TECHNISCHE UNIVERSITÄT WIEN
Vienna | Austria

DIPLOMARBEIT

Raum-zeitliche Hotspot Analyse für Bewegungsdaten

zur Erreichung des akademischen Grades

Diplomingeneurin

im Rahmen des Masterstudiums

Technische Mathematik

durch

Miriam Schwebler

Matrikelnummer: 01000421

Ausgeführt am Institut für Stochastik und Wirtschaftsmathematik, an der Fakultät für Mathematik und Geoinformation an der Technischen Universität Wien, in Kooperation mit dem Austrian Institute of Technology.

Betreuer:
Univ.Prof. Dipl.-Ing. Dr.techn. Peter Filzmoser, TU Wien
Dr. Anita Graser, Austrian Institute of Technology

Wien, Jänner 2023

Abstract

The cornerstone for cluster analysis was already laid in the 19^th century, when John Snow investigated the cause of a Cholera outbreak in Soho, London. The COVID crisis has shown, that today, more than 150 years later, the need for spatial cluster analysis and hotspot detection is as relevant as ever. Rooted in geography, hotspot analysis is often focused merely on the spatial aspect of the problem. But as the characteristics and intensity of a hotspot may vary over time, it is just as important to include the temporal aspect as well. Therefore, the need for spatio-temporal methods to investigate clustering problems arises. Since there have been first approaches to tackle the problem on the plane, the present thesis investigates these methods and their applicability to research problems with data associated to networks.

To investigate the existing methodological toolkit, various approaches to spatial, temporal and spatio-temporal statistics and methods are introduced that have primarily been designed for research problems on the plane. To see if they can be validly transferred to a network shaped problem, notions such as point patterns on networks, network spatial autocorrelation and distance concepts on networks are examined. The proposed approach to investigate spatio-temporal problems on networks is a two-stage process. First, the time series for every link is split into time slices of equally sized time intervals. Then, for each of those slices a spatial hotspot analysis is performed. This results in a time series for each link, characterizing every time slice as no point of interest or hotspot / coldspot with significance levels 1%, 5% or 10%. After that, for every link its corresponding time series is subjected to a pattern detection process that includes a trend test to identify diminishing, persistent and intensifying hotspots. In total, nine different patterns are defined that can be detected this way. For the convenience of reducing the workload of investigation, during this thesis a `Python` tool was developed, that handles the above-described process. Another benefit, that results from the established tool is, that it offers to possibility for visualisation of the results.

Finally, the proposed approach and the corresponding `Python` implementation were tested in form of a case study on taxi movement data in the city of Vienna to investigate a possible shift of hotspots caused by the launch of the new main train station Wien Hauptbahnhof. The data was provided by the data owner 'Taxi 31300' and the Austrian Institute of Technology. The results showed that the implemented approach indeed identified points of interested in terms of hotspot detection. On the other hand, the formulated hypothesis of a hotspot shift could only be confirmed to some extent. Unfortunately, a couple of flaws in the `Python` packages that were used for the implementation were discovered during the

evaluation process. These were reported to and acknowledged by the developers.
In conclusion, it became apparent during the analysis of the case study, that the spatio-temporal hotspot analysis tool can only assist and support the user by providing a framework for their analysis. There are still certain parameters that have to be decided on by the researcher and therefore are exposed to subjectivity. This may consequently lead to a distortion of the results. The implemented tool for spatio-temporal hotspot analysis represents a first prototype that aims at the analysis of a specific data source. Additional work is necessary to enhance this prototype to a more sophisticated and user-friendly implementation that can be made publicly available and be effectively used for spatio-temporal hotspot analysis on networks.

Kurzfassung

Der Grundstein zur Clusteranalyse wurde bereits im 19. Jahrhundert gelegt, als John Snow nach der Ursache eines Cholera Ausbruchs in Soho, London, suchte. Die COVID Krise hat gezeigt, dass die räumliche Clusteranalyse sowie die Aufspürung von Hotspots, heute so aktuell ist wie vor über 150 Jahren. Die Hotspot Analyse ist tief im Forschungsbereich der Geographie verwurzelt und untersucht daher oft lediglich den räumlichen Aspekt eines Forschungsproblems. Die Art und Intensität eines Hotspots kann sich jedoch im Lauf der Zeit verändern, daher ist es von maßgeblicher Bedeutung auch die zeitliche Achse des Problems mitzudenken. Darin begründet sich die Notwendigkeit zur Erforschung von raum-zeitlichen Analysemethoden. Erste Versuche das Problem anzugehen, werden in der vorliegenden Arbeit untersucht und auf deren Anwendbarkeit auf Forschungsprobleme untersucht, deren Datenpunkte auf einem Netzwerk und nicht in der Ebene liegen.

Um einen Überblick über das existierende Methodenset zu bekommen, werden zu Beginn räumliche, zeitliche und raum-zeitliche Statistiken und Methoden vorgestellt, welche für die Anwendung auf der Ebene definiert wurden. Um eine valide Aussage zu treffen, ob diese Methoden auch auf Netzwerke übertragen werden können, werden Konzepte wie Punktmustern auf Netzwerken, räumliche Autokorrelation auf Netzwerken und Abstandsdefinitionen auf Netzwerken betrachtet. Der vorgeschlagene Ansatz untersucht raum-zeitliche Fragestellungen auf Netzwerken in einem zweiphasigen Prozess. Zuerst wird die Zeitreihe einer jeden Kante in Zeitscheiben von gleichgroßen Zeitintervallen geteilt. Für jede dieser Zeitscheiben wird dann eine räumliche Analyse des Netzwerks vorgenommen. Dadurch ergibt sich für jede Kante eine Zeitreihe, welche für jeden Zeitschritt die Kennzeichnung bekommt, ob es sich um einen Punkt von Interesse handelt bzw. werden Hotspots und Coldspots von Signifikanzniveau 1%, 5% und 10% unterschiedlich bewertet. Die resultierende Zeitreihe wird einem Prozess zur Musterfindung unterzogen, dieser beinhaltet einen Trendtest, mit dessen Hilfe abnehmende, anhaltende und sich verstärkende Hotspots kategorisiert werden können. Insgesamt können so neun verschiedene Muster unterschieden werden. Um die Durchführung der obigen Analyse zu vereinfachen, wurde im Rahmen der Arbeit ein entsprechendes `Python` Tool entwickelt. Dieses bietet außerdem den Vorteil einer möglichen Visualisierung der Ergebnisse.

Der vorgeschlagene Ansatz, sowie dessen Implementierung wurde in Form einer Fallstudie zu Taxi-Bewegungsdaten in Wien getestet, um eine etwaige Verschiebung von Hotspots vom Wiener Westbahnhof zum 2015 neu eröffneten Hauptbahnhof zu überprüfen. Die Taxidaten wurden von ihrem Eigentümer ‚Taxi 31300' und dem Austrian Institute of Technology bereitgestellt. Die Ergebnisse der Fallstudie zeigten zum Einen, dass der

vorgeschlagene Ansatz tatsächlich plausible Punkte als Hotspots identifiziert. Andererseits konnte die aufgestellte Hypothese nur zum Teil bestätigt werden. Leider wurden im Zuge der Auswertungen auch einige Defizite in den verwendeten `Python` Bibliotheken aufgedeckt. Abschließend kann gesagt werden, dass durch die Analyse der Fallstudie erkannt wurde, dass das entwickelte `Python` Tool zur raum-zeitlichen Analyse von Hotspots den Forschenden zwar bei seinen Auswertungen unterstützen kann. Allerdings gibt es bei dem vorgeschlagenen Ansatz Parameter, welche subjektiv vom Anwender festgelegt werden müssen und daher zu einer Verzerrung der Ergebnisse führen können. Außerdem kann das implementierte Tool lediglich als ein erster Prototyp verstanden werden, welcher für die Analyse einer bestimmten Datenquelle entwickelt wurde. Um diesen Prototypen weiter zu verbessern und öffentlich für die Anwendung zur raum-zeitlichen Hotspot Analyse auf Netzwerken freizugeben, bedarf es noch zusätzlicher Arbeit und Weiterentwicklung.

Danksagung

An dieser Stelle möchte ich mich bei all denjenigen bedanken, die mich während der Anfertigung dieser Diplomarbeit unterstützt und motiviert haben.

Besonderer Dank gilt Herr Univ.Prof. Dipl.-Ing. Dr.techn. Peter Filzmoser, der meine Diplomarbeit betreut hat und mir unermüdlich mit hilfreichen Anregungen und konstruktiver Kritik zur Seite stand.

Außerdem bedanke ich mich bei meiner Betreuerin am Austrian Institute of Technology, Dr. Anita Graser, welche mich während meines Praktikums geleitet und beraten hat. Insbesondere danke ich meinem Kollegen Hannes Koller, der mir durch seinen wertvollen Input zu Beginn der Implementierung, Tür und Tor geöffnet hat um mit Freude an der Programmierung zu arbeiten und im Endeffekt dazu verholfen hat schöneren Code zu schreiben.

Zudem möchte ich mich bei meiner Familie und meinen Freunden bedanken. In erster Linie Bernadette Fina, die mich nach langer Flaute dazu motiviert hat diese Diplomarbeit endlich in einem Zug fertig zu schreiben. Außerdem Sarah Fanta, mit der ich während des Studiums für die meisten Prüfungen gemeinsam gelernt habe und ohne die ich vermutlich nicht bis zu diesem Punkt gekommen wäre. Abschließend bedanke ich mich auch bei meinen Eltern, Erwin und Andrea, meinen Geschwistern Bernadette, Katharina, Simon und Matthias, und meinem Partner Franz, die mein Leben einfach generell zu einem Besseren machen und ohne die ich nicht der Mensch wäre, der ich heute bin.

Dankeschön!

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die vorliegende Diplomarbeit selbstständig und ohne fremde Hilfe verfasst, andere als die angegebenen Quellen und Hilfsmittel nicht benutzt bzw. die wörtlich oder sinngemäß entnommenen Stellen als solche kenntlich gemacht habe.

Wien, Jänner 2023

Miriam Schwebler

∞

Contents

Chapter 1

Problem Analysis

The focus of traditional hotspot analysis is on identifying spatial clusters of a given phenomenon. There is a wide range of its utilisation in various research fields. Hotspot analysis finds its application for example in detecting hotspots in crime activity (Chainey and Ratcliffe, 2005), disease outbreaks (Getis et al., 2003, Hinman, Blackburn, and Curtis, 2006) or tumor detection (Khan Niazi et al., 2014, McIntire et al., 2019). With that, it is covering research fields such as geography, biology, medical science and criminology, to name just a few. Still, the common trait among the above mentioned applications is that a spatial clustering is to be detected.

In the present thesis hotspot detection in a geographical context is engaged. In specific, hotspots in movement data on networks and corresponding spatio-temporal patterns will be examined. Thanks to the development of GIS, the geographic information system, problems of data scarcity mostly belong to the past. Further advances in positioning technology led to large amounts of data collected every day with increasing accuracy. The downside of this quick advance is, that it resulted in a lack of time to develop appropriate methods to deal with and analyse this vast amount of available data (Dodge et al., 2016). While hotspot detection is usually seen in spatial context, the present thesis intends to make use of the huge quantity of available data and incorporate not only the spatial but also the temporal aspect into the hotspot and pattern finding process. In hotspot analysis approaches in geographical context, data of the phenomena to be examined are traditionally aggregated over pre-defined polygons representing countries, districts or roadblocks. But aggregating data that originally corresponds to a network structure into polygons and using spatial methods for hotspot detection that were originally designed for analysing events on a plane, can easily lead to false conclusions (Okabe and Sugihara, 2012). Therefore, the present thesis investigates an effective approach for spatio-temporal

hotspot analysis on networks. Furthermore, the implementation of a corresponding tool using well-established open-source software, is proposed.

After presenting the motivation for the conduct of hotspot analysis, this chapter submits the current state of the art. Thereafter, the research gap is outlined and research questions are defined. Then, the scope and further structure of the present thesis is discussed.

1.1 Motivation

In 1854, the English physician and epidemiologist John Snow investigated a cholera outbreak in Soho, London. He marked the residence of each casualty on a map of the corresponding area, as well as plotting every water source. With this approach, he was able to trace the outbreak back to its source, namely by identifying the water source within the cluster that was responsible for the outbreak. He removed the handle of the water pump and with that contributed extensively to the curtailment of the disease. The approach of using the tools of cartography not only for visualisation of events such as disease casualties but also to conduct a cluster analysis of these geographically dependent phenomena was unique for that time. By this, Snow provided a cornerstone in modern spatial analysis and the uses of geographic methodology in epidemiology (Stamp, 1964).

Today, more than 150 years later, the COVID-19 pandemic shows that the need for spatial cluster analysis and hotspot detection is as relevant as ever. With the initial outbreak in Wuhan, China, in November 2019, incidents soon spread over the whole world causing the WHO, the World Health Organization, to categorise the outbreak as a pandemic. Since the COVID-19 symptoms can range from none to life-threatening, countermeasures to curtail the spread of the disease had to be taken all over the world. One of the most important things when taking measures, is to convey the reasons that led to certain decisions properly and verify them with the help of tools that make them intelligible to the vast majority of the population. The usage of concise figures, maps and other illustrations make a significant contribution to the clear communication of the severity of the situation. Using spatial hotspot analysis on a more local level, say on district or community level, would give on the one hand a lot of insight on whether a region with high number of cases is simply that, a region with a high number of cases, or if the numbers are statistically significant. On the other hand, the correlation of newly recorded cases in neighbouring regions would be illuminated. Adding the temporal component would give a statistically verified insight in whether a location is for example a new hotspot with lower incidence rates in the past or whether it is an intensifying/diminishing hotspot, i.e. a hotspot with increasing/decreasing incidence rates. So, doing a spatio-temporal

hotspot analysis would lead to a much more profound conclusion than simply comparing the number of emerging cases of the disease by hand. Therefore, it could be used as a tool for further decision making, e.g. starting or ending a lockdown period.

So, while the tool that will be developed in the course of the present thesis includes the possibility do do spatio-temporal analysis on networks using suitable methods, it will also be applicable to data that needs the traditional approach of aggregating data points in neighbouring polygons. In the second case, the tool will provide the possibility of bringing only the added element of including the temporal aspect in its hotspot detection process.

1.2 State-of-the-Art

Hotspot analysis can be associated with many fields of study such as geography, biology, epidemiology and many more. It is rooted in spatial analysis and aims at the identification of statistically significant clusters of events scattered over the research area. The key element in spatial hotspot analysis is that of spatial autocorrelation. It examines whether the observation of some quality in one area is related to the occurrence of the same quality its neighbouring regions. The study of spatial autocorrelation dates back to the 1950's when methods such as the Moran's index (Moran, 1950) or Geary's c (Geary, 1954) were introduced. Both of these approaches offered a global measure of spatial autocorrelation that gave insight on whether the examined data is autocorrelated as a whole. It wasn't until the publication of Cliff and Ord, 1973, that the subject gained in popularity. They published a comprehensive work on spatial autocorrelation and introduced the idea of the detection of spatial patterns. They further engaged in this topic in their dissemination 'Spatial Processes: Models & Applications' (Cliff and Ord, 1981).

With the growth of data availability not only the need for appropriate theory for analysing existing data but also suitable tools for illustrating results increased. At first, a good deal of techniques that followed the ideas of Tukey, 1977, on *exploratory data analysis* (EDA) were introduced. For example box plots, Chernoff faces, Tukey star diagrams or scatterplot matrices were often used in spatial analysis. The downside of these newly discovered proceedings was that they ignored the characteristics of spatial data such as spatial autocorrelation. So, there was a growing call for an appropriate set of spatial statistical methods (Cliff and Ord, 1973, Cliff and Ord, 1981, Cressie, 1993).

That's when Getis and Ord first introduced a statistic that handled spatial autocorrelation on a local scale, the Getis-Ord G_i statistic, and therefore offered a variety of new possibilities for application (Getis and Ord, 1992). With the publication of their second paper on the topic (Ord and Getis, 1995), the idea of a localized version of spatial

autocorrelation methods was also addressed by Anselin, 1995, who introduced the notion of a 'local indicator of spatial association - LISA'. In his paper, Anselin offered formulas of a localized version of Moran's index and Geary's c. Taking the Morans's index as a starting point, Anselin, 1996, developed a simple tool for visualisation to easily assess the degree of spatial association: the Moran scatter plot. With that a cornerstone for the development of *exploratory spatial data analysis* (ESDA) was laid.

The 21^{st} century brought many technological advances causing unprecedented amounts of data being collected, for example through cheap geospatial technology embedded in mobile phones, as well as the affordable computing power to process it. However, the methodological tools employed by researchers often remain the same as in times when available data was still scarce and with low temporal resolution (Arribas-Bel and Tranos, 2018). One such issue would be the incorporation of time into the toolkit of spatial data analysis. In their paper 'Space-Time Analysis: Concepts, Quantitative Methods, and Future Directions' An et al., 2015, deal with the fact that in geographical context the concepts of space and time tend to be treated as separate fields and not in a joint way (Prasannakumar et al., 2011, Qi, Yang, and Jin, 2013, Gebru et al., 2019).

In recent years, there have been various attempts on modifying the spatial weight matrix, that is needed for calculation of the Moran's index and the Getis-Ord G_i statistic, on the behalf of integrating the temporal component. The purpose of this modification is that the newly designed spatio-temporal weight matrix includes not only the spatial but also the temporal aspects of neighbourhood. One of the more popular approaches is to define spatio-temporal distances that will form the spatio-temporal matrix needed for further calculation (Huang, Wu, and Barry, 2010, Yu, 2014, Tang, Tseng, and Chan, 2019). Lee and Li, 2017, noticed that this approach would lead to scaling effects when locations and time where defined differently. They suggested another approach for modifying the spatial weight matrix. They proposed to define a spatial weight matrix as well as a temporal weight matrix and combine them via multiplication. So, the Hadamard product of both matrices will result in a spatio-temporal weight matrix. In a later statement, Lee and Li, 2018, emphasize the fact that space and time differ significantly in their structure and the problem of space-time interaction has to be addressed very carefully. Based on Lee and Li's definition of a spatio-temporal Moran's index, Arribas-Bel and Tranos, 2018, defined a novel approach that they called 'Space-Time Calendar'. After applying the spatio-temporal Moran's index to the study area, one can pick one region that is of particular interest and inspect it more closely with help of the space-time calendar. It offers a way to visually assess slow and fast dynamics of the given area within one map. That way patterns could be revealed that would otherwise stay undetected.

Other approaches include Kulldorff's Scan statistic (Kulldorff, 1997) and its extension to a space-time scan statistic, STSS (Kulldorff et al., 2005, Kulldorff, Huang, and Konty, 2009). The spatial scan statistic is a window statistic, that is designed to detect clusters in multi-dimensional point processes within shapes like circles or ellipses. For the integration of time, the detection window is extended to cylinders. It was originally designed for the research field of epidemiology and was mainly used in this context (Rao, Shi, and Zhang, 2017, Kiani et al., 2021), but it has also found its way into hotspot detection on various topics such as forest fires (Hudjimartsu, Djatna, Ambarwari, et al., 2017), event detection within a Twitter dataset (Cheng and Wicks, 2014), Kang, 2010, used the space-time scan statistic to detect agglomeration processes in economies and Cheng and Adepeju, 2014, included the time component into the modifiable area unit problem, MAUP, by using the space-time scan statistic.

