comparison, see Hüttenrauch (2016)). Some public-use surveys, such as the European Social Survey (ESS), already include a large number of contextual variables at different geographic levels (*e. g.*, national and regional migration, or unemployment rates) in their data distributions, and make these data readily available to researchers. Furthermore, various kinds of geospatial data have become publicly available, for instance, authoritative topographic data (*e. g.*, OpenStreetMap). The following sub-sections describe possible applications of these contextual data in research.

II.4.2.1 Socioecological Psychology

The emerging field of *socioecological psychology*, also known as *geographic(al) psychology* (for reviews, see Rentfrow (2013) and Oishi (2014)), utilises the new possibilities of integrating contextual information and traditional survey data. Socioecological psychology directs attention to how objective (as opposed to perceived) features of macro-level social ecologies (*i. e.*, physical, interpersonal, economic, or political environments) shape human behaviour, cognition, and emotion—and how human behaviour, in turn, gives rise to changing social ecologies ("niche construction"). Extant socioecological studies mostly relate individual-level psychological outcomes to socioecological variables that are measured at (*not aggregated to*) the national or regional level and that assume the role of a predictor of individual-level variability (*e. g.*, Talhelm et al. (2014)) or a moderator of individual-level relationships (*e. g.*, Jokela et al. (2015) and Lechner et al. (2017)). For example, Talhelm et al. (2014) were able to show how the agricultural legacy of regions in China shapes the cultural and psychological traits of these regions' inhabitants until the present day; they found that a history of farming rice was linked to more interdependent traits, whereas a history of farming wheat was linked to more independent cultural patterns. Other socioecological studies extend this focus, investigating the spatial distribution of psychological constructs and their contextual-level correlates. The level of analysis in the latter strand of studies is thus a geographic one: Individual survey responses (such as answers to a Big Five personality battery) per geographic unit are aggregated in order to map them to spatial contexts and link them to each other (Rentfrow et al. 2015) or, alternatively, to external data sources such as health statistics (Kitchen et al. 2012) or entrepreneurship rates (Obschonka et al. 2013).

Although socioecological psychology is still in a nascent state, it already exerts a noticeable influence on psychological theorising. Socioecological studies are contributing to a gradual shift in the traditional focus on the individual toward a more environmentally informed understanding of the discipline's key phenomena. (Arguably, this marks a veritable "spatial turn", especially in the fields of personality psychology and social psychology.) While this development opens up new avenues for collaboration with disciplines at the interface of human behaviour and geography (Rentfrow 2013; Oishi 2014), it also brings methodological challenges, which will be discussed later.

II.4.2.2 Survey Responses as a Function of Georeferenced Indicators

Geographic context can also be used to identify (and correct for) sources of variance in survey responses. Depending on one's focus, such variance may represent either explained variability or nuisance variance in survey responses. For example, participants' life satisfaction scores might be influenced by (a) aspects of the natural environment, such as the greenness of neighbourhoods (Leslie et al. 2010), (b) the built environment (McGinn et al. 2007), or (c) circumstances of the survey location, such as indoor versus outdoor interviewing (Iosa et al. 2012).

A prime application of this approach is the influence of weather on mood and well-being. With governments increasingly adopting well-being as a policy target, subjective ratings of life satisfaction and happiness are often important indicators that complement panel data on regional and macroeconomic factors (Schyns 1998) (for a recent Eurobarometer analysis, see Brulé and Veenhoven (2014)). For well-being to inform public policy choices, one would like to be sure that any regional differences in average well-being ratings are truly related to economic prosperity and other policy-relevant factors, rather than being driven by "nuisance" factors such as the climate at survey locations (Rehdanz and Maddison

2005; Brulé and Veenhoven 2015) or by transient weather conditions during interviews (Schimmack et al. 2002).

Laboratory and field evidence has shown that judgements of life satisfaction are influenced by the reported weather conditions (Schwarz and Clore 1983), and that ambient temperature ratings, in turn, depend on people's current mood (Messner and Wänke 2011). Although this cross-sectional evidence was challenged by panel data (Lucas and Lawless 2013; Schmiedeberg and Schröder 2014), more recent panel data providing detailed information on all relevant weather variables at the precise location and time of an interview have, in fact, revealed variation in life satisfaction scores as a function of weather (Feddersen et al. 2016). Beyond well-being, studies have shown that Big Five trait ratings can also be influenced by contextual factors such as weather (Rammstedt et al. 2015). Increasing spatial granularity yields better evidence on climatic and weather influences on survey responses. It may become possible to purge respondents' mood, well-being, or life satisfaction ratings of unwarranted nuisance variance and to obtain unbiased scores that offer a more solid ground for policy decisions.

II.4.2.3 Experience Sampling in Dynamic Contexts

Methods of studying individuals in their natural settings—often in real time, on repeated occasions, and free of retrospective biases—offer tremendous potential for survey research. One such method—and one that has recently gained some popularity—is the experience sampling, or event sampling, method (ESM; Reis and Gable (2000)), which allows respondents to be surveyed in their natural environments on repeated occasions (Larson and Csikszentmihalyi 1983; Hektner et al. 2007). ESM prompts participants (*e. g.*, via mobile devices) to take a survey at fixed time intervals or randomly throughout the day. In this way, the likelihood of events, the base rate of behaviours, or the prevalence of feelings can be surveyed amidst temporal fluctuations of experiences and dynamic transitions between places. The recent emergence of mobile electronic devices allows even large crowds to be observed at nearly any time and place so as to investigate relationships with increased ecological validity (Shiffman 2007).

There are three typical ESM procedures—signal-contingent (survey after notification via pager or SMS text message), event-contingent (recording data after predefined events have occurred), and interval-contingent (data acquisition after periods of time have passed)—whose respective (dis-)advantages have been described elsewhere (Conner and Barrett 2012). Here, we would like to stress that adding a *spatial layer* to this threefold distinction allows context-aware ESM to be used. Augmenting ESM data with location data (*e. g.*, Global Positioning System (GPS) coordinates gathered by users' mobile devices) offers a convenient way of conducting surveys at predetermined locations (which allows further data to be gathered about these locations *as socially relevant places*). Location data might help to explain individuals' attitudes and behaviours. These data include not only static factors, such as types of buildings or population density, or rather stable influences such as unemployment rates, but also each individual's exposure to noise at specific workplaces, stressful traffic encounters at specific intersections, etc.

So far, traditional population surveys mostly abstract from the dynamic contexts in which respondents generate their responses, or in which they have experiences that they report only later. From this perspective, survey samples must first and foremost mirror the population. However, it is worthwhile reflecting on the fact that any interview represents a mere snapshot of a respondent's state of mind generated within a specific spatiotemporal slice of the environment. Population surveys typically leave such short-term volatility and spatial dynamics of survey responses unmonitored. By linking ESM data to rich contextual information such as location and time, survey research proceeds to the next stage, where human characteristics are explained as a function of idiosyncratic events, personal contexts, and participants' spatial transitions. Research in health-related and occupational fields has started to incorporate these new possibilities (Sonnentag et al. 2012; Richardson et al. 2013)—for instance, by using ESM to investigate whether environmental factors, such as rare exposure to nature, might influence mental health (Reichert et al. 2016).

II.4.2.4 Objective Data Capture by Means of Wearables

While ESM focuses on the *subjective* experiences that respondents have over any pre-specified time span, these data can be amended with *objective* data on the same individuals. There has recently been a rapid rise in the use of wearable sensors to measure a number of physiological parameters (*e. g.*, heart rate variability, blood pressure, or skin conductance; Swan (2012)). These sensors, together with the increasing penetration rate of smartphones across age groups, have paved the way toward virtually ubiquitous data acquisition and have opened up new opportunities for obtaining information about the environment (Triantafyllidis et al. 2017).

For example, the so-called *quantified-self* movement promotes the use of sensor technology for acquiring data about one's own daily life, ranging from concrete physiological parameters to rather abstract parameters such as physical performance and associated affective consequences (*e. g.*, emotional states). This movement is reinforced by the rapid development of wearable sensors that allow for continuous surveillance of everyday activities and daily routines (Swan 2013). Although people are joining the quantified-self movement mainly to achieve *self-awareness through self-monitoring* (Ayobi et al. 2016), it has also led to rising awareness of physiological sensor devices among the wider public. As a result, citizens' familiarity with the use of sensors has dramatically increased. This is of particular importance for survey research, as most of the sensor-based quantified-self applications are explicitly geolocated, which allows survey data to be complemented with additional data from wearables at high temporal and spatial resolution, thereby yielding information that cannot be obtained by simply asking survey questions. Physiological signals obtained from wearable sensors can then be used to make inferences about individual experiences that are associated with, or can be mapped to, events and places in the environment. For instance, one could compare a survey intended to identify dangerous traffic intersections in a city with a study that captures heart rate and blood pressure data from drivers, cyclists, and pedestrians. This may help to identify city areas or spots of increased stress levels other than those identified by participants' subjective ratings. Sufficiently rich data may allow the emotional experiences of future pedestrians, cyclists, or motorists at the same location to be predicted, thus enabling a more citizen-centric planning of city infrastructure (Resch et al. 2015c).

II.4.2.5 Humans as Proactive Sensors (Rather Than as Respondents)

User-generated data are by no means limited to physiological data from wearable sensors that are collected for a specific purpose and under the researcher's control. A number of new approaches elicit, observe, or analyse information generated by individuals (and groups) that are hardly under the control of a researcher. Instead, participants act more and more as researchers of their own affairs, and thus control over the data-generation process is increasingly left to them. GIScience capitalises on this trend.

Citizen sensing describes a unique measurement approach in which persons do not merely deliver reports but rather act as non-technical, context-aware sensors with situational intelligence and extensive background knowledge about their present location (Resch 2013). Specifically, citizens are asked to provide their impressions, perceptions, and observations about a well-defined issue with explicit reference to geographic space. Akin to ESM, people provide their subjective recordings through eDiaries, which are designed to be *context-aware*. Contextualised reports can be gathered through dedicated smartphone apps (Triantafyllidis et al. 2017).

A recent example is the collection of citizens' subjective feelings and emotions about different places in the city (Resch et al. 2015a). Participants who move in geographic space are equipped with a smartphone app for reporting sensations and impressions—for instance, about traffic safety or public safety. Each dataset is associated with location and timestamp, which enables spatiotemporal analysis of the data. Apart from an immediate glimpse of a geographic context, this allows for an analysis of changes in ecosystems as continuously monitored through citizen-sensing technologies. Rather than acting as mere respondents to questionnaires, this approach empowers participants to proactively report not only on their spatial transitions but also on changes in ecosystems themselves.

One methodological implication of this survey method is that data are unlikely to be fully reproducible (Sagl and Resch 2014). From a survey research perspective, data reproducibility may not even be a goal (the data always present slices of information tied to time and context); yet from a GIScience perspective, the aim is to obtain (stable or reliable) information about the environment. Moreover, given users' proactive role in generating responses, the sampling and data generation processes are not necessarily controlled, and the observations are highly idiosyncratic. The reliability of such a measurement procedure is a far cry from that of representative population surveys with regular waves, or calibrated wearable sensors that produce measurements in well-defined physical settings. With high volatility in the sampling process, one problem is how to generalise to a whole population from self-selected samples who themselves determine what snapshots in time to deliver and when. In Section II.4.3, we elaborate on challenges common to survey research and GIScience.

II.4.2.6 Spatiotemporally Distributed Information in Social Media

Social media represent a useful resource for complementing survey data (Hill et al. 2013; Murphy et al. 2014). In contrast to citizen sensing, the analysis of data from social media does not require additional survey infrastructure (eDiary apps, digital surveys, etc.). Rather than surveying individuals about specific locations, this approach analyses aggregated, anonymised data from collective sources such as Flickr, Twitter, Foursquare, or the mobile phone network (Resch 2013). In this manner, information can be gained about the situational awareness of human environments and temporal dynamics on the basis of human communication, without attributing data to specific individuals. Social media posts reveal people's thoughts, emotions, or activities in geographic space, time, and linguistic space (Steiger et al. 2016b).

Although unprompted social media posts cannot be considered to be interviews in the formal sense of the word, people still provide "answers" (to questions that are not asked by an interviewer) by stating their perceptions and opinions. Yet, in contrast to classic surveys, it is more difficult to correctly map the target population. It may be difficult (albeit possible) to gauge opinions among specific subpopulations (Pötzschke and Braun 2016). However, it is almost impossible to get a representative picture of the entire general population. Therefore, social media data cannot replace targeted and structured surveys. Yet they are a useful extension, and they can yield additional insights (*e. g.*, via content analysis and text mining) that are not bound to pre-formulated questions and researcher-determined response categories. Instead, the focal topic can be determined by the social media user; the information obtained there is not elicited in a "synthetic situation" of a formal interview; and with regard to both the amount of content and its format, the user can express him- or herself freely. To provide an application example, the topic of sentiment analysis is currently gaining momentum. It deals with (the strength of) positive, negative, or neutral sentiments as conveyed by the polarity of words, sentences, or documents chosen by social media users (Liu and Zhang 2012). Newer approaches (Resch et al. 2016) automatically extract from Twitter tweets and posts from other social network sites affective content that corresponds to the fundamental model of basic emotions (anger, disgust, fear, happiness, sadness, and surprise; Ekman and Friesen (1971)) or the refined model of four basic emotions (happiness, sadness, fear, and anger; Jack et al. (2014)).

Leveraging user-generated social media data has one major advantage over traditional surveys: the possibility of near real-time analysis. Analysing user-generated data allows large-scale environmental, social, and geographic developments to be investigated "in the now", rather than after they occur. This kind of continuous cross-sectional monitoring—with unknown changes in the population that produces the data—is far from the quality of surveying a panel repeatedly in waves. However, it partly mitigates some shortcomings of traditional surveys, such as their low temporal resolution. Recent examples demonstrate the suitability of social media data in applications such as earthquake detection (Sakaki et al. 2010; Sakaki et al. 2013; Crooks et al. 2013) or the analysis of political sentiment (Wang et al. 2012a; Caldarelli et al. 2014; Vasiliu et al. 2016).

II.4.3 Challenges and Recommendations

Although the applications of geodata and GIS tools discussed thus far open up promising new research avenues, there are a number of challenges and pitfalls that survey researchers interested in applying these applications in their own research must bear in mind. Some are well known among GIScientists, but less so among survey researchers (and vice versa). In this section, we discuss these challenges and pitfalls and, where possible, suggest some remedies.

II.4.3.1 Spatial Scale

Spatial scale is a central issue for GIScientists, and thus for spatial data acquisition and analysis. *Scale* may refer to different components of a geographic analysis, such as the level of geographic detail at which observations are made ("sampling scale"), the spatial range at which processes of interest operate ("phenomenon scale"), or the degree of abstraction of a spatial analysis ("analysis scale") (Dungan et al. 2002; Ruddell and Wentz 2009). In a more technical sense, scale comprises *grain* (the smallest distinguishable parts possible) and *extent* (size of the study area) (Turner et al. 1989). This technical use has a geometric interpretation of scale. It prevails in physical geography, but is often inappropriate when investigating social processes through surveys. Socially meaningful spatial scales, such as neighbourhood, city, region, and nation, are often better suited for surveys (McMaster and Sheppard 2004).

Spatial scale is not only an objective frame of reference for spatial phenomena but also a property of people's subjective perceptions of space that has a strong bearing on their answers to questions about local geographic phenomena. People perceive their spatial surroundings in unique ways and imbue them with individual meaning (see Dangschat (2007)). Respondents also use idiosyncratic spatial scales that are limited by their spatial perception capabilities (Wender et al. 2002). For example, when asked about their "local community", voters in the British Election Study thought of completely different areas, ranging from streets and suburbs, through regions, to whole countries (Fieldhouse et al. 2014). Different mental systems are involved in perceiving phenomena at different spatial scales (Montello and Golledge 1998; Tversky et al. 1999; Hegarty et al. 2006), and a number of intrinsic and extrinsic factors influence the idiosyncratic spatial scale that people use (Witt et al. 2010), with well-established differences in spatial perception along the lines of gender (after puberty, males tend to perform better at spatial cognition; Weiss et al. (2003)), age (younger people tend to underestimate distances) (Sugovic and Witt 2013), emotions (impacting on perception) (Zadra and Clore 2011), and properties of the physical environment (visual/acoustic cues) (Iosa et al. 2012).

To illustrate, imagine an interviewer asking about an areal region such as an urban green space, a residential neighbourhood, or a local community. Respondents will use their subjective representations of the region based on their idiosyncratic conception of space. Using their imaginations, they will mentally construe the region in question. Hence, any information gained when looking at space through the eyes of survey respondents is potentially susceptible to scale differences, because the location, shape, and size of any perceived areas will influence respondents' answers. For example, whether there are enough early childhood education and care centres in a suburb might crucially depend on the correct or incorrect inclusion of an institution into the referenced area of interest, necessitating an accuracy check of respondents' mental representations. One can also try to exploit respondents' expertise. For instance, citizens may include areas in their answers that have not been considered by experts, which may be beneficial in natural hazard analysis when the goal is to identify areas prone to urban floods (Klonner et al. 2016).

The fact that different respondents use different (and highly idiosyncratic) spatial scales when thinking about the physical environment—and that even one and the same respondent may resort to different spatial scales when thinking about his or her surroundings—implies that respondents' answers in any survey on the physical and social environments do not refer to a fixed, objective geographic frame of reference. Spatial heterogeneity manifests itself as nuisance variance in the data, which increases the total survey error (Groves and Lyberg 2010). More specifically, heterogeneity in respondents' spatial scales causes instabilities of estimated means of quantitative data (due to spatial trends or discrete spatial regimes) and variances (spatial heteroscedasticity) (Ord and Getis 2012). Moreover, mixing highly different individual representations of arbitrary regional conceptions may not only render inferences based on such responses unreliable, or even bias-prone, but may even make numerical aggregates of respondents' answers difficult to interpret. This spatial-scale-related heterogeneity contributes to another form of (non-spatial) heterogeneity well-known in survey research, namely variability due to differing respondent and interviewer characteristics, or due to specific interactions between interviewers and respondents (Gabler and Lahiri 2009; Schaeffer et al. 2010; West et al. 2013).

There are different ways to address such scale-related issues: First, survey design requires careful construction of questions and questionnaires. All items that refer to a spatial phenomenon (*e. g.*, "your neighbourhood") should be as explicit as possible in order to lower the risk of ambiguities. One possible solution is to assist interviewees by providing a map of the area of interest whenever possible (*i. e.*, standardising the geographic presentation). However, this cannot always be smoothly integrated into the interview process. Moreover, it does not fully rule out the problem of different subjective geographic representations, and the maps provided to respondents restrict their answers to what is displayed on the map. As an alternative solution, mental maps and sketch maps can be used to document respondents' representations of geographic space (Boschmann and Cubbon 2014). Mental maps are free-form drawings, and sketch maps are accurate maps augmented by the respondents, allowing the researcher to get a clearer picture of the respondent's inherent scale use (Coulton et al. 2010). Using mental maps minimises the risk of accidentally mixing different scales during analysis and interpretation.

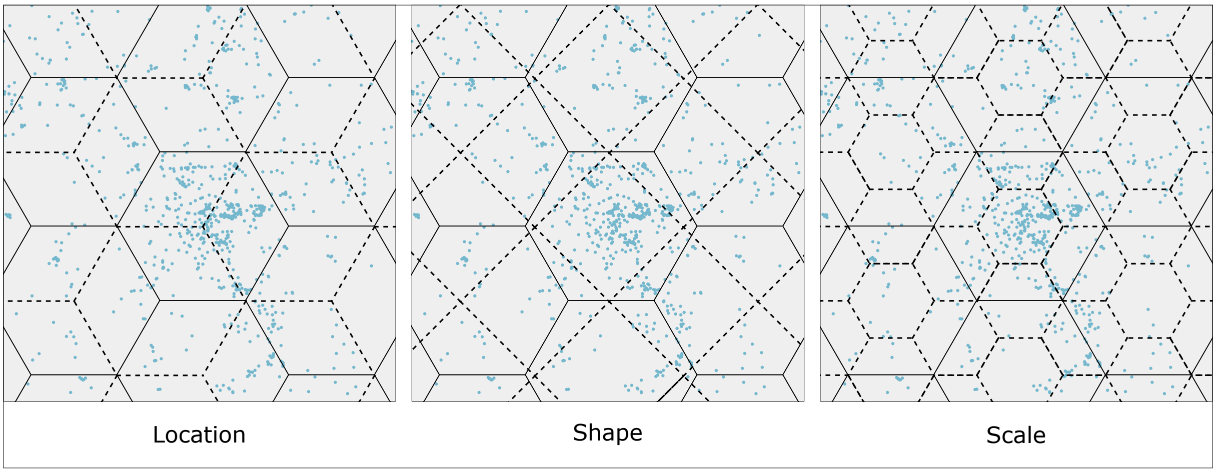


Figure II.4.1: Three researcher-dependent aspects illustrating the influence of MAUP on data analysis: different locations resulting from shifted polygon positions, different distributions as a function of distinct polygon forms, and different scale resolutions due to diverse polygon sizes (resulting in different aggregation effects).