Another way to work around the MAUP problem is to use kernel density estimation, KDE. Originally, the kernel density estimation was a tool for estimating a smooth empirical probability function (Silverman, 1986). Nowadays, it is a commonly employed spatial analysis technique that is used to create heat maps to visually assess the results. By regarding the relation between space and time as orthogonal, Brunsdon, Corcoran, and Higgs, 2005, expanded the notion of the kernel density estimation to a space-time kernel density estimation. In combination with point patterns, the kernel density estimation can be applied for hotspot detection. A series of estimations is made over a grid that is placed on the entire point pattern and detects high and low density areas. The disadvantage is, that a bandwidth has to be set by the researcher. So, it is highly subjective but at the same time will influence whether a hotspot is detected or categorised as arbitrary. That's why hotspots detected by the kernel density function are not statistically significant, therefore, the KDE has to be used in combination with other hotspot detection techniques such as the Getis-Ord G_i statistic (Kalinic and Krisp, 2018). Other researchers investigating the usage of kernel density estimation in hotspot detection are Nakaya and Yano, 2010, who investigated crime clusters and compared the results of the space-time kernel density estimation and scan statistics. Hart and Zandbergen, 2014, examined the influence of interpolation methods, grid cell size and bandwith on the kernel density estimation and the corresponding hotspot mapping. Hu et al., 2018, suggested a different method to extend the kernel density estimation to a space-time version than Brunsdon, Corcoran, and Higgs, 2005. Milic et al., 2019 examined the influence of data classification methods on the predictive accuracy of kernel density estimation hotspot maps.

Pre-existing Tools

After having seen the current state-of-the-art methodology to deal with spatial and spatio-temporal hotspot analysis, a view to existing open source implementation of such is needed. Both programming languages `R` and `Python` offer packages for dealing with spatial hotspot analysis. In case of `R`, a collection of methods for basic spatial point pattern analysis and data exploration exists. It includes tools that can aggregate point data in areas, x-y charts and histograms, heat maps created by kernel density estimation or the Getis-Ord G_i^* statistic for hot- and coldspot detection (Minn, 2000–2021). All of these methods and approaches to point pattern analysis have in common that they, for once, are only considering the spatial component of the data not including the temporal aspect. Furthermore, all of them aggregate points in polygon shaped areas which can lead to misinterpretation if working with data that is collected on a network (Okabe and Sugihara, 2012). Actually, `R` also contains the package `DRHotNet` which stands for 'Differential Risk Hotspots in a Linear Network', that was released in July last year. It tackles the problem of aggregation of events into areas, but its main focus is onto finding differential risk hotspots, i.e. it offers procedures to detect hotspots of events that are closely linked to other factors (Briz-Redón, 2021). For example factors such as age, obesity, blood pressure, etc. being linked to the differential risk of having a stroke.

`Python` has its own assembly of tools in various packages. In the present context, the most important one would be the community driven library `PySAL`, the Python spatial analysis library, and in particular the packages `ESDA`, which stands for 'exploratory spatial data analysis' and `spaghetti`, a composition of '**Spa**tial **G**raph**h**s: N**et**works, **T**opology, & **I**nference'. `ESDA` deals with subjects such as global and local measures of spatial autocorrelation, including the implementation of Moran's I, Geary's c and Getis-Ord statistics. Furthermore, it provides joint count statistics for binary attributes. For all these measures it offers analytical and permutation based approaches. `spaghetti` on the other hand, addresses the problem of spatial analysis on networks. First of all it allows to read in line data and generate a network/graph representation from it. It is possible to extract contiguity weights and identify connected components. It is possible to snap point pattern observations onto this network, calculate an all neighbour distance matrix, calculate the Moran's I with the help of observation counts on network segments as well as network spatial weights. Additionally, point patterns can be simulated for cluster analysis via the K function attribute. Apart from that, it is possible to do graph theoretic analysis such as calculating the shortest paths, find the minimum/maximum spanning trees, etc (Rey et al., 2015, Gaboardi et al., 2018, Gaboardi, Rey, and Lumnitz, 2021, Rey et al., 2021).

1.3 Research Gap

Dealing with hotspot analysis, the majority of studies still focuses merely on the spatial aspect of the problem. Existing methods for spatio-temporal analysis and their suitability to hotspot detection have to be examined more closely. Various approaches that include the temporal component into former solely spatial methods, have been proposed. These range from the integration of time as an additional dimension to forming the product of spatially and temporally defined distances. Furthermore, these methods have in common that they are supposed to be applied to surfaces. The applicability of existing methods to data collected on a network still has to be explored.

So, the approach proposed in the present thesis has to acknowledge the importance of both spatial and temporal aspects of the examined data. At the same time it has to be applicable to a network shaped research area. Furthermore, there is also a practical requirement, namely that an implementation of the approach using the well-established open-source programming language `Python` is possible.

As seen in the previous chapter (1.2), the state-of-the-art approaches meet some of the requirements but there is no technique that covers all of them. Due to these considerations, the objective of the present thesis is to develop a suitable approach for spatio-temporal hotspot analysis that is applicable to networks. To be more specific, the following research questions are to be addressed by this work:

(i). Which method has to be chosen for spatio-temporal hotspot analysis that fulfills the following requirements? It

- integrates the spatial and temporal aspects of the data,
- is applicable to data collected on a network,
- is based on methods implemented in well-established open-source programming language `Python`.

(ii). How does the proposed approach perform compared to state-of-the-art approaches from a conceptual perspective?

(iii). How can the proposed approach be realized and implemented as a `Python` tool for spatio-temporal hotspot detection? Such that the proposed tool

- enables the output as dataframe,
- produces plots for visual assessment of the results,
- is transferable from network based data to surface based data.

1.4 Scope and Structure of this Thesis

The present thesis evaluates existing approaches for spatial and spatio-temporal hotspot analysis and their applicability to network structures. With that it contributes to fill the gap in appropriate methods for spatio-temporal hotspot analysis on networks. Futhermore, a `Pyhton` tool will be developed that realises the developed approach and will then be published as an open-source `Python` package. The scope of this thesis is focused on data represented in the form of point patterns within their geographic surroundings presented in network shape.

The thesis is organised as follows: chapter 2 offers a conceptual view on spatio-temporal hotspot analysis. Previously used state-of-the-art methods are discussed in more detail to provide a sound basis for further discussion of the approach chosen for this thesis. Section 3 discusses analysis approaches for networks and with that offering the tools for development of a spatio-temporal hotspot analysis approach that can be applied to networks. In chapter 4 the pre-existing methodology is reviewed, compared and analysed concerning their suitability for spatio-temporal hotspot analysis as well as their applicability on network shaped data. Further, the implementation of a corresponding tool using `Python` is discussed. In section 5 the approach is applied to a case study using taxi movement data in Vienna to investigate the hotspot shift from Vienna's former central railway station Wien Westbahnhof to the restructured station Wien Hauptbahnhof in 2015. Last, chapter 6 concludes the present thesis by revisiting the defined research question and by giving an outlook to potential future work.

Chapter 2

A Conceptual View on Spatio-Temporal Hotspot Analysis

To perform spatio-temporal hotspot analysis, tools of spatial and temporal analysis need to be united. As mentioned before, in the literature concerning geographical analysis frequently only spatial analysis is performed, neglecting the aspect of time. There have been a couple of approaches on spatio-temporal hotspot analysis but often for the purpose of adding the temporal component, spatial analysis is carried out at two or more points in time and then the particular results are compared. See, for example, GeoDa's "Time Editor" or "Map Movie" (Anselin and Li, 2022). From a statistical point of view this approach isn't a valid solution in terms of a time series analysis. Therefore, in the following chapter not only former approaches to spatial and spatio-temporal hotspot analysis will be examined but also trend analysis methods that could be combined with spatial analysis tools for providing a valid two stage spatio-temporal analysis approach. So, the following chapter will consist of three major topics, namely spatial hotspot analysis, trend analysis and spatio-temporal hotspot analysis.

2.1 Spatial Hotspot Analysis

The concept of space is a central aspect in regional sciences. If the presence of some property in a specific region makes it more or less likely that the same feature also appears in the neighbouring regions we say that the phenomenon exhibits so-called *spatial autocorrelation*. This quality might be a point-like object like plants in a field or cases of a certain disease, but it is also possible to test for spatial autocorrelation among variate values collected at certain points, such as rainfall values at various meteorological stations

(Cliff and Ord, 1973). The analysis of spatial data and the necessary techniques to detect spatial autocorrelation have received a lot of attention in the literature. By now, research in the area of spatial dependency between observations has been conducted for more than forty years (Lee and Li, 2017). Starting with a rather theoretical approach to spatial science and because of the lack of effective ways to incorporate spatial aspects of data, empirical work on this subject was rather limited. The inclusion of spatial aspects and their applicability in data analysis was only achieved due to technological progress in the field of geographic information system - GIS (Anselin and Getis, 2010).

2.1.1 Spatial Weight Matrix

By measuring spatial autocorrelation one assumes that all events are related to some extent. Therefore, an approach is necessary that is able to quantify the relationship between observations. This is done by defining a *spatial weight matrix* which can be based on various criteria, most commonly distance or boundary conditions. The following definitions are based on a section in the Encyclopedia of GIS by Smith, 2008.

Definition. The *distance* d_{ij} between two events i and j is defined as the Euclidean distance, i.e.

$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

for (x_i, y_i) being the location of event i and (x_j, y_j) being the location of event j.

Definition. To define the *spatial weight matrix* W based on distance there are two possible ways:

1) By pre-defining a distance threshold d_θ. Then

$$w_{ij} = \begin{cases} 1 & d_{ij} \leq d_\theta \\ 0 & d_{ij} > d_\theta. \end{cases}$$

2) By using the *First Law of Geography:* "Everything is related to everything else, but near things are more related than distant things." (Tobler, 1970). Then

$$w_{ij} = \frac{1}{d_{ij}}.$$

Then the *spatial weights matrix* is defined as

$$W = \sum_i \sum_j w_{ij}.$$

Distances are usually easily computed and therefore pose the advantage of distance based spatial weights. On the other hand, it is also possible that shared boundaries between spatial units are important. Hence, the spatial weight matrix can also be defined based on boundaries.

Definition. Let $bnd(i)$ be the set of boundary points of unit i. To define the *spatial weight matrix* based on boundaries there are two most commonly used ways:

1) *Queen contiguity weights*: Here it is possible that the boundaries between unit i and unit j share only a single point. Then the queen contiguity weights can be calculated as follows

$$w_{ij} = \begin{cases} 1 & bnd(i) \cap bnd(j) \neq \emptyset \\ 0 & bnd(i) \cap bnd(j) = \emptyset. \end{cases}$$

2) *Rook contiguity weights*: Is a stronger version of the queen contiguity weights where the length of the shared boundary between i and j has to be positive and is denoted by l_{ij}. Then the rook contiguity weights are defined by

$$w_{ij} = \begin{cases} 1 & l_{ij} > 0 \\ 0 & l_{ij} = 0. \end{cases}$$

Then the *spatial weights matrix* is defined as

$$W = \sum_i \sum_j w_{ij}.$$

For further elaboration on the subject, see for example Cliff and Ord, 1981 or Cliff and Ord, 1973, where the importance of choosing the correct weights to avoid incorrect correlation is stressed. Getis, 2009, illustrates the three types of spatial weight matrices, namely from the theoretical, the topological and the empirical point of view.

In the following passages the most commonly used approaches to spatial autocorrelation analysis will be introduced.

2.1.2 Moran's Index

Moran's Index (Moran, 1950), often simply called Moran's I, for spatial autocorrelation has found its application in many research fields. It is mainly used in geography and geographic information science as a first step to detect spatial dependency in a set of events. Usually these geographic objects are represented as polygons or points. Moran's

index measures the degree to which these geographic events show patterns like clustering or dispersion. It represents the ratio of the covariance of attribute values of neighbouring geographic objects and the covariance between all geographic objects in the study area (Lee and Li, 2017).

Definition. Let n be the number of observations being analyzed, x_i be the attribute value of geographic object i, $\overline{x}$ be the mean of all x_i and let w_{ij} be the spatial weight between observation i and j. Then the *global Moran's I* can be calculated as

$$I = \frac{n \sum_i \sum_j w_{ij}(x_i - \overline{x})(x_j - \overline{x})}{\sum_i \sum_j w_{ij} \sum_i (x_i - \overline{x})^2}.$$

Originally, the Moran index was designed as a global measure of spatial association only. In 1995, Luc Anselin defined the so-called *local indicator of spatial association - LISA*, which was a new category for measures of spatial autocorrelation. To be categorised as a LISA, a measure of spatial association has to satisfy the following two requirements:

1) the LISA for each observation gives an indication of the extent of significant spatial clustering of similar values around that observation;

2) the sum of LISAs for all observations is proportional to a global indicator of spatial association.

That means on the one hand that a LISA is a measure of local spatial autocorrelation and on the other hand there has to be a global spatial autocorrelation coefficient that decomposes into its constituents of local indicators. Hence, it is possible to see how much individual locations constitute to the global autocorrelation. For the Moran's I to become a local indicator of spatial association, Anselin defined a localized version (Anselin, 1995).

Definition. Based on Anselin (1995), the *localized Moran's I* for the geographic object i can be defined as

$$I_i = \frac{z_i \sum_j w_{ij} z_j}{\sigma^2}$$

where $z_i = (x_i - \overline{x})$ and σ is the standard deviation.

With this definition, Moran's index can be classified as a LISA with the sum of the local Moran being proportional to the global Moran's I.

Anselin further elaborated on and extended the Moran's I, in 1996 he defined the Moran scatterplot that has the form of a bivariate scatterplot that shows the linear association between the original variable and its spatially lagged counterpart. It can be augmented with a linear regression which has the slope of Moran's I (Anselin, 1980, Anselin, 1996).

The Moran's I as well as its localized version found its application in many research areas. Sokal, Oden, and Thomson, 1998, tested the applicability of the Moran's I in spatially distributed biological data. Griffith, 2005, discussed the effective geographic sample size in presence of spatial autocorrelation. Chainey and Ratcliffe, 2005, worked with the GIS and crime data, while Braithwaite and Li, 2007, explored the identification of terrorism hotspots. Bucher et al., 2021, investigated ways to integrate reliability of observations into Moran's I estimates.

2.1.3 Geary's c Statistic

Geary's c statistic was originally suggested as a global measure of spatial autocorrelation. It was designed to determine if neighbouring observations of the same phenomenon are correlated (Geary, 1954). Basically, Geary's c shares the same form of an autocorrelation coefficient as Moran's I by having a measure of covariance among the x_i as a nominator and a measure of variance as the denominator. But by calculating the squared difference of the x_i it opens room for new applications and interpretations (Cliff and Ord, 1981).

Definition. Let n be the number of observations, x_i the attribute value at location i, $\overline{x}$ the mean of all x_i and $W = (w_{ij})$ the spatial weight matrix. Then the ratio of contiguity c, typically called *Geary's c*, is defined as

$$c = \frac{(n-1)\sum_i \sum_j w_{ij}(x_i - x_j)^2}{2\sum_i \sum_j w_{ij} \sum_i (x_i - \overline{x})^2}.$$

At first, Geary's c statistic experienced a lack of applications. This was mainly due to time-consuming computations that had to be done by hand. With the introduction with desk calculators and later on computers Geary's c gained popularity (Jeffers, 1973).

In 1995, Anselin introduced the local version of Geary's c and elaborated on it further in his paper published in 2019. Analogous to the global form local Geary focuses on the squared differences (Anselin, 1995, Anselin, 2019).

Definition. Let x_i be the attribute value for observation i and $W = (w_{ij})$ the spatial weight matrix. Based on Anselin (1995) the *local Geary* statistic for each observation i is defined as

$$c_i = \sum_j w_{ij}(x_i - x_j)^2.$$

In contrast to local Moran's I and due to the approach via squared differences there is no need to standardize x. The squared differences represent either similarity or dissimilarity. Small values imply similarity and therefore positive spatial autocorrelaten while large value

imply dissimlarity and hence negative spatial autocorrelation. As a result to the squared differences approach Geary's c statistic is able to detect patterns that go unnoticed by local Moran and vice versa (Anselin, 2019).

Furthermore, local Geary's c satisfies the two defining requirements of a LISA. Namely, being a local measure of spatial association and having a global statistic that is proportional to the sum of its local constituents.

Sokal, Oden, and Thomson, 1998, tested the applicability of Geary's c to spatially distributed biological data, Anselin, 2019, extended local Geary's c to a multivariate context, Magnussen and Fehrmann, 2019, and Magnussen et al., 2020, compared various variance estimators, among them Geary's c, under systematic sampling and evaluated their results in a simulation study.

2.1.4 Getis-Ord G_i Statistic

The Getis-Ord G_i statistics was first introduced in 1992 by A. Getis and J.K. Ord and further elaborated upon in their paper in 1995. At a time where geologists commonly used globally applied statistics like Moran's I and Geary's c, they introduced a new statistics for spatial autocorrelation on a local scale. With that Getis and Ord were among the first ones to introduce a tool that provides the possibility to detect local "pockets" of spatial dependence that wouldn't display when only using global statistics (Getis and Ord, 1992).

Typically when measuring spatial dependency, statistical tests detect the similarities between neighbouring regions. But it is possible that not only positive but also negative dependence is apparent in the data. Therefore, Getis and Ord revised and enhanced their G_i statistic such that it is able to detect both kinds of patterns. On the other hand, the G_i statistic is, in contrast to other local statistics like local Moran or local Geary statistics, not considering spatial outliers (Ord and Getis, 1995).

Getis and Ord defined two versions of the G_i statistic, one that considers only the location's neighbours in its calculation while the other does take the value of the specified location into account.

Definition. Let n be the number of regions and $i = 1, 2, \ldots, n$. Each region i is associated with its attribute value x_i. Then the G_i *statistic* for region i is defined as

$$G_i = \frac{\sum_j w_{ij}x_j - W_i\overline{x}(i)}{\sigma(i)\sqrt{\frac{(n-1)\sum_j w_{ij}^2 - W_i^2}{(n-2)}}}, \; j \neq i.$$

where $W_i = \sum_{j \neq i} w_{ij}$, $\overline{x}(i) = \frac{\sum_j x_j}{(n-1)}$ for $i \neq j$ and $\sigma^2(i) = \frac{\sum_j x_j^2}{(n-1)} - [\overline{x}(i)]^2$.

And the G_i^* *statistic* for region i is defined as

$$G_i^* = \frac{\sum_j w_{ij}x_j - W_i^*\overline{x}}{\sigma\sqrt{\frac{n\sum_j w_{ij}^2 - W_i^{*2}}{(n-1)}}}, \ \forall j.$$

where $W_i^* = W_i + w_{ii}$ and $w_{ii} \neq 0$, $\overline{x}$ and σ denote the usual sample mean and standard deviation.

The G_i and G_i^* statistic respectively measure the degree of spatial dependence that is determined by the attribute value at location i and by the attribute values in its neighbourhood (Hinman, Blackburn, and Curtis, 2006).

Furthermore, Getis and Ord defined a global version of the G_i statistic, the *general G statistic*. But since the sum of local G_i statistics is not proportionally related to its global counterpart it can not be categorised as a LISA (Anselin, 1995). The G_i statistic has found various applications, to name just a few, Sokal, Oden, and Thomson, 1998, tested its applicability in spatially distributed biological data, Getis et al., 2003, investigated the spatial patterns of the yellow fever mosquito, Chainey and Ratcliffe, 2005, applied the G_i statistic to crime data, Hinman, Blackburn, and Curtis, 2006, examined typhoid outbreaks and Braithwaite and Li, 2007, detected terrorism hotspots with help of the G_i statistic.

Even though all three of the presented local statistics test the null hypothesis of spatial randomness, each of them uses different criteria to detect deviations from the null. While local Geary's c calculates the squared difference, local Moran's I is a product-moment coefficient and the G_i and G_i^* statistics use a sum (Anselin, 2019). This provokes a different interpretation of the respective results. Moran's I_i measures the joint co-variation of neighbouring areas. In case of a clustering of values of the same sign that deviate significantly from the mean, the I_i is positive representing a positive spatial autocorrelation. In case that the value of the i-th area has the opposite sign of its neighbouring regions but still deviates strongly from the mean, the I_i is negative implying negative spatial autocorrelation. As Geary's c works with the squared differences between the values of the pivot and its neighbourhood, high values of c denote negative spatial autocorrelation. On the other hand, when the data values of an area and its neighbours are close to the mean Geary's c indicates positive spatial autocorrelation. At the same time this would be marked as weak or zero autocorrelation by Moran's I_i (Sokal, Oden, and Thomson, 1998). When working with the G_i or G_i^* statistic, a postitive value implies a spatial clustering of

high values while a negative value denotes a clustering of low values (Anselin, 1995). It is important to keep in mind that while all three local statistics may reject the hypothesis of spatial randomness, the interpretation differs between them. Therefore, the main purpose of these statistics remains in data exploration (Anselin, 2019).

2.2 Trend Analysis

The central object of interest when doing trend analysis is a time series. A time series is a sequence of observations that are monitored at equally distanced points in time. They occur very frequently even in our every day life, for example a daily series of the maximum degree in one's hometown or a weekly series of the number of road accidents. They are used in various fields such as economics, business, engineering, the natural sciences (especially geophysics and meteorology), and the social sciences (Box, Jenkins, and Reinsel, 2008). There are various stochastic and dynamic models that were developed for the analysis of time series. We will focus on trend analysis which is part of the forecasting field of research. Here, the goal is to model statistical dependence of time and previous values of a temporally ordered metric feature to be able to make predictions on future values (Eckstein, 2016). But trend analysis can not only be used for forecasting but also to detect patterns in the data, which is the goal of the present thesis.

2.2.1 Autocorrelation

Autocorrelation in the context of time series analysis measures, as the term itself already indicates, the relationship between lagged values of a time series (Hyndman and Athanasopoulos, 2018). It is an intrinsic feature of a time series suggesting that successive observations are dependent on each other and therefore are of considerable interest (Box, Jenkins, and Reinsel, 2008). For a process y_t the *autocorrelation coefficient* ρ of lag k can be defined as the k-th autocovariance divided by the variance

$$\rho_k = Corr(y_t, y_{t-k}) = \frac{Cov(y_t, y_{t-k})}{\sqrt{\mathbb{V}(y_t)}\sqrt{\mathbb{V}(y_{t-k})}}.$$

The autocorrelation coefficients can then be visualised by an *autocorrelation function* $\{\rho_k\}$ which is a useful device to assess the behavior of a stationary process. The autocorrelation function is dimensionless and independent of the scale of the measurement of the time series (Box, Jenkins, and Reinsel, 2008).

If a trend is present in the data the autocorrelation coefficient will be large and positive

for small lags in time and will decrease with the increase of the lag (Hyndman and Athanasopoulos, 2018). Having a positive autocorrelation means that an unusually large value of a process y_t is expected to be succeeded by a value that is larger than the average value of y_{t+1}. On the other hand, a negative autocorrelation implies that a large value of y_t is likely to be followed by a small value at y_{t+1} (Hamilton, 2020).

A process that deserves special mention is the *white noise process* or often simply called *white noise*. It is a stationary process $\{\epsilon\}_{t=-\infty}^{\infty}$ that is a sequence of independent and identically distributed random variables for which all autocorrelation coefficients are zero (Lütkepohl, Krätzig, and Phillips, 2004).

2.2.2 Autoregressive Model

The autoregressive model belongs to the family of difference equations. It is a process where the current value of a variable is a finite, linear combination of its past values and a random shock (white noise) ϵ_t (Box, Jenkins, and Reinsel, 2008). As the term *auto*regressive indicates, that means a new value is forecast based on a linear combination of previous values of the variable (Hyndman and Athanasopoulos, 2018).

Definition. An *autoregressive process of order p*, short AR(p), is a process y_t that is a finite weighted sum of p previous deviations $y_{t-1}, \ldots, y_{t-p}$ of the process at equally distanced points in time, plus a random shock ϵ_t. It can be written as

$$y_t = \alpha_1 y_{t-1} + \cdots + \alpha_p y_{t-p} + \epsilon_t$$

where ϵ_t is an unobservable zero mean white noise process and $\alpha_i, i = 1, \ldots, p$ are fixed coefficients (Lütkepohl, Krätzig, and Phillips, 2004, Box, Jenkins, and Reinsel, 2008).

For better readability one can define a *lag operator* L, as proposed by Lütkepohl, Krätzig, and Phillips, 2004, as $L^j y_t = y_{t-j}$ with $L^0 y_t = y_t$. So the lag operator specifies the shift of time periods. Then the process can be written as

$$(1 - \alpha_1 L - \cdots - \alpha_p L^p) y_t = \epsilon_t \text{ or simply } \alpha(L) y_t = \epsilon_t.$$

An alteration of the parameters α_i, $i = 1, \ldots, p$ will produce different time series patterns. At the same time the variance of the error term ϵ_t will change the scale of the series but not the patterns itself (Hyndman and Athanasopoulos, 2018).

As mentioned above if a trend is present in the data then the autocorrelation for small lags will be positive and large while it decreases with greater lags. That's why the AR(1)

process is often considered when doing trend analysis (e.g. Hamed and Rao, 1998, Hamed, 2009):

$$y_t = \alpha_1 - y_{t-1} + \epsilon_t \text{ or } (1 - \alpha_1 L)y_t = \epsilon_t \text{ with the lag operator.}$$

Then the autocorrelation function is given by

$$\rho_k = \begin{cases} 1 & k = 0 \\ \alpha_1^k & k > 0. \end{cases}$$

2.2.3 Moving Average Model

In comparison to the autoregressive model, a time series can also be represented as a process that is linearly dependent on a finite number q of previous white noise error terms ϵ_t and can be used to forecast future values (Box, Jenkins, and Reinsel, 2008).

Definition. A process y_t that consists of the weighted sum of the elements of a white noice process and can be represented as

$$y_t = \epsilon_t + m_1\epsilon_{t-1} + \cdots + m_q\epsilon_{t-q}$$

with ϵ_t being a zero mean white noise and m_i being the coefficients of the model, is called a *moving average process of order* q or short MA(q) (Lütkepohl, Krätzig, and Phillips, 2004).

The moving average process is a stationary process, that means that its properties do not depend on the time of observation. Again, as seen previously for the AR(p) model, it that can be written more efficiently by defining a *lag operator* L as $L^j y_t = y_{t-j}$ with $L^0 y_t = y_t$. Then the process can be written as

$$y_t = (1 + m_1 L + \cdots + m_q L^q)\epsilon_t \text{ or } y_t = m(L)\epsilon_t.$$

For uniqueness of the representation one has to make restrictions on the coefficients $m_i,\ i = 1, \ldots, q$ or m respectively (Lütkepohl, Krätzig, and Phillips, 2004). A change in the parameters m_i will result in different time series patterns. On the other hand the variance of the error terms ϵ_t will change the scale of the series but not its patterns (Hyndman and Athanasopoulos, 2018).

Following Hamed and Rao, 1998 who used the MA(1) model for trend detection we will state its definition explicitly:

$$y_t = \epsilon_t + m_1\epsilon_{t-1} \text{ or } y_t = (1 - m_1 L)\epsilon_t \text{ with help of the lag operator.}$$

Then the autocorrelation function is given by

$$\rho_k = \begin{cases} -\dfrac{m_1}{1+m_1^2} & k=1 \\ 0 & k>1. \end{cases}$$

ARMA Model

Combing both autoregressive and moving average model, Whittle, 1951, described the so-called *ARMA* model. These processes offer a very effective tool for describing the dynamics of a time series.

Definition. The *autoregressive moving average model of order p,q* or short ARMA(p,q) is a univariate process that is defined as

$$y_t = \alpha_1 y_{t-1} + \cdots + \alpha_p y_{t-p} + \epsilon_t - m_1\epsilon_{t-1} - \cdots - m_q\epsilon_{t-q}.$$

Following the above notation, $\alpha_i,\ i = 1,\ldots,p$, are the coefficients of the AR(p) process, $m_i,\ i = 1,\ldots,q$, are the coefficients of the MA(q) process and ϵ_t is zero mean white noise.

Again a simplification with help of the lag operator is possible, therefore the ARMA model can be represented more compactly as

$$(1 - \alpha_1 L - \cdots - \alpha_p L^p)y_t = (1 - m_1 L - \cdots - m_q L^q)\epsilon_t$$

or even simpler as

$$\alpha(L)y_t = m(L)\epsilon_t.$$

There are two possible ways to interpret the ARMA model. For once it can be seen as a AR(p) process $\alpha(L)y_t = \epsilon_t$ with the white noise ϵ_t following a MA(q) process $\epsilon_t = m(L)a_t$. The second possibility is that the ARMA model can be described as a MA(q) process $y_t = m(L)b_t$ with b_t following the AR(p) process $\alpha(L)b_t = a_t$ such that $\alpha(L)y_t = m(L)\alpha(L)b_t = m(L)a_t$. Furthermore a stationary and invertible ARMA(p,q) process can be represented both as an infinite moving average process or as an infinite autoregressive process (Box, Jenkins, and Reinsel, 2008).

In the course of time, various variations of the ARMA model were created. Among them are the autoregressive integrated moving average model (ARIMA), the seasonal autoregressive moving averagel mode (SARIMA) and the vectorautoregressive moving average model (VARIMA) for multivariate analysis and many hybrid versions of these. The ARMA model as well as its various variants have been widely used for modeling and

simulation of meteorological phenomena such as precipitation data (Babu et al., 2011, Valipour, 2015), water quality (Zetterqvist, 1991), stream flow (Carlson, MacCormick, and Watts, 1970), wind speed (Liu, Tian, and Li, 2012) and many more. But it's been used in other fields as well such as the prediction of prices (Barrett, 2021), energy comsumption (Barak and Sadegh, 2016, Yuan, Liu, and Fang, 2016), electricity forecasting (Bouzerdoum, Mellit, and Pavan, 2013, Wesolowska, 2018) or motor vehicle crashes (Dezman et al., 2016).

2.2.4 Mann-Kendall Trend Test

The Mann-Kendall trend test is a rank correlation test which measures the degree of similarity between two observations and is therefore used to assess the significance of the relation between them (Kendall, 1962).

In 1938, Kendall suggested a rank correlation coefficient τ for the purpose of having an objective way to measure a subjectively given ranking by an individual. The coefficient offers a way to compare two different rankings of a set. There are two possible scenarios where the coefficient can be used. For once, if there is a natural ranking, a quality like weight or height, and an observer who ranks the individuals or objects according to his personal perception. In this case, the rank correlation coefficient gives information on how reliable the observer's judgment of some quality is. The second possibility is that there is no objective ranking of some quality, so two observers rank it in order of their own liking. For example, ranking a set of musical compositions according to individual preference. Here, the coefficient τ measures the closeness of correspondence between the two given rankings, in other words their compatibility. In the above example the coefficient would give insight if the two observers share the same taste or if they are rather incompatible within assigned limits of probability (Kendall, 1938).

Originally designed for psychological work, Mann extended the rank correlation coefficient τ given by Kendall. He picked up the idea of the statistic to confirm its usability for tests of randomness against trend. Mann showed its practicability in case of independent observations such as analysis of trends in time series (Mann, 1945).

Definition. Let $X = x_1, x_2, \ldots, x_n$ be a set of observations captured at times $t = 1, \ldots, n$. Then the statistic S given by Kendall, 1962, is defined as

$$S = \sum_{i<j} a_{ij},$$

where

$$a_{ij} = sgn(x_j - x_i) \;=\; \begin{cases} 1 & x_i < x_j \\ 0 & x_i = x_j \\ -1 & x_i > x_j. \end{cases}$$

That means, whenever the observation's value does correlate with its rank according to its natural order (from our perspective: time), the value 1 is assigned and whenever they aren't ordered ascendingly, $a_{ij} = -1$.

The Mann-Kendall trend test is commonly used as a distribution-free test of trend in time series. It is a non-parametric test which means that in comparison to parametric tests it only needs the data to be independent, but the distribution of the data may remain unknown. So, neither the power nor the significance of this test are affected by the distribution of the data, this constitutes one of the main advantage of the Mann-Kendall trend test. On the other hand, the test follows the basic assumption that the data is random and identically distributed. That's why, usually it is necessary to subject the data a pre-whitening process to meet this requirement or to modify the test to admit serial correlation in the data (Hamed and Rao, 1998, Hamed, 2009).

Given the null hypothesis that the observations $x_1, \ldots, x_n$ are randomly ordered, the statistic S tends to normality for large n. Then the mean value and variance of S are given by

$$\begin{aligned} \mathbb{E}(S) &= 0 \\ \mathbb{V}(S) &= \frac{n(n-1)\ (2n+5)}{18}. \end{aligned}$$

Furthermore, Kendall points out the possible existence of ties in the data. If two observations are indistinguishable, i.e. they share the same value, they are said to be *tied*. In this case the variance requires some modification. If ties of the extent t are given in the observations then the variance is given by

$$\mathbb{V}(S) = \frac{n(n-1)(2n+5) - \sum_t t(t-1)(2t+5)}{18}$$

(Kendall, 1962).

The Mann-Kendall trend test has been subject to many discussions because of its condition of independent data. The effects of autocorrelation on it and its validity in case of autocorrelated data are assessed by various researchers. The technique of reducing the effect of autocorrelation, the pre-whitening process, was introduced by Von Storch,

1999, and has among others been applied by Douglas, Vogel, and Kroll, 2000, Hamilton, Whitelaw, and Fenech, 2001, Yue et al., 2002, Partal and Kahya, 2006. One of the earliest studies on modifying the Mann-Kendall trend test itself to balance out autocorrelation was done by Lettenmaier, 1976, who enlarged the variance of the Mann-Kendall trend test by a factor depending on the first order autocorrelation parameter. Others who modified the variance to decrease the impact of autocorrelation were Hamed and Rao, 1998, although Yue et al., 2002, showed that even after the correction proposed by Hamed and Rao, the rejection rate was still higher than it was supposed to be. Yue and Wang, 2004, used the approach of effective sample sizes to eliminate the influence of serial correlation. Hirsch and Slack, 1984, proposed a seasonal Mann-Kendall trend test that compensates for autocorrelation due to seasonality in the data. This approach was among others applied by Zetterqvist, 1991, Gocic and Trajkovic, 2013, Zhai, Xia, and Zhang, 2014, Ashraf et al., 2021. The field of study where the Mann-Kendall trend test is most commonly applied is in meteorological and hydrological settings such as water quality, precipitation or temperature data and water flow data (Hamed, 2008, Gocic and Trajkovic, 2013).

2.3 Spatio-Temporal Analysis

With the development of GIS - the geographic information system - a massive amount of spatio-temporal data was generated that is responsible for many geographic research initiatives (Richardson, 2013). By now, many geographers and scholars in related fields of study have recognized the importance of integration of time for understanding a wide range of human experiences. Still, the majority of research methods are formulated in static spatial terms and the variety of methods to assess spatio-temporal relations are scarce (Kwan, 2013). Even when both, space and time, are considered, they are usually treated like separate aspects. Statistical methods addressing space and time in the same model and in geographic context are rather limited (Law, Quick, and Chan, 2014). In the following a few examples on spatio-temporal approaches are given.

2.3.1 Spatio-temporal Moran's Index

The central aspect when calculating the Moran's I is the exogenous specification of a spatial weights matrix. Yet, this matrix is usually defined in a spatial context only, even though the explored data is collected over time. In 2013, Dubé and Legros introduced a way of integrating the variable of time into the existing methods of geographical hotspot analysis. They presented an extended version of Moran's I for measuring spatio-temporal

clustering of geographic events on a global scale. By integrating the temporal component into the former spatial weights matrix and therefore transforming it to a spatio-temporal weights matrix (Dubé and Legros, 2013). Attempts of formulating a space-time weight matrix have previously been conducted by various researchers, e.g. Huang, Wu, and Barry, 2010, Wu, Li, and Huang, 2014. All of them have in common that they need estimates for spatio-temporal autocorrelation which are computationally intensive. In 2017, Lee and Li picked up Dubé and Legros' idea of a spatio-temporal weights matrix and added a localized version of spatio-temporal Moran's I (Lee and Li, 2017).

Definition. Let $W = (w_{ij})$ be the spatial weight matrix as defined in 2.1.1 and let $T = (t_{ij})$ be a temporal weight matrix that is defined analogous to the spatial weight matrix with d_{ij} being the temporal distance between two observations. Then the *spatio-temporal weight matrix* V is defined as the Hadamard product of W and T, i.e. $V = (v_{ij})$ with $v_{ij} = w_{ij}t_{ij}$.

Having the definition of a spatio-temporal weight matrix, global and local Moran's I can be calculated as defined in 2.1.2 with the spatio-temporal weight matrix taking the place of the spatial weight matrix:

$$I = \frac{n \sum_i \sum_j v_{ij}(x_i - \overline{x})(x_j - \overline{x})}{\sum_i \sum_j v_{ij} \sum_i (x_i - \overline{x})^2} \quad \text{and} \quad I_i = \frac{z_i \sum_j v_{ij} z_j}{\sigma^2}.$$

Analogous considerations of modifying the spatial weight matrix to a spatio-temporal weight matrix have been made for the Getis-Ord G_i statistic. See for example Tang, Tseng, and Chan, 2019 or Wang and Lam, 2020.

In their paper "Characterizing the Spatial Structure(s) of Cities 'on the fly': the Space-Time Calendar" Arribas-Bel and Tranos picked up this extension of Moran's index to establish the so-called 'Space-Time Calendar'. This novel approach manages to visualise slow and fast dynamics in one map. Their goal was to be able to reveal patterns that would usually be hidden by time aggregation. With the Space-Time Calendar it is possible to visually assess the clustering within each day over a long period (months or even years) at the same time. Spatio-temporal analysis is performed every day over the study period and then recorded in temporal units on the y-axis. As can be seen in figure 2.1, hotspots and coldspots are colored red and blue, respectively, while grey stands for non-significant clustering, and spots without available data are left blank. Then several runs over a longer period of time are stacked on the x-axis (Arribas-Bel and Tranos, 2018). The disadvantage of the Space-Time Calendar though, is that only one location of the study area can be inspected at a time. Furthermore, the temporal aspect is only considered within each day itself and then those days are stacked together over a longer period without assessing time over weeks or months in a statistical way.

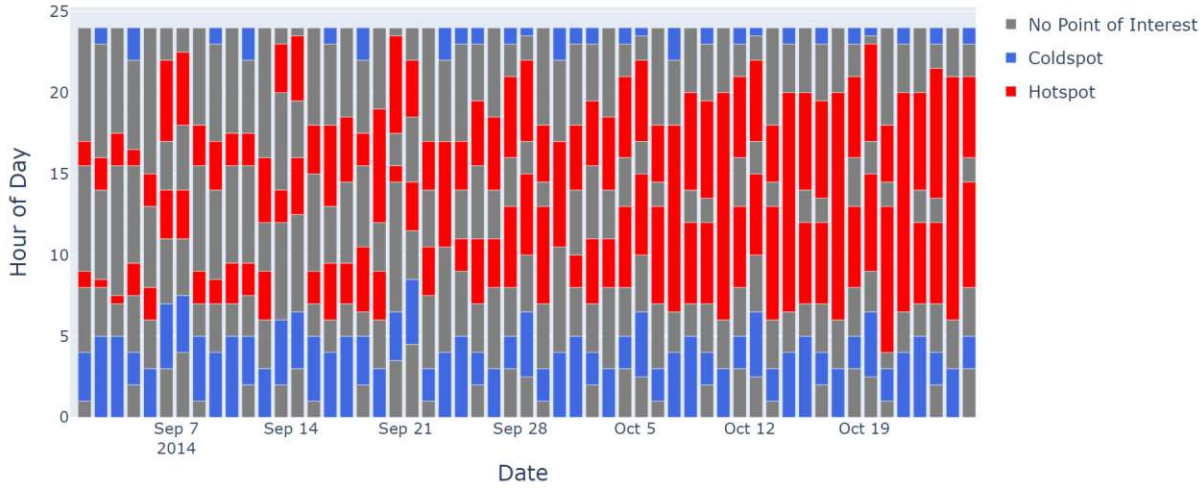

Figure 2.1: Example of a Space-Time Calendar for an intensifying Hotspot.