II.4.3.2 The Modifiable Areal Unit Problem

The modifiable areal unit problem (MAUP) (Openshaw 1984) is a well-known issue that occurs when researchers aggregate data to reflect areal units. MAUP describes the fact that the choice of an—often arbitrary—spatial unit for an analysis can influence the outcomes of that analysis. Figure II.4.1 illustrates how the three key characteristics of spatial units—position, shape, and scale—affect the analysis of underlying data points (*e. g.*, from georeferenced surveys). For example, obesity rates can be meaningfully analysed at the country level or at the state level. Depending on the level, we might see a different statistical pattern, either A or B, and draw the respective conclusions. Yet, even though the shape of geographic units (*e. g.*, state borders) can certainly carry meaning for political and administrative bodies, it is still arbitrary and does not necessarily best reflect the aggregated data and the associated data-generating processes (from a causal or associative point of view). Given that the geographic units are arbitrary, so, too, is their position (encompassing specific locations) and the resulting distribution of data points to be aggregated. Even if the lattice of administrative units were transformed only *marginally*, the substantive conclusions that a researcher arrives at might change *drastically*.

MAUP is one of the long-standing and still unresolved issues in GIScience, and its ramifications are vividly discussed throughout different academic fields. Recent examples include investigations of human mobility (Mitra and Buliung 2012; Xu et al. 2014b), criminology (Vogel 2016; Gerell 2017), and forestry (Mas et al. 2015; Kozak and Szwagrzyk 2016). For instance, Mitra and Buliung (2012) related properties of the built environment to children's mode of active/passive school transportation (*i. e.*, whether they walk or cycle to school rather than taking the bus). Testing six different spatial configurations, they found that the sign as well as the size of the regression coefficients varied across scales and polygon forms used for defining the built-environment variables. Similarly, Vogel (2016) investigated the relationships between environmental factors and violence. Respondents were aggregated to reflect census tracts as well as units at the city block level. While the analyses revealed significant associations of the environmental factors, the effect of the geographic neighbourhood did not exist at the block level, but only at the level of census tracts. As these examples show, the effect of MAUP on survey outcomes can be severe. MAUP should be taken into account by testing the replicability across different spatial units.

For survey researchers, MAUP matters (a) for the answers given by respondents on the basis of subjective representations of geography (mental representations of spatial phenomena), and (b) for the objective scale of georeferenced external data. MAUP is thus an important issue when it comes to augmenting classic survey data with external data such as census variables. External data are often in aggregated form, not free from geometric arbitrariness. The choice of the geographic level of analysis in many studies to date appears to have been driven largely by data availability rather than by a priori theoretical considerations of what constitutes meaningful context information. For instance, when analysing the impact of covariates on respondents' answers, as in the previous example taken from Vogel (2016), it is clear that some information is available only through the census. In such cases, MAUP is essentially inevitable, but it should be kept in mind when drawing any inferences.

Social science studies often use countries, or somewhat more fine-grained administrative units available in the data distribution, as their level of analysis, without further detailing whether this choice is conceptually meaningful. This is, to some extent, understandable, given that contextual data provided by public institutions (*e. g.*, unemployment rates) are often limited to these levels. However, many of these studies still address fairly distal macro-contexts rather than respondents' more proximal ecosystems (*e. g.*, a city, or a district within a city) in which the (supposedly crucial) person–environment transactions that are constitutive of individual development take place (Bronfenbrenner 1979; Lerner 1991). While the national level may be appropriate for many research questions, it would be desirable to devote greater attention to the choice and justification of the geographic level of analysis. Ideally, researchers would consider using the geographic level that appears most appropriate from a theoretical point of view, rather than the level for which data happen to be available. Moreover, they should report whether their substantive findings are robust across different geographic levels of analysis (Saib et al. 2014).

II.4.3.3 Maps, Distortion, Meaning, and Visualisation

Another perennial issue in GIScience is the cartographic representation of the results generated through geospatial analysis. Unlike typical charts, maps can be used to bias communication in ways that survey researchers might be less familiar with. As (Monmonier 1996) states, cartography has the power to bias the presentation of spatial information by generating "selective truth". Cartographic styles may strongly influence which information is ultimately perceived by the respondent. Ways of biasing maps include, inter alia, the choice of spatial aggregation and scale levels (related to MAUP), but also the selection of suggestive colour ramps, the creation of categories and classes according to different criteria (natural statistical breaks, quantiles or units of standard deviations, etc.), the presentation of relative or absolute numbers, the influence of different coordinate reference systems (geographic vs. projected), or the choice of icons that represent geographic features. Figure II.4.2 illustrates these effects by showing the same piece of information (crude U.S. birth rate in 2000) in different ways as originally described by (Monmonier 2005). Thus, the information obtained from maps may be subject to arbitrariness, regardless of whether they are used for survey sampling, as visualisation aids in a survey, or to draw inferences from an analysis. The complex questions involved have given rise to the scientific endeavour known as "critical cartography" (Crampton 2010; Crampton and Krygier 2005).

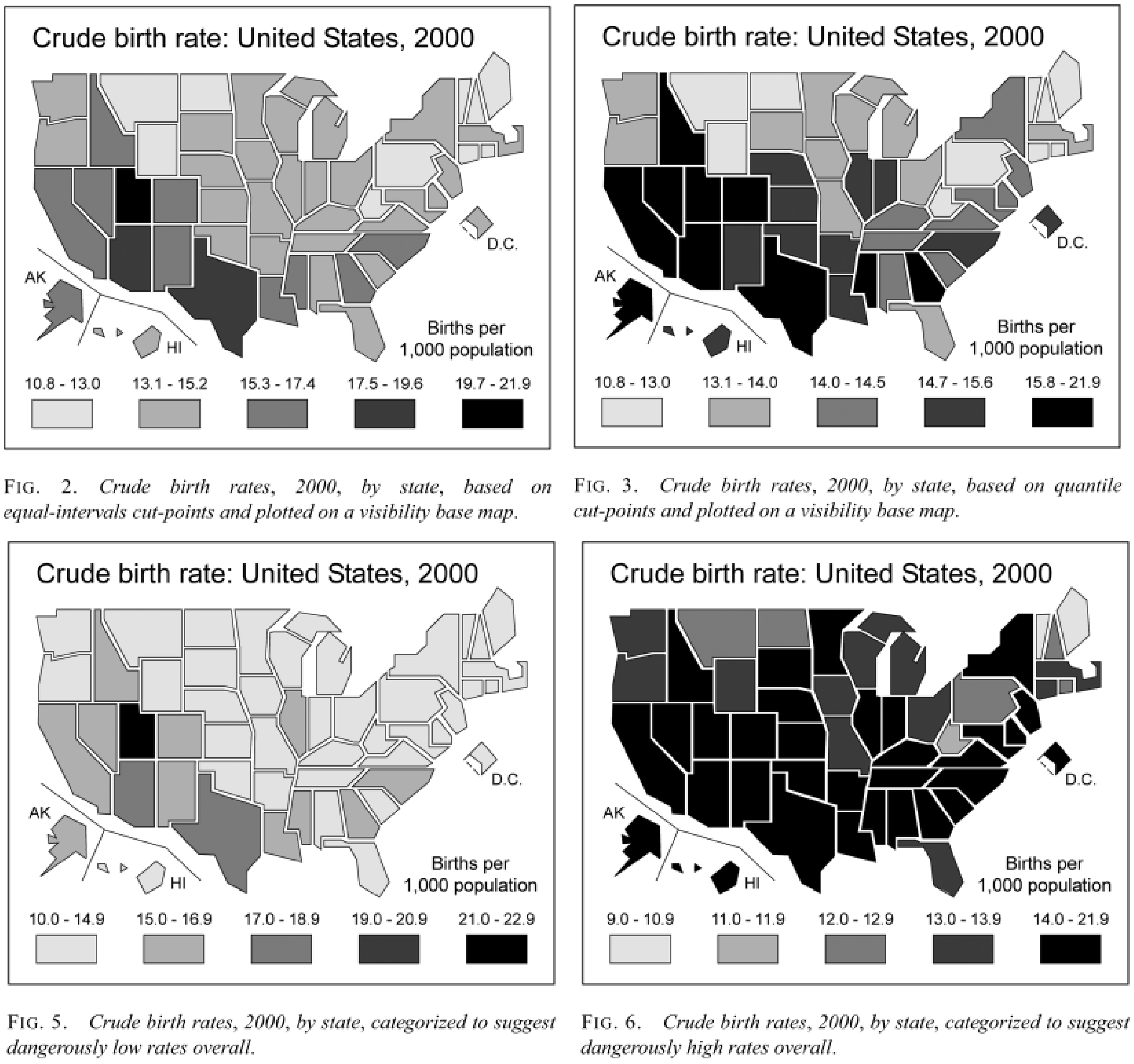

Crude birth rate: United States, 2000
AK
HI
D.C.
Births per 1,000 population
10.8 - 13.0
13.1 - 15.2
15.3 - 17.4
17.5 - 19.6
19.7 - 21.9
Crude birth rate: United States, 2000
AK
HI
D.C.
Births per 1,000 population
10.8 - 13.0
13.1 - 14.0
14.0 - 14.5
14.7 - 15.6
15.8 - 21.9
Crude birth rate: United States, 2000
AK
HI
D.C.
Births per 1,000 population
10.0 - 14.9
15.0 - 16.9
17.0 - 18.9
19.0 - 20.9
21.0 - 22.9
Crude birth rate: United States, 2000
AK
HI
D.C.
Births per 1,000 population
9.0 - 10.9
11.0 - 11.9
12.0 - 12.9
13.0 - 13.9
14.0 - 21.9

FIG. 2. *Crude birth rates, 2000, by state, based on equal-intervals cut-points and plotted on a visibility base map.*

FIG. 3. *Crude birth rates, 2000, by state, based on quantile cut-points and plotted on a visibility base map.*

FIG. 5. *Crude birth rates, 2000, by state, categorized to suggest dangerously low rates overall.*

FIG. 6. *Crude birth rates, 2000, by state, categorized to suggest dangerously high rates overall.*

Figure II.4.2: A classic example of “how to lie with maps”: Arbitrary choices influence which information is being communicated with, and obtained from, maps (Fig. 2, 3, 5, and 6 from Monmonier (1996) and Monmonier (2005)).

II.4.3.4 Analysing Context and Using Georeferenced Contextual Data

Several problems exist that relate to the concept of “context”. According to Dey (2001), context is defined as implicit or explicit information that is useful to characterise a situation. External, physical contexts are strongly associated with the objective physical environment, typically measured by physical sensors (*e. g.*, room temperature). However, as noted in sections II.4.2.4 and II.4.2.5 above, contexts can also be described through respondents’ subjective impressions at an individual level (Hong et al. 2009) or by aggregating respondent data from wearables and tracking devices (Bettini et al. 2010; Sagl et al. 2015). The spectrum of available technologies for capturing contextual information allows situational features to be quantified comprehensively and in unprecedented detail. These features include geographic aspects such as current environmental conditions (weather, air quality, etc.), the human perception of urban spaces, and the individual and collective behavioural responses to a range of functional settings including traffic infrastructures, open spaces, neighbourhoods, or residential areas. All these settings are of considerable importance for human-environment interactions and citizens’ quality of life, yet the number of characteristics with which to describe (and analyse) the impact of these contexts is manifold.

One limitation of most socioecological research to date is that—again due to data availability—it adopts a rather static view of contexts. The contextual information in these studies is often confined to cross-sectional snapshots, with the result that the dynamic nature of contexts goes unnoticed. We would like to challenge survey researchers to aim for a more dynamic conceptualisation of contexts. Environments change (as do people). Once chosen for analysis, geographic variables may not represent the *same* context a few days, months, or years later. For instance, contextual factors, such as weather conditions, traffic density, air pollution, vegetation, etc., are characterized by high spatial and temporal variability. Especially if longitudinal data on individual survey respondents are available, there may be ample opportunity to also treat contextual information as time-varying. Linking *changes* in ecological variables to variation in individual-level outcomes may stimulate new research questions and also aid in identifying the direction of causal influence.

Another challenge arises when individual survey data are aggregated to a geographic level in order to map them into a spatial context and infer something about the target population or the context. When participating in a survey, individuals may provide answers about their current environment as indicated by a GPS location, and they may appear to be knowledgeable about the reference object. However, their true degree of expertise may be concealed due to the complexities of the question-answering process. For instance, participants might be living in different environments during the week than at the weekend (*e. g.*, commuters). Simply assuming that data reflect information about some location just because a location happens to be available can introduce error of unknown magnitude, especially if respondents’ answers are mapped to geographic units to which their answers do not actually belong (*e. g.*, due to imprecise question wording, or participating in a survey on a mobile device at an unintended location that differs from what is reported as the place of residence, etc.).

Finally, any mapping of survey data to geographic units is done on the assumption that geographic units can be *validly* characterised by individual survey responses. Only then can statistical aggregates across respondents legitimately describe the specified geographic unit—for instance, by means, variances, and other indices of heterogeneity such as fractionalisation and polarisation (Chakravarty 2015). For this assumption to be valid, two minimum requirements must be met: (a) For reliable estimates, the number of data points from individual response units (cases) per geographic unit must be large enough in relation to the unit population that the aggregate measure intends to describe; if not, a higher aggregation level (and spatial scale on which conclusions can be drawn) may be required. (b) The survey sample must be sufficiently representative of the target population characteristics in each geographic unit, lest bias arise in characterising the unit by aggregate measures. Great care must be exerted in testing the extent to which the data can be considered representative of the target population and the contextual variables inferred from them (Doff 2010), especially if participant self-selection and participants' selective mobility are not controlled for (Rentfrow et al. 2015; Jokela et al. 2015).

II.4.3.5 Fallacies and Statistical Issues

With the increasing availability of big data and geocoded databases, there is also an increased risk of inferring ecological relationships (based on aggregate data) that may lead to misleading causal inferences if these are extended to the level of individual agents. For instance, some geographic areas may be inhabited by groups of different sizes (majority/minorities), and the same areas may have different likelihoods of showing other characteristics (*i. e.*, skewed base rates reflecting a different prevalence of specific attributes in these areas). But that does not mean that the statistical relationship between aggregate properties supplies the right clue to the underlying causal pathways. For example, it is possible to obtain area data on crime and correlate them with other area-level information (*e. g.*, ethnic composition). The temptation is to naïvely use the aligned skewed base rates of attributes to infer a relationship between them where none actually exists—a so-called pseudo-contingency (Kutzner and Fiedler 2017).

Higher crime rates may be observable in areas with a higher prevalence of a minority group. And yet the relationship at the aggregate level cannot hold the minority accountable for the crime rates observed in the areas in question. Based on the observed aggregate-level information, pseudo-contingencies provide a legitimate proxy for inferences at the ecological level. However, they are, at best, a heuristic for individual-level inferences. They may reflect genuine contingencies under various conditions, yet pseudo-contingencies are also at risk of inviting the wrong inference level. As the access to databases with high geographic resolution increases, survey researchers and GIScientists have a responsibility to ensure the correct interpretation of their data. We may face an increasing ethical obligation to correct blatant misuse of data (*e. g.*, for political or ideological purposes).

We caution readers that, irrespective of the specific aggregation level chosen, there is always the risk of an *aggregation bias*, which refers to the difference between results established at the level of the units of analysis (say, states or groups) and results established for lower levels of analysis (say, counties or individuals) when using aggregated data. Making inferences from higher to lower levels runs the risk of committing an ecological fallacy (Piantadosi et al. 1988), whereby an observed association between variables is erroneously taken to operate at a lower aggregation level than the one actually studied (Robinson 1950; Robinson 2009; Robinson 2011). Conversely, an individualistic fallacy may result when relationships observed at a micro-level are erroneously extrapolated to a macro-level (Clark and Avery 1976). For example, the outcome of encounters between social groups cannot be predicted on the basis of how individual group members from different groups interact with each other (Doerr et al. 2011; Lichter et al. 2012). Drawing conclusions about individuals (the survey units) through aggregated quantities (to match the geographic units)—and vice versa—requires drawing inferences carefully and properly (Grotenhuis et al. 2011), which usually necessitates multi-level data analysis (Nezlek 2008).

Furthermore, when using georeferenced data, many statistical methods are no longer suitable. *Spatial autocorrelation*—that is, the degree to which one object is similar to other spatially nearby objects (Goodchild 2009)—jeopardises the independence requirement of many statistical techniques. The phenomenon refers to the common finding that observations with a higher proximity in geographic space tend to be more similar to each other than those at a greater distance; this often results in patterns such as gradients or clusters. Such patterns may also be found among survey data. Using spatially distributed data that are either externally linked to, or gathered from, surveys requires methods of data analysis that detect, describe (*i. e.*, quantify), and, if necessary, adjust for the presence of spatial autocorrelation (Assuncao and Reis 1999; Banerjee et al. 2014; Getis 2010; Oden 1995; Waldhör 1996). We refer interested readers to an online introduction[1] and to recent accessible treatises of applied spatial analysis (Fischer and Getis 2010a; Ward and Gleditsch 2008).

II.4.3.6 Analysing User-Generated and Spatiotemporally Distributed Data

Another set of challenges arises when integrating new methods of data collection such as citizen sensing (*i. e.*, acquiring people's feedback through dedicated technologies such as smartphone apps) or linking collective data sources (*e. g.*, mobile phone or social networks) with traditional survey approaches. Traditional ways of analysing geospatial data mostly presume a well-defined data acquisition process and follow the first law of geography (Tobler 1970), according to which processes happening close to each other have a stronger influence than distant ones. However, Tobler's law may not hold for most user-generated data. For example, social media posts about a large-scale sports event or a national election may be related in time (when they are posted) and semantics (the content of the posts), but not in geographic space (as they may be sent from users in different places throughout the world). The reason is that the data-generating process for social media posts (as opposed to "traditional" spatial data like demographic data or transportation infrastructure data) is not standardised, nor is it under the control of a researcher. Instead, the mechanisms generating user-driven data are unpredictable and technically arbitrary; user motivations are often hidden, but they are likely to be context-bound. This non-standardised, uncontrolled data-generating process also implies that representativeness for the whole population may be impossible to achieve with data from wearables, citizen sensing, or social media (although targeting more specific populations may be realistic; Pötzschke and Braun (2016)). This issue has been largely neglected in previous research, and it constitutes a potentially high-impact research gap, even though first attempts at overcoming it within spatial analyses are being made (for the case of social media data, see Westerholt et al. (2015) and Westerholt et al. (2016)).

Another still largely unresolved question is how participants' responses are influenced by *repeatedly* interacting with technical devices (smartphones), especially if they *frequently* encounter dedicated survey questions. From a psychological viewpoint, besides typical memory errors, this may induce several kinds of biases. First, conditioning effects may occur such that people become conditioned to specific locations and provide pre-determined answers that they have learned to automatically associate with the location when prompted for responses. Given the frequency of recurring situations, they may not be motivated to engage with the question as expected. Due to the cooperative principle that governs effective communication (Grice 1975), some respondents may alter their statements when answering questions repeatedly—although their opinions have not really changed—because they think that new information must be provided; other respondents may stick with what they answered earlier in order to appear consistent and not contradict themselves. Second, it may not always be possible to gain reports immediately at the location of interest. However, delayed responding may introduce retrospective bias (*e. g.*, inaccurate recall, recency effects, false memories) into respondents' cognitive representations (see, *e. g.*, Steffens and Mecklenbräuker (2007)). Survey researchers can offer specific advice on how to minimise the impact of such biases on survey quality.

[1] https://docs.aurin.org.au/portal-help/analysing-your-data/spatial-statistics-tools/introduction-to-spatial-autocorrelation

Still, as the availability of data from wearables, social media, and other new data sources increases, greater research efforts will be necessary to resolve questions of survey designs, data quality, representativeness, and potential biases, and to link these new data to traditional surveys. Here, GIScientists can benefit from the expertise of psychologists and other social scientists with regard to traditional surveys, and we call on these disciplines to jointly tackle the aforementioned issues. Compared to traditional surveys, surveys in the domain of GIScience often encompass user-generated data, or they comprise a strong technological component (*e. g.*, GPS receivers, physical assessments, advanced spatial analysis techniques). A possibly hidden assumption among some researchers may be that surveys must invariably entail the gathering of subjective data. Yet, with GIS tools, surveys may be based not on a single questionnaire at all, but rather only on objective data capture. Hence, the boundaries of surveys become fuzzy.

Whatever survey concept applies, full documentation of the survey design and its quality is required because only this permits estimating, and potentially correcting for sources of, sampling related error (Dever et al. 2008; Gabler and Quatember 2013). This includes intended and actual populations under study (to determine over- and under-coverage) as well as the sampling design, the obtained sample size, and any missing data (*e. g.*, non-response, drop-out; Little and Rubin (2002)). With new forms of data, such as data from social media, this information may be unavailable, so that the quality of the data collection cannot be assessed (Brickman Bhutta 2012). However—depending on the study goal—this information may be an indispensable requirement (Rothman et al. 2013; Pötzschke and Braun 2016). Moreover, most current approaches in geospatial analysis rely on well-defined data structures with known degrees of uncertainty and small error margins, although these requirements are not met by vast portions of user-generated data (Steiger et al. 2015b). Guidelines may help researchers to minimise total survey error (Groves and Lyberg 2010) and improve total survey quality under budgetary constraints (Biemer 2010), for instance, those published by the German Data Forum (RatSWD 2015), AAPOR[2], and ESOMAR[3].

II.4.3.7 Privacy Concerns and Data Protection Issues

The last challenge we highlight is the use of any personal—including geocoded—data and researchers' ethical obligation to protect users' privacy (Goebel et al. 2010a; Goebel et al. 2010b). Typical privacy risks are *presence leakage* (an attacker might identify individuals present in, or absent from, the database) and *association leakage* (an attacker might unambiguously associate individuals with sensitive information). The risk of *deductive disclosure*—identifying a person by a combination of personal characteristics—is a challenge for GIS research (due to the inclusion of geocodes, tracking of individuals, and data linkage). This issue calls for technologies and legal frameworks to protect data against deductive disclosure of participants' identities, unintended transfer, or other misuse by third parties (Barcena et al. 2014).

Legislation that ensures a degree of data safety (keeping data available in the future) and data security (limiting access to data) varies from country to country. Consequently, researchers sharing sensitive data in international collaborations may have to deal with diverging legal requirements and policies for raw and derived data across various countries. Moreover, respondents' willingness to voluntarily share highly personal data with scientists differs across individuals and settings. However, support for research is usually closely linked to trust in the security of data and their protection against misuse. Ironically, many users willingly share private information in other places and do not actively try to conceal or protect it,

[2] http://www.aapor.org/Standards-Ethics/Best-Practices.aspx

[3] http://www.esomar.org/knowledge-and-standards/codes-and-guidelines/guideline-on-opinion-polls-and-published-surveys.php

even if they claim that they are concerned about their privacy (Acquisti et al. 2013). A striking example is the vast amount of sensitive data (including rich location data) that people share on social media platforms such as Facebook, where they typically have little influence on data collection and processing policies. Likewise, estimates show that one-third of the free smartphone apps collect location information, yielding numerous possibilities of analysing geographic data and extracting information from them (Kersten and Klett 2012). Apparently, operators and service providers—whose business models often rest on collecting and selling customer/user data (*e. g.*, Google)—effectively insinuate that less privacy is the new social norm, and that it means better services for the user (Johnson 2010).

On the researcher's side, several means exist to protect participants and their rights, including privacy (Resch et al. 2015a). First, all participants should participate voluntarily in data-rich scientific studies through an opt-in agreement, after a thorough briefing (informed consent)—something that is rather self-evident from the perspective of survey research. Principal investigators and researchers must enter into a data-sharing agreement about which data will be collected, analysed and stored, where and for how long, and who will have access to them.

When data are to be stored and made available to other researchers, a possible way of allaying concerns about privacy is to restrict access to sensitive data. For instance, it might be feasible to use different levels of access privileges to sensitive datasets in a data archive ("data enclave"; Lane (2014)). However, such archives usually involve increased levels of burden. Sometimes, only aggregate query results can be obtained, or access might be limited to eligible researchers in a controlled, secure environment with high-security data storage facilities (*e. g.*, GESIS' Secure Data Center[4] with an on-site safe room). Sharing multi-site research data safely (via the cloud) during collection requires technical solutions that are still in their infancy, and new standards have be developed and enforced (Palanisamy and Liu 2015; Veena and Devidas 2014).

For applications that require the collection and storage of personal information, the Electronic Frontier Foundation and others recommend anonymising data and using strong cryptographic protocols at various stages of data transmission and handling. However, trajectories of people moving through space (at specified times) can still undermine anonymity. In the case of spatial information, more specifically, previously anonymous users can be re-identified relatively easily by their spatial profiles because personal geodata are highly unique to an individual. Indeed, Montjoye et al. (2013) showed that only four random positions of a track may be needed to identify individuals. In this context, the concept of *location privacy* describes the ability of an individual to move in public space without their geographic location being collected or stored. The most restrictive way to achieve *location privacy* and to prevent misuse of personal data is to opt out of research altogether and to prohibit the collection of any data (Blumberg and Eckersley 2009)—which is not usually a scientifically viable option. If possible, trajectory data should be analysed and shared at an aggregated rather than an individual level. Furthermore, privacy should also be protected by splitting trajectories into sub-paths so that they cannot be reconstructed. Although this involves a certain amount of information loss, the restoration of identities is prevented (Sarowar Sattar et al. 2013; Wang 2010).

It is often necessary to georeference the survey data so that they can be mapped to relevant spatial units in order, for example, to link individual respondents' data to contextual data. Several methods of georeferencing exist. Direct georeferencing requires that exact locations be collected via spatial coordinates (*e. g.*, 2D or 3D, GPS). Indirect georeferencing assumes that relevant spatial units are inferred from postal codes, administrative units, etc. However, the use of online geocoding services for converting

[4] http://www.gesis.org/en/services/data-analysis/data-archive-service/secure-data-center-sdc

terms such as ZIP codes to locations such as latitude, longitude, and elevation by means of direct georeferencing (*e. g.*, via Google services) should not, in our view, be the first choice. Geocoding involves the risk that non-aggregated scientific use files might become de-anonymised, as geocodes potentially undermine a user's location privacy. Even though single locations are not revealing in themselves, complete time-stamped location patterns may be used to identify an individual (especially if a company knows more about that individual than the information contained in the scientific data). Moreover, reverse geocoding—back coding of latitude and longitude to a comprehensible address—involves the risk that an individual's identity will be leaked—even from mere dots representing individuals on a published map. Identity leakage from maps can be prevented by aggregating data points prior to drawing the map or by skewing the presentation of individual data points (Brownstein et al. 2006).

Instead of using geocoding services from companies with commercial interests, we suggest using public geocoding services such as that provided by the German Federal Agency of Cartography and Geodesy (BKG). This service allows users to tag any geographically identifiable object (*e. g.*, on the basis of available address information) with precise geographic coordinates (reverse geocoding is possible, too). It is usually available only to federal authorities. However, under an agreement between GESIS and the BKG, it may also be used by other institutions (Schweers et al. 2016) (see GESIS Georefum[5]). Some specialized centres provide software, services, and support for linking databases while observing privacy-preserving record linkage (*e. g.*, German Data Linkage Center[6]; Schnell (2013a)).

Note that bias can be introduced later on when linking the datasets (Sakshaug et al. 2012). Respondents who are used to being interviewed about sensitive issues (*e. g.*, political attitudes) may not be willing to consent to their data being linked to additional databases, and those who are willing may not be representative of the population. GIScience participants may agree to be tracked (*e. g.*, resulting in trajectories), yet this may be due to previous self-selection (which would be accompanied by overall lower response rates). Initial self-selection and subsequent selective dropout may introduce bias into a combined dataset.

We conclude this discussion by encouraging researchers to reflect on the ethical implications and the long-term societal impact of fine-grained spatial analyses. For example, terms such as "air quality" or "pollutant dispersion" are only surrogates for more direct and far-reaching influences on individuals, such as life expectancy, respiratory diseases, or quality of life (Resch et al. 2012). Knowledge about these phenomena at high geospatial resolutions may affect relevant aspects in people's lives, such as health insurance rates or real estate prices. Researchers' ethical responsibility to find the appropriate spatial granularity level when providing information and communicating research findings has never been more acute. The scientific drive to provide ever more accurate, possibly finer-grained, and complete information competes with other ethical principles surrounding privacy concerns and prevention of misleading conclusions.

II.4.4 Conclusion

In this article, we have discussed the promising new opportunities of integrating GIScience tools into survey research in general, and psychological survey research in particular, and the challenges associated with these opportunities. In so doing, we have focussed mainly on how survey research can profit from incorporating recent advances in GIScience. We highlight, however, that GIScientists can also profit

[5] http://www.gesis.org/forschung/drittmittelprojekte/projektuebersicht-drittmittel/georefum

[6] http://www.record-linkage.de

greatly from the accumulated wisdom in survey research methodology, for example, when thinking about measurement and assessment, data quality, or representativeness. We are certain that intensified interdisciplinary dialogue holds great potential for future research. In our view, both survey researchers and GIScientists would benefit from incorporating each others' traditions into their own theorising and methodologies. In this process, survey researchers can act as consultants to GIScientists just as much as GIScientists can inspire survey researchers with new advancements.

A stronger integration of the research traditions will also enable highly inspirational interdisciplinary research. Future research at the intersection between survey research methodology and GIScience may even blur the very boundaries of the survey concept and bring us closer to studying the person–context transactions that are deemed crucial in shaping individual behaviour and development (Bronfenbrenner 1979; Lerner 1991). Thanks to the progress that has been made in GIScience, the study of the current environment that Lewin (1936) once envisioned *can* now include precise temporal and spatial aspects. Context information can increasingly be incorporated in real time, and it may be based on subjective as well as objective contextual characteristics of individual situations.

Obviously, the fruitfulness of future research enterprises depends on the engagement of researchers from both sides, their growing awareness of the tools, methods, and concepts they offer each other, and of the goals and challenges associated with each of them. We hope that this overview will be instrumental in fostering dialogue between survey research and GIScience.

References (Chapter II.4)

Acquisti, A, L John and G Loewenstein (2013). 'What Is Privacy Worth?' *The Journal of Legal Studies* 42 (2), pp. 249–274. DOI: 10.1086/671754.

Agnew, J and D Livingstone (2011). 'Space and Place'. In: *Handbook of Geographical Knowledge*. Ed. by J Agnew and D Livingstone. London, UK: SAGE, pp. 316–330. DOI: 10.4135/9781446201091.n24.

Assuncao, R and E Reis (1999). 'A New Proposal to Adjust Moran's I for Population Density'. *Statistics in Medicine* 18 (16), pp. 2147–2162. DOI: 10.1002/(SICI)1097-0258(19990830)18:16<2147::AID-SIM179>3.0.CO;2-I.

Ayobi, A, P Marshall and A Cox (2016). 'Reflections on 5 Years of Personal Informatics: Rising Concerns and Emerging Directions'. In: *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. Ed. by C Lampe, D Morris and J Hourcade. Santa Clara, CA: ACM Press, pp. 2774–2781. DOI: 10.1145/2851581.2892406.

Banerjee, S, A Gelfand and B Carlin (2014). *Hierarchical Modeling and Analysis for Spatial Data*. 2nd ed. Boca Raton, FL: CRC Press.

Barcena, M, C Wueest and H Lau (2014). *How Safe is your Quantified Self?* Tech. rep. Mountain View, CA: Symantec Corporation.

Bettini, C, O Brdiczka, K Henricksen, J Indulska, D Nicklas, A Ranganathan and D Riboni (2010). 'A Survey of Context Modelling and Reasoning Techniques'. *Pervasive and Mobile Computing* 6 (2), pp. 161–180. DOI: 10.1016/j.pmcj.2009.06.002.

Biemer, P (2010). 'Total Survey Error: Design, Implementation, and Evaluation'. *Public Opinion Quarterly* 74 (5), pp. 817–848. DOI: 10.1093/poq/nfq058.

Blumberg, A and P Eckersley (2009). *On Locational Privacy, and How to Avoid Losing it Forever*. Tech. rep. San Francisco, CA: Electronic Frontier Foundation.

Boschmann, E and E Cubbon (2014). ‘Sketch Maps and Qualitative GIS: Using Cartographies of Individual Spatial Narratives in Geographic Research’. *The Professional Geographer* 66 (2), pp. 236–248. DOI: 10.1080/00330124.2013.781490.

Brickman Bhutta, C (2012). ‘Not by the Book: Facebook as a Sampling Frame’. *Sociological Methods & Research* 41 (1), pp. 57–88. DOI: 10.1177/0049124112440795.

Bronfenbrenner, U (1979). *The Ecology of Human Development: Experiments by Nature and Design*. Cambridge, MA: Harvard University Press.

Brownstein, J, C Cassa and K Mandl (2006). ‘No Place to Hide - Reverse Identification of Patients from Published Maps’. *New England Journal of Medicine* 355 (16), pp. 1741–1742. DOI: 10.1056/NEJMc061891.

Brulé, G and R Veenhoven (2014). ‘Average Happiness and Dominant Family Type in Regions in Western Europe around 2000’. *Advances in Applied Sociology* 04 (12), pp. 271–288. DOI: 10.4236/aasoci.2014.412031.

— (2015). ‘Geography of Happiness: Configurations of Affective and Cognitive Appraisal of Life Across Nations’. *International Journal of Happiness and Development* 2 (2), pp. 101–117.

Caldarelli, G, A Chessa, F Pammolli, G Pompa, M Puliga, M Riccaboni and G Riotta (2014). ‘A Multi-Level Geographical Study of Italian Political Elections from Twitter Data’. *PLOS ONE* 9 (5). Ed. by Matjaz Perc, e95809. DOI: 10.1371/journal.pone.0095809.

Chakravarty, S (2015). *Inequality, Polarization and Conflict*. Economic Studies in Inequality, Social Exclusion and Well-Being. New Delhi: Springer India. DOI: 10.1007/978-81-322-2166-1.

Clark, W and K Avery (1976). ‘The Effects of Data Aggregation in Statistical Analysis’. *Geographical Analysis* 8 (4), pp. 428–438. DOI: 10.1111/j.1538-4632.1976.tb00549.x.

Clifford, N, S Holloway, S Rice and G Valentine (2009). *Key Concepts in Geography*. 2nd ed. Thousand Oaks, CA: SAGE.

Conner, T and L Barrett (2012). ‘Trends in Ambulatory Self-Report: The Role of Momentary Experience in Psychosomatic Medicine’. *Psychosomatic Medicine* 74 (4), pp. 327–337. DOI: 10.1097/PSY.0b013e3182546f18.

Coulton, C, T Chan and K Mikelbank (2010). *Finding Place in Making Connections Communities Applying GIS to Residents’ Perceptions of Their Neighborhoods*. Tech. rep. Washington, DC: The Urban Institute.

Crampton, J (2010). *Mapping: A Critical Introduction to Cartography and GIS*. New York, NY: Wiley-Blackwell.

Crampton, J and J Krygier (2005). *An Introduction to Critical Cartography*. Vol. 4. 1. Okanagan University College, Dept. of Geography, pp. 11–33.

Crooks, A, A Croitoru, A Stefanidis and J Radzikowski (2013). ‘#Earthquake: Twitter as a Distributed Sensor System’. *Transactions in GIS* 17 (1), pp. 124–147. DOI: 10.1111/j.1467-9671.2012.01359.x.

Curini, L, St Iacus and L Canova (2015). ‘Measuring Idiosyncratic Happiness Through the Analysis of Twitter: An Application to the Italian Case’. *Social Indicators Research* 121 (2), pp. 525–542. DOI: 10.1007/s11205-014-0646-2.

Dangschat, J (2007). ‘Raumkonzept zwischen struktureller Produktion und individueller Konstruktion’. *Ethnologie und Raum* 9 (1), pp. 24–44.

Dever, J, A Rafferty and R Valliant (2008). 'Internet Surveys: Can Statistical Adjustments Eliminate Coverage Bias?' *Survey Research Methods* 2 (2), pp. 47–62. DOI: 10.18148/srm/2008.v2i2.128.

Dey, A (2001). 'Understanding and Using Context'. *Personal and Ubiquitous Computing* 5 (1), pp. 4–7. DOI: 10.1007/s007790170019.

Diekmann, A and R Meyer (2010). 'Demokratischer Smog? Eine Empirische Untersuchung zum Zusammenhang Zwischen Sozialschicht und Umweltbelastungen'. *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 62 (3), pp. 437–457. DOI: 10.1007/s11577-010-0108-z.

Doerr, C, Ashby P, J Kunstman and D Buck (2011). 'Interactions in Black and White: Racial Differences and Similarities in Response to Interracial Interactions'. *Group Processes & Intergroup Relations* 14 (1), pp. 31–43. DOI: 10.1177/1368430210375250.

Doff, W (2010). *Puzzling Neighbourhood Effects*. Amsterdam: IOS Press.

Dungan, J, J Perry, M Dale, P Legendre, J Fortin, A Jakomulska, M Miriti and M Rosenberg (2002). 'A Balanced View of Scale in Spatial Statistical Analysis'. *Ecography* 25 (2), pp. 626–640. DOI: 10.1034/j.1600-0587.2002.250510.x.

Ekman, P and W Friesen (1971). 'Constants Across Cultures in the Face and Emotion'. *Journal of Personality and Social Psychology* 17 (2), pp. 124–129. DOI: 10.1037/h0030377.

Feddersen, J, R Metcalfe and M Wooden (2016). 'Subjective Wellbeing: Why Weather Matters'. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 179 (1), pp. 203–228. DOI: 10.1111/rssa.12118.

Fieldhouse, E, J Green, H Schmitt, G Evans and C Eijk (2014). *The 2015 British Election Study: Voters in Context*. Berlin.

Fischer, M and A Getis (2010a). *Handbook of Applied Spatial Analysis*. Heidelberg: Springer.

Gabler, S and P Lahiri (2009). 'On the Definition and Interpretation of Interviewer Variability for a Complex Sampling Design'. *Survey Methodology* 35, pp. 85–99.

Gabler, S and A Quatember (2013). 'Repräsentativität von Subgruppen bei Geschichteten Zufallsstichproben'. *AStA Wirtschafts- und Sozialstatistisches Archiv* 7 (3-4), pp. 105–119. DOI: 10.1007/s11943-013-0132-3.

Gerell, M (2017). 'Smallest is Better? The Spatial Distribution of Arson and the Modifiable Areal Unit Problem'. *Journal of Quantitative Criminology* 33 (2), pp. 293–318. DOI: 10.1007/s10940-016-9297-6.

Getis, A (2010). 'Spatial Autocorrelation'. In: *Handbook of Applied Spatial Analysis*. Ed. by M Fischer and A Getis. Heidelberg: Springer, pp. 255–278.

Goebel, J, G Wagner and M Wurm (2010a). 'Exemplarische Integration Raumrelevanter Indikatoren auf Basis von "Fernerkundungsdaten" in das Sozio-Ökonomische Panel (SOEP)'. Berlin.

Goebel, J, M Wurm and G Wagner (2010b). 'Exploring the Linkage of Spatial Indicators from Remote Sensing Data with Survey Data - the Case of the Socio-Economic Panel (SOEP) and 3D City Models'. Berlin.

Goodchild, M (1992). 'Geographical Information Science'. *International Journal of Geographical Information Systems* 6 (1), pp. 31–45. DOI: 10.1080/02693799208901893.

— (2009). 'What Problem? Spatial Autocorrelation and Geographic Information Science'. *Geographical Analysis* 41 (4), pp. 411–417. DOI: 10.1111/j.1538-4632.2009.00769.x.

— (2010). 'Twenty Years of Progress: GIScience in 2010'. *Journal of Spatial Information Science* 6 (1), pp. 31–45. DOI: 10.5311/JOSIS.2010.1.2.

Goodchild, M and K Kemp (1992). 'NCGIA Education Activities: The Core Curriculum and Beyond'. *International Journal of Geographical Information Systems* 6 (4), pp. 309–320. DOI: 10.1080/02693799208901915.

Grice, H (1975). 'Logic and Conversation'. In: *Syntax and Semantics*. Ed. by P Cole and J Morgan. New York, NY: Academic Press, pp. 41–58.

Grotenhuis, M te, R Eisinga and S Subramanian (2011). 'Robinson's Ecological Correlations and the Behavior of Individuals: Methodological Corrections'. *International Journal of Epidemiology* 40 (4), pp. 1123–1125. DOI: 10.1093/ije/dyr081.

Groves, R and L Lyberg (2010). 'Total Survey Error: Past, Present, and Future'. *Public Opinion Quarterly* 74 (5), pp. 849–879. DOI: 10.1093/poq/nfq065.

Hegarty, M, D Montello, A Richardson, T Ishikawa and K Lovelace (2006). 'Spatial Abilities at Different Scales: Individual Differences in Aptitude-Test Performance and Spatial-Layout Learning'. *Intelligence* 34 (2), pp. 151–176. DOI: 10.1016/j.intell.2005.09.005.

Hektner, J, J Schmidt and M Csikszentmihalyi (2007). *Experience Sampling Method: Measuring the Quality of Everyday Life*. Thousand Oakes, CA: SAGE.

Hill, C, E Dean and J Murphy (2013). *Social Media, Sociality, and Survey Research*. Hoboken, NJ: Wiley.

Hoffmeyer-Zlotnik, J (2013). *Regionalisierung Sozialwissenschaftlicher Umfragedaten*. Wiesbaden: VS Verlag. DOI: 10.1007/978-3-322-90525-3.

Hong, J, E Suh, J Kim and S Kim (2009). 'Context-Aware System for Proactive Personalized Service Based on Context History'. *Expert Systems with Applications* 36 (4), pp. 7448–7457. DOI: 10.1016/j.eswa.2008.09.002.

Hüttenrauch, B (2016). *Targeting Using Augmented Data in Database Marketing: Decision Factors for Evaluating External Sources*. Wiesbaden: Springer Fachmedien. DOI: 10.1007/978-3-658-14577-4.

Iosa, M, A Fusco, G Morone and S Paolucci (2012). 'Walking There: Environmental Influence on Walking-Distance Estimation'. *Behavioural Brain Research* 226 (1), pp. 124–132. DOI: 10.1016/j.bbr.2011.09.007.