2.3.2 Space-Time Scan Statistic

Originally, the scan statistic was developed by Naus, 1965. It was a window statistic that aims at finding the maximum size of a cluster of points on a line. In 1997, Kulldorff proposed an extended and therefore more generally applicable version of this attempt for the usage in spatial cluster analysis. While the prior approach focused on one-dimensional analysis, Kulldorff extended the notion in various directions, namely by proposing a multi-dimensional point process, by allowing the scanning window to vary in size while moving along the observation area and by using either an inhomogeneous Poisson process or Bernoulli process as a baseline for calculation (Kulldorff, 1997). The spatial scan statistic is used for the detection of spatial clusters as well as their approximate location. For the calculation a circular window is placed on the map. Then the center of the window is moved over the observation area such that various locations include different sets of neighbouring census areas. The observation window does not only change its position but also its size. For each center of circles, the radius of it is varied from zero to a maximum such that it never includes more that 50 percent of the total count of observations. So, an extensive amount of different windows is created of which each is possibly containing a cluster. With the help of maximum likelihood ratio, a Poisson or Bernoulli process and a Monte Carlo hypothesis testing, the window that most likely contains a cluster can be identified. By sorting the circles by their maximum likelihood it is possible to identify secondary clusters as well (Kulldorff, 2001).

Kulldorff developed the spatial scan statistic mainly for epidemiological uses. In 2001, he addressed the problem that doing solely spatial analysis over a longer period will make

it less likely to detect recently emerging clusters of diseases. Therefore, he suggested a space-time scan statistic. He proposed, that instead of using two-dimensional windows on a planar study area, a cylinder is observed. So, the original circular window presents the base of the cylinder and therefore still represents the original and merely spatial aspect of the data. Time is represented as the third dimension and therefore constitutes the height of the cylinder. Now, not only the base is moving and changing in size as in the two-dimensional case, the starting date can vary as well. Both variations are independent of each other. So, for each location not only the window size is varied but also the height of the cylinder, considering various starting dates for possible clusters. Since in epidemiology and disease outbreaks active clusters are most important, only cylinders that reach to the end of the study period are considered in the final calculation. That way, historical clusters are excluded (Kulldorff, 2001).

So, the space-time scan statistic by Kulldorff, 2001, is defined as follows:

Definition. Let $[Y_1, Y_2]$ be the time interval for which data exists and let s and t be the start and end dates of the cylinder, respectively. Then all cylinders Z with $Y_1 \leq s \leq t = Y_2$ are considered. Conditioned on the total number of cases N, the space-time scan statistic is defined as the maximum likelihood ratio over all possible cylinders

$$S = \frac{max_Z\{L(Z)\}}{L_0} = max_Z\Big\{\frac{L(Z)}{L_0}\Big\},$$

where $L(Z)$ is the maximum likelihood for cylinder Z and L_0 is the likelihood function under the null hypothesis of randomness.

Let n_Z be the number of cases within cylinder Z. For the Poisson model, let $\mu(Z)$ be the expected number of cases under the null hypothesis such that $\mu(A) = N$ if A is the whole study area. Then $\frac{L(Z)}{L_0}$ can be calculated as

$$\frac{L(Z)}{L_0} = \Big(\frac{n_Z}{\mu(Z)}\Big)^{n_Z} \cdot \Big(\frac{N - n_Z}{N - \mu(Z)}\Big)^{N-n_Z},$$

if $n_Z > \mu(Z)$ and $\frac{L(Z)}{L_0} = 1$ otherwise.

Analogous to the mere spatial case, Monte Carlo hypothesis testing and a ranking due to the cylinder's maximum likelihood can be executed for cluster detection.

Kulldorff also points out that in case of time periodic surveillance, for example for disease surveillance, a highlighting of active clusters is needed. Therefore, he suggested another adaption of the space-time scan statistic such that p-values of clusters, detected after the time periodic surveillance began, are higher (Kulldorff, 2001). Kulldorff et al.,

2005, developed a space-time permutation scan statistic. With the help of an underlying probability model it should be possible to account for the lack of population-at-risk data for epidemiological surveillance purposes. In 2006, Kulldorff et al. proposed a transformation from a circular scanning window to an elliptic one. They suggested it as a special case of the original spatial scan statistic, since the shape of the observation window has no influence on the further calculations of the statistic. Unfortunately, they found that the elliptic scan statistic of its own performed rather poorly. That's why one has either to restrict the calculation to small ellipses or when using a large number of elliptic shapes, one has to combine them with an eccentricity penalty (Kulldorff et al., 2006). Kulldorff, Huang, and Konty, 2009, proposed a scan statistic for continuous data based on the normal probability model.

All versions of Kulldorff's scan statistics were originally defined for epidemiological purposes like detection of disease clusters, early detection of disease outbreaks, detection of geographical clusters of continuous variables such as low birth weight (Dahly et al., 2013, Rao, Shi, and Zhang, 2017, Kiani et al., 2021). But there have been applications for hotspot detection in other research fields as well, such as hotspot detection of agglomeration processes in economies (Kang, 2010), events on Twitter (Cheng and Wicks, 2014) or forest fires (Hudjimartsu, Djatna, Ambarwari, et al., 2017).

2.3.3 Space-Time Kernel Density Estimation

The kernel density estimation (KDE) was originally defined by the statisticians Rosenblatt, 1956, and Parzen, 1962, as a non-parametric tool for estimating a smooth empirical probability function of a random variable. It is commonly used in the fields of signal processing and econometrics. In his book 'Density Estimation For Statistics And Data Analysis' Silverman, 1986, elaborated on density estimations in general before this method made its way to spatial analysis. When using spatial data, a two-dimensional grid surface is created covering the study area, then a small kernel matrix is systematically run over it. The effects of observation points over their adjacent space is often visualised by heat maps. That way, locations with a cluster of events are highlighted and hotspots can be detected (Minn, 2000–2021). Brunsdon, Corcoran, and Higgs, 2005, picked up the idea of using the kernel density estimation as a visualisation tool. They examined various visualisation techniques as well as their effectiveness for spatio-temporal pattern and trend detection. They analysed three different techniques and in the course of doing that, extended the notion of the KDE to a space-time kernel density estimation.

Definition. Let n be the number of observations of the form (x_i, y_i, t_i) for $i = 1, \ldots, n$.

Let $k_1(\cdot,\cdot)$ be a probability function defined over a two-dimensional space (x_i, y_i) where h_1 is the bandwidth of the estimate and let $k_2(\cdot)$ be a probability function defined over time with bandwidth h_2. Then the estimate of the probability density at point (x_i, y_i, t_i) is given by

$$f(x,y,t) = \frac{1}{nh_1^2h_2}\sum_{i=1}^{n} k_1\Big(\frac{x-x_i}{h_1}, \frac{y-y_i}{h_1}\Big)k_2\Big(\frac{t-t_i}{h_2}\Big).$$

The space-time kernel density estimation offers a new approach to examining space-time events that are represented as point patterns. The value of the probability density function $f(x,y,t)$ describes the likelihood of an event happening at location (x,y) ad time t. Unfortunately, the visualisation of the space-time kernel density estimation would require a four-dimensional space, namely two dimensions for representing the geographical space (x_i, y_i), a third dimension to display the time component and a fourth dimension to map the density function. Brunsdon, Corcoran, and Higgs, 2005, suggested using an isosurface as a way of visualising at least some aspects of the density function in three dimensions. With that, patterns extending over a day or even a week can be seen at once.

Li and Racine, 2007, suggested an alternative and more general definition of a space-time kernel density estimation, they called it the *generalised product kernels*:

Definition. Let n be the number of observations of the form (x_i, y_i, t_i) for $i = 1, \ldots, n$. Let $k(\cdot)$ be an univariate kernel function that may vary with a specific dimension $\in (1, \ldots, q)$. Then

$$f(x) = \frac{1}{nh_1 \ldots h_q}\sum_{i=1}^{n} K\Big(\frac{x-X_i}{h}\Big),$$

where

$$K\Big(\frac{x-X_i}{h}\Big) = k\Big(\frac{x_1-X_{i1}}{h_1}\Big) \times \cdots \times \Big(\frac{x_q-X_{iq}}{h_q}\Big).$$

So, while Brunsdon et al. combine a bivariate kernel function $k_1(\cdot,\cdot)$ for space with an univariate kernel function $k_2(\cdot)$ for time, the definition of Li and Racine use the product of a univariate kernel function $k(\cdot)$ and is therefore expandable up to q dimensions.

Working with point patterns, the kernel density estimation can be used for hotspot detection. The downside is that the bandwidth that has to be set by the researcher brings in a highly subjective element to the calculation. Depending on how the bandwidth is defined, hotspots are detected or categorised as arbitrary. That's why hotspots detected by the kernel density function are not considered statistically significant, therefore, the KDE has to be used in combination with other hotspot detection techniques such as the Getis-Ord Gi statistic (Kalinic and Krisp, 2018). Addressing the problem of subjectivity on the bandwidth definition, Hart and Zandbergen, 2014, examined the influence of

interpolation methods, grid cell size, and bandwidth on hotspot mapping via the kernel density estimation.

Using the definition of a space-time kernel density estimation by Brunsdon et al., Nakaya and Yano, 2010, created a space-time cube for visualising crime clusters. Delmelle et al., 2014, used it visualising the impact of space-time uncertainties on dengue fever patterns. Li et al., 2020, analysed traffic violation behaviour at urban intersections. Hu et al., 2018, used the definition of a spatio-temporal kernel density estimation by Li and Racine to develop a framework for predictive crime hotspot mapping and evaluation.

Chapter 3

Exploring Spatial Analysis on Networks

Taking a closer look at the approaches introduced in chapter 2.1 and 2.3, all of them share the fact that they were indeed defined for doing spatial analysis. Though, they are meant to be applied to aggregated data across subareas of the study area, such as districts, postal zones or road blocks. Since usually census-related data (like population counts etc.) is widely available, this subarea based spatial analysis has become one of the most popular tools for empirical spatial analysis. In most theoretical work on spatial analysis an unbounded homogeneous space with Euclidean distance is assumed to be the study area. This assumption offers the optimal conditions for the development of new theories of spatial analysis or spatial stochastic processes (Okabe and Sugihara, 2012). Yet, assuming an ideal space is far from real world problems that are associated with and whose data is collected on network-shaped areas such as streets, railways, pipeline systems, rivers or communication networks.

So, tools for spatial analysis on networks are needed. But what does it actually mean to do spatial analysis on networks? What problems may arise and which obstacles need to be considered? The following chapter is inspired by the book "Spatial Analysis Along Networks: Statistical and Computational Methods" by Okabe and Sugihara, 2012, and will deal with the above questions. In particular, we will focus on events that occur on or alongside networks in form of point patterns. Among others, point patterns can represent the occurrence of traffic accidents, crime incidences, the contamination of rivers or, as we will see in the case study later on (see chapter 5), taxi stops in the sense of entry and exit points.

3.1 Point Patterns On and Alongside Networks

Doing hotspot analysis on an ideal space seems fanciful. However, many real world problems are closely connected to a network-shaped research area. These network events can be divided into two categories, namely events happening directly on the network and events that happen close to, particularly alongside networks. Events that happen on networks include car accidents, street crime sites, beaver lodges in watercourses, leakages in gas pipelines, breaks in a wiring network or blood clots in a vascular network and many more. As for alongside networks events, the entrances to almost all buildings in a city are adjacent to streets, thus, all events constrained to facilities next to a network arc can be categorised as an alongside network event.

The main difference between doing spatial analysis on a plane and spatial analysis on a network is, that while distance on the plane can easily be measured by the Euclidean distance, the concept of distance can not be assessed that easily on a network. Unfortunately, *planar spatial analysis* is often applied to networks treating it as if it were a plane. Though, this easily leads to false conclusions as can be straightforwardly deducted from the following figure 3.1. The first part (a) shows points on a plane that seem to be clustered to some extent. On the other hand, when adding a network structure and considering the points to be network events (b), they are simply randomly distributed on the network.

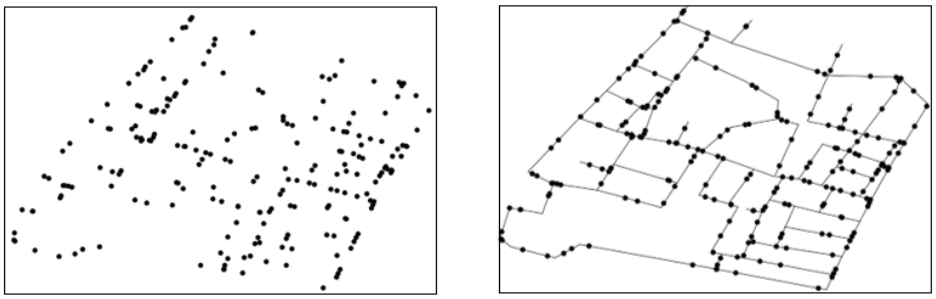

Figure 3.1: While there seems to be a clustering apparent in (a), the points aren't clustered in (b) which are the same points embedded in a network structure. This Figure was recreated from Okabe and Sugihara, 2012.

The point distributions are the same for the first and second part of the figure and were randomly generated using the uniform distribution on the network. It demonstrates that applying spatial analysis tools that were developed for the analysis of a planar study area to network shaped data will most likely result in misleading findings. In fact, not only

Okabe and Sugihara, 2012, warned from wrong results by using planar spatial analysis methods on a network. Yamada and Thill, 2004, conducted a study on traffic accidents in Buffalo showing that the K function method overestimates occurring clusters. Lu and Chen, 2007 studied crime distributions along streets and also warned from the effect when using planar methods instead of network spatial analysis tools.

Closely related to the issue of using planar spatial analysis methods on networks, is the Euclidean distance assumption. It is often chosen because of its straightforward calculation and on the other hand because it is assumed to approximate the shortest path distance on a network over a large study region. Still, it remains problematic when studying a small area like a city. That is why it is important to introduce *network spatial analysis*. In the following it is assumed that:

- The considered events occur on or alongside a network.
- If a method is chosen that includes distance variables, the shortest path distance will be assumed.

It should be noted that this cannot be a generalisation for all events that are probable to occur on a network shaped research area. Take for example the service area of a cell phone antenna. The antenna itself may be stationed alongside a network but its service area is characterised by Euclidean distance. So, it is important to take the properties of ones study object into account.

When using the above assumptions for network spatial analysis, it is easy to adapt to situations like directed networks when for example taking the current of a river into account or considering delivery routes in a city without neglecting one-way streets. Furthermore, one could handle non-planar networks like bridges or tunnels crossing other streets. The downside on the other hand of using the shortest-path distance on a network is, that it takes several steps to be computed. The most apparent obstacle is that the data sources of the network and point pattern usually differ. So, the typical case to encounter would be that the points aren't exactly mapped onto the network but rather slightly off. Therefore, the points have to be assigned to the network and only then an algorithm for calculating the shortest path distance can be applied (Okabe and Sugihara, 2012).

3.1.1 Conceptualising the Real World

When facing a problem in the real world, the first challenge is to describe it in detail and qualitative terms such that subsequently it can be put into a logical, quantitative model. Dealing with spatial phenomena, there are two main ways to put them into an abstract

model, namely the *object-based model* and the *field-based model*. In the following, both of them will be described in more detail.

Object-Based Model

When modeling real world phenomena, the first task to arise is to find a way of representing entities in a more abstract form. In case of the object-based model they can be qualified as discrete, distinct and well-demarcated abstract things termed *objects*. They can be represented by various geometric shapes such as points, lines, areas or solids that are representing the entities' abstract form as well as their abstract characteristics. These characteristics are the attributes of the entities that include spatial as well as non-spatial properties (see figure 3.2). While the non-spatial attributes usually cover the identifiers of the object, the spatial features cover geometric as well as topological features. The geometric properties are the geometric elements that form an object and are therefore quantitative characteristics. Typical geometric properties would be the length of a street or the area of a forest. Topological features on the other hand characterise the qualitative properties of the geometric elements forming the object. These are the characteristics that are preserved under topological transformation of the space and include attributes such as connectivity or intersections.

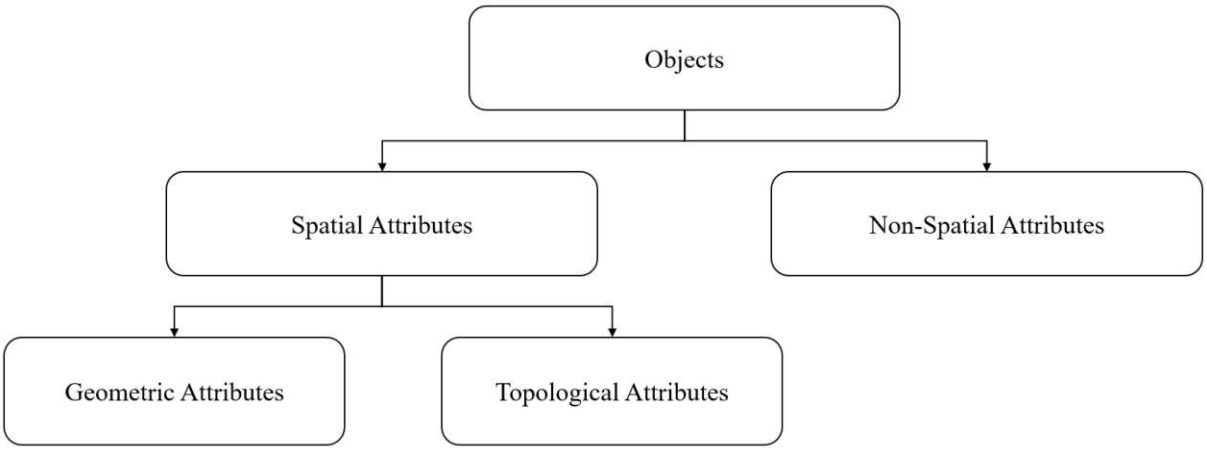

Figure 3.2: Types of attributes of objects in the object-based model. This figure was recreated from Okabe and Sugihara, 2012.

After putting the real world into the concept of an object-based model, the next step is to define the corresponding data model. A data model is needed to describe the conceptual model in terms of numeric values and therefore being able to process it with a computer. In case of the object-based approach, the most commonly used data model is the *vector*

data model.

In case of the object based model, the main component of the concept model are objects which can be described by spatial and non-spatial attributes, both of them need a data model of their own. The non-spatial characteristics can be assembled in a relational database in the form of tables. Let's take a closer look at the data model describing the spatial attributes. Again, there are two sub-sections of the spatial attributes, namely the geometric and the topological properties. The data model for the geometric features describes a finite number of points or rather the coordinates of a finite number of points forming the studied geometric objects. Typically, these points represent the vertices forming a polygon or the endpoints of a line segment. Note, that a curved surface can be approximated by a collection of connected polygons, while a curved line can be described by a sequence of connected straight line segments. These geometric elements (points, lines and polygons) can be described in terms of vectors, therefore this data model is called vector data model.