Jack, R, O Garrod and P Schyns (2014). 'Dynamic Facial Expressions of Emotion Transmit an Evolving Hierarchy of Signals over Time'. *Current Biology* 24 (2), pp. 187–192. DOI: 10.1016/j.cub.2013.11.064.

Johnson, B (2010). *Privacy no longer a social norm, says Facebook founder*.

Jokela, M, W Bleidorn, M Lamb, S Gosling and P Rentfrow (2015). 'Geographically Varying Associations Between Personality and Life Satisfaction in the London Metropolitan Area'. *Proceedings of the National Academy of Sciences* 112 (3), pp. 725–730. DOI: 10.1073/pnas.1415800112.

Kersten, H and G Klett (2012). *Mobile Device Management*. Heidelberg: MITP.

Kitchen, P, A Williams and J Chowhan (2012). 'Sense of Community Belonging and Health in Canada: A Regional Analysis'. *Social Indicators Research* 107 (1), pp. 103–126. DOI: 10.1007/s11205-011-9830-9.

Klonner, C, S Marx, T Usón and B Höfle (2016). 'Risk Awareness Maps of Urban Flooding via OSM Field Papers-Case Study Santiago de Chile'. In: *Proceedings of the ISCRAM 2016 Conference*. Ed. by A Tapia, P Antunes, V Banuls, K Moore and J de Albuquerque. Rio de Janeiro.

Kozak, J and M Szwagrzyk (2016). 'Have There Been Forest Transitions? Forest Transition Theory Revisited in the Context of the Modifiable Areal Unit Problem'. *Area* 48 (4), pp. 504–512. DOI: 10.1111/area.12267.

Kutzner, F and K Fiedler (2017). 'Stereotypes as Pseudocontingencies'. *European Review of Social Psychology* 28 (1), pp. 1–49. DOI: 10.1080/10463283.2016.1260238.

Lane, J (2014). *Privacy, Big Data, and the Public Good: Frameworks for Engagement*. New York, NY: Cambridge University Press.

Lange, N de (2013). *Geoinformatik in Theorie und Praxis*. 3rd ed. Heidelberg: Springer Spektrum.

Larson, R and M Csikszentmihalyi (1983). 'The Experience Sampling Method'. *New Directions for Methodology of Social and Behavioral Science* 15, pp. 41–56.

Lechner, C, M Obschonka and R Silbereisen (2017). 'Who Reaps the Benefits of Social Change? Exploration and Its Socioecological Boundaries'. *Journal of Personality* 85 (2), pp. 257–269. DOI: 10.1111/jopy.12238.

Leeuw, E de, J Hox, D Dillman and European Association of Methodology. (2008). *International Handbook of Survey Methodology*. New York, NY: Lawrence Erlbaum Associates, p. 549.

Lerner, R (1991). 'Changing Organism-Context Relations as the Basic process of Development: A Developmental Contextual Perspective'. *Developmental Psychology* 27 (1), pp. 27–32. DOI: 10.1037/0012-1649.27.1.27.

Leslie, E, T Sugiyama, D Ierodiaconou and P Kremer (2010). 'Perceived and Objectively Measured Greenness of Neighbourhoods: Are They Measuring the Same Thing?' *Landscape and Urban Planning* 95 (1-2), pp. 28–33. DOI: 10.1016/j.landurbplan.2009.11.002.

Lewin, K (1936). *Principles of Topological Psychology*. New York, NY: McGraw-Hill, p. 260.

Lichter, D, D Parisi and M Taquino (2012). 'The Geography of Exclusion: Race, Segregation, and Concentrated Poverty'. *Social Problems* 59 (3), pp. 364–388. DOI: 10.1525/sp.2012.59.3.364.

Little, Roderick J and D Rubin (2002). *Statistical Analysis With Missing Data*. 2nd ed. New York, NY: Wiley.

Liu, B and L Zhang (2012). 'A Survey of Opinion Mining and Sentiment Analysis'. In: *Mining Text Data*. Ed. by C Aggarwal and C Zhai. New York, NY: Springer US, pp. 415–463. DOI: 10.1007/978-1-4614-3223-4_13.

Lucas, R and N Lawless (2013). 'Does Life Seem Better on a Sunny Day? Examining the Association Between Daily Weather Conditions and Life Satisfaction Judgments'. *Journal of Personality and Social Psychology* 104 (5), pp. 872–884. DOI: 10.1037/a0032124.

Mas, J, V Pérez, A Andablo Reyes, M Castillo Santiago and A Flamenco Sandoval (2015). 'Assessing Modifiable Areal Unit Problem in the Analysis of Deforestation Drivers Using Remote Sensing and Census Data'. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XL-3 (W3), pp. 77–80. DOI: 10.5194/isprsarchives-XL-3-W3-77-2015.

McGinn, A, K Evenson, A Herring, S Huston and D Rodriguez (2007). 'Exploring Associations Between Physical Activity and Perceived and Objective Measures of the Built Environment'. *Journal of Urban Health : Bulletin of the New York Academy of Medicine* 84 (2), pp. 162–184. DOI: 10.1007/s11524-006-9136-4.

McMaster, R and E Sheppard (2004). 'Introduction: Scale and Geographic Inquiry'. In: *Scale and Geographic Inquiry: Nature, Society, and Method*. Ed. by E Sheppard and R McMaster. Oxford, UK: Blackwell, pp. 1–22.

Messner, C and M Wänke (2011). 'Good Weather for Schwarz and Clore'. *Emotion* 11 (2), pp. 436–437. DOI: 10.1037/a0022821.

Meyer, R and H Enzler (2013). 'Geographische Informationssysteme (GIS) und ihre Anwendung in den Sozialwissenschaften am Beispiel des Schweizer Umweltsurveys'. *Methoden, Daten, Analysen* 7 (3), pp. 317–346. DOI: 10.12758/mda.2013.016.

Mitra, R and R Buliung (2012). 'Built Environment Correlates of Active School Transportation: Neighborhood and the Modifiable Areal Unit Problem'. *Journal of Transport Geography* 20 (1), pp. 51–61. DOI: 10.1016/j.jtrangeo.2011.07.009.

Monmonier, M (1996). *How to Lie with Maps*. 2nd ed. Chicago, IL: University of Chicago Press.

— (2005). 'Lying with Maps'. *Statistical Science* 20 (3), pp. 215–222. DOI: 10.1214/088342305000000241.

Montello, D and R Golledge (1998). *Scale and Detail in the Cognition of Geographic Information*. Tech. rep. Santa Barbara, CA: Varenius Specialist Meeting, University of California.

Montjoye, Y de, C Hidalgo, M Verleysen and V Blondel (2013). 'Unique in the Crowd: The Privacy Bounds of Human Mobility'. *Nature Scientific Reports* 3, p. 1376. DOI: 10.1038/srep01376.

Murphy, J, M Link, J Childs, C Langer, E Dean, M Stern, J Pasek, J Cohen, M Callegaro, P Harwood, Trent D Buskirk and Michael F Schober (2014). *Social Media in Public Opinion Research: Report of the AAPOR Task Force on Emerging Technologies in Public Opinion Research*. Tech. rep. American Association for Public Opinion Research.

Nezlek, J (2008). 'An Introduction to Multilevel Modeling for Social and Personality Psychology'. *Social and Personality Psychology Compass* 2 (2), pp. 842–860. DOI: 10.1111/j.1751-9004.2007.00059.x.

Obschonka, M, E Schmitt-Rodermund, R Silbereisen, S Gosling and J Potter (2013). 'The Regional Distribution and Correlates of an Entrepreneurship-Prone Personality Profile in the United States, Germany, and the United Kingdom: A Socioecological Perspective'. *Journal of Personality and Social Psychology* 105 (1), pp. 104–122. DOI: 10.1037/a0032275.

Oden, N (1995). 'Adjusting Moran's I for Population Density'. *Statistics in Medicine* 14 (1), pp. 17–26. DOI: 10.1002/sim.4780140104.

Oishi, S (2014). 'Socioecological Psychology'. *Annual Review of Psychology* 65 (1), pp. 581–609. DOI: 10.1146/annurev-psych-030413-152156.

Okner, B (1972). 'Constructing A New Data Base From Existing Microdata Sets: The 1966 Merge File'. In: *Annals of Economic and Social Measurement*. Ed. by S Berg. Cambridge, MA: NBER, pp. 325–362.

Openshaw, S (1984). *The Modifiable Areal Unit Problem*. Norwich, UK: Geobooks.

Ord, J and A Getis (2012). 'Local Spatial Heteroscedasticity (LOSH)'. *The Annals of Regional Science* 48 (2), pp. 529–539. DOI: 10.1007/s00168-011-0492-y.

Palanisamy, B and L Liu (2015). 'Privacy-Preserving Data Publishing in the Cloud: A Multi-level Utility Controlled Approach'. In: *2015 IEEE 8th International Conference on Cloud Computing*. New York, NY: IEEE, pp. 130–137. DOI: 10.1109/CLOUD.2015.27.

Piantadosi, S, D Byar and S Green (1988). 'The Ecological Fallacy'. *American Journal of Epidemiology* 127 (5), pp. 893–904.

Pötzschke, S and M Braun (2016). 'Migrant Sampling Using Facebook Advertisements: A Case Study of Polish Migrants in Four European Countries'. *Social Science Computer Review* pending (pending). DOI: 10.1177/0894439316666262.

Rammstedt, B, M Mutz and R Farmer (2015). 'The Answer Is Blowing in the Wind'. *European Journal of Psychological Assessment* 31 (4), pp. 287–293. DOI: 10.1027/1015-5759/a000236.

RatSWD (2012). *Georeferenzierung von Daten: Situation und Zukunft der Geodatenlandschaft in Deutschland*. Tech. rep. Berlin: Rat für Sozial- und Wirtschaftsdaten.

— (2015). 'Quality Standards for the Development, Application, and Evaluation of Measurement Instruments in Social Science Survey Research'. Berlin.

Rehdanz, K and D Maddison (2005). 'Climate and Happiness'. *Ecological Economics* 52 (1), pp. 111–125. DOI: 10.1016/j.ecolecon.2004.06.015.

Reichert, M, T Törnros, A Hoell, H Dorn, H Tost, H Salize, A Meyer-Lindenberg, A Zipf and U Ebner-Priemer (2016). 'Using Ambulatory Assessment for Experience Sampling and the Mapping of Environmental Risk Factors in Everyday Life'. *Die Psychiatrie* 13 (2), pp. 94–102.

Reis, H and S Gable (2000). 'Event-Sampling and other Methods for Studying Everyday Experience'. In: *Handbook of Research Methods in Social and Personality Psychology*. Ed. by H Reis and C Judd. Cambridge, UK: Cambridge University Press, pp. 190–222.

Rentfrow, P (2013). *Geographical Psychology: Exploring the Interaction of Environment and Behavior*. Washington, DC: American Psychological Association.

Rentfrow, P, M Jokela, M Lamb, J Potter, L Goldberg and R McCrae (2015). 'Regional Personality Differences in Great Britain'. *PLOS ONE* 10 (3). Ed. by R Latzman, e0122245. DOI: 10.1371/journal.pone.0122245.

Resch, B (2013). 'People as Sensors and Collective Sensing - Contextual Observations Complementing Geo-Sensor Network Measurements'. In: *Lecture Notes in Geoinformation and Cartography: Progress in Location-Based Services*. Ed. by J Krisp. Heidelberg: Springer, pp. 391–406. DOI: 10.1007/978-3-642-34203-5_22.

Resch, B, R Britter and C Ratti (2012). 'Live Urbanism - Towards SENSEable Cities and Beyond'. In: *Sustainable Environmental Design in Architecture*. Ed. by S Rassia and P Pardalos. New York, NY: Springer, pp. 175–184. DOI: 10.1007/978-1-4419-0745-5_10.

Resch, B, M Sudmanns, G Sagl, A Summa, P Zeile and J Exner (2015a). 'Crowdsourcing Physiological Conditions and Subjective Emotions by Coupling Technical and Human Mobile Sensors'. In: *GI_Forum 2015*. Ed. by A Car, T Jekel, J Strobl and G Griesebner. Salzburg: Wichmann, pp. 514–524. DOI: 10.1553/giscience2015s514.

Resch, B, A Summa, G Sagl, P Zeile and J Exner (2015c). 'Urban Emotions - Geo-Semantic Emotion Extraction from Technical Sensors, Human Sensors and Crowdsourced Data'. In: *Lecture Notes in Geoinformation and Cartography: Progress in Location-Based Services 2014*. Ed. by G Gartner and H Huang. Heidelberg: Springer, pp. 199–212. DOI: 10.1007/978-3-319-11879-6_14.

Resch, B, A Summa, P Zeile and M Strube (2016). 'Citizen-Centric Urban Planning through Extracting Emotion Information from Twitter in an Interdisciplinary Space-Time-Linguistics Algorithm'. *Urban Planning* 1 (2), pp. 114–127. DOI: 10.17645/up.v1i2.617.

Richardson, D, N Volkow, M Kwan, R Kaplan, M Goodchild and R Croyle (2013). 'Spatial Turn in Health Research'. *Science* 339 (6126), pp. 1390–1392. DOI: 10.1126/science.1232257.

Robinson, W (1950). 'Ecological Correlations and the Behavior of Individuals'. *American Sociological Review* 15 (3), pp. 351–357. DOI: 10.2307/2087176.

— (2009). 'Ecological Correlations and the Behavior of Individuals'. *International Journal of Epidemiology* 38 (2), pp. 337–341. DOI: 10.1093/ije/dyn357.

— (2011). 'Erratum: Ecological correlations and the behavior of individuals'. *International Journal of Epidemiology* 40, p. 1134. DOI: 10.1093/ije/dyr082.

Rothman, K, J Gallacher and E Hatch (2013). 'Why Representativeness Should be Avoided'. *International Journal of Epidemiology* 42 (4), pp. 1012–1014. DOI: 10.1093/ije/dys223.

Ruddell, D and E Wentz (2009). 'Multi-Tasking: Scale in Geography'. *Geography Compass* 3 (2), pp. 681–697. DOI: 10.1111/j.1749-8198.2008.00206.x.

Sagl, G and B Resch (2014). *Mobile Phones as Ubiquitous Social and Environmental Geo-Sensors*.

Sagl, G, B Resch and T Blaschke (2015). 'Contextual Sensing: Integrating Contextual Information with Human and Technical Geo-Sensor Information for Smart Cities'. *Sensors* 15 (7), pp. 17013–17035. DOI: 10.3390/s150717013.

Saib, M, J Caudeville, F Carre, O Ganry, A Trugeon and A Cicolella (2014). 'Spatial Relationship Quantification between Environmental, Socioeconomic and Health Data at Different Geographic Levels'. *International Journal of Environmental Research and Public Health* 11 (4), pp. 3765–3786. DOI: 10.3390/ijerph110403765.

Sakaki, T, M Okazaki and Y Matsuo (2013). 'Tweet Analysis for Real-Time Event Detection and Earthquake Reporting System Development'. *IEEE Transactions on Knowledge and Data Engineering* 25 (4), pp. 919–931. DOI: 10.1109/TKDE.2012.29.

Sakshaug, J, M Couper, M Ofstedal and D Weir (2012). 'Linking Survey and Administrative Records: Mechanisms of Consent'. *Sociological Methods & Research* 41 (4), pp. 535–569. DOI: 10.1177/0049124112460381.

Sarowar Sattar, A, J Li, X Ding, J Liu and M Vincent (2013). 'A General Framework for Privacy Preserving Data Publishing'. *Knowledge-Based Systems* 54, pp. 276–287. DOI: 10.1016/j.knosys.2013.09.022.

Schaeffer, N, J Dykema and D Maynard (2010). 'Interviewers and Interviewing'. In: *Handbook of Survey Research*. Ed. by P Marsden and J Wright. 2nd ed. Bingley, UK: Emerald Publishing, pp. 437–470.

Scheuren, F (2004). *What is a Survey?* 2nd ed. Washington, DC: American Statistical Association.

Schimmack, U, E Diener and S Oishi (2002). 'Life-Satisfaction is a Momentary Judgment and a Stable Personality Characteristic: the Use of Chronically Accessible and Stable Sources'. *Journal of Personality* 70 (3), pp. 345–384. DOI: 10.1111/1467-6494.05008.

Schmiedeberg, C and J Schröder (2014). 'Does Weather Really Influence the Measurement of Life Satisfaction?' *Social Indicators Research* 117 (2), pp. 387–399. DOI: 10.1007/s11205-013-0350-7.

Schnell, R (2013a). 'Getting Big Data but Avoiding Big Brother'.

— (2013b). 'Linking Surveys and Administrative Data'.

Schwarz, N and G Clore (1983). 'Mood, Misattribution, and Judgments of Well-Being: Informative and Directive Functions of Affective States'. *Journal of Personality and Social Psychology* 45 (3), pp. 513–523. DOI: 10.1037/0022-3514.45.3.513.

Schweers, S, K Kinder-Kurlanda, S Müller and P Siegers (2016). 'Conceptualizing a Spatial Data Infrastructure for the Social Sciences: An Example from Germany'. *Journal of Map & Geography Libraries* 12 (1), pp. 100–126. DOI: 10.1080/15420353.2015.1100152.

Schyns, P (1998). 'Crossnational Differences in Happiness: Economic and Cultural Factors Explored'. *Social Indicators Research* 43 (1/2), pp. 3–26. DOI: 10.1023/A:1006814424293.

Shiffman, S (2007). 'Designing Protocols for Ecological Momentary Assessment'. In: *The Science of Real-Time Data Capture: Self-Reports in Health Research*. Ed. by A Stone, S Shiffman, A Atienza and L Nebeling. New York, NY: Oxford University Press, pp. 27–53.

Sonnentag, S, C Binnewies and S Ohly (2012). 'Event-Sampling in Occupational Health Psychology'. In: *Research Methods in Occupational Health Psychology : Measurement, Design and Data Analysis*. Ed. by R Sinclair, M Wang and L Tetrick. New York, NY: Routledge, pp. 208–228.

Steffens, M and S Mecklenbräuker (2007). 'False Memories: Phenomena, Theories, and Implications'. *Journal of Psychology* 215, pp. 12–24. DOI: 10.1027/0044-3409.215.1.12.

Sugovic, M and J Witt (2013). 'An Older View on Distance Perception: Older Adults Perceive Walkable Extents as Farther'. *Experimental Brain Research* 226 (3), pp. 383–391. DOI: 10.1007/s00221-013-3447-y.

Swan, M (2012). 'Sensor Mania! The Internet of Things, Wearable Computing, Objective Metrics, and the Quantified Self 2.0'. *Journal of Sensor and Actuator Networks* 1 (3), pp. 217–253. DOI: 10.3390/jsan1030217.

— (2013). 'The Quantified Self: Fundamental Disruption in Big Data Science and Biological Discovery'. *Big Data* 1 (2), pp. 85–99. DOI: 10.1089/big.2012.0002.

Talhelm, T, X Zhang, S Oishi, C Shimin, D Duan, X Lan and S Kitayama (2014). 'Large-Scale Psychological Differences Within China Explained by Rice Versus Wheat Agriculture'. *Science* 344 (6184), pp. 603–608. DOI: 10.1126/science.1246850.

Tobler, W (1970). 'A Computer Movie Simulating Urban Growth in the Detroit Region'. *Economic Geography* 46, pp. 234–240. DOI: 10.2307/143141.

Triantafyllidis, A, C Velardo, D Salvi, S Shah, V Koutkias and L Tarassenko (2017). 'A Survey of Mobile Phone Sensing, Self-Reporting, and Social Sharing for Pervasive Healthcare'. *IEEE Journal of Biomedical and Health Informatics* 21 (1), pp. 218–227. DOI: 10.1109/JBHI.2015.2483902.

Tröndle, M, S Greenwood, V Kirchberg and W Tschacher (2014). 'An Integrative and Comprehensive Methodology for Studying Aesthetic Experience in the Field: Merging Movement Tracking, Physiology, and Psychological Data'. *Environment and Behavior* 46 (1), pp. 102–135. DOI: 10.1177/0013916512453839.

Tuan, Y (1977). *Space and Place: The Perspective of Experience*. Minneapolis, MN: University of Minnesota Press.

Turner, M, V Dale and R Gardner (1989). 'Predicting Across Scales: Theory Development and Testing'. *Landscape Ecology* 3 (3-4), pp. 245–252. DOI: 10.1007/BF00131542.

Tversky, B, J Bauer Morrison, N Franklin and D Bryant (1999). 'Three Spaces of Spatial Cognition'. *The Professional Geographer* 51 (4), pp. 516–524. DOI: 10.1111/0033-0124.00189.

Vasiliu, L, A Freitas, F Caroli, S Handschuh, R McDermot, M Zarrouk, M Hürlimann, B Davis, T Daudert, M Khaled, D Byrne, S Fernández and A Cavallini (2016). 'In Or Out? Real-Time Monitoring of BREXIT Sentiment on Twitter'. In: *SEMANTICS 2016*. Leipzig.

Veena, K and D Devidas (2014). 'Data Anonymization Approaches for Data Sets Using Map Reduce on Cloud: A Survey'. *International Journal of Science and Research* 3 (4), pp. 308–311.

Visser, P, J Krosnick and P Lavrakas (2000). 'Survey Research'. In: *Handbook of Research Methods in Social and Personality Psychology*. Ed. by H Reis and C Judd. New York, NY: Cambridge University Press, pp. 223–252.

Vogel, M (2016). 'The Modifiable Areal Unit Problem in Person-Context Research'. *Journal of Research in Crime and Delinquency* 53 (1), pp. 112–135. DOI: 10.1177/0022427815597039.

Waldhör, T (1996). 'The Spatial Autocorrelation Coefficient Moran's I Under Heteroscedasticity'. *Statistics in Medicine* 15 (7-9), pp. 887–892. DOI: 10.1002/(SICI)1097-0258(19960415)15:7/9<887::AID-SIM257>3.0.CO;2-E.

Wang, H (2010). ‘Privacy-Preserving Data Sharing in Cloud Computing’. *Journal of Computer Science and Technology* 25 (3), pp. 401–414. DOI: 10.1007/s11390-010-9333-1.

Wang, H, D Can, A Kazemzadeh, F Bar and S Narayanan (2012a). ‘A System for Real-Time Twitter Sentiment Analysis of 2012 U.S. Presidential Election Cycle’. In: *Proceedings of the ACL 2012 System Demonstrations*. Ed. by M Zhang. Jeju Island: Association for Computational Linguistics, pp. 115–120.

Ward, M and K Gleditsch (2008). *Spatial Regression Models*. Thousand Oaks, CA: SAGE, p. 99.

Warf, B and S Arias (2009). *The Spatial Turn : Interdisciplinary Perspectives*. London, UK: Routledge.

Weiss, E, G Kemmler, E Deisenhammer, W Fleischhacker and M Delazer (2003). ‘Sex Differences in Cognitive Functions’. *Personality and Individual Differences* 35 (4), pp. 863–875. DOI: 10.1016/S0191-8869(02)00288-X.

Wender, K, D Haun, B Rasch and M Blümke (2002). ‘Context Effects in Memory for Routes’. In: *Spatial Cognition III*. Ed. by C Freksa, W Brauer, C Habel and K Wender. Tutzing: Springer, pp. 209–231. DOI: 10.1007/3-540-45004-1_13.

West, B, F Kreuter and U Jaenichen (2013). ‘"Interviewer" Effects in Face-to-Face Surveys: A Function of Sampling, Measurement Error, or Nonresponse?’ *Journal of Official Statistics* 29 (2), pp. 277–297. DOI: 10.2478/jos-2013-0023.

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 (5), pp. 868–887. DOI: 10.1080/13658816.2014.1002499.

Westerholt, R, E Steiger, B Resch and A Zipf (2016). ‘Abundant Topological Outliers in Social Media Data and Their Effect on Spatial Analysis’. *PLOS ONE* 11 (9), e0162360. DOI: 10.1371/journal.pone.0162360.

Witt, J, D Proffitt and W Epstein (2010). ‘When and How are Spatial Perceptions Scaled?’ *Journal of Experimental Psychology: Human Perception and Performance* 36 (5), pp. 1153–1160. DOI: 10.1037/a0019947.

Xu, P, H Huang, N Dong and M Abdel-Aty (2014b). ‘Sensitivity Analysis in the Context of Regional Safety Modeling: Identifying and Assessing the Modifiable Areal Unit Problem’. *Accident Analysis & Prevention* 70, pp. 110–120. DOI: 10.1016/j.aap.2014.02.012.

Zadra, J and G Clore (2011). ‘Emotion and Perception: The Role of Affective Information’. *Wiley Interdisciplinary Reviews: Cognitive Science* 2 (6), pp. 676–685. DOI: 10.1002/wcs.147.

II.4.5 Appendix

II.4.5.1 Glossary of Terms

Table II.4.1: Glossary of Terms

Label	Explanation	Reference
Geography	The discipline dealing with the interactions between humans (or natural systems) and space. The endeavor is limited to the humanly comprehensible scale.	Clifford et al. (2009)
Geographic Information Science (GIScience)	The discipline that investigates theoretical issues regarding the nature, acquisition, storage, analysis, and presentation of geospatial information and data, while abstracting these from specific geographic questions.	Goodchild (1992) and Goodchild (2010)
Geoinformatics	Largely overlaps with GIScience; preferred among German-speaking scholars; stronger technological focus, as it accentuates the development and application of methods and technology.	Lange (2013)
Geographic Information System (GIS)	A GIS is a system of hardware, software, and procedures to support the capture, management, manipulation, analysis, modelling, and display of spatially-referenced data for solving complex planning and management problems.	Goodchild and Kemp (1992)

II.5 Research on Social Media Feeds—A GIScience Perspective

II.5.1 Introduction

During the last two decades, the role of internet users changed dramatically. While they were mostly passive content consumers before, they are now considered proactive data producers. This phenomenon is summarized by the term "Prosumer" (Ritzer and Jurgenson 2010) and gets facilitated through major technological advancements such as ubiquitous access to the mobile Internet and a widespread use of smartphones equipped with positioning and sensing capabilities. These outlined developments do not just happen recently, but trace back to the much older development around the so called "Web 2.0" (ITU 2014). In geospatial terms, these developments are well reflected by Mike Goodchild's popular definition of "Citizens as Sensors" (Goodchild 2007), where ordinary people capture and disseminate "Volunteered Geographic Information (VGI)". Haklay further puts this development into broader context and rather coined the term "GeoWeb" (Haklay et al. 2008). OpenStreetMap (OSM) is probably the most prominent example of VGI.

Projects like OSM provide a well-defined data capturing protocol as well as a clear mission regarding their contributed contents. In contrast, data originating from online social networks (another source of VGI) is way more heterogeneous and diverse. At the same time, however, it may also provide high levels of semantic detail and is generated by a larger number of users. Consequently, it gained the interest of various research disciplines. These range from sociology toward linguistics, and of course geography and GIScience. The latter one is facilitated by the fact that a great deal of information contributed to social media is geotagged. Thus, the remainder of this chapter focuses on the spatial aspects of social media, and potential applications that can be derived from this kind of data.

Section II.5.2 highlights the general potential of social media analysis for investigating social phenomena. We do that by outlining selected case studies from the exemplary field of human mobility analysis. These have demonstrated the usefulness of social media for investigating mobility patterns as well as human spatial behaviour. Section II.5.3 then provides an overview of several different application domains of social media analyses, with a particular focus on Twitter. Finally, Section II.5.4 discusses some technical issues of established spatial analysis methods when these are applied to social media data. We conclude the chapter by summarizing its different parts. We further provide recommendations regarding future GIScience research on social media data.

II.5.2 Utilisation of Social Media Data for Investigating Urban Environments

The spatial and social structures of a city as well as the dynamic nature of human activities result in certain collective and individual human behaviour patterns. Social media data can help to "sense" this type of information from urban environments in an in-situ manner. GIScience research thereby is focused on the overall question how corresponding spatiotemporal patterns from ubiquitous sensor networks and heterogeneous data streams can be explored, extracted, validated and aggregated. In turn, such information might enable us to sense everyday spatial processes and to gain knowledge about urban environments, especially with respect to collective human dynamics. The study of these issues has become one of the primary objectives of GIScience (Giannotti and Pedreschi 2008).

The information originating from social media messages (*e. g.*, tweets in case of Twitter) may contain spatial, temporal and semantic attributes. Considering these dimensions, social media can be considered as a (partial) proxy of real-world happenings. However, space, time as well as semantics are influenced by each user's individual perception of the surrounding space. It is thus important to figure out ways to circumvent these issues for gaining trustworthy and objective information from these data sources.

The following short paragraphs outline case studies in which a range of GIScience researchers has drawn human mobility and urban study related knowledge from Twitter. We group these studies in accordance to their underlying research goals. The listed paragraphs thus provide the reader a quick overview of both the types of studies that have been conducted as well as methods and outcomes.

II.5.2.1 Mobility and Social Behaviour

Studying the social dynamics of a city remains a challenging endeavour, which has recently been carried out in a qualitative manner. Thus, social media might be a promising source of information in order to provide a better understanding of social dynamics within urban environments and resulted in various research efforts. Regarding the analysis of collective human mobility and activity patterns from social media, Cho et al. (2011) investigate social ties and their influence on human mobility patterns by comparing social media check-in data and cellphone location data. They found a stronger association of social network ties influencing long-distance travel than short range spatially and temporally periodic movements. Within the observed Twitter user pattern, Lee and Sumiya (2010) study user behaviour by measuring geographic regularities and detecting geo-social events through identifying Regions of Interest (RoI). Another approach conducted by Noulas et al. (2011), Cranshaw et al. (2012) and Kafsi et al. (2015) is the identification of characteristic neighbourhoods, collective movement patterns and social ties within certain user communities from Foursquare and other social media data. In a similar approach for Twitter, Li et al. (2014) measure the spatial dispersion of users in a community and their trajectories. Hawelka et al. (2014) aim to further empirically validated the observed human behaviour patterns and found a correlation between the conducted Twitter census and economic key figures. Furthermore, Li et al. (2013) explore spatiotemporal pattern of Twitter and Flickr data and investigated a relationship between socioeconomic characteristics of people who are generating social media posts in the US.

II.5.2.2 Mobility and Underlying Urban Structures

The exploration of the relationships and the impact of urban structures on human mobility is an interesting study area for social media researcher. Wakamiya et al. (2011) investigate temporal patterns of crowd behaviour over Japan by spatial partitioning tweets in order to extract urban characteristics. On a smaller scale several studies investigate the connection with extracted urban activities from social media and their connection with the underlying urban structure. Kling and Pozdnoukhov (2012) were able to detect spatiotemporal clusters of frequently occurring urban topics in New York. Furthermore, Ferrari et al. (2011) also work with georeferenced tweets and a semantic probabilistic topic modelling approach to automatically extract urban patterns from location-based social networks. The study concluded that extracted urban motion patterns and identified hotspots in the city allow the inference of crowd behaviours that recur over time and space. A similar approach by using Foursquare data by Cheng et al. (2011) and Hasan et al. (2013) also resulted in the characterisation of urban human mobility and activity patterns. Andrienko et al. (2013) correlated the spatiotemporal clusters of keyword based filtered georeferenced tweets of places where people tweet with US population densities. The results have shown strong correlations between the observed Twitter distribution and census data, suggesting that social media is a reliable proxy for the inference of mobility patterns. One further application is to derive intra-urban events showing distinct mobility patterns over time. This spatiotemporal movement has proven to reflect typical mobility behaviour in the underlying urban structures (Steiger et al. 2015b).

II.5.2.3 Mobility and Human Activities

Several studies infer individual and collective human daily activity patterns by analysing crowdsourced information, such as taxi trip records (Liang et al. 2012), GPS traces (Azevedo et al. 2009; Jiang et al. 2009) or mobile phone records (Candia et al. 2008; Gao 2015). Consequently, a large literature body also focus on studying human mobility and activity pattern from social media data. (Krumm et al. 2013) estimate individual home locations of heavy Twitter users and apply machine learning algorithms to classify and predict individual travel behaviour. Jin et al. (2016) developed a method to infer users' mobility patterns from check-ins in Foursquare. Coffey and Pozdnoukhov (2013) go one step further and semantically annotate mobility flow datasets with activity information and trip purposes from tweets. Similarly, Wu et al. (2015) utilise social media to annotate the location history of mobile phone users for the characterisation of certain social activities. Focusing on the content of tweets, Grinberg et al. (2013) proposed a method to detect semantic patterns to infer clusters of users' real world activity. Gao (2015) developed a probabilistic approach to make place recommendations based on the users' geo-social circles, as extracted from Foursquare. In another study, the authors estimate spatiotemporal mobility flows from Twitter for the area of greater Los Angeles to infer origin- and destination trips (Gao et al. 2014). Results have shown similar pattern when comparing with community survey data. In a previous study we introduced a semantic and spatial analysis method (Steiger et al. 2014b), through which we were able to extract geographic features from uncertain Twitter data and have shown that observed clusters correspond to landmarks, such as highly frequented squares and major transportation hubs. A further investigation revealed similar semantic layers that represent collective human mobility flows in co-occurrence with underlying social activity (Steiger et al. 2014a) and could thus lead to new insights in characterising urban mobility.

II.5.2.4 Future Research Recommendations

Further research needs to be conducted to assess the reliability of social media datasets. It also must be noted that the data collected from wireless devices are influenced by GPS/WIFI inaccuracy issues (Zandbergen and Barbeau 2011). Moreover, users can individually choose to share their precise location to a tweet or just a general location information (such as a city or neighbourhood). This resulting location uncertainty leads to imprecise location information of geotagged tweets (Li et al. 2011).

Within the semantic attribute one must consider that the containing information may relate to events in the past, present or even future (Sengstock and Gertz 2012). Principally the text corpora as such in social media posts are relatively sparse and vague. It may also be fairly ambiguous and hence featuring only a weak indicator of a real world event. This uncertain semantic knowledge is a result of the fact that people using Twitter have individual motivations to post information and their main intention is to primarily serve their own communication needs. One further typical characteristic of social media is that users do not post equally distributed in geographic space and time leading to a heterogeneous dispersion of posts. Jatowt et al. (2015) further assess these varying temporal patterns and dynamics within social media. Furthermore, georeferenced social media posts only represent a small fraction of the overall available data. Not all user groups use all types of social media platforms similarly, which produces a potentially strong socio-demographic bias (Longley and Adnan 2016). Last, the application of spatial and semantic methods themselves creates uncertainties, since the distribution of specific geographic phenomena and their semantic complexities within tweets are not known beforehand (Westerholt et al. 2015). Hence, it is important to compare and validate results with other acquired sensor data.

Conducting further research in this area however will be worthwhile, since study results may provide new additional insights into the complex human-sensor-city relationship at a much more fine-grained spatial and temporal level than before. New knowledge gained from this research will provide a better understanding of individual and collective human behavior within urban environments and may assist stakeholders and decision makers in their planning processes.

II.5.3 Application Domains of Social Media Analyses

Location-based social networks (LBSN) (Roick and Heuser 2013) offer a vast amount of voluntary content. The investigation of human activities in location-based social networks is one promising example of exploring spatial structures in order to infer underlying spatiotemporal patterns. Twitter for example is more and more recognised by numerous research domains. In particular it provides an opportunity for GIScience to understand geographic processes and spatial relationships comprised in social networks. Summarising the current state of research concerning the application for spatiotemporal analyses, one outcome of a previously conducted systematic literature (Steiger et al. 2015a) revealed that Twitter analyses are mainly focused on the spatiotemporal classification and detection of events. Principal investigated application domains are:

II.5.3.1 Event Detection

To detect events, researchers are currently looking for spatial, temporal and semantic patterns within Twitter. In this respect people act as a social sensors for events (Yardi and Boyd 2010; Chae et al. 2012). Disaster- and emergency management as one event detection subfield has been the primarily identified application in nearly a third of all reviewed studies (Sakaki et al. 2010; Murthy and Longwell 2013; Crooks et al. 2013). Further research has been conducted on utilising Twitter in traffic management. This can be found in 14 % of reviewed studies (Kosala and Adi 2012; Wakamiya et al. 2012; Lenormand et al. 2014). Another area which seems to be quite popular is research on Twitter data for disease/health management adding up to another 5 % of the reviewed studies (Lampos and Cristianini 2010; Gomide et al. 2011; Sofean and Smith 2012). A famous example is the derivation and prediction of information on infection sources and the spreading of an illness from Twitter messages (Culotta 2010; Collier et al. 2011). One prominent example is earthquake detection from Twitter data (Longueville et al. 2010; Zook et al. 2010). This has been successfully accomplished in a number of studies correlating results with official earthquake sensor data (Tapia et al. 2011; Thomson et al. 2012). Sakaki et al. (2010) have developed an algorithm that uses Twitter to calculate earthquakes' epicenters and the typhoons' trajectories. Moreover, situational information can be derived from location-related short messages to coordinate emergency responses (Vieweg et al. 2010). Also in the context of disease and health management similar outcomes have been derived. Tweets showing disease incidents have shown similar spatiotemporal distributions as those in with official reports. With these studies research has proven the trustworthiness and a high level of representativeness of tweets throughout different application domains (Albuquerque et al. 2015).

II.5.3.2 Location Inference

Locations of users within social networks can be inferred or even predicted with the help of direct or indirect geolocation information derived from the provided metadata or from the semantic content (Kinsella et al. 2011; Hong et al. 2012; Hiruta et al. 2012). The geographic accuracy could be increased by extracting the textual information from the tweet or from the metadata itself. For example, Lamprianidis and Pfoser (2011) have extracted locations and their names from Flickr pictures by clustering user-generated data points associated with geo-referenced pictures. Kelm et al. (2013) discusses various methods to extract place names from textual data from articles, posts or tags in geo-social networks, including place name gazetteer and statistical language modelling. Some methods follow an opposite approach and infer the location of a feature from implicit location information. Serdyukov et al. (2009) model the probability that a group of tags be assigned to a location. Similarly, (Gallagher et al. 2009) used location probability maps generated from tags for the same purpose. Van Laere et al. (2010) have pursued the same goal using k-medoids and Naïve Bayes clustering methods. Some approaches focus on inferring a user's or a group of users' location. Cheng et al. (2010) have proposed a probabilistic method to determine users' location from the content of their Twitter messages. Other authors have proposed to use the location of users' friends to achieve the same goal (Backstrom et al. 2010). Stefanidis et al. (2013) have proposed a framework to harvest ambient geospatial information from social media feeds to locate social hotspots or to map social networks in a given geographical area. Ajao et al. (2015) summarise the broad range of available techniques applied to infer direct and indirect location from Twitter messages and social media users.

II.5.3.3 Geo-Social Network Analysis

Another important domain of research is social analysis which investigates relationships of individual users within a social network (Wu et al. 2011; Cranshaw et al. 2012). Geo-social network analysis seeks to identify the structure of social networks and their distribution in geographic space (Scellato et al. 2010; Lee and Sumiya 2010). Social ties may feature distinct spatial distributions enabling spatiotemporal analyses. These distributions can help finding collective social activities and ultimately understanding geographical processes. A subfield of geo-social network analysis are sentiment and emotion analysis (Wang et al. 2012a; Quercia et al. 2012). This field of research also offers a great potential for GIScience in the context of extracting contextual emotional information within urban and rural environments. One promising further field of research within social analysis which should be mentioned is urban planning and management which also could benefit from the rich data found in location based social networks such as Twitter. In the context of disaster management, several studies aim to infer the social dimensions within certain geo-located communities in twitter during disaster events (Conover et al. 2013; Bakillah et al. 2015).

II.5.3.4 Future Research Recommendations

Social Media data for research has proven to be a valuable source, as it not only comes for free, but also features a high spatiotemporal resolution. This kind of data especially enables possibilities to find spatial patterns and events which can help validating existing information sources. One identified main research gap is the exploration of human spatial behaviour (Miller and Goodchild 2015) in order to gain knowledge about the underlying geographic processes and dynamics. Furthermore, the current research foci allow to transfer established methods from various disciplines (*e. g.* Computer- and Information Science, Social Science etc.) into other disciplines and enhancing new applications. As one example, more use of computer linguistic approaches to leverage knowledge from textual information, combined with methods for spatiotemporal analysis from computational sciences could lead to new insights within specific geographic application domains, such as disaster management or human mobility analysis.

II.5.4 Spatial Analysis of Social Media Feeds – Challenges and Approaches

The primary goal of spatial analysis is to explore structures within spatial data. This typically involves tasks like finding clusters on a map or figuring out distributional characteristics of data. One theoretical field underlying spatial analysis is spatial statistics. This field provides the basic principles that are underlying many spatial analysis problems. Key to this field is identifying spatial correlations, and thus hints on systematic patterns in geographic data (Fischer and Getis 2010b). Respective methods and techniques are thus useful tools for gaining geographic insight into social media data.

The spatial analysis of social media data is typically conducted in an exploratory manner. This is due to lacking knowledge about potential underlying spatial processes, and thus about social media messages and their dispersal in geographic space in general. Useful tools on that regard are the K-Function (Ripley 1976) (purely geometric) and the mark correlation function (Stoyan and Stoyan 1994) (attribute values), both originating from spatial point pattern analysis. These methods allow identifying significant geometric clustering and regularity within stochastic point patterns. When the geometry is fixed (or rather treated as such) spatial autocorrelation statistics like Moran's I (Moran 1950; Cliff and Ord 1973) and hot spot statistics like Getis-Ord's G statistics (Getis and Ord 1992; Ord and Getis 1995) are suitable alternatives. These assess the degree of randomness within georeferenced attributes associated to units on a fixed geographic layout. In fact, many of the latter are essentially identical to different variants of the mark correlation function (see, *e. g.*, Shimatani (2002)). Thereby, Moran's I tests for correlations between neighboured observations across space, while G separates between extremal values (*i. e.*, high and low).

As mentioned earlier, thorough spatial knowledge about social media datasets is typically lacking. Consequently, analysts oftentimes proceed with a trial-and-error approach when parametrising the methods mentioned above. It is common practice to apply these techniques to different scales. The goal then is to sort out that scale at which patterning seems to be most pronounced. However, the techniques mentioned so far were designed long before the appearance of social media and similar kinds of user-generated data. The idea of the following two sections is thus to briefly reflect differences between social media and more traditional data, and to give some recommendations with respect to the spatial analysis of these.