Concerning the topological attributes, they describe the relationship between geometric elements. It will be discussed in more detail in the context of a network shaped research area, where the extracted graph covers the representation of topological attributes.

Definition. A *link* is a line segment satisfying the following conditions:

- There is only a finite number of lines whose length is also finite. Therefore a line is actually a line segment.
- Every line segment is a straight line segment. Curved lines are therefore approximated by a connected sequence of straight line segments.
- All line segments are pairwise connected. That means that whichever pair of points on any line segment is selected, there is always at least one path in the network from one point to the other.

The endpoints of the above defined links are called *nodes*.

Definition. Let $V = (v_1, \ldots, v_{n_V})$ be a set of nodes and $L = (l_1, \ldots, l_{n_L})$ be a set of links. A *geometric network* is a pair of sets of nodes and links and is denoted by $N = (V, L)$ or simply N.

Remark. Note, that in the present thesis only finite networks are regarded. That means every considered network consists of a finite number of links of finite length and a finite number of nodes.

Apart from the fact that the geometric network displays the geometrical and topological attributes of a network object, it also functions as the underlying space for spatial analysis on a network. So, the geometric network can also be called *network space*. The network space can not only be described as a pair of sets of nodes and links $N = (V, L)$, it can also be seen as the union of its links $\tilde{L} = \bigcup_{i=1}^{n_L} l_i$ representing the set of points constituting N.

Even though the topological attributes are indirectly suggested by the geometric network, the corresponding graph can be extracted to make those properties explicit. By regarding the set of nodes V as a set of vertices V and the set of links L as a set of edges L, the graph $G = (V, L)$ can be extracted. That way, the quantitative properties, namely the geometrical attributes, are neglected (such as the location of nodes and length of links) while the qualitative, namely topological attributes are exhibited. So the graph can be seen as the abstract representation of the network where the main focus lies with pairs of vertices that are connected by edges (Okabe and Sugihara, 2012).

Note, that even though the network nodes are now considered to be the graph vertices and analogously the network links are taken as graph edges, there is a subtle difference between the network and graph representation. The network $N = (V, L)$ usually refers to real world systems that are represented in terms of a network, such as friendship networks on Facebook. The corresponding graph $G = (V, L)$ is the mathematical representation of the defined network. So, one can say that while in graph theory the interest lies mostly in questions about the graphs itself, network theorists are more concerned about the real world situations that are simply modeled by a network (Barabási and Pósfai, 2016). Alternatively, one can say that a graph does not contain information about for example the length of edges. To include this information a weighted graph is needed which can be referred to as a network (Sierksma and Ghosh, 2010).

Field-Based Model

In case of the object based model, entities were considered to be discrete, distinct and well-demarcated abstract things. In case of the field-based model the entities are described by a function of an attribute value or a vector of attribute values for every possible point. Here, the term 'field' is to be understood as vector field and not as a field in abstract algebra. The functions representing the entities of the research area are called *field functions*. The field-based model is mostly used in the context of temperature, precipitation, elevation, soil wetness, slope of terrains, etc. Though it is primarily used when the attribute values have interval or ratio scales, it is also applicable when they have nominal or ordinal scales. An example would be a binary function defined on every point over the research area,

indicating whether a certain criterion applies or doesn't apply at the considered point, e.g. categorising a region into cultivated and non-cultivated land. Similarly, one can define a function with multiple values such that the land usage can be classified into various categories instead of only two.

The data model most often associated to the field-based model is the *raster data model*. While the field function as defined above can easily map an infinite number of points on the continuum of space, for the raster data model one has to restrict the field function to a finite number of points. If the necessity arises, attribute values on non-selected sample points can be spatially interpolated. There are various suggested methods for sampling the attribute values:

(i) A lattice can be used to identify equally spaced points and take those attribute values into consideration. This is the most commonly used method, the data identified by the grid lattice is called *raster data*.

(ii) Attribute values of irregularly spaced points on the research area are examined.

(iii) The space is considered to be tessellated in a regular way. Then a representative value of the attribute values in the tessellated cell is used, for example the mean, median or maximum.

(iv) The space is considered to be tessellated in an irregular way. As in the previous method, a representative value of the attribute values in the tessellated cell is used.

For application of the field-based model for a network, a field function on the network space has to be defined. As seen before, the attribute values can take ratio, interval, ordinal and nominal scales. Although, ratio and interval scales are more frequently used since ordinal and nominal scales are usually covered by an object-based model (Okabe and Sugihara, 2012).

Note, that basically the object-based model can also be used with the raster data model as well as the field-based model with the vector data model. Though, in practice there is a strong association between the models as suggested above. Note also, that the raster data model is rarely used for networks.

3.1.2 Essential Computational Methods for Network Spatial Analysis

In the previous chapter we introduced the notion of the object-based model as well as the field-based model and their corresponding data models in terms of conceptualizing phenomena of the real world. These data models have the purpose of representing the conceptual models in form of numerical values to make them accessible by computational methods. Since the goal of the present thesis is to develop a spatio-temporal analysis toolkit for networks, the introduction of statistical methods applicable to networks is needed. Unfortunately, these statistical methods necessitate a lot of geometric computation which rapidly becomes very time consuming when working with usually very large spatial data sets. So, computational methods that master to find the thin line between acceptable computation time, required memory and accuracy of computation are needed. In the following section the basic computational methods are introduced before moving on to the more specialized topics of network spatial autocorrelation, nearest neighbour distance methods, hotspot analysis on networks and kernel density estimation.

Planar and non-planar Networks

When doing analysis on networks one has to distinguish two major types of networks, namely *planar* and *non-planar* networks.

Definition. A network is called *planar* if it can be embedded in the plane. That means that there exists a drawing of it, such that its links intersect only at their endpoints: the network's nodes.

A network is called *non-planar* if it doesn't meet the above definition. That means that every possible drawing of the network contains at least one link that intersects with at least one other link at a different point than their shared node.

Putting these definitions in real world context, one can consider a city map. When talking about road crossings, as presented in figure 3.3a, then they will usually be presented in a planar way. That means that the intersection of the roads will be represented as an intersection of two links within a node, as demonstrated in figure 3.3b. So, when reaching the crossing, one is free to continue down any of the other three roads, independent of the road that is chosen to reach the crossing. In case there is a bride or a tunnel within the street network (figure 3.3c), they will be mapped in form of a non-planar crossing, as can be seen in figure 3.3d. An intersection of two links without a node at the crossing implies

that it is not possible to cross from one road to the other at the point of the intersection because they are located on different layers of the network (Okabe and Sugihara, 2012).

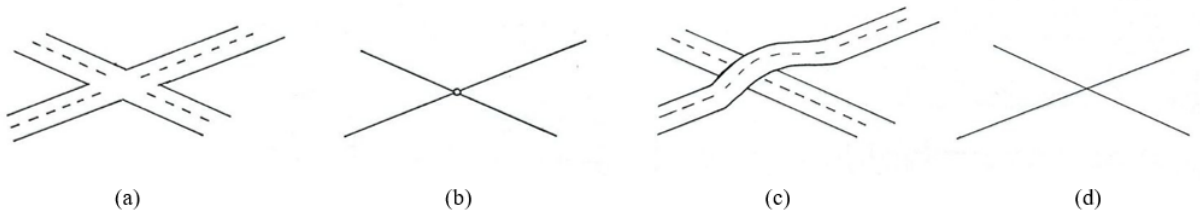

Figure 3.3: Two types of crossings: planar (a) and non-planar (b), as well as their geometric representation. This Figure was recreated from Okabe and Sugihara, 2012.

Typical examples of non-planar networks would include public transportation networks including tram- and subways, a simple road network involving bridges or tunnels for crossing other streets, or people's paths in multi-level buildings.

Shortest Path Distance

Many times when doing spatial analysis on networks, the resolution of a problem will need the calculation of the shortest path from a given point to another. Let's first elaborate on what finding the shortest path actually means. As we have seen before (Chapter 3.1.1), a network can be defined as a weighted graph. This implies that crossing from one node to the next comes with a certain cost depending on the edge that is passed: $c(e)$. These costs can be interpreted in various ways such as actual transportation costs when moving goods from one point to the next or when looking for a shortest path one can interpret these costs as the length of the edge one is moving along. A *path* is an alternating sequence of vertices and edges $(v_0, e_1, v_1, e_2, \ldots, e_n, v_n)$ with e_i being the edge from v_{i-1} to v_i and $0 \leq i \leq n$, to move from the starting vertex v_0 to the endpoint v_n. Then the *shortest path* from one vertex to another is the path that causes the minimal costs.

One of the most efficient algorithms for finding the single-source shortest path, is the *Dijkstra Algorithm*. It is a greedy algorithm which means that at each iteration stage the algorithm chooses the best option for its next step without taking future moves into account. Even though this problem-solving heuristic usually doesn't produce an optimal outcome, it is possible to produce a locally optimal solution (Encyclopedia of Mathematics, 2014). The Dijkstra algorithm achieves that for directed graphs with non-negative edge costs. So, henceforth it is assumed that $c(e_{ij}) \geq 0 \ \forall i, j$.

Finding the shortest path with the Dijkstra algorithm is an iterative process. Every vertex is labeled with a 2-dimensional vector. The first element presents the length of

the shortest path to reach the corresponding vertex, while the second element states the predecessor contributing to the shortest path to reach the current vertex. These labels can be updated in every iteration step. In a first step, the source vertex s is defined and its label is set to $(0, s)$ since the predecessor is s itself and it takes no edge to reach it (which corresponds to an edge of length 0). Every other label is set to $(\infty, *)$. Since their vertices haven't been considered yet, their shortest path is set to infinity and their predecessor is unknown. Then the iteration starts: in every step the set of vertices is partitioned into two sets V_1, V_2, namely the vertices that have already been processed: V_1, and the set of vertices V_2 that have to be considered yet. Due to the fact that the length for every edge is greater than or equal to zero, $c(e_{ij}) \geq 0$, the set V_1 is never empty since it always contains the source vertex. In the first iteration every vertex that is adjacent to the starting vertex s is taken into account and their label is updated according to the shortest path to it (in this case merely the distance from the source) and their predecessor (in this case s). In every iteration step the algorithm tests every vertex in set V_2 to check if the detected path can be improved. That means for every $v \in V_2$ the algorithm checks if there exists a $u \in V_1$ such that a path (u, v) from u to v exists such that $label_v > label_u + c(e_{uv})$. In case such a path doesn't exist, the present label remains, in case such a path is found, $label_v$ is updated accordingly. At the end of every iteration step the shortest path to exactly one of the vertices is obtained, the corresponding node v_0 in V_2, the one with the smallest label, is transferred to V_1. That way, the algorithm proceeds until every vertex was handled and V_2 is empty, then the shortest path from the starting point s to every other vertex was found (Dijkstra, 1959, Sierksma and Ghosh, 2010).

Example. Figure 3.4 and table 3.1 demonstrate the Dijkstra algorithm. Here, the source vertex is vertex A. In the initialisation the label of A is set to $(0, A)$ while all other labels remain at $(\infty, *)$. In the initialisation the set V_1 contains only vertex A, the other vertices are in set V_2. Now, all vertices of set V_2 are inspected to see if there is an edge from A to any of the vertices. This is the case for vertices B, C and D, their labels are updated accordingly as is highlighted in figure 3.4 and can be seen in table 3.1. Note, that B, C and D remain in set V_2 since there is still the possibility of a shorter path than the direct path from the source vertex. In the next iteration step the vertex with the shortest distance from A is chosen - vertex D - and again the labels of vertices in V_2 that are successor to D are updated. Then, D is moved from set V_2 to set V_1 since its shortest path was found. This way the algorithm continues until the set of V_2 is empty. In every iteration step the shortest path to exactly one vertex is found.

One of the downsides of using the shortest path distance for network spatial analysis is that it is not as simple as using the Euclidean distance between two points on a plane.

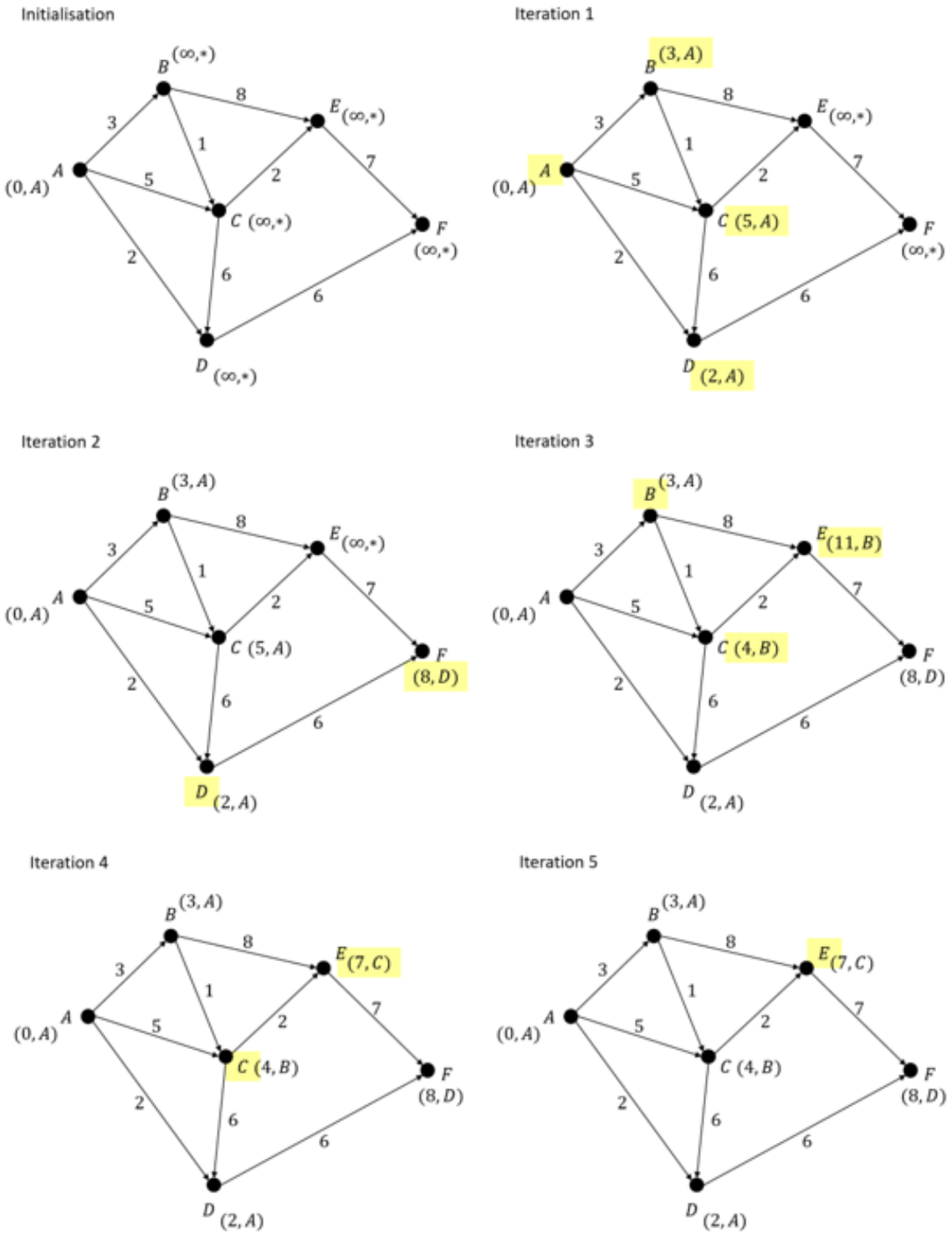

Figure 3.4: The first graph shows the initialisation phase. Then the following iteration steps are given with the considered vertex and updated labels highlighted.

Iteration		A	B	C	D	E	F	
0	Costs Predecessor	0 -	∞ -	∞ -	∞ -	∞ -	∞ -	$V_1 = \{A\}$ $V_2 = \{B, C, D, E, F\}$
1	Costs Predecessor	0 -	3 A	5 A	2 A	∞ -	∞ -	$V_1 = \{A\}$ $V_2 = \{B, C, D, E, F\}$
2	Costs Predecessor	0 -	3 A	5 A	2 A	∞ -	8 D	$V_1 = \{A, D\}$ $V_2 = \{B, C, E, F\}$
3	Costs Predecessor	0 -	3 A	4 B	2 A	11 B	8 D	$V_1 = \{A, B, D\}$ $V_2 = \{C, E, F\}$
4	Costs Predecessor	0 -	3 A	4 B	2 A	7 C	8 D	$V_1 = \{A, B, C, D\}$ $V_2 = \{E, F\}$
5	Costs Predecessor	0 -	3 A	4 B	2 A	7 C	8 D	$V_1 = \{A, B, C, D, E\}$ $V_2 = \{F\}$
6	Costs Predecessor	0 -	3 A	4 B	2 A	7 C	8 D	$V_1 = \{A, B, C, D, E, F\}$ $V_2 = \{\}$

Table 3.1: Initialisation as well as iteration steps of the Dijkstra algorithm for the graph presented in figure 3.4

Several steps need to be taken: usually when doing a case study, the point pattern and network data will be obtained from different sources. That means that the points will most likely not be exactly on but rather slightly off the network and therefore must be assigned onto the edges of the network. Only after that an algorithm for calculating the shortest path can be applied (Okabe and Sugihara, 2012).

Voronoi Diagrams

The network Voronoi diagram is an extension of the Voronoi diagram defined on the Cartesian plane or Cartesian space in higher dimensions. It was named after the Russian mathematician Georgy Voronoi (1868-1908). Nowadays, the Voronoi diagram is directly and indirectly part of many spatial analysis methods (Okabe and Sugihara, 2012).

Definition. Let p be an arbitrary point and let $P = \{p_1, \ldots, p_n\}$ be a set of points on the Cartesian plane. $d_E(\cdot, \cdot)$ denotes the Euclidean distance of two points in the plane. Then we can define a set of points as

$$V_{VD}(p_i) = \{p \mid d_E(p, p_i) \leq d_E(p, p_j), j = 1, \ldots, n\}.$$

Then the set V_{VD} is called *Voronoi polygon* of p_i and the points $p_1, \ldots, p_n$ are called *generator points*. Thus, we can define

$$\mathcal{V}(P) = \{V_{VD}(p_1), \ldots, V_{VD}(p_n)\}$$

as the *ordinary planar Voronoi diagram* generated by P.

The ordinary planar Voronoi diagram is characterised by three aspects, namely space, generators and distance. These can be used to extend the notion to networks by replacing the Cartesian space by a network space and the Euclidean distance by the shortest path distance $d(\cdot, \cdot)$. Then

$$V_{VD}(p_i) = \{p \mid d(p, p_i) \leq d(p, p_j), j = 1, \ldots, n\}$$

is called the *Voronoi subnetwork* of p_i or *Voronoi cell* of p_i. Analogously, the *ordinary network Voronoi diagram* is defined as

$$\mathscr{V}(P) = \{V_{VD}(p_1), \ldots, V_{VD}(p_n)\}.$$

This is the standard definition for network Voronoi diagrams for undirected networks. Of course, it can be generalised such that it can be used on directed or weighted networks.

For the directed network Voronoi diagram, the shortest path distance has to be divided into the *outward shortest path distance* and the *inward shortest path distance*. The former denotes the shortest path distance from the node p_i to p: $d(p_i, p)$, the latter denotes the shortest path distance to the node p_i from p: $d(p, p_i)$. Therefore, there are two ways of defining the Voronoi subnetwork of p_i, namely

$$V_{VD}(p_i) = \{p \mid d(p_i, p) \leq d(p_j, p), j = 1, \ldots, n\},$$
$$V_{VD}(p_i) = \{p \mid d(p, p_i) \leq d(p, p_j), j = 1, \ldots, n\}.$$

Together the outward and inward network Voronoi diagrams constitute the *directed network Voronoi diagram* $\mathscr{V}(P)$. Note, that there is also a change in the properties between the undirected network Voronoi diagram and the directed network Voronoi diagram. For example, while undirected network Voronoi diagrams were mutually exclusive and collectively exhaustive, this property doesn't always hold for directed network Voronoi diagrams. It holds only if additionally there are no nodes with only leaving links and no nodes that are only incident to links pointing towards the node (Okabe and Sugihara, 2012).