II.5.4.1 Potential Issues and Pitfalls

The issues presented in the following are likely to occur when analysing social media feeds with established methods from spatial analysis. It is important to note that social media feeds provide a mixture of indications from different real-world (and also some solely virtual) phenomena. This is due to the autonomous manner in which the data is being collected. Users can contribute any type of content from any place at any time. Such a mixture might be beneficial in terms of the wealth of contained information about the users' everyday lives. However, it also imputes some critical problems when it comes to spatial analysis. Probably the most trivial yet critical among these is the mere mixture of information as such. Any attribute which is derived from social media is highly likely to include information from several different real-world phenomena. Analysing social media therefore comes at the risk of drawing conclusions about a mixture population that might not exist in reality. In most circumstances this is not desirable, since it does not lead to reasonable insight about any real-world process. One way to overcome this problem would be an accurate a priori semantic separation. However, that is a non-trivial task on its own right given the colloquial language used in corresponding messages.

Another issue with social media data is the implicit subjectivity that is per se introduced by the notion of "humans as sensors" (Goodchild 2007). One implication from that concept is the diversity at which people perceive environments (see also Section A). Similar phenomena might lead to varying responses among different users. This inevitably leads to an increased difficulty in analysing the semantics (*i. e.*, the attribute value) of the observations; and thus to a potential misclassification of phenomena. The implication of that for spatial analysis is crucial: techniques such as measures of spatial autocorrelation or spatial regression techniques are based on both, spatial characteristics as well as the attribute values. Consequently, spatial analysis techniques might end up in spurious results when the analyst fails controlling such effects.

The analysis of social media can also lead to an artificial increase in the number of type I/type II errors. This problem is likely to occur whenever testing hypotheses about spatial patterns with social media datasets. One might be interested in assessing spatial heterogeneity by means of local statistics like local Moran's I (Anselin 1995) or G_i^* (Ord and Getis 1995). It is common sense that these methods lead to an increase in type I errors due to alpha error inflation (Nelson 2012). Thus, it is important to control the alpha level accordingly (*e. g.*, through techniques such as False-Discovery-Rate (Benjamini and Hochberg 1995)). With social media datasets, however, phenomena operating at smaller scales than the adjusted analysis scale might be considered by accident; and inadvertently influence the analysis. This is due to the mixture described above which is leading to spatially overlapping representations of different phenomena. The result is an increased amount of spurious indications of significant spatial effects.

Another critical implication of the scale-mixture outlined above is a potential creation of wrong and misleading relationships across scale levels. Recall that observations from smaller scale levels are prone to inherently being included in analyses at larger scales due to potential geometric mixture. Effects from smaller scales are therefore likely to be propagated towards analyses at larger scales. Due to this effect, some results become impossible, *e. g.*, in scenarios where one wants to assess spatial autocorrelation at some large scale that is influenced by highly autocorrelated observations from smaller scales. If there is spatial autocorrelation present at some small scale (*e. g.* one "heavy" Twitter user recurrently posting from a particular location), it will be carried through to all larger scales being observed in the same geographic neighbourhood.

Further discussion of these and related problems (including some empirical results) can be found in Westerholt et al. (2015) (including a discussion of a multi-scale modification of the local G statistic) and Lovelace et al. (2016). The presented list of effects is of course not exhaustive. There might be many more effects, some of which are still about to be discovered. The subsequent section provides some hints and recommendations about how to precede with the spatial analysis of social media data.

II.5.4.2 Some Recommendations

Spatial autocorrelation is the core principle underlying a great deal of spatial analysis methodology. Therefore, it is crucial to accurately assess this characteristic in order to design applicable methods, and for drawing reasonable geographic conclusions. This is not just important for exploratory tests on spatial clustering and heterogeneity, but also crucial for model-driven spatial regression scenarios such as Geographically Weighted Regression (GWR) (Fotheringham et al. 2002) and for assessing model misspecification (Cliff and Ord 1981). Unfortunately, in case of social media analysis, the assessment of spatial autocorrelation is strongly affected by the problems depicted in the previous section. Therefore, one recommendation in terms of future research is to work on appropriate adaptations of corresponding measures and techniques in order to account for multi-scale (or rather: "mixed-scale") and multi-categorical effects. As long as these are not available, one should carefully parametrise respective techniques. Another (aspatial) approach might be to decompose social media datasets a priori, probably based on some other characteristic such as the Tweets' semantics. The worst option of all, however, would be to neglect the specific spatial characteristics of social media data when conducting spatial analysis. That would lead to a wrong evaluation of spatial effects; and thus to wrong analysis results.

Another recommendation is related to one of the promising opportunities that come with social media datasets: their wealth of information. We can obtain an array of valuable and potentially interrelated properties from social media data. These include temporal, semantic and spatial information. Correspondingly, one should try to analyse all these dimensions simultaneously instead of considering them in a separated fashion. This might unveil a much deeper understanding of social phenomena that are reflected in such datasets. Recent research efforts like, *e. g.*, Steiger et al. (2016b) reflect this idea. However, it yet remains a challenge to find measures to incorporate these different kinds of information in joint methodology in a reasonable way.

II.5.4.3 Conclusion and an Outlook on Future Work

We outlined some potential pitfalls when analyzing social media data spatially. These are caused by the inherent characteristics of the data, *i. e.*, the way in which the data is collected and what such services are used for. Potential problems include geometric mixtures of differently scaled data; semantic mixtures that get blurred in joint attributes derived from the data; and (more generally) spurious assessments of spatial correlations and thus pattern in the data.

The previous paragraphs are clearly biased towards the concept of spatial autocorrelation. On the one hand this focus is due to the research focus of the authors. On the other hand this is due to the central role which spatial autocorrelation plays throughout the entire field of spatial analysis. However, there are of course other important characteristics and pitfalls that might also influence the spatial analysis of social media data. The observations come, for instance, with considerable uncertainties with respect to relevant dimensions: The text snippets are colloquial and oftentimes difficult to interpret (semantics), the time stamp is sometimes not in line with real-world happenings (temporal) and the geographic coordinates are prone to positioning inaccuracies (spatial). The intensities of all these uncertainties appear to be varying across different users, devices, regions, etc. All these uncertainties indeed have impact on the results of spatial analysis.

Future methodological research should focus on the specific spatial characteristics of social media data (that are not yet known to a full extent). For now, across all disciplines and domains, it is common sense to apply established standard methodology to social media data. Relatively little emphasis is put on purely methodological research on the background of the special characteristics of these datasets. Thus, there is still plenty of room for improvement. The discipline of GIScience could play a vital role in these developments. Beyond purely empirical research, the impact of the spatial disciplines has been quite small so far. However, given that many research questions around social media are distinctive spatial ones, we should put much more emphasis on specialized spatial analysis techniques for social media.

II.5.5 Conclusion

On the one hand, social media data offers an array of new perspectives regarding many research questions and applications. On the other hand, however, these datasets also come with a set of issues that need to be taken into account, in particular when it comes to spatial analysis. GIScience can contribute to the development of new spatial analysis methods for social media data. Current major issues from a GIScience perspective include:

- the need of spatial analysis methods to be adapted towards uncertain and unstructured data types from LBSN;
- the handling of geographic scale effects when analysing social media data;
- the need for combining different methods across disciplinary boundaries (e.g. social network analysis, semantic analysis, spatiotemporal analysis), in order to better utilise all available information dimensions;
- the development of data fusion and information extraction methods that take several different data sources simultaneously into account.

This would support exploring latent patterns and sensing geographical processes from social media data in a more realistic manner. GIScience could thus contribute to answering these important geographic questions and may play a major role in the further exploration of social media data.

References (Chapter II.5)

Ajao, O, J Hong and W Liu (2015). ‘A Survey of Location Inference Techniques on Twitter’. *Journal of Information Science* 41 (6), pp. 855–864. DOI: 10.1177/0165551515602847.

Albuquerque, J de, B Herfort, A Brenning and A Zipf (2015). ‘A Geographic Approach for Combining Social Media and Authoritative Data Towards Identifying Useful Information for Disaster Management’. *International Journal of Geographical Information Science* 29 (4), pp. 667–689. DOI: 10.1080/13658816.2014.996567.

Andrienko, G, N Andrienko, H Bosch, T Ertl, G Fuchs, P Jankowski and D Thom (2013). ‘Thematic Patterns in Georeferenced Tweets through Space-Time Visual Analytics’. *Computing in Science & Engineering* 15 (3), pp. 72–82. DOI: 10.1109/MCSE.2013.70.

Anselin, L (1995). ‘Local Indicators of Spatial Association - LISA’. *Geographical Analysis* 27 (2), pp. 93–115. DOI: 10.1111/j.1538-4632.1995.tb00338.x.

Azevedo, T, R Bezerra, C Campos and L de Moraes (2009). ‘An Analysis of Human Mobility Using Real Traces’. In: *2009 IEEE Wireless Communications and Networking Conference*. Budapest: IEEE, pp. 1–6. DOI: 10.1109/WCNC.2009.4917569.

Backstrom, L, E Sun and C Marlow (2010). 'Find Me If You Can: Improving Geographical Prediction with Social and Spatial Proximity'. In: *Proceedings of the 19th International Conference on World Wide Web*. Ed. by J Freire and S Chakrabarti. Raleigh, NC: ACM Press, pp. 61–70. DOI: 10.1145/1772690.1772698.

Bakillah, M, R Li and S Liang (2015). 'Geo-Located Community Detection in Twitter with Enhanced Fast-Greedy Optimization of Modularity: The Case Study of Typhoon Haiyan'. *International Journal of Geographical Information Science* 29 (2), pp. 258–279. DOI: 10.1080/13658816.2014.964247.

Benjamini, Y and Y Hochberg (1995). 'Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing'. *Journal of the Royal Statistical Society: Series B (Methodological)* 57 (1), pp. 289–300.

Candia, J, M González, P Wang, T Schoenharl, G Madey and A Barabási (2008). 'Uncovering Individual and Collective Human Dynamics from Mobile Phone Records'. *Journal of Physics A: Mathematical and Theoretical* 41 (22), p. 224015. DOI: 10.1088/1751-8113/41/22/224015.

Chae, J, D Thom, H Bosch, Y Jang, R Maciejewski, D Ebert and T Ertl (2012). 'Spatiotemporal Social Media Analytics for Abnormal Event Detection and Examination Using Seasonal-Trend Decomposition'. In: *2012 IEEE Conference on Visual Analytics Science and Technology*. Washington, DC: IEEE, pp. 143–152. DOI: 10.1109/VAST.2012.6400557.

Cheng, Z, J Caverlee and K Lee (2010). 'You Are Where You Tweet: A Content-Based Approach to Geo-Locating Twitter Users'. In: *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*. Ed. by N Koudas, G Jones, X Wu, K Collins-Thompson and A An. Toronto: ACM Press, pp. 759–768. DOI: 10.1145/1871437.1871535.

Cheng, Z, J Caverlee, K Lee and D Sui (2011). 'Exploring Millions of Footprints in Location Sharing Services'. In: *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*. Barcelona, pp. 81–88.

Cho, E, S Myers and J Leskovec (2011). 'Friendship and Mobility'. In: *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Ed. by J Gosh and P Smyth. New York, NY: ACM Press, pp. 1082–1090. DOI: 10.1145/2020408.2020579.

Cliff, A and J Ord (1973). *Spatial Autocorrelation*. London, UK: Pion.

— (1981). *Spatial Processes: Models & Applications*. London, UK: Pion.

Coffey, C and A Pozdnoukhov (2013). 'Temporal Decomposition and Semantic Enrichment of Mobility Flows'. In: *Proceedings of the 6th ACM SIGSPATIAL International Workshop on Location-Based Social Networks*. Ed. by A Pozdnoukhov. New York, New York, USA: ACM Press, pp. 34–43. DOI: 10.1145/2536689.2536806.

Collier, N, N Son and N Nguyen (2011). 'OMG U Got Flu? Analysis of Shared Health Messages for Bio-Surveillance'. *Journal of Biomedical Semantics* 2 (Suppl. 5), S9. DOI: 10.1186/2041-1480-2-S5-S9.

Conover, M, C Davis, E Ferrara, K McKelvey, F Menczer and A Flammini (2013). 'The Geospatial Characteristics of a Social Movement Communication Network'. *PLoS ONE* 8 (3), e55957. DOI: 10.1371/journal.pone.0055957.

Cranshaw, J, R Schwartz, J Hong and N Sadeh (2012). 'The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City'. In: *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media*. Dublin.

Culotta, A (2010). ‘Towards Detecting Influenza Epidemics by Analyzing Twitter Messages’. In: *Proceedings of the First Workshop on Social Media Analytics*. Washington, DC: ACM Press, pp. 115–122. DOI: 10.1145/1964858.1964874.

Ferrari, L, A Rosi, M Mamei and F Zambonelli (2011). ‘Extracting Urban Patterns from Location-Based Social Networks’. In: *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks*. Ed. by Y Zheng and M Mokbel. New York, NY: ACM Press, pp. 9–16. DOI: 10.1145/2063212.2063226.

Fischer, M and A Getis (2010b). ‘Introduction’. In: *Handbook of Applied Spatial Analysis*. Ed. by M Fischer and A Getis. Heidelberg: Springer, pp. 1–24. DOI: 10.1007/978-3-642-03647-7_1.

Fotheringham, A, C Brunsdon and M Charlton (2002). *Geographically Weighted Regression : The Analysis of Spatially Varying Relationships*. Chichester, UK: Wiley.

Gallagher, A, D Joshi, Y Jie and L Jiebo (2009). ‘Geo-location Inference from Image Content and User Tags’. In: *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. Miami, FL: IEEE, pp. 55–62. DOI: 10.1109/CVPRW.2009.5204168.

Gao, S (2015). ‘Spatio-Temporal Analytics for Exploring Human Mobility Patterns and Urban Dynamics in the Mobile Age’. *Spatial Cognition & Computation* 15 (2), pp. 86–114. DOI: 10.1080/13875868.2014.984300.

Gao, S, J Yang, B Yan and G McKenzie (2014). ‘Detecting Origin-Destination Mobility Flows From Geotagged Tweets in Greater Los Angeles Area’. In: *Proceedings of the Eighth International Conference on Geographic Information Science*. Vienna.

Getis, A and J Ord (1992). ‘The Analysis of Spatial Association by Use of Distance Statistics’. *Geographical Analysis* 24 (3), pp. 189–206. DOI: 10.1111/j.1538-4632.1992.tb00261.x.

Giannotti, F and D Pedreschi (2008). ‘Mobility, Data Mining and Privacy: A Vision of Convergence’. In: *Mobility, Data Mining and Privacy*. Heidelberg: Springer, pp. 1–11. DOI: 10.1007/978-3-540-75177-9_1.

Gomide, J, A Veloso, W Meira, V Almeida, F Benevenuto, F Ferraz and M Teixeira (2011). ‘Dengue Surveillance Based on a Computational Model of Spatio-Temporal Locality of Twitter’. In: *Proceedings of the 3rd International Web Science Conference*. Koblenz: ACM Press. DOI: 10.1145/2527031.2527049.

Goodchild, M (2007). ‘Citizens as Sensors: the World of Volunteered Geography’. *GeoJournal* 69 (4), pp. 211–221. DOI: 10.1007/s10708-007-9111-y.

Grinberg, N, M Naaman, B Shaw and G Lotan (2013). ‘Extracting Diurnal Patterns of Real World Activity from Social Media’. In: *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*. Ed. by N Ellison, B Hogan, P Resnick and I Soboroff. Cambridge, MA: AAAI Press.

Haklay, M, A Singleton and C Parker (2008). ‘Web Mapping 2.0: The Neogeography of the GeoWeb’. *Geography Compass* 2 (6), pp. 2011–2039. DOI: 10.1111/j.1749-8198.2008.00167.x.

Hasan, S, X Zhan and S Ukkusuri (2013). ‘Understanding Urban Human Activity and Mobility Patterns Using Large-Scale Location-Based Data from Online Social Media’. In: *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*. Ed. by Y Zheng. New York, NY: ACM Press. DOI: 10.1145/2505821.2505823.

Hawelka, B, I Sitko, E Beinat, S Sobolevsky, P Kazakopoulos and C Ratti (2014). ‘Geo-Located Twitter as Proxy for Global Mobility Patterns’. *Cartography and Geographic Information Science* 41 (3), pp. 260–271. DOI: 10.1080/15230406.2014.890072.

Hiruta, S, T Yonezawa, M Jurmu and H Tokuda (2012). ‘Detection, Classification and Visualization of Place-Triggered Geotagged Tweets’. In: *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. Pittsburgh, PA: ACM Press, pp. 956–963. DOI: 10.1145/2370216.2370427.

Hong, L, A Ahmed, S Gurumurthy, A Smola and K Tsioutsiouliklis (2012). ‘Discovering Geographical Topics in the Twitter Stream’. In: *Proceedings of the 21st International Conference on World Wide Web*. Ed. by M Rabinovich and S Staab. Lyon: ACM Press, pp. 769–778. DOI: 10.1145/2187836.2187940.

ITU (2014). *Measuring the Information Society*. Tech. rep. International Telecommunication Union (ITU).

Jatowt, A, E Antoine, Y Kawai and T Akiyama (2015). ‘Mapping Temporal Horizons’. In: *Proceedings of the 24th International Conference on World Wide Web*. Ed. by A Gangemi, S Leonardi and A Panconesi. New York, NY: ACM Press, pp. 484–494. DOI: 10.1145/2736277.2741632.

Jiang, B, J Yin and S Zhao (2009). ‘Characterizing the Human Mobility Pattern in a Large Street Network’. *Physical Review E* 80 (2), p. 021136. DOI: 10.1103/PhysRevE.80.021136.

Jin, L, X Long, K Zhang, Y Lin and J Joshi (2016). ‘Characterizing Users’ Check-In Activities Using their Scores in a Location-Based Social Network’. *Multimedia Systems* 22 (1), pp. 87–98. DOI: 10.1007/s00530-014-0395-8.

Kafsi, M, H Cramer, B Thomee and D Shamma (2015). ‘Describing and Understanding Neighborhood Characteristics through Online Social Media’. In: *Proceedings of the 24th International Conference on World Wide Web*. Ed. by A Gangemi, S Leonardi and A Panconesi. New York, NY: ACM Press, pp. 549–559. DOI: 10.1145/2736277.2741133.

Kelm, P, V Murdock, S Schmiedeke, S Schockaert, P Serdyukov and O van Laere (2013). ‘Georeferencing in Social Networks’. In: *Social Media Retrieval*. Ed. by N Ramzan, R van Zwol, J Lee, K Clüver and X Hua. London: Springer, pp. 115–141. DOI: 10.1007/978-1-4471-4555-4_6.

Kinsella, S, V Murdock and N O’Hare (2011). ‘"I’m eating a sandwich in Glasgow": Modelling Locations with Tweets’. In: *Proceedings of the 3rd International Workshop on Search and Mining User-Generated Contents*. Ed. by I Cantador, F Carrero, J Cortizo, P Rosso, M Schedl and J Troyano. Glasgow, UK: ACM Press, pp. 61–68. DOI: 10.1145/2065023.2065039.

Kling, F and A Pozdnoukhov (2012). ‘When a City Tells a Story’. In: *Proceedings of the 20th International Conference on Advances in Geographic Information Systems*. Ed. by P Kröger, E Tanin and P Widmayer. New York, NY: ACM Press, pp. 482–485. DOI: 10.1145/2424321.2424395.

Kosala, R and E Adi (2012). ‘Harvesting Real Time Traffic Information from Twitter’. *Procedia Engineering* 50, pp. 1–11. DOI: 10.1016/j.proeng.2012.10.001.

Krumm, J, R Caruana and S Counts (2013). ‘Learning Likely Locations’. In: *International Conference on User Modeling, Adaptation, and Personalization*. Ed. by S Carberry, S Weibelzahl, A Micarelli and G Semeraro. Rome: Springer, pp. 64–76. DOI: 10.1007/978-3-642-38844-6_6.

Lampos, V and N Cristianini (2010). ‘Tracking the Flu Pandemic by Monitoring the Social Web’. In: *2nd International Workshop on Cognitive Information Processing*. Elba: IEEE, pp. 411–416. DOI: 10.1109/CIP.2010.5604088.

Lamprianidis, G and D Pfoser (2011). ‘Jeocrowd: Collaborative Searching of User-Generated Point Datasets’. In: *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. Chicago, IL: ACM Press, pp. 509–512. DOI: 10.1145/2093973.2094061.

Lee, R and K Sumiya (2010). ‘Measuring Geographical Regularities of Crowd Behaviors for Twitter-Based Geo-Social Event Detection’. In: *Proceedings of the 2nd ACM SIGSPATIAL International*

Workshop on Location Based Social Networks. Ed. by W Peng and X Xing. New York, NY: ACM Press, pp. 1–10. DOI: 10.1145/1867699.1867701.