The weighted network Voronoi diagram can be extended from the weighted planar Voronoi diagram. We replace the shortest path distance by a *weighted shortest path distance* $d_w(\cdot, \cdot) = \alpha_i d(\cdot, \cdot) + \beta_i$, where α_i, β_i are positive constants that reflect the weights for a certain path such as transportation costs. As before, we can define the *weighted network Voronoi diagram* $\mathscr{V}(P)$ via

$$\begin{aligned} V_{VD}(p_i) &= \{p \mid d_w(p, p_i) \leq d_w(p, p_j), j = 1, \ldots, n\} \\ &= \{p \mid \alpha_i d(p, p_i) + \beta_i \leq \alpha_i d(p, p_j) + \beta_i, j = 1, \ldots, n\}. \end{aligned}$$

Note, that as defined above larger weights α_i, β_i will result in smaller Voronoi cells. Depending on the objective of the study these can be adapted in such a way that larger constants will lead to larger Voronoi cells (Okabe and Sugihara, 2012).

3.2 Network Spatial Autocorrelation

As we have seen in the previous chapter 2, the concept of spatial autocorrelation, namely the correlation between the same attribute values at different locations, needs to be considered when doing spatial analysis. As the original statistics for spatial autocorrelation were developed for areal data, the concept of network autocorrelation needs to be examined in more detail. Over time, multiple ways of exploring network autocorrelation were designed and used in different contexts.

White, Burton, and Dow, 1981, were the first ones to extend the notion of spatial autocorrelation as suggested by Cliff and Ord, 1973, to a more abstract level. As their main focus was on social networks, their suggestion of a network autocorrelation and the applications of studies based on this formulation was later called *social autocorrelation*.

Later, Goodchild, 1987, explored the autocorrelation between the attribute values of links in a general network instead of social networks alone. Building on this idea Black, 1992, applied this *network autocorrelation* to transport networks and flow systems. While spatial autocorrelation focuses on variables at given location that are influenced by variables at nearby locations, he understood network autocorrelation as the dependence of variable values of given links on a network to similar values on links that are connected to the link under examination (Okabe and Sugihara, 2012).

3.2.1 Classifying Autocorrelation

In the present thesis, the approach suggested by Okabe and Sugihara, 2012, will be examined in more detail. They suggested a definition of *network spatial autocorrelation* that is formulated in a more general way and therefore covers a broader range of application.

Definition. Let $\{e_1, \ldots, e_n\}$ be a set of entities with $x_i,\ i = 1, \ldots, n$, being the attribute values corresponding to e_i. And let $r_{ij},\ i \neq j,\ i, j = 1, \ldots, n$ be the *relations* between e_i and e_j with $w_{ij} \in \mathbb{R},\ i \neq j,\ i, j = 1, \ldots, n$, being the corresponding *relational value* of r_{ij}. Then $N_R = (\{e_1, \ldots, e_n\}, \{r_{ij} \mid i \neq j,\ i, j = 1, \ldots, n\})$ is called a *relational network*. The

entities e_i are embedded in a space S in which their relations are described by N_R. The autocorrelation between the attribute values x_i of the corresponding entities e_i in space S, with their relations being described by the relational network N_R with relational values w_{ij} is called *abstract autocorrelation*.

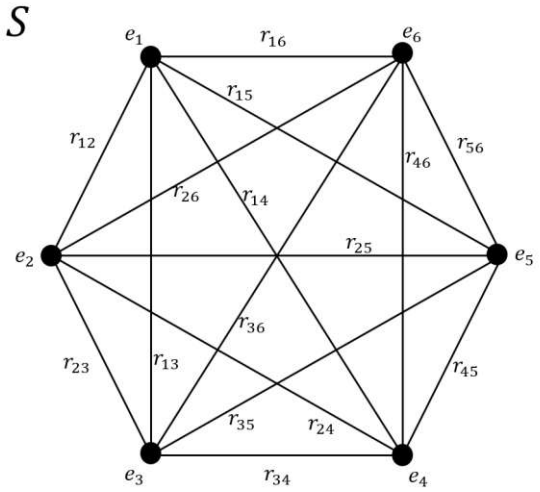

Figure 3.5: A relational network for abstract autocorrelation: the entities $e_i,\ i = 1, \ldots, n$, represent the nodes of the network, while the relations $r_{ij},\ i \neq j,\ i, j = 1, \ldots, n$, represent the links. The relational network is located in a space S.

Remark. A few remarks need to be made on the above definition.

(i) For the abstract autocorrelation all possible relations between entities are considered. Therefore, the relational network is represented by a complete graph K_n, as can be seen in figure 3.5. Of course, when illustrating a real-world situation some relations are insignificant and can therefore be neglected. But still, conceptually all possible relations are taken into consideration.

(ii) The possibility of neglecting insignificant links can be done with help of the relational values. For example, $w_{ij} = 1$ in case there actually is a connection between two entitites and $w_{ij} = 0$ in case there isn't.

(iii) The relational values w_{ij} are usually called *spatial weights* in spatial analysis (introduced in section 2.1.1).

(iv) Usually, the relational network N_R is a directed network with $r_{ij} \neq r_{ji}$ and $w_{ij} \neq w_{ji}$.

As a relational network is embedded in a space S, it has to be noted that there are different kinds of spaces. Among others, White, Burton, and Dow, 1981, primarily investigated social networks and therefore the studied space was mainly in the *non-physical* domain. When spatial analysis is conducted and spatial statistics to be applied, the only space considered is the real world, meaning the *physical* space.

Apart from the space the relational network is situated in, also the entities and relations between them can be categorised as physical and non-physical, for example:

- Entities: stores and districts vs. societies and cultures
- Relations: distance between two stores vs. political groups with opposing policies
- Space: regional and architectural space vs. social and religious space

So, there are 2^3 possible combinations of the above characteristics which can lead to mixed classes of autocorrelation due to the physical / non-physical classification, see figure 3.6. If entities, relations and space can be categorised as physical, we call the resulting autocorrelation *strictly spatial autocorrelation* or simple *spatial autocorrelation*. Spatial autocorrelation in classic spatial analysis, where most studies consider entities as subareas of the study area and distances are measured by the Euclidean distance, is henceforth called *planar spatial autocorrelation*. From a mathematical point of view its structure corresponds to the above definition of abstract autocorrelation. This is due to the fact that even if it seems like attribute values are considered over a continuous plane, the plane is actually treated like the composition of multiple discrete subareas. Though, spatial autocorrelation is not restricted to the planar case but entities can be embedded in a physical network space. The correpsonding spatial autocorrelation is therefore called *network spatial autocorrelation*.

When dealing with network spatial autocorrelation further categorisations according to the geometrical form of the observed entities can be made. For example, if the entities are taxistops on a city street network (as we will see in chapter 5), they are points located on the links of a network. Therefore, the network spatial autcorrelation can more precisely be referred to as *points-on-network autocorrelation* or in case those points happen to be the nodes of the network one can call it *nodes-on-network autocorrelation*. The geometric form of the entities e_i can also be that of a line segment. For example, one could count the number of accidents on any 100m road segment (see Black, 1991), then the corresponding network spatial autocorrelation can be called *lines-on-network autocorrelation*. If all links of the network have attribute values one can refer to it as *links-on-network autocorrelation*. If the entities are a set of connected line segments that constitute a subnetwork, for

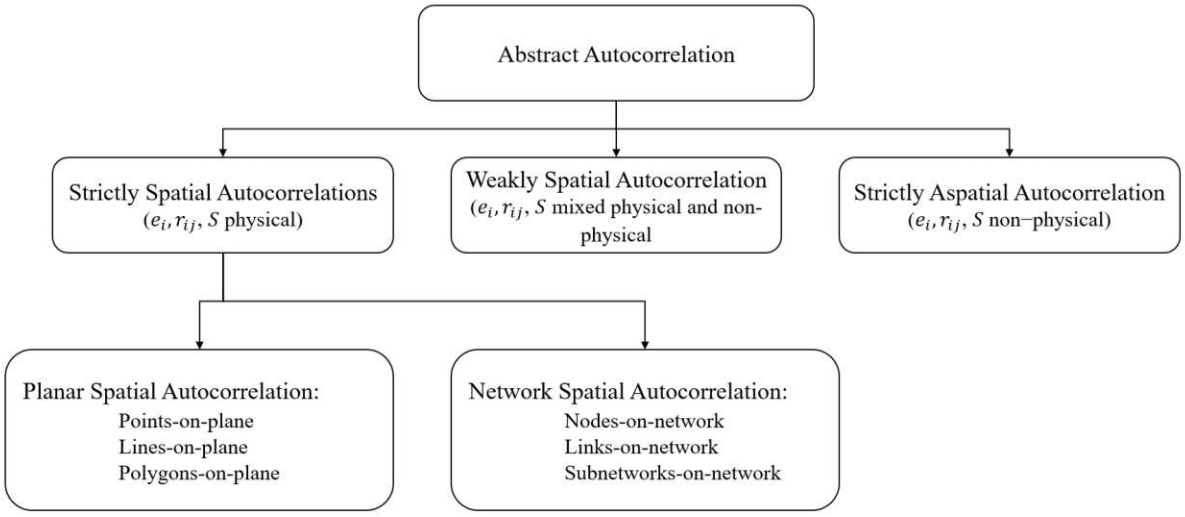

Figure 3.6: Classifications of abstract autocorrelation. This Figure was recreated from Okabe and Sugihara, 2012. Note, that both *weakly spatial autocorelation* and *strictly apsatial autcorrelation* may have further classifications as well but they are of no importance and therefore out of scope for the present thesis.

example streets of single districts in a city, then the network spatial autocorrelation can be addressed as *subnetworks-on-network autocorrelation*. The matching counterparts for planar spatial autocorrelation are called *points-on-plane autocorrelation*, *lines-on-plane autocorrelation* and *polygons-on-plane autcorrelation* (Okabe and Sugihara, 2012).

3.2.2 Spatial Randomness

There are two types of spatial randomness that can be defined on networks, namely *permutation spatial randomness* and *normal variate spatial randomness*. As seen before, entities can take various geometrical forms on a network such as points, line seqments or subnetworks. Associated with these entities $e_i,\ i = 1, \ldots, n$, are their attribute values $x_i,\ i = 1, \ldots, n$. Those spatial data units will henceforth be referred to as *network cells* $L_i,\ i = 1, \ldots, n$.

Definition. For the *permutation spatial randomness* the attribute values $x_1, \ldots, x_n$ are fixed but they are randomly assigned to the n network cells $L_1, \ldots, L_n$. Further, permutation spatial randomness assumes that every permutation of attribute values $x_i,\ i = 1, \ldots, n$, occurs with the same probability.

Since all permutations occur with the same probability, the expected attribute values are the same for all network cells and therefore uniform. That's why, permutation spatial

randomness is often used in spatial analysis as null hypothesis for statistical tests.

Normal variate spatial randomness doesn't assume the attribute values to be fixed but rather to be derived from a discrete probability distribution or a continuous probability density function:

Definition. *Normal variate spatial randomness* assumes that the attribute values $x_i,\ i = 1, \ldots, n$, assigned to network cells $L_1, \ldots, L_n$ are independently and identically generated from a normal distribution.

As for the permutation spatial randomness, the attribute values' expected value is the same and therefore uniform across all network cells (Okabe and Sugihara, 2012).

3.3 Network Distance Methods

A central element when doing spatial analysis is the definition of *distance* and closely related to it is the concept of *neighbourhood*. In chapter 2.1.1 the topic of a spatial weights matrix in planar spatial analysis was discussed. There, the weights were defined based on various concepts of distance: Euclidean distance, distance threshold, inverse distance or based on boundaries (queen and rook criteria).

Now, the focus is on network spatial analysis and therefore the concept of distance has to be investigated in more detail. The aforementioned distance concepts are based on the Euclidean distance. As stated before, the Euclidean distance is often chosen for networks because of its straightforward calculation. It is assumed to approximate the shortest path distance on a network over a large study area. Still, it remains problematic when studying a small area or a city because it can lead to a significant overestimation of clusters. That's the reason why it is advisable to use the shortest path distance when working with networks. On its basis, the concept of *nearest neighbour distance* and the *K function method* will be presented.

3.3.1 Nearest Neighbour Distance

The concept of *network nearest neighbour distance* is an extension of the planar nearest neighbour distance, which belongs to the most traditionally used methods in planar spatial analysis. The extension to networks was already broadly discussed in the 1950s (Cottam and Curtis, 1949, Clark and Evans, 1954, Pielou, 1959).

Network nearest neighbour distance is based on the shortest path distance introduced in a previous chapter (see 3.1.2). Okabe and Sugihara, 2012, introduced various kinds of

the network nearest neighbour distance methods. In the following we will have a look at the *network auto nearest neighbour distance*, in literature this method is conventionally referred to as the *network nearest neighbour distance*. It examines points of the same kind, for example taxi stops within a city, and therefore investigates the spatial patterns based on the shortest path distance.

Let $N = (V, L)$ be a network consisting of nodes $V = \{v_1, \ldots, v_{n_V}\}$ and links $L = \{l_1, \ldots, v_{l_L}\}$. Let $\tilde{L} = \bigcup_{i=1}^{n_L} l_i$ be the union of the links, that means a set of points forming the links of the network. And let $P = \{p_1, \ldots, p_n\}$ be a set of points created by a stochastic process on $\tilde{L}$ with a fixed n. We assume that all entities on the network are known and therefore represented by p_i. We assume that the points in P are independently and identically distributed and follow the null hypothesis of complete spatial randomness. To test this hypothesis a statistic test is needed that measures the distance from every point $p_i \in P$ to its next nearest point in P. In doing so, we can focus on two different objectives, namely

- Focus on one point p_i and check if the next nearest point is significantly close - *network local nearest neighbour distance method*
- Consider all points and test if the average distance for every point to their next nearest point is significantly short - *network global nearest neighbour distance method*

For the local nearest neighbour distance method we consider a subnetwork of $\tilde{L}$ for p_i, namely $\tilde{L}(t|p_i) = \{p \mid d(p_i,\ p) \leq t, p \in \tilde{L}\}$. So, the subnetwork $\tilde{L}(t|p_i)$ consists of all points that are within a distance threshold of length t. The distance $d(\cdot, \cdot)$ is the shortest path distance.

Definition. The *nearest neighbour distance* from p_i is the distance $d(p_i, p^*)$, where p^* is the next nearest point of $p_i \in P$.

For a statistical test the probability distribution function of $d(p_i, p^*)$ under the null hypothesis of complete spatial randomness is needed. It is derived from the probability that the next nearest point p^* of p_i lies within the distance threshold of t:

$$\begin{aligned}\mathbb{P}(d(p_i, p^* \leq t) &= 1 - \mathbb{P}(\forall p_j \in P \backslash \{p_i\} : d(p_i, p_j) > t \\ &= 1 - \mathbb{P}(p_1, \ldots, p_{i-1}, p_{i+1}, \ldots, p_n \in \tilde{L} \backslash \tilde{L}(t|p_i)),\end{aligned}$$

where $\tilde{L} \backslash \tilde{L}(t|p_i))$ describes the complement network of $\tilde{L}(t|p_i))$.
Then the probability distribution function is given by

$$1 - \left(\frac{|\tilde{L}| - |\tilde{L}(t|p_i)|}{|\tilde{L}|} \right)^{n-1},$$

where $|\tilde{L}|$ is the length of $\tilde{L}$.
The exact calculation of the probability distribution function can be looked up in Okabe and Sugihara, 2012 (chapter 5.4). It provides the expected value $\mu(p_i)$ and variance $\sigma(p_i)$. Then one can define and local index as the ratio of the value $d(p_i, p_i^*)$ to its expected value as $I_L = \frac{d(p_i, p_i^*)}{\mu(p_i)}$.

The global nearest neighbour distance method is derived from the local version, so the local index has to be extended to a global index. Instead of only for a single p_i, the nearest neighbour distance has to be calculated for all $p \in P$ and then the average is taken:

$$\frac{1}{n}\Big(\sum_{i=1}^{n} d(p_i, p_i^*) \Big).$$

Then the expected value of the random variable $\frac{1}{n}\Big(\sum_{i=1}^{n} d(p_i, p_i^*) \Big)$ under the null hypothesis of complete spatial randomness can be calculated by integration:

$$\mu = \frac{1}{|\tilde{L}|} \int_{p_i \in \tilde{L}} \mu(p_i) dp_i,$$

where the integral is the integration of $\mu(p_i)$ along $\tilde{L}$. Since an analytical solution for the integral is not easily found, Monte Carlo simulation is employed, then the global index is defined as the ratio of the average of $d(p_i, p_i^*),\ i = 1, \ldots, n$ to its expected value:

$$I_G = \frac{1}{\mu}\Big(\frac{1}{n} \sum_{i=1}^{n} d(p_i, p_i^*) \Big).$$

Again, the probability density function of the random variable $\frac{1}{n} \sum_{i=1}^{n} d(p_i, p_i^*)$, is needed for statistical tests. However, here as well the analytical solution is not easily obtained, therefore as n gets large one can apply the central limit theorem and conclude that the null hypothesis can be tested using the normal distribution with the expected value and variance received by Monte Carlo simulation.

Note, that the global nearest neighbour distance usually is the one applied in literature (Okabe and Sugihara, 2012).

3.3.2 K Function Methods

Besides the nearest neighbour distance method, the K function method is the second main method for point pattern analysis based on distance. Both of them answer similar research questions but in a different way. While the nearest neighbour distance method corresponds

to questions concerning the fact if the distance to the next nearest point is particularly short, the K function method asks if within a given distance from each point there are significantly many other points.

As we have seen for the network nearest neighbour distance method, Okabe and Sugihara, 2012, distinguish various kinds of the K function method. In the following, the *network auto K function method* will be introduced, which is the one usually referred to as *network K function method* in literature. It measures the number of points within a given distance for points of the same kind.

Let $N = (V, L)$ be a network consisting of nodes $V = \{v_1, \ldots, v_{n_V}\}$ and links $L = \{l_1, \ldots, v_{l_L}\}$. Let $\tilde{L} = \bigcup_{i=1}^{n_L} l_i$ be a set of points forming the links of the network. Let $P = \{p_1, \ldots, p_n\}$ be a set of points of the same kind that are created by a stochastic process on $\tilde{L}$ with a fixed n. The points in P are independently and identically distributed, therefore they follow the null hypothesis of complete spatial randomness. For testing this null hypothesis two statistics are defined:

- Focus on one point p_i and count the number of points $p_j \in P, j \neq i$, within a predefined distance t. This statistic is called the *network local K function method*.
- Calculate the average number of points within a given distance t from every point in P to the other points in P. This statistic is called the *network global K function method*.

For the local K function method we consider for a $p_i \in P$ a subnetwork $\tilde{L}(t|p_i) \subset \tilde{L}$. It is defined as $\tilde{L}(t|p_i) = \{p \mid d(p_i, p) \leq t, p \in \tilde{L}\}$, therefore it contains all points within a shortest path distance of t. This is the network equivalent of a disk with radius t centered at p_i in the plane. Then, let $n(t|p_i)$ be the number of points in P (excluding p_i) that are incorporated in $\tilde{L}(t|p_i)$. Then we define

$$K(t|p_i) = \tfrac{1}{\rho} n(t|p_i),$$

where $\rho = \dfrac{n-1}{|\tilde{L}|}$ is the density of points and $|\tilde{L}|$ is the length of $\tilde{L}$.