Lenormand, M, M Picornell, O Cantu-Ros, A Tugores, T Louail, R Herranz, M Barthelemy, E Frias-Martinez and J Ramasco (2014). 'Tweets on the Road'. *PLoS ONE* 9, e105407. DOI: 10.1371/journal.pone.0105184.

Li, Chao, Zhongying Zhao, Jun Luo, Ling Yin and Qiming Zhou (2014). 'A Spatial-Temporal Analysis of Users? Geographical Patterns in Social Media: A Case Study on Microblogs'. In: *Database Systems for Advanced Applications. DASFAA 2014.* Ed. by W Han, M Lee, A Muliantara, N Sanjaya, B Thalheim and S Zhou. Heidelberg: Springer, pp. 296–307. DOI: 10.1007/978-3-662-43984-5_22.

Li, L, M Goodchild and B Xu (2013). 'Spatial, Temporal, and Socioeconomic Patterns in the Use of Twitter and Flickr'. *Cartography and Geographic Information Science* 40 (2), pp. 61–77. DOI: 10.1080/15230406.2013.777139.

Li, W, P Serdyukov, A de Vries, C Eickhoff and M Larson (2011). 'The Where in the Tweet'. In: *Proceedings of the 20th ACM International conference on Information and Knowledge Management*. Ed. by B Berendt, A de Vries, W Fan, C Madconald, I Ounis and I Ruthven. Hyderabad: ACM Press, pp. 2473–2476. DOI: 10.1145/2063576.2063995.

Liang, X, X Zheng, W Lv, T Zhu and K Xu (2012). 'The Scaling of Human Mobility by Taxis is Exponential'. *Physica A: Statistical Mechanics and its Applications* 391 (5), pp. 2135–2144. DOI: 10.1016/j.physa.2011.11.035.

Longley, P and M Adnan (2016). 'Geo-temporal Twitter Demographics'. *International Journal of Geographical Information Science* 30 (2), pp. 369–389. DOI: 10.1080/13658816.2015.1089441.

Longueville, B de, A Annoni, S Schade, N Ostlaender and C Whitmore (2010). 'Digital Earth's Nervous System for Crisis Events: Real-Time Sensor Web Enablement of Volunteered Geographic Information'. *International Journal of Digital Earth* 3 (3), pp. 242–259. DOI: 10.1080/17538947.2010.484869.

Lovelace, R, M Birkin, P Cross and M Clarke (2016). 'From Big Noise to Big Data: Toward the Verification of Large Data Sets for Understanding Regional Retail Flows'. *Geographical Analysis* 48 (1), pp. 59–81. DOI: 10.1111/gean.12081.

Miller, H and M Goodchild (2015). 'Data-Driven Geography'. *GeoJournal* 80 (4), pp. 449–461. DOI: 10.1007/s10708-014-9602-6.

Moran, P (1950). 'Notes on Continuous Stochastic Phenomena'. *Biometrika* 37 (1/2), pp. 17–23. DOI: 10.2307/2332142.

Murthy, D and S Longwell (2013). 'Twitter and Disasters'. *Information, Communication & Society* 16 (6), pp. 837–855. DOI: 10.1080/1369118X.2012.696123.

Nelson, T (2012). 'Trends in Spatial Statistics'. *The Professional Geographer* 64 (1), pp. 83–94. DOI: 10.1080/00330124.2011.578540.

Noulas, A, S Scellato, C Mascolo and M Pontil (2011). 'Exploiting Semantic Annotations for Clustering Geographic Areas and Users in Location-based Social Networks'. In: *Fifth International AAAI Conference on Weblogs and Social Media*. Ed. by L Adamic, R Baeza-Yates and S Counts. Barcelona: AAAI Press.

Ord, J and A Getis (1995). 'Local Spatial Autocorrelation Statistics: Distributional Issues and an Application'. *Geographical Analysis* 27 (4), pp. 286–306. DOI: 10.1111/j.1538-4632.1995.tb00912.x.

Quercia, D, L Capra and J Crowcroft (2012). ‘The Social World of Twitter: Topics, Geography, and Emotions’. In: *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*. Dublin: AAAI Press.

Ripley, B (1976). ‘The Second-Order Analysis of Stationary Point Processes’. *Journal of Applied Probability* 13 (2), pp. 255–266. DOI: 10.2307/3212829.

Ritzer, G and N Jurgenson (2010). ‘Production, Consumption, Prosumption: The Nature of Capitalism in the Age of the Digital ’Prosumer’’. *Journal of Consumer Culture* 10 (1), pp. 13–36. DOI: 10.1177/1469540509354673.

Roick, O and S Heuser (2013). ‘Location Based Social Networks - Definition, Current State of the Art and Research Agenda’. *Transactions in GIS* 17 (5), pp. 763–784. DOI: 10.1111/tgis.12032.

Sakaki, T, M Okazaki and Y Matsuo (2010). ‘Earthquake Shakes Twitter Users : Real-time Event Detection by Social Sensors’. In: *Proceedings of the 19th International Conference on World Wide Web*. Ed. by J Freire and S Chakrabarti. Raleigh, NC: ACM Press, pp. 851–860. DOI: 10.1145/1772690.1772777.

Scellato, S, C Mascolo, M Musolesi and V Latora (2010). ‘Distance Matters: Geo-Social Metrics for Online Social Networks’. In: *Proceedings of the 3rd Wonference on Online Social Networks*. Boston, MA: USENIX Association, p. 8.

Sengstock, C and M Gertz (2012). ‘Latent Geographic Feature Extraction from Social Media’. In: *Proceedings of the 20th International Conference on Advances in Geographic Information Systems*. New York, NY: ACM Press, pp. 149–158. DOI: 10.1145/2424321.2424342.

Serdyukov, P, V Murdock and R van Zwol (2009). ‘Placing Flickr Photos on a Map’. In: *Proceedings of the 32nd International ACM SIGIR conference on Research and Development in Information Retrieval*. Ed. by M Sanderson, C Zhai and J Zobel. Boston, MA: ACM Press, pp. 484–491. DOI: 10.1145/1571941.1572025.

Shimatani, K (2002). ‘Point Processes for Fine-Scale Spatial Genetics and Molecular Ecology’. *Biometrical Journal* 44 (3), pp. 325–352. DOI: 10.1002/1521-4036(200204)44:3<325::AID-BIMJ325>3.0.CO;2-B.

Sofean, M and M Smith (2012). ‘A Real-Time Architecture for Detection of Diseases Using Social Networks’. In: *Proceedings of the 23rd ACM Conference on Hypertext and Social Media*. Milwaukee, WI: ACM Press, pp. 309–310. DOI: 10.1145/2309996.2310048.

Stefanidis, A, A Crooks and J Radzikowski (2013). ‘Harvesting Ambient Geospatial Information from Social Media Feeds’. *GeoJournal* 78 (2), pp. 319–338. DOI: 10.1007/s10708-011-9438-2.

Steiger, E, J de Albuquerque and A Zipf (2015a). ‘An Advanced Systematic Literature Review on Spatiotemporal Analyses of Twitter Data’. *Transactions in GIS* 19 (6), pp. 809–834. DOI: 10.1111/tgis.12132.

Steiger, E, T Ellersiek and A Zipf (2014a). ‘Explorative Public Transport Flow Analysis from Uncertain Social Media Data’. In: *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information*. Ed. by R de By and C Wenk. New York, NY: ACM Press. DOI: 10.1145/2676440.2676444.

Steiger, E, J Lauer, T Ellersiek and A Zipf (2014b). ‘Towards a Framework for Automatic Geographic Feature Extraction from Twitter’. In: *Proceedings of the Eighth International Conference on Geographic Information Science*. Vienna.

Steiger, E, B Resch and A Zipf (2016b). 'Exploration of Spatiotemporal and Semantic Clusters of Twitter Data Using Unsupervised Neural Networks'. *International Journal of Geographical Information Science* 30 (9), pp. 1694–1716. DOI: 10.1080/13658816.2015.1099658.

Steiger, E, R Westerholt, B Resch and A Zipf (2015b). 'Twitter as an Indicator for Whereabouts of People? Correlating Twitter with UK Census Data'. *Computers, Environment and Urban Systems* 54, pp. 255–265. DOI: 10.1016/j.compenvurbsys.2015.09.007.

Stoyan, D and H Stoyan (1994). *Fractals, Random Shapes, and Point Fields: Methods of Geometrical Statistics*. Chichester, UK: Wiley.

Tapia, A, K Bajpai, J Jansen, J Yen and L Giles (2011). 'Seeking the Trustworthy Tweet: Can Microblogged Data Fit the Information Needs of Disaster Response and Humanitarian Relief Organizations'. In: *Proceedings of the 8th International ISCRAM Conference*. Lisbon.

Thomson, R, N Ito, H Suda, F Lin, Y Liu, R Hayasaka, R Isochi and Z Wang (2012). 'Trusting Tweets: The Fukushima Disaster and Information Source Credibility on Twitter'. In: *Proceedings of the 9th International ISCRAM Conference*. Vancouver.

Van Laere, O, S Schockaert and B Dhoedt (2010). 'Towards Automated Georeferencing of Flickr Photos'. In: *Proceedings of the 6th Workshop on Geographic Information Retrieval*. Ed. by R Purves, P Clough and C Jones. Zurich: ACM Press. DOI: 10.1145/1722080.1722087.

Vieweg, S, A Hughes, K Starbird and L Palen (2010). 'Microblogging During Two Natural Hazards Events'. In: *Proceedings of the 28th International conference on Human Factors in Computing Systems*. Ed. by G Fitzpatrick, S Hudson, K Edwards and T Rodden. Atlanta, GA: ACM Press, pp. 1079–1088. DOI: 10.1145/1753326.1753486.

Wakamiya, S, R Lee and K Sumiya (2011). 'Crowd-Based Urban Characterization: Extracting Crowd Behavioral Patterns in Urban Areas from Twitter'. In: *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks*. Ed. by Y Zheng and M Mokbel. New York, NY: ACM Press, pp. 77–84. DOI: 10.1145/2063212.2063225.

— (2012). 'Crowd-sourced Urban Life Monitoring: Urban Area Characterization Based Crowd Behavioral Patterns from Twitter'. In: *Proceedings of the 6th International Conference on Ubiquitous Information Management and Communication*. Kuala Lumpur: ACM Press.

Wang, H, D Can, A Kazemzadeh, F Bar and S Narayanan (2012a). 'A System for Real-Time Twitter Sentiment Analysis of 2012 U.S. Presidential Election Cycle'. In: *Proceedings of the ACL 2012 System Demonstrations*. Ed. by M Zhang. Jeju Island: Association for Computational Linguistics, pp. 115–120.

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 (5), pp. 868–887. DOI: 10.1080/13658816.2014.1002499.

Wu, F, Z Li, W Lee, H Wang and Z Huang (2015). 'Semantic Annotation of Mobility Data using Social Media'. In: *Proceedings of the 24th International Conference on World Wide Web*. Ed. by A Gangemi, S Leonardi and A Panconesi. New York, NY: ACM Press, pp. 1253–1263. DOI: 10.1145/2736277.2741675.

Wu, S, J Hofman, W Mason and D Watts (2011). 'Who Says What to Whom on Twitter'. In: *Proceedings of the 20th International Conference on World Wide Web*. Ed. by E Bertino and R Kumar. Hyderabad: ACM Press, pp. 705–714. DOI: 10.1145/1963405.1963504.

Yardi, S and D Boyd (2010). 'Tweeting from the Town Square: Measuring Geographic Local Networks'. In: *Proceedings of the Fourth International Conference on Weblogs and Social Media*. Washington, DC: AAAI Press.

Zandbergen, P and S Barbeau (2011). 'Positional Accuracy of Assisted GPS Data from High-Sensitivity GPS-enabled Mobile Phones'. *Journal of Navigation* 64 (3), pp. 381–399. DOI: 10.1017/S0373463311000051.

Zook, M, M Graham, T Shelton and S Gorman (2010). 'Volunteered Geographic Information and Crowdsourcing Disaster Relief: A Case Study of the Haitian Earthquake'. *World Medical & Health Policy* 2 (2), pp. 6–32. DOI: 10.2202/1948-4682.1069.

II.6 The Impact of Different Statistical Parameter Values between Point Based Datasets when Assessing Spatial Relationships

Abstract

User-generated datasets like those extracted from geosocial media are challenging for spatial analysis. These kinds of data are collected through unmoderated modes of acquisition offering the users a great deal of freedom in terms of content and other features. Further, the data are influenced by idiosyncratic spatial concepts of the users. The resulting datasets are therefore heterogeneous and comprise different (often inseparable) statistical populations in a spatially and temporally superimposed form. As a consequence, traditional notions of stationarity, which are often required in spatial analysis, are frequently violated and drawn conclusions about disclosed spatial structures are then misleading. This paper examines the influence of different statistical parameter values on the exemplary case of Moran's I estimation, a popular measure of spatial autocorrelation. The pattern investigated consists of two partially, spatially overlapping sub-patterns each interacting on different geometric scales. Normal variates drawn from populations with different means and variances are repeatedly assigned to these sub-patterns and Moran's I is calculated for 20.000 overall configurations. The results show strong influences of discrepancies in statistical parameter values on the characterization of the evaluated spatial patterns. Scales and combinations of different orders of magnitude of mean and variance differences also play a role. The results indicate that the spatial analysis of geosocial media posts must take into account different superimposed statistical populations to ensure meaningful results.

Keywords: Spatial Analysis, Spatial Autocorrelation, Spatial Statistics, Stationarity, Geosocial Media

II.6.1 Introduction and Background

Spatial analysis techniques like hot-spot estimators, spatial autocorrelation measures and spatial regression models (Getis 2008) are applied to investigate the interaction behaviour within spatial random variables (Fischer and Getis 2010b). One important assumption when using these techniques is the notion of stationarity, describing different forms of homogeneity with varying degrees of intensity (Zimmermann and Stein 2010). Spatial autocorrelation techniques like Moran's *I* are based on second-order (or weak) stationarity (Cliff and Ord 1981; Aldstadt 2010) which imply constant means and variances. This assumption is important to assure the validity of auxiliary parameters and to simplify randomisation procedures for constructing null models. Many recent user-generated and ambient datasets like those extracted from Twitter infringe traditional stationarity conditions. These kinds of data are obtained from unmoderated acquisition schemes that allow users to choose freely the locations, time stamps and contents of their posts. This leads to a noisy dataset featuring few observations about many simultaneous phenomena (Lovelace et al. 2016). Further ambiguity is added by the idiosyncratic spatial perceptions of the users (Wender et al. 2002) and by demographic characteristics like age or gender (Weiss et al. 2003;

Sugovic and Witt 2013). The resulting non-identical random variables are thus spatially and temporally mixed, because not all of these complex differences can be sorted out a priori. Using these data in the vein of the humans-as-sensors concept (Goodchild 2007) thus requires a treatment of their inherent heterogeneity, affecting stationarity assumptions.

This paper examines the influence of varying statistical parameter values within co-located but non-identical random variables on the spatial autocorrelation measure Moran's *I*. Related work has been carried out recently by (Westerholt et al. 2015; Westerholt et al. 2016), who investigated superimposed scale characteristics and the effect of inappropriately positioned but highly cross-linked observations on spatial analysis results. By analogy, it was shown in earlier works that Moran's *I* requires a minimum degree of variability within the analysed attributes (Walter 1992b), whereas variability in the connectivity degrees of the random variables is a major nuisance affecting the validity of analysis results (Tiefelsdorf and Boots 1997; Tiefelsdorf et al. 1999). It was further found that unstable variance ("heteroscedasticity") leads to problematic randomizations and thus to wrong inferences (Oden 1995; Waldhör 1996; Assuncao and Reis 1999). Griffith (2010) recently investigated effects of attribute value deviations from normality, which is a prerequisite for a sufficiently fast convergence of Moran's *I* to a normal distribution. He conjectured that deviations are unproblematic as long as the distribution of the data resembles a bell curve, or is at least symmetric in shape. Most outlined results have been achieved under the premise of spatially disjoint random variables. This paper supplements these findings with the case of varying means and variances under the assumption of spatially superimposed random variables.

The presented work analyses a range of possible mean-variance combinations resembling different kinds of overlapping but eventually indistinguishable phenomena. One-thousand synthetic points are generated mimicking two processes, each of which is operating at a specific interaction scale. These are populated with normal attributes featuring specific mean-variance combinations between the two sub-patterns. Two populations are thus involved in each studied case, one for the larger-scale, and another for the smaller-scale one of the overlapping processes. In addition, these cases are studied under the premises that (i) both involved sub-patterns are themselves spatially uncorrelated or (ii) that both patterns are spatially structured. Indications are given for systematic behaviours in these combinations. Further, influences of the differing means and variances on the magnitude and range of Moran's *I* are revealed. The achieved insights facilitate a better understanding of spatial analysis results obtained from geosocial media and related data.

II.6.2 Methods

II.6.2.1 Pattern Construction

Synthetic data is used to have full control over parameters and to achieve interpretable results. The geometric setup of two overlapping point patterns is generated by placing an initial random point first. Additional 500 points are added iteratively and conditional on the respective preceding point by drawing random directions and distances from uniform distributions. A second pattern that was created in the same way is then moved so that it overlaps about 25 % of the first pattern. The continuous uniform distributions used for drawing directions and distances on two interaction scales are given by $\mathcal{U}(0, 360)$, and $\mathcal{U}(40, 50)$ ("small-scale") or $\mathcal{U}(70, 80)$ ("large-scale").

The generated synthetic point locations are assigned normal attribute values which are randomly assigned for spatially uncorrelated cases (Figure II.6.1a). In contrast, the values are allocated to the points

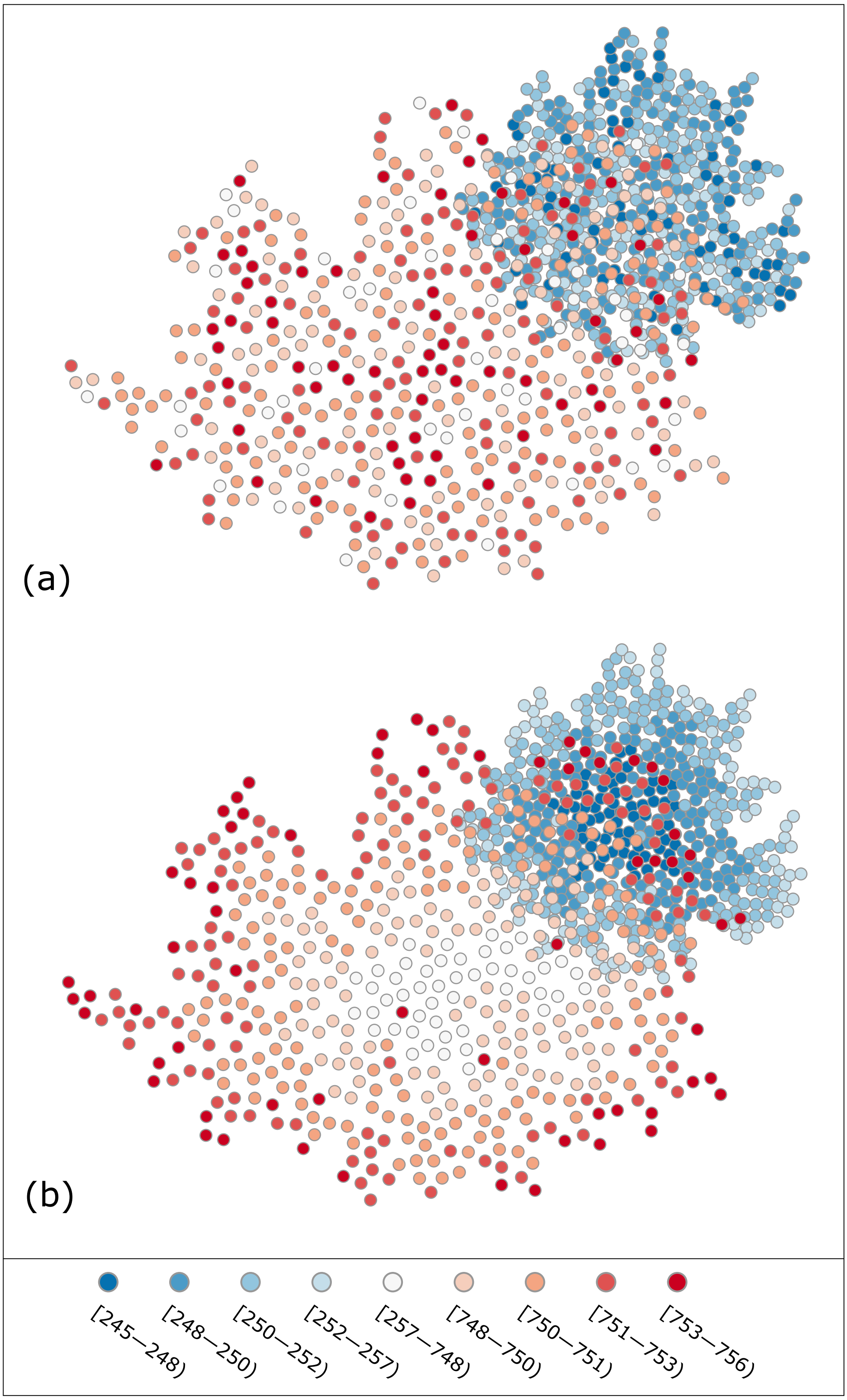


Figure II.6.1: Illustration of the investigated overlapping patterns for $\mu_1 = 250, \mu_2 = 750, \sigma_1 = \sigma_2 = 1$. (a) Spatially random patterns, (b) spatially autocorrelated patterns.

in a radial manner when patterns are spatially structured (Figure II.6.1b). In the interior there are lower values, which increase towards the edges of the respective sub-pattern. The outline of the actual means and standard deviations used is found in Section II.6.3.

II.6.2.2 Moran's *I*

The estimator studied, Moran's *I*, is a measure of spatial autocorrelation. It measures the degree of correspondence between structures in geographic space and those found in an attribute. It reads as (Cliff and Ord 1981; Getis 2010)

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \cdot \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{II.6.1}$$

where $x_1, \ldots, x_n$ represent n attribute values with mean $\bar{x}$ indexed over spatial units s_i. The w_ij denote pairwise positive spatial weights. Moran's *I* is the most frequently used estimator of spatial autocorrelation. It is typically preferred over alternative measures like Geary's *c* for its superior power characteristics and because it is less prone to statistical and configurational outliers (Chun and Griffith 2013). The applied spatial weights have a distance cut-off at 80 distance units (the upper bound of the large-scale interaction) and follow an inverse distance weighting scheme given by

$$w_{ij} = \begin{cases} |s_i - s_j|^{-2} & |s_i - s_j| \leq 80, \\ 0 & \text{otherwise.} \end{cases} \tag{II.6.2}$$

This scheme is chosen for resembling the distance-based rules that are used for constructing the patterns (see Section II.6.2.1).

II.6.3 2.3 Heat Maps of *I* with Differing Configurations

Moran's *I* is estimated from 20.000 different random statistical configurations on the overlapping point pattern. Two heat maps are generated from these: one for the case of uncorrelated attributes (Figure II.6.1a) and another map for the spatially-structured sub-patterns (Figure II.6.1b). Each grid cell in these heat maps represents a specific statistical configuration. This makes it possible to examine the role of the relationship of different means and variances of a process to multiples of the same values on the other process. Each parameter value of one process is thereby adjusted to a multiple of the same parameter of the other process. The heat maps are centred, meaning that the mean and variance for both processes are the same (1:1) in the central grid cell. A ratio of 1:3 in the left x-direction then means that the mean value of the small-scale pattern is 3 times that of the large-scale pattern. This applied scheme is illustrated in Figure II.6.2.

II.6.4 Results

For all results obtained, the initial means and variances, multiples of which are taken, start at $\mu = 25$ and $\sigma^2 = 400$. Depending on which side the heat map is viewed, integer multiples of these values are adapted either for the small-scale (left and up) or for the large-scale pattern (right and down). The multiplication factor thereby corresponds to the number of shifted grid cells.

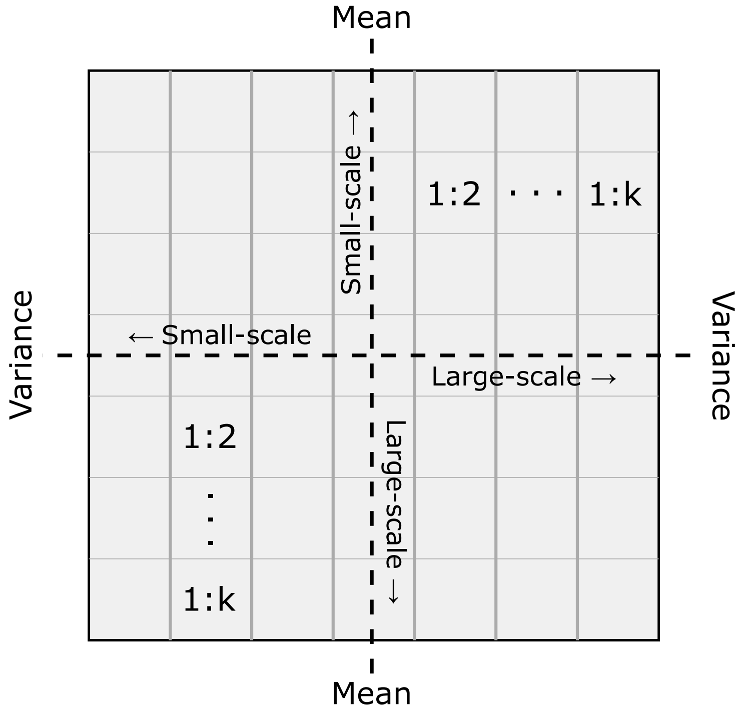


Figure II.6.2: Illustration of the applied heat maps. Variable k denotes the maximum number of multiples of the statistical parameters from the respective other investigated pattern.

II.6.4.1 Statistical Influences with Superimposed Spatial Random Patterns

The results for the case of spatially uncorrelated overlapping patterns are given in Figure II.6.3. The Moran's *I* values in the heat map in Figure II.6.3a appear noisy. This is caused by the randomness introduced by the lack of spatial structure in the two overlapping patterns.

The means involved must be almost identical to observe Moran's *I* values close to its expected value of $E[I] = -0.001$. This is supported by the box plots given in Figure II.6.3b showing that, as soon as one of the means is more than three times that of the other, the spatial pattern in the data appears too negatively autocorrelated. Further, high positive outliers that indicate clustering are only found on the same interval of nearly identical means. These outliers are caused by similar values from the different patterns, which are arbitrarily arranged next to each other by the spatial randomness in the attributes. However, this cannot happen when the means become too different, because all values are then too far away from the overall joint mean value, prohibiting positive autocorrelation.

Mean ratios determine the magnitude of Moran's *I*. When the means are very different the overall pattern tends to be underestimated. The degree of underestimation converges to an almost constant level after the ratio of the means exceeds a factor of 10. Beyond this mark, further differences in the means have only a minor impact on the magnitude of Moran's *I*. The box plots in Figure II.6.3b reveal this effect by the absence of a common trend line. The mean-induced effects are symmetric indicating that it does not matter whether the mean of the small-scale process exceeds the large-scale mean or vice versa.

The ratio of the attribute variances dominates the range of Moran's *I*. Figure II.6.3c shows that the variability in the estimated *I* values is small when the variances are roughly identical. In contrast, the dispersion of Moran's *I* increases when the variances of the two populations become more different. Moran's *I* then shows a wider range of values with more outliers, both positive and negative. These effects are again symmetric, which shows that the scale of the overlapping patterns is not crucially important for a characterisation of spatial autocorrelation when random attribute patterns overlap.

II.6.4.2 Effects of Spatially Autocorrelated Patterns

The heat map shown in Figure II.6.4 provides the Moran's *I* values for the case of spatially structured superimposed patterns. The spatial structuring causes a smoother transition of Moran's *I* over the grid cells of the heat map, meaning that the estimation of the statistic is more predictable with respect to statistical parameters than with spatially random attributes.

Differences in mean values determine the magnitude of Moran's *I*. In contrast to the symmetric behaviour observed with spatially random patterns, larger means in the small-scale process lead to higher Moran's *I* estimates than vice versa (Figure II.6.4b). The reason is that, because of the applied weighting scheme, more values above the global combined mean value are being related with a relatively high weight, in turn leading to higher *I* values. This demonstrates a strong interaction between the type of applied spatial weights and the involved superimposed geometric scales.

The rate at which differing means become effective is not symmetrical. While a relative increase in the mean of the smaller-scale process takes effect slowly, a sharper decrease in Moran's *I* is observed when the large-scale process becomes prominent. Clearly, there is a strong interaction between geometric and statistical parameters in the spatial analysis of spatially structured, partially overlapping patterns.

Differing variances play a minor role in comparison to the effects induced by mean differences. One notable observation is made in the case of dominant small-scale variances when the means are held almost identical at the same time. A large number of more pronounced positive autocorrelations is found on this interval, and that is caused by the generally larger number of points in the outer parts of the patterns. These feature higher attribute values than the interior parts. When the variance increases, the differences between interiors and outer parts become more pronounced, meaning that more and higher attribute values from one pattern interact with similar ones from the other. This effect vanishes once the small-scale means exceed those of the large-scale pattern by a factor of approximately 15. Further, when the radial attribute pattern is reversed, the same effect appears in reversed form (*i. e.*, the red grid cells in the heat map are then mirrored on the X-axis).

Another variance effect is that the range of Moran's *I* is smallest when the variances of the involved attributes are almost identical. The affected interval is narrow, and there is a sharp but symmetric increase in both magnitude and range of Moran's *I* as soon as either of the variances dominates.

II.6.5 Discussion and Conclusions

This paper examines the effects of different spatially superimposed statistical populations as those found in geosocial media data. The results are obtained on a synthetic spatial layout that mimics a partial geometric overlap of different phenomena. The following key insights are obtained:

1. Differing means determine the magnitude of Moran's *I*.
2. Differing variances determine the range of Moran's *I*.
3. Differences in the means and variances are only marginally related to their associated scales when the overlapping patterns are themselves spatially random.
4. When superimposed patterns are spatially structured, the scale of the pattern associated with the dominant mean value is stronger related with changes in the interpretation of Moran's *I*.

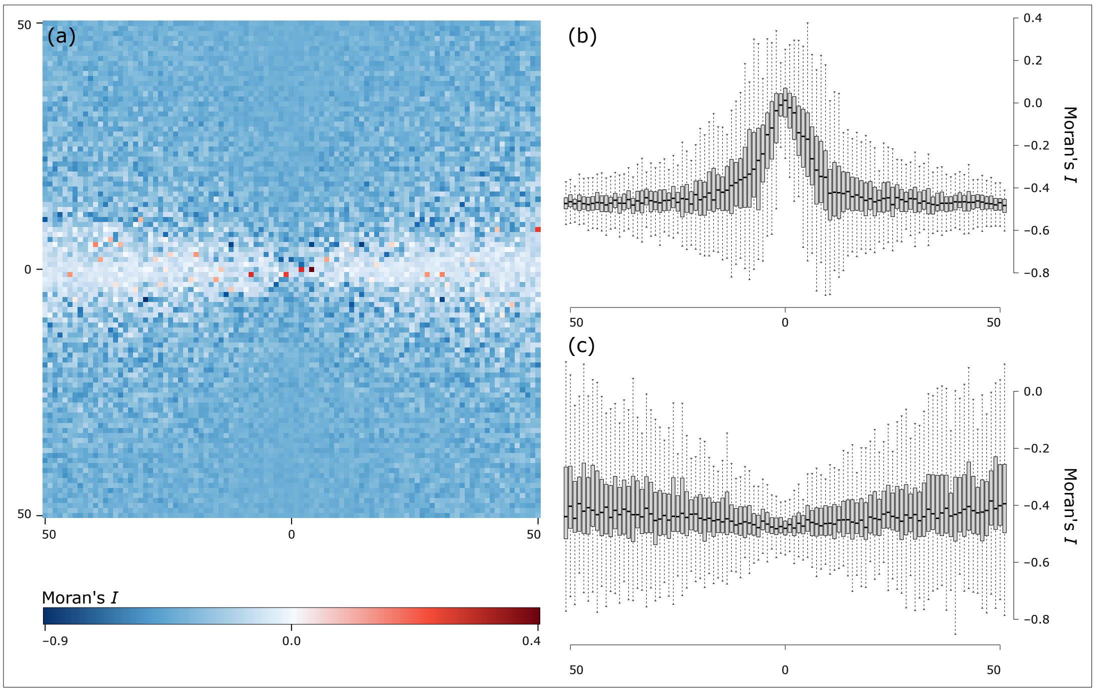


Figure II.6.3: Moran's *I* with superimposed spatially random patterns. (a) Heat map of Moran's *I* values with different mean-variance combinations in the attributes; (b) Box plots summarizing the influences of mean differences (*i. e.*, the rows); (c) Box plots summarizing the influences of differing variances (*i. e.*, the columns).

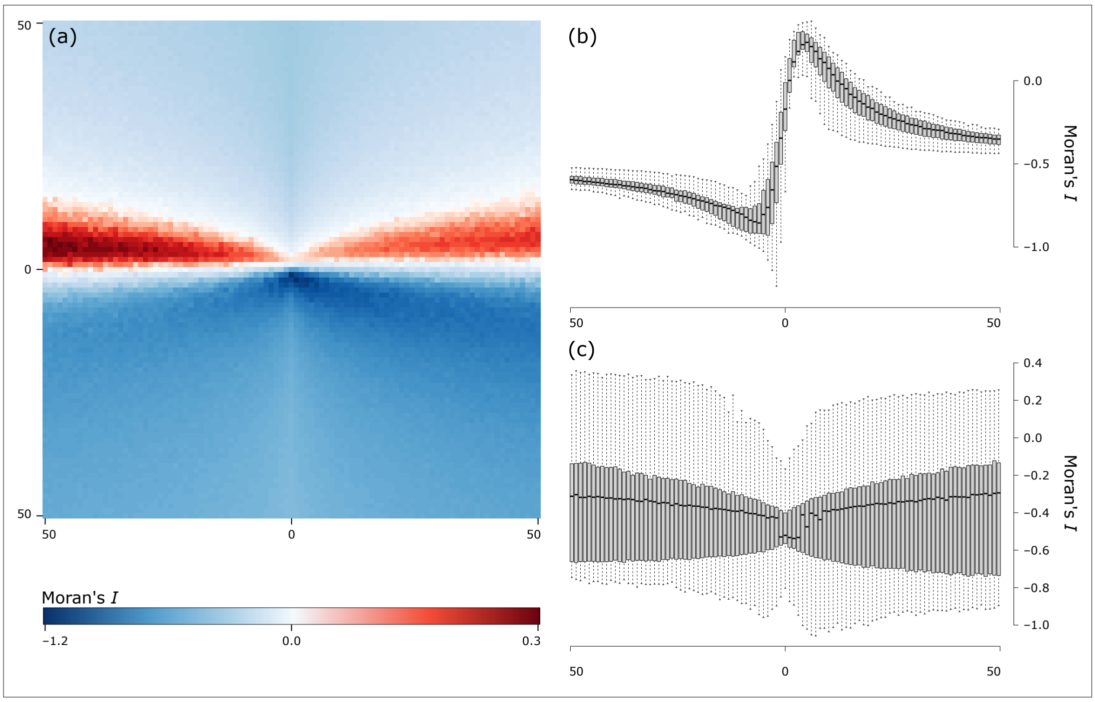


Figure II.6.4: Moran's *I* with superimposed spatially autocorrelated patterns. (a) Heat map of Moran's *I* values with different mean-variance combinations in the attributes; (b) Box plots summarizing the influences of mean differences (*i. e.*, the rows); (c) Box plots summarizing the influences of differing variances (*i. e.*, the columns).

Limitations exist in both the chosen layout as well as the applied spatial weighting scheme. Other geometric forms and interaction types exist, as well as further relevant weighting schemes that are not investigated in this paper. Further, the drawn variates are taken from normal distributions only. Count data or rates are beyond the scope of this paper and deserve treatment in future research. This is especially the case when the overlapping attributes form mixtures not non-symmetric random variables (*cf.* Griffith 2010).

Despite being spatial nature, the research carried out in this paper contributes to the recent efforts to develop a GIScience theory of platial analysis. The focus on spatial superposition is thereby interesting, because, other than in traditional GIS, places are spatially overlapping and co-located places must not be mutually related (Goodchild 2015). This work further supports efforts in other related disciplines facing similar technical issues. The event-sampling method (ESM) from psychology, which collects survey responses in situ, is one such example (Bluemke et al. 2017) for which the obtained results are useful with respect to the design of appropriate analytical approaches and to the interpretation of the collected survey responses.

Future research should consider other geometric setups combined with other types of attributes and dispersal mechanisms. Further, related measures like Geary's *c* or G_i^* might lead to slightly different results, as these combine statistical information in different ways. For instance, unlike Moran's *I*, Geary's *c* estimates covariance through calculating squared attribute differences, which could change the results obtained in this paper.

Acknowledgements

This work was supported by the German Research Foundation (DFG) through the priority programme 'Volunteered Geographic Information: Interpretation, Visualisation and Social Computing' (SPP 1894). I am grateful to Dr Adam Rousell for his comments on improving the title of the paper.

References (Chapter II.6)

Aldstadt, J (2010). 'Spatial Clustering'. In: *Handbook of Applied Spatial Analysis*. Ed. by M Fischer and A Getis. Heidelberg: Springer, pp. 279–300.

Assuncao, R and E Reis (1999). 'A New Proposal to Adjust Moran's I for Population Density'. *Statistics in Medicine* 18 (16), pp. 2147–2162. DOI: 10.1002/(SICI)1097-0258(19990830)18:16<2147::AID-SIM179>3.0.CO;2-I.

Bluemke, M, C Lechner, B Resch, R Westerholt and J Kolb (2017). 'Integrating Geographic Information into Survey Research: Current Applications, Challenges, and Future Avenues'. *Survey Research Methods* 11 (3), pp. 307–327. DOI: 10.18148/srm/2017.v11i3.6733.

Chun, Y and D Griffith (2013). *Spatial Statistics and Geostatistics*. London, UK: SAGE.

Cliff, A and J Ord (1981). *Spatial Processes: Models & Applications*. London, UK: Pion.

Fischer, M and A Getis (2010b). 'Introduction'. In: *Handbook of Applied Spatial Analysis*. Ed. by M Fischer and A Getis. Heidelberg: Springer, pp. 1–24. DOI: 10.1007/978-3-642-03647-7_1.

Getis, A (2008). 'A History of the Concept of Spatial Autocrrelation: A Geographer's Perspective'. *Geographical Analysis* 40 (3), pp. 297–309. DOI: 10.1111/j.1538-4632.2008.00727.x.

Goodchild, M (2007). 'Citizens as Sensors: the World of Volunteered Geography'. *GeoJournal* 69 (4), pp. 211–221. DOI: 10.1007/s10708-007-9111-y.

— (2015). 'Space , Place and Health'. *Annals of GIS* 21 (2), pp. 97–100. DOI: 10.1080/19475683.2015.1007895.

Griffith, D (2010). 'The Moran Coefficient for Non-Normal Data'. *Journal of Statistical Planning and Inference* 140 (11), pp. 2980–2990. DOI: 10.1016/j.jspi.2010.03.045.

Lovelace, R, M Birkin, P Cross and M Clarke (2016). 'From Big Noise to Big Data: Toward the Verification of Large Data Sets for Understanding Regional Retail Flows'. *Geographical Analysis* 48 (1), pp. 59–81. DOI: 10.1111/gean.12081.

Oden, N (1995). 'Adjusting Moran's I for Population Density'. *Statistics in Medicine* 14 (1), pp. 17–26. DOI: 10.1002/sim.4780140104.

Sugovic, M and J Witt (2013). 'An Older View on Distance Perception: Older Adults Perceive Walkable Extents as Farther'. *Experimental Brain Research* 226 (3), pp. 383–391. DOI: 10.1007/s00221-013-3447-y.

Tiefelsdorf, M and B Boots (1997). 'A Note on the Extremities of Local Moran's Iis and Their Impact on Global Moran's I'. *Geographical Analysis* 29 (3), pp. 248–257. DOI: 10.1111/j.1538-4632.1997.tb00960.x.

Tiefelsdorf, M, D Griffith and B Boots (1999). 'A Variance-Stabilizing Coding Scheme for Spatial Link Matrices'. *Environment and Planning A* 31 (1), pp. 165–180. DOI: 10.1068/a310165.

Waldhör, T (1996). 'The Spatial Autocorrelation Coefficient Moran's I Under Heteroscedasticity'. *Statistics in Medicine* 15 (7-9), pp. 887–892. DOI: 10.1002/(SICI)1097-0258(19960415)15:7/9<887::AID-SIM257>3.0.CO;2-E.

Walter, S (1992b). 'The Analysis of Regional Patterns in Health Data. II. II. The Power to Detect Environmental Effects'. *American Journal of Epidemiology* 136 (6), pp. 742–759.

Weiss, E, G Kemmler, E Deisenhammer, W Fleischhacker and M Delazer (2003). 'Sex Differences in Cognitive Functions'. *Personality and Individual Differences* 35 (4), pp. 863–875. DOI: 10.1016/S0191-8869(02)00288-X.

Wender, K, D Haun, B Rasch and M Blümke (2002). 'Context Effects in Memory for Routes'. In: *Spatial Cognition III*. Ed. by C Freksa, W Brauer, C Habel and K Wender. Tutzing: Springer, pp. 209–231. DOI: 10.1007/3-540-45004-1_13.

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 (5), pp. 868–887. DOI: 10.1080/13658816.2014.1002499.

Westerholt, R, E Steiger, B Resch and A Zipf (2016). 'Abundant Topological Outliers in Social Media Data and Their Effect on Spatial Analysis'. *PLOS ONE* 11 (9), e0162360. DOI: 10.1371/journal.pone.0162360.

Zimmermann, D and M Stein (2010). 'Classical Geostatistical Methods'. In: *Handbook of Spatial Statistics*. Ed. by A Gelfand, P Diggle, M Fuentes and P Guttorp. Boca Raton, FL: CRC Press, pp. 29–44.