To test the hypothesis of complete spatial randomness $(n-1)$ points are uniformly and independently generated over the whole network $\tilde{L}$. Thus, the probability distribution function of the random variable $n(t|p_i)$ is a binomial distribution with parameters $(n-1)$ and $\dfrac{|\tilde{L}(t|p_i)|}{|\tilde{L}|}$. That's why, the random variable $K(t|p_i)$ follows the binomial distribution under the null hypothesis. Its expected value and variance are given by

$$\mathbb{E}(K(t|p_i)) = \mu_i(t) = |\tilde{L}(t|p_i)|,$$

$$\mathbb{V}(K(t|p_i)) = \sigma_i(t) = \tfrac{1}{n-1}\ |\tilde{L}|\ |\tilde{L}(t|p_i)|\ \left(1 - \frac{|\tilde{L}(t|p_i)|}{|\tilde{L}|}\right).$$

Given the random variable $K(t|p_i)$ then let $K^*(t|p_i)$ and $K^{**}(t|p_i)$ be its upper and lower critical values with significance level α. If an observed value of $K(t|p_i)$ is smaller or greater than $K^*(t|p_i)$ or $K^{**}(t|p_i)$ respectively, it leads to the conclusion that with a confidence level of $1-\alpha$ points are significantly clustered around p_i.

The global K function method is derived from the local K function method. For the local method $n(t|p_i)$ was defined for a specific point p_i. To generalise this to a global statistic $K(t)$, we average $n(t|p_i)$ over $i = 1, \ldots, n$, i.e.

$$K(t) = \frac{1}{\rho}\ \Big(\frac{1}{n}\ \textstyle\sum_{i=1}^{n} n(t|p_i)\Big).$$

In order to test the null hypothesis of complete spatial randomness, it is necessary to compute the expected value and variance of $K(t)$. Since an analytical solution is hard to find, Monte Carlo simulation can be used (Okabe and Sugihara, 2012).

3.4 Hotspot Analysis on Networks

While the spatial hotspot analysis investigated in section 2.1 was focused on hotspot analysis on the Cartesian plane, now the possibilities for spatial hotspot analysis on network space will be examined in more detail. Besag and Newell, 1991, sounded a note of caution that while the terms of hotspot analysis and cluster analysis are often used interchangeably, one has to keep in mind that there exists a subtle difference. While *tests of clustering* try to answer the question if an observed pattern has arisen by chance or if there actually are significant clusters present, the tests for the *detection of clusters* monitor a large area divided into smaller subareas (network and subnetworks respectively) for evidence of individual hot- or coldspots. In contrast to the test of clustering, the tests for the detection of clusters have no biased opinion on where these clusters are to be expected. So the detection of a hotspot would lead to further investigations and a more sophisticated study of the results.

In the following chapter, techniques for spatial clustering and hotspot analysis on networks are discussed in more detail: point cluster analysis, kernel density estimation and the Moran's I statistic. A central element of spatial analysis on networks is the shortest path distance which is the core element of every introduced method. Note also, that the

nearest neighbour distance method described in chapter 3.3.1 and the K function method discussed in chapter 3.3.2 can be categorised as cluster analysis.

3.4.1 Point Cluster Analysis

Network point cluster analysis is a point pattern analysis method for networks. The method checks for similarities between the attribute values of objects and groups those objects accordingly. In case of *spatial* cluster analysis, this similarity refers to the spatial proximity between objects which is measured by the shortest path distance between them on the network.

Network point cluster analysis is rooted in the spatial cluster analysis for point patterns on the plane. Yiu and Mamoulis, 2004, recognized clustering as one of the most important spatial analysis tasks and also that the importance of using shortest path distance instead of Euclidean distance for networks. They proposed variants of partitioning, density-based and hierarchical methods. Jin et al., 2006, studied distance-based outliers by partitioning each edge of the spatial network into a set of segments of the same length. Clauset, Moore, and Newman, 2008, developed a method for hierarchical point cluster analysis on a network.

There are two types of clustering methods, namely

- Hierarchical clustering: objects are united in hierarchically structured clusters.
- Non-hierarchical clustering: objects are allocated to predefined points that form the center of the resulting cluster.

In the following, a method for hierarchical clustering will be introduced. Let $N = (V, L)$ be a planar network with nodes $V = \{v_1, \ldots, v_{n_V}\}$ and links $L = \{l_1, \ldots, l_{n_L}\}$. Let $\tilde{L} = \bigcup_{i=1}^{n_L} l_i$ be the union of the links, that means a set of points forming the links of the network. And let $P = \{p_1, \ldots, p_n\}$ be point-like objects on $\tilde{L}$. The goal is to partition these points into subsets of P such that *clusters* are formed, i.e. $P_i \subset P,\ P_i \neq \emptyset,\ P_i \cap P_j = \emptyset\ \forall i \neq j$ and $\bigcup_{i=1}^{n_L} P_i = P$. This procedure is called *point clustering*.

The initial step of the hierarchical clustering procedure consists of creating an *initial level cluster set* C_0. This is done by defining every point $p_i \in P$ as a cluster of its own:

$$C_0 = \{P_1^{(0)}, \ldots, P_n^{(0)}\} = \{\{p_1\}, \ldots, \{p_n\}\}.$$

To be able to begin the point clustering, a suitable intercluster distance $d(P_i^{(0)}, P_j^{(0)})$ has to be defined for the procedure. A possible candidate would be the shortest path distance. Hence, the pair $(P_*^{(0)}, P_{**}^{(0)})$ with the minimal intercluster distance within every pair of

clusters in C_0 can be identified. This pair will be removed from C_0, instead the union of them will be added to the set and the *first level cluster set* C_1 is found:

$$C_1 = (C_0 \backslash \{P_*^{(0)}, P_{**}^{(0)}\}) \cup \{P_*^{(0)} \cup P_{**}^{(0)}\}.$$

This process is continued iteratively until in the n-th step $C_{n-1} = \{P_1^{n-1}\} = \{P\}$ will eventually be obtained. This way, spatially agglomerated points on a network can be found (Okabe and Sugihara, 2012).

3.4.2 Kernel Density Estimation

The network kernel density estimation is an extension of the kernel density estimation in the plane which is a non-parametric density estimation. Though, the extension is not trivial. A possible approach would be to assume that since the network is embedded in the plane one could treat the points on a network the same as points on a plane. Though, using the kernel density estimation on networks the same as on planes produces bias (Okabe and Sugihara, 2012).

Let q be an arbitrary point in an undirected, connected network $\tilde{L}$ that is embedded in the plane. Let $L_q \subset \tilde{L}$ be a subnetwork such that the shortest path distance between q and any other point on L_q is less than or equal to h. L_q is called the *buffer network* of q with width h.

Okabe and Sugihara, 2012, suggested a construction of an unbiased kernel density estimator using the shortest path distance from q to p on $\tilde{L}$. Let $k(d(q,p)$ be a function called *base kernel density function* that is characterised by

(i) $\int_{p \in \tilde{L}} k(d(q,p))dp = 1$,

(ii) $\exists h \geq 0 : \begin{cases} k(d(q,p)) = 0 & d(q,p) \geq h \\ k(d(q,p)) > 0 & d(q,h) < h \end{cases}$,

(iii) $k(d(q,p))$ is not increasing with respect to $d(q,p)$,

(iv) $k(d(q,p))$ is continuous with respect to $d(q,p)$.

Okabe and Sugihara, 2012, further introduced an *equal-split discontinuous kernel density function* and an *equal-split continuous kernel density function*. But generally speaking, the kernel density function may be defined in various ways as long as the above conditions are satisfied to avoid bias.

3.4.3 Moran's I Statistics

The Moran's I is a spatial autocorrelation statistic, therefore it measures the correlation between nearby attribute values. This concept of distance is integrated via the spatial weights matrix w_{ij}, as we have seen in chapter 2.1.2. In case of a spatial network there are two possible types of spatial weights:

- Topological weights: here the adjacency relationship between network cells is measured, for example $w_{ij} = 1$ if network cells L_i and L_j are adjacent and $w_{ij} = 0$ if they aren't.

- Metric weights: measure the distances between representative points, in the literature this is usually done by using a decreasing function of the shortest path distance such as $\alpha e^{-\beta d(p_i,p_j)}$ or $\frac{\alpha}{d(p_i,p_j)^\beta}$, where $d(\cdot,\cdot)$ is the shortest path distance and $\alpha, \beta > 0$ are constants.

As we have seen before there is a local and global version of the Moran's I statistic. Apart from using either topological or metric weights for the spatial weight matrix, the local and global Moran's I can be calculated for networks exactly the same way as in the planar case. Only three additional steps have to be taken for the network spatial weight calculation, namely the network has to be tessellated into network cells, a representative point has to be defined for each network cell and then the spatial weights between all pairs of network cells have to be calculated.

Let $N = (V, L)$ be a network with nodes $V = \{v_1, \ldots, v_{n_V}\}$ and links $L = \{l_1, \ldots, l_{n_L}\}$. Let $\tilde{L} = \bigcup_{i=1}^{n_L} l_i$ be the union of the links, that means a set of points forming the network.

At first, the representative set $\tilde{L}$ has to be tessellated into network cells $L_1, \ldots, L_n$. This tessellation can result automatically from the data structure. For example, by using the postal codes for a city street network then each district would be a network cell and only the particular boundary points have to be added to N. In case the data doesn't suggest a specific division, a network Voronoi diagram can be constructed. For this, define $P = \{p_1, \ldots, p_n\} \subset V$ and construct a Voronoi diagram for P as discussed in section 3.1.2. Then the Voronoi subnetworks can be used as network cells with representative points p_i.

Second, representative points for the network cells have to be chosen. This can be either done arbitrarily or by taking the center of each network cell. In case the network cells were constructed via a Voronoi subnetwork, the generating points p_i can be used as representative points. Either way, this results in a network $N' = (N', L')$ with $V' = V \cup P$ and L' being the adapted set of links according to the inserted points of P in L.

The third step covers the calculation of the spatial weights w_{ij}. Using the shortest path distance, the weight for every $p_i \in P$ can be easily calculated. Put together, they result in the spatial weights matrix (Okabe and Sugihara, 2012).

Analogously, this procedure can be used to calculate the spatial weights matrix for the Getis-Ord G_i statistic or Geary's c statistic.

Chapter 4

Development of a Python Tool for Spatio-Temporal Hotspot Analysis

In the course of this thesis a `Python` tool was developed to do spatio-temporal hotspot analysis for geographical data in form of point patterns. The goal was to create a tool that can process not only regular maps (for example of countries) but a tool that is also able to handle network shaped maps like a rail system or a street map. Furthermore, it was important that in this development process only open source tools were utilised.

The starting point is that of exploratory data analysis which was originally promoted by John Tukey, 1977. The ambition is to explore the data further than hypothesis testing and model fitting. The approach of exploratory data analysis could possibly lead to additional insights on the examined data. In the course of the underlying thesis, the focus is on the analysis of entities and objects in form of point patterns on networks. Therefore, the initial subsection will discuss the appropriate conceptual foundations and methodology. Then the most important available `Python` packages are introduced, followed by the implementation procedure and the evaluation of the process.

4.1 Conceptual Foundations

First of all statistical methods of spatio-temporal analysis had to be chosen. The starting point of finding an appropriate approach was that of geographical analysis. As stated before, in geographical analysis the focus of study was usually entirely on spatial analysis. Only through technical achievements such as the global positioning system (GPS) and the development of geographic information systems (GIS), it became possible to collect and work with huge amounts of geographical and time referenced data. The need to be able to

handle this data was further pushed by the fact that everybody was carrying electronic devices such as mobile phones and therefore producing endless amounts of geo-referenced data with it.

So, the initial approach to finding a suitable method for spatio-temporal analysis on networks was to explore merely spatial analysis methods. As seen in chapter 2.1 there are three most commonly used methods in spatial cluster analysis, namely Geary's c, Moran's index and Getis and Ord's G_i statistics. First of all, when doing hotspot analysis it is important to use localized spatial statistics. When using global statistics one receives information on signs of stationarity or non-stationarity in the data that is explored. While stationarity stands for 'no clustering' in the data, non-stationarity means that there is a clustering in the data. The disadvantage of global statistics is, that one will indeed know that there is at least one cluster somewhere, but not where it is to be found. That's why localized statistics have to be used when spatial clusters and hotspots are to be identified (Braithwaite and Li, 2007). All three of the introduced spatial statistics have localized versions that are able to identify various types of clusters including hotspots in the data. As seen in chapter 2.1.2, Moran's index is able to detect a clustering of similar values as well as to identify an area that has values of the opposite sign than its neighbours. As the intention of this thesis is to do hotspot analysis and with that finding hotspots as well as coldspots, i.e. areas of clusterings of high values and areas with a clustering of low values, Moran's index would lead only half the way. Namely in finding areas with similar values but it would not give any information on if it's a clustering of high or low values per se. Still, Moran's index could be a possibility to progress since further investigation of the data will reveal if it's a clustering of particularly high or low values. Another advantage of Moran's index would be that there is already a first approach on a spatio-temporal Moran's index that could be an option (see chapter 2.3.1).

As Geary's c works with squared differences, it measures how similar neighbouring values are. While an outcome of a high value would mean that there's a big difference between neighbours and therefore negative spatial autocorrelation, a low value simply means that neighbours have similar values. This defines no clustering of high or low values yet, Moran's index would categorise it probably simply as weak or even zero autocorrelation. That's why Geary's c will not be the approach chosen for hotspot detection in the present thesis.

The Getis-Ord G_i and G_i^* statistic identifies clusters of high values as well as clusters of low values by producing a significantly high or low outcome with reference to their average. Therefore, it seems like a good approach for the detection not only of hotspots

but also of coldspots within one step. So, there are two possible candidates as the starting point for our spatio-temporal hotspot analysis. Both of them have in common that they are highly dependent on a spatial weights matrix. As we have seen in section 3.4.3, the spatial weights matrix can be adapted to be applicable on network space. Therefore, this wouldn't pose a complication either. Consequently, a closer look on both methods is needed. While either procedure needs three items for their analysis, namely a study area, features with values and a neighbourhood for each feature, Moran's index and the G_i statistic approach the scan for clusters differently. The Getis-Ord G_i (or G_i^*) statistic takes a look at each feature's neighbourhood excluding i (including i) to determine if their values are significantly different than those from the study area. If so, feature i is part of a hotspot/coldspot (ESRI, 2021c). If continuing with the Getis-Ord statistic, the G_i^* method would be chosen because it includes feature i in its calculation. On the other hand, local Moran's index first compares feature i's neighbourhood (excluding feature i) to the study area to see if they differ significantly. Then i is checked against its own neighbourhood. That way outliers within hotspots can be found, i.e. high values within a coldspot or low values within a hotspot. This method is of good use when looking for abnormal or unexpected trends (ESRI, 2021a).

The Getis-Ord G_i^* statistic can be preferred over its alternatives in identifying hotspots for several reasons. First, it best approximates the typical definition of a hotspot by identifying areas where local averages are significantly higher than global averages (Chainey and Ratcliffe, 2005). Unlike local Moran's I, the G_i^* statistic does not require that in a geographic neighbourhood both feature i and its neighbours display significantly high or low values in comparison to the study area itself. Furthermore, G_i^* has the ability to identify features with high values even if they are located within a neighbourhood of high values (Braithwaite and Li, 2007). In the plane, the Getis-Ord G_i^* statistic has been used for hotspot detection in various contexts. For example for analysing spatial clustering in patterns of voting (O'Loughlin, 2002), clustering of incidents of criminal activity (Chainey, Reid, and Stuart, 2002) or for identifying transnational terrorism hotspots (Braithwaite and Li, 2007).

Since it transpires, that the introduced state-of-the-art spatial analysis techniques are applicable to networks, it needs to be examined if they are (in combination with a time series analysis) preferable over the introduced spatio-temporal methods. Lee and Li, 2017, suggested an approach for a spatio-temporal Moran's I statistic by modifying the spatial weights matrix for it to include not only the spatial but also the temporal component. In a reply to a comment on their paper "Extending Moran's Index for Measuring Spatiotemporal Clustering of Geographic Events", Lee and Li, 2018, stressed the fact that space and time

have different units that are not compatible and can therefore not be integrated directly. Both the binary adjacency approach as well as the multiplicative principle are probable ways that are currently available for addressing the problem of space and time interaction. Still, the way events are related to each other spatio-temporally are totally different than simply spatially or merely temporally. On a spatial basis, events can influence each other mutually, that means that it is possible that two events A and B influence each other to the same extent or probably that event A influences event B to a greater extent than event B influences event A. If considered with a temporal point of view, the situation is not at all similar. While past events can influence what is happening in the present, current events would not change incidences of the past. That's why, their integration of time into a spatio-temporal weights matrix has to be considered only as a first step to a systematic treatment of spatio-temporal analysis (Lee and Li, 2018).

The space-time scan statistic is a window statistic that identifies the maximum size of clusters. The original method was extended by Kulldorff to be applicable to the plane at first and to space later on. It was developed for epidemiological use cases such as the detection of disease clusters or early detection of disease outbreaks. For the space-time scan statistic an observation window (or cylinder in the 3-dimensional space) is created for scanning the study area. Therefore, the Euclidean distance is entangled with the core of the statistic. As we have seen, the Euclidean distance leads to severe misconceptions when used in network spatial analysis. Therefore, this spatio-temporal statistic is abandoned as an option for the development of a spatio-temporal analysis method for networks.

The third method that was discussed is the space-time kernel density estimation. As we have seen in chapter 3.4.2, the planar kernel density estimation is transferable to network space. For the space-time kernel density estimation a probability density function for every point in space and time is calculated that represents the likelihood of an observation being made at the corresponding location. In case of using the kernel density function for hotspot detection, Kalinic and Krisp, 2018, noted that these can't be considered statistically significant because the parameter of bandwidth has to be specified by the practicioner and therefore results in a problem of subjectivity. That's why the method would have to be used in combination with other hotspot detection methods such as the Getis-Ord G_i statistic. With this all three introduced state-of-the-art approaches have to be dismissed in case of spatio-temporal hotspot analysis on networks.

As a result, the choice for the spatio-temporal analysis in the present thesis will be a two stage process. The international supplier of geographic information system software ESRI (Environmental Systems Research Institute) proposes a two level approach for emerging hotspot analysis. Their suggestion is to do a spatial analysis via the G_i^* statistic and then

continue with a trend analysis with help of the Mann-Kendall trend test (ESRI, 2021b). Since we are doing an exploratory data analysis this could be a valid option. An additional advantage of a two stage process is, that is suffices to apply the merely spatial analysis technique to the network which significantly reduces the effort for implementing adaptions to existing statistics in `Python`.

As we've seen in chapter 2.2.4 the Mann-Kendall trend test is a rank correlation test. Usually, when doing a time series analysis or trend analysis the distribution of data needs to be known or estimated. One of the main advantages of a rank test is that it is distribution-free. That means, that its power and significance is not biased by the actual distribution of the data. That's the reason why the Mann-Kendall trend test has been widely used for testing trends in natural phenomena such as temperature, rainfall or water quality series where the actual distribution is unknown (Hamed, 2009). Another advantage the Mann-Kendall trend test poses, is that due to the ranking procedure the effect of outliers is diminished (Kendall, 1962). The null hypothesis that is tested by the Mann-Kendall trend test is that the data is independent and randomly ordered. However, in case the data is autocorrelated the probability of detecting trends, that actually don't exist or not detecting trends when they exist, increases (Hamed and Rao, 1998). So the disadvantage of the Mann-Kendall trend test is that the data has to be subjected to a pre-whitening process to remove autocorrelation from the data. Unfortunately, this process is problematic of its own, Yue et al., 2002, showed that using pre-whitening to remove positive autocorrelation also resulted in weakening the magnitude of an existing trend and therefore it 'potentially leads to inaccurate assessment of the significance of a trend'. Hamed and Rao, 1998, introduced a second method to mitigate the effect of autocorrelation and leave the data intact at the same time by modifiying the variance. Moreover, when the data is not autocorrelated the altered formula reduces to the original version of calculation. In case the Mann-Kendall trend test is chosen, this will be the approach taken.

Both, the first order autoregressive model, AR(1), as well as the first order moving average model, MA(1), have been widely used to describe natural processes. The AR(1) model is characterised by an exponentially decaying correlation function and is therefore often categorised as a short-term correlation process (Hamed, 2009). The AR(1) model as well as the MA(1) model and therefore also the ARMA(1,1) model are linear processes that are defined by linear equations with constant coefficients (Brockwell and Davis, 2016). Hamed and Rao, 1998, explored the effect of autocorrelation on the AR(1) and the MA(1) model. They found that for both models the variance is increased by impacts of autocorrelation. In case of the AR(1) model though, this effect is much larger than for

the MA(1) model. This is due to the fact that for the AR(1) model the autocorrelation extends beyond the first lag. On the other hand, Yue et al., 2002, showed that their impact on the estimates of the slope of the trend is almost identical to that of the Mann-Kendall statistic. While the variance of the estimators of the slope are altered, the mean and distribution type stays the same. Since there is an approach for the Mann-Kendall trend test to assess and reduce the effects of autocorrelation in the data, that at the same time reduces to the original version of the Mann-Kendall trend test in case no autocorrelation is present, the Mann-Kendall statistic will be chosen over the AR(1) and MA(1) model.

Following ESRI's approach, a two stage method with a spatial hotspot analysis via Getis-Ord's G_i^* statistic followed by a trend analysis with help of the Mann-Kendall trend test is pursued. The next step is to review existing `Python` packages to see if some statistics or mechanics are already implemented and could be enhanced and adapted for the proposed approach. In any case, a network based implementation for the Getis-Ord G_i^* statistic is needed as well as an implementation for the Mann-Kendall trend test with adaptions made to the variance calculations to incorporate the possibility of autocorrelated data.

4.2 Availability of Existing Python Packages

`Python` is a programming language that incorporates a great variety of libraries and packages to cover a wide range of research fields and analysis methods. The following chapter will first introduce the basic packages and libraries that will be relevant for the tool implementation. Then, the `Python` spatial analysis library `PySAL` will be introduced in more detail. Here, the included packages that could be of interest when proceeding with the implementation will be presented. Note though, that not all of them were used in the course of the tool implementation but they were rather a first collection of possible starting points.

4.2.1 Basic Packages for Data Analysis

In this first subsection, the most important libraries for basic computations and data analysis are introduced. Note, that this list is far from complete and their importance and utility can vary depending on the programmer.

NumPy

`NumPy` stands for numerical mathematics, the library is specialised in working with multi-dimensional arrays and matrices. For those, it contains many numerical computing tools such as universal functions (e.g. random number generators), linear algebra routines, tensors, nearest neighbour search, etc. Using the `NumPy` library gives `Python` a lot of functionalities previously known from `MATLAB` ("NumPy", 2022).

SciPy

`SciPy` is a broadly applicable library for scientific computing that extends `NumPy` and adds even more `MATLAB` functionalities. The SciPy library includes among others hierarchical clustering, modules for optimisation, numerical integration and linear algebra routines, algorithms for spatial structures or statistical functions ("SciPy", 2022).

matplotlib

`Matplotlib` is a plotting library that is strongly connected to `NumPy`. It is possible to create not only static but also animated and interactive visualisations such as line plots, histograms, scatter plots, 3D plots, image plots, contour plots and polar plots ("matplotlib", 2022).

pandas and geopandas

The package `pandas` was developed to be able to do spatial analysis in `Python`. It is based on or more specifically enhances the libraries `SciPy` and `matplotlib` but in particular `NumPy`. In `pandas`, large datasets can be handled more efficiently and intuitively in form of DataFrames. Therefore, many functionalities of the above packages are already included, though, since not every feature was improved, especially `NumPy` is often used together with `pandas`.

`Geopandas` is a further extension of pandas, thanks to this module it is possible to process inputs for spatial data in the form of Shapefiles or a PostGIS database. Many spatial analysis tools are included such as overlay analysis, geocoding, spatial aggregation methods and additional GIS-functionalities (Tenkanen, 2017).

NetworkX

The `NetworkX` package was developed for the construction and manipulation of networks. It includes tools to study their structure and dynamics as well as functions for complex

networks. It incorporates data structures for graphs, digraphs, and multigraphs, standard graph algorithms and network structure and analysis measures such as finding subgraphs, cliques, k-cores or exploring adjacency, degree, diameter and betweenness. Furthermore, edges can have attributes such as weights or a time series ("NetworkX", 2022).

4.2.2 PySAL

`PySAL` is a `Python` library that contains a great variety of packages concerning geo-computation and spatial data science. Their main functionalities can be categorised into: weights, spatial econometrics, spatial dynamics, ESDA (exploratory spatial data analysis), clustering and computational geometry (Rey and Anselin, 2010). Because of its extensive growth and the great activity in further development, the library was restructured to facilitate the continuous advancement and refinement. `PySAL` 2.0 was released in 2019 (Rey et al., 2021).

Because of its great extent, in the following only the `PySAL` packages that could be of interest for the implementation in the course of this thesis will be introduced. The following descriptions are based on the paper 'The PySAL Ecosystem: Philosophy and Implementation' by Rey et al., 2021.

libpysal

The `libpysal` package contains the fundamental algorithms that are underpinning all other packages. It includes the *input/output* module that offers the possibility to work with geospatial file formats. Further, it contains a *computational geometry* module that includes several algorithms to transform geometric shapes, for example Voronoi tessellations. Additionally, it offers *example data sets*. But most interesting for the present thesis is the *weights* module, it provided the possibility to store spatial weight matrices and tools for manipulating and operating on them.

esda

`esda` - exploratory spatial data analysis - investigates spatial patterns in the data. It includes the exploration of *spatial dependence* of random spatial processes on their nearby realisations, *spatial heterogeneity* of them, i.e. when a nearby process exhibits different behaviour and *spatial autocorrelation*, the statistical dependence of the attribute values of a given variable in context with other nearby measurements.

The methods included in the `esda` package are applicable to continuous or binary

data and can be conducted either on local or global scale. Furthermore, statistics about boundary strength and measures of aggregation errors in statistical analyses are available.

Note, that the `esda` is primarily designed for employment on the plane and includes implementation of the Moran's I, Moran's scatterplot, Getis-Ord G_i and G_i^* statistic on a global and local scale.

giddy

`giddy` stands for geospatial distribution dynamics. It functions as an extension to the `esda` package and offers analysis tools for spatio-temporal data. The role of space is examined in the light of its dynamic over time. `giddy` mainly contains Markov chain models such as spatial Markov, LISA Markov, Full Rank Markov and Geographic rank Markov models.

spaghetti

To overcome the fact that the Euclidean based framework is not suitable for the analysis of networks, the package `spaghetti` - spatial graphs: networks, topology and inference - was developed. For the study of networks and corresponding statistical processes, it offers the appropriate data structures and analytical methods. Most importantly, it includes the Dijkstra algorithm for the shortest path determination which represents the basis for all distance based statistical methods on networks. Furthermore, it provides functionalities such as the possibility to map near-network observations onto the next nearest link or high-performance geometric and spatial computations using `geopandas` for high-resolution interpolation along networks.

splot

`splot` is a visualisation package for spatial analysis. It includes the visualisation of global and local spatial autocorrelation via the Moran scatterplot or cluster maps. Further, it is possible to visualise the temporal analysis of cluster dynamics through heatmaps or rose diagrams. `splot` is also capable of creating multivariate choropleth maps. It can be assessed via the package `matplotlib` or `bokeh` to allow for static visualisations for publications and interactive visualisations for the purpose of further investigation.

4.3 Implementation

In this section, the implementation of the `Python` tool for spatio-temporal hotspot analysis is described in more detail. The tool aims at the processing of taxi movement data in form

of point patterns and evaluating them on the OpenStreetMap (OSM, "OpenStreetMap", 2022). The data is provided by the data owner 'Taxi 31300' and the Austrian Institute of Technology,

The framework of the implementation is described in more detail in section 4.3.1. This is followed by a section on data handling (4.3.2) and details on the execution of the implemented tool (4.3.3).

4.3.1 Framework

Since a two stage approach to the spatio-temporal hotspot analysis on networks was chosen, it is primarily necessary to have a look at spatial analysis on networks. As we have seen in chapter 3.4.3, the statistics that were introduced for the planar spatial analysis (section 2.1: Moran's I, Geary's c and Getis-Ord G_i^*) can be transferred to network space relatively easy. This is done by modifying the spatial weights matrix, two ways of doing so were discussed in the corresponding chapter. The first possibility is to define topological weights that reflect whether two links are connected. The second possibility suggests a decreasing function of the shortest path distance, this is the option chosen for the implementation of the spatio-temporal hotspot analysis tool.

To add the temporal scale, the data for investigation is split into multiple time slices. The size of those time slices can be chosen freely by the user according to the inspected data. So in case, a week has to be analysed, the size of these time slices can be set to hours. In case years have to be analysed, it is probably more useful to set the time slices' size to days or weeks. This is of course an incomplete list of the possible sizes to be chosen from, to name just the most common ones, see table 4.1.

Alias	Description	Alias	Description
D	Calendar Days	H	Hours
W	Weeks	min	Minutes
M	Months	S	Seconds
Q	Quarters		
Y	Years		

Table 4.1: Most commonly used frequency aliases. Note, that the aliases can be modified as well, for example '30S' for thirty seconds. For a complete list have a look at "Offset Aliases", 2022.

Then, the Getis-Ord G_i^* statistic is performed for each of these time slices. That way, a time series is created for the individual links, labeling them as hotspot, coldspot or

no point of significance. The hotspot detection with the Getis-Ord G_i^* statistic can be performed based on different confidence levels. In the proposed analysis process, hotspots and coldspots are categorised in three classes depending on their significance. So, there is a grading of three levels of detected hotspots and coldspots within a significance level α of 1%, 5% or 10%. Based on that, a time series forms that can be ranked based on this categorisation, it can then function as an input for the Mann-Kendall trend test.

In section 2.2.4, it was mentioned that the Mann-Kendall trend test is a rank correlation test. That implies that the given time series is ranked based on its individual values and then it is evaluated if any trend can be detected within this ranking compared to the natural ranking $1, \ldots, n$. That way, it can be assessed if an upward trend, a downward trend or no trend at all exists within the series. In the proposed approach, the Mann-Kendall trend test is only applied if within the time series of one link more than 90% of the time steps are categorised as demonstrating a point of significance, i.e. more than 90% of the values identify as a hotspot or more than 90% of the values mark a coldspot.

As the approach of exploratory data analysis was chosen, it is possible that more interesting patterns can be detected than just an increasing or decreasing trend. Therefore, eight patterns have been defined to categorise the results of analysis. These patterns were inspired by ESRI, 2021b, they are described in the following list in terms of hotspots. Of course, for the implementation they are defined for coldspots analogously. An illustration of the various patterns can be examined in figure 4.1.

(0) *No pattern detected*: that means that none of the following definitions apply.

(1) *New hotspot*: The only time a hotspot was detected in this area was in the last time step.

(2) *Consecutive hotspot*: There is a sequence of significant hotspots in the last time steps, though it has never been a hotspot before the start of that sequence. Furthermore, there are less than 90% of the values categorised as hotspot or coldspot. In the implementation it will be possible to specify a value or percentage on how many of the last time steps in the time series have to be categorised as a significant hotspot to end up being labeled as a consecutive hotspot.

(3) *Intensifying hotspot*: is detected by the Mann-Kendall trend test, the significance level for the detection of a trend is set to 5%. Here, at least 90% of the time steps a hotspot of the three intensity categories has to be detected for the time series to be analysed by the Mann-Kendall trend test.

(4) *Persistent hotspot*: is identified by the Mann-Kendall trend test in case there are more than 90% of the time steps categorised as a hotspot by the Getis-Ord G_i^* statistic, but the Mann-Kendall trend test can not find an increasing or decreasing trend within the significance level of 5%.

(5) *Diminishing hotspot*: functions analogous to the intensifying hotspot. It is detected by the Mann-Kendall trend test within a significance level of 5%. The Mann-Kendall trend test will be applied when at least 90% of the time steps are categorised as a hotspot of the three intensity categories.

(6) *Sporadic hotspot*: in case of the sporadic hotspot it means that at some point one time step has been categorised as a hotspot but then again no hotspot was detected. At the same time in the whole series no time step was ever labeled as a coldspot.

(7) *Oscillating hotspot*: in case of the oscillating hotspot the corresponding time series shows a hotspot in the last step but also a coldspot in at least one of the previous time steps. Furthermore, the user has the possibility to fix a value or percentage of preceding label switches from hotspot to coldspot.

(8) *Historical hotspot*: is found if indeed more than 90% of the time steps were categorised as hotspots but at least the last time step was no hotspot.

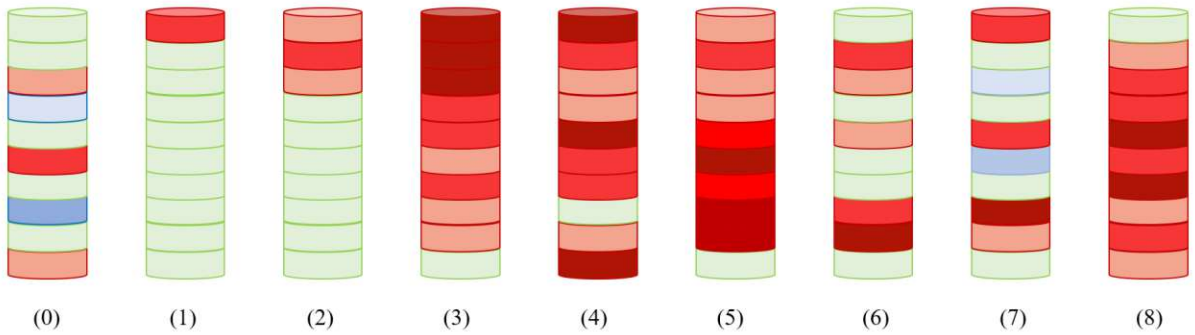

Figure 4.1: This graphic illustrates a possible realisation of the various patterns. Hotspots are colored red and coldspots are colored blue. The intensity of the significant spots can be deduced from the given color.

Note, that not every pattern requires that in more than 90% of the time steps a significant hot- or coldspot was detected. Figure 4.2 shows the possible ways of finding the individual patterns described above. While patterns 0, 1, 2 and 6 are detected when less than 90% of the time steps show significant spots, pattern number 7 can be detected

in any case. On the other hand labels 3, 4, 5 and 8 need a categorisation of the individual time steps as hotspots or coldspots for more than 90% of the time. If the last time step wasn't categorised accordingly, the pattern is ruled out as a historical hotspot. So, only in cases 3, 4, 5 the Mann-Kendall trend test is applied. All other labels are based on the fact that the time series didn't qualify for the Mann-Kendall trend test or only less than 90% of the time steps are categorised as significant.

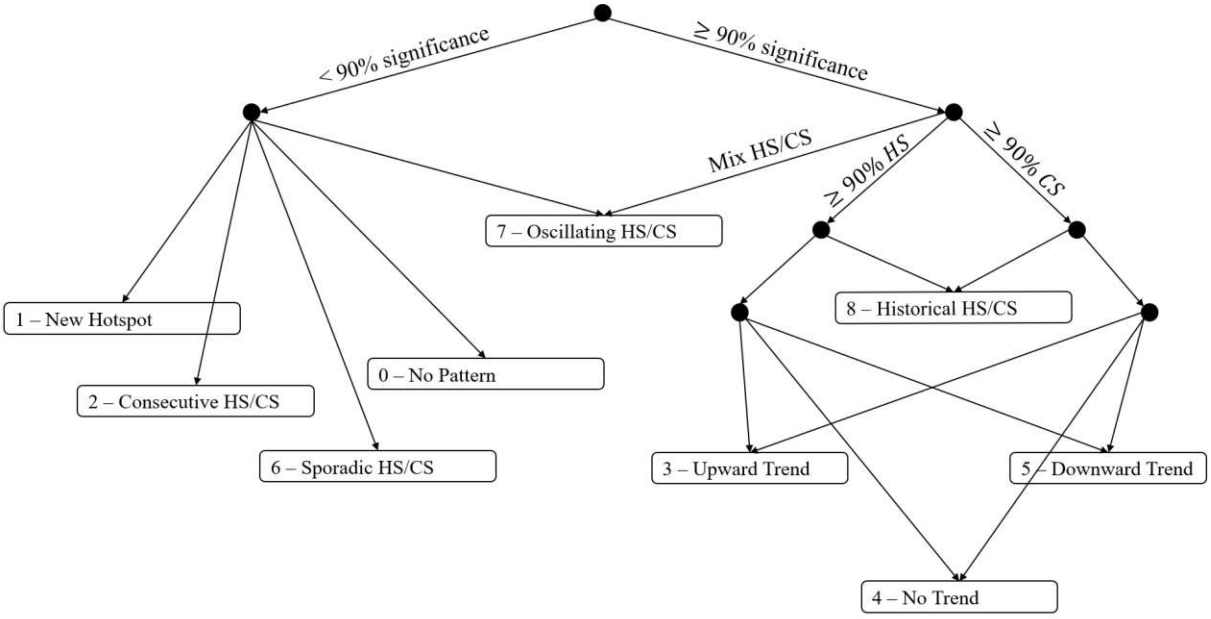

Figure 4.2: This graphic shows the possible categorisation and gives an overview on the corresponding requirements.

4.3.2 Data Handling

For the `Python` tool for spatio-temporal hotspot analysis it is assumed that the data is provided in form of shapefiles. `.shp` is a non-topological file format for the purpose of storing geometric location and attribute information of geographic objects, these can be points, lines or polygons (ESRI, 2021d). Therefore, both the links for the network creation and the point patterns can be stored in the form of shapefiles.

The `Python` package `spaghetti` is able to process these shapefiles and map the observations in form of point patterns onto the nearest link of the provided network. That way, for both the network and the point pattern a network/point pattern object and a corresponding `geopandas` DataFrame is created to store the data of the links and points. These can be used to start the analysis process, for further details see the next section

4.3.3.

The results of the analysis consist of a table of links with the corresponding patterns. This table can be used for further analysis and processing or it can easily be extracted to a `.xlsx` or `.csv` file. Furthermore, the network map and point patterns can be plotted. To make the different patterns clearly visible and pass on a visual impression, the network map can automatically be plotted with the edges coloured according to their patterns. For an example have a look at chapter 5.

4.3.3 Application

For the implementation, three main `Python` classes were defined that form the core. These include the `NetworkStatCalculator`, the `TimeSlicer` and the `PatternFinder`.

NetworkStatCalculator

As can be deducted from the chosen name, the `NetworkStatCalculator` contains all operations that are directly related to the network itself. This includes among others the construction of the network object and filtering the longest component, the mapping of the point patterns onto the nearest link as well as computing their count, this count can be regarded as the attribute value of the corresponding link. In addition, the associated dataframes for both links and points can be constructed. Furthermore, the Getis-Ord G_i^* realisation as well as a function for plotting its evaluation is included in this class. Note, that the G_i^* statistic is already implemented for planar spatial analysis in the package `esda` and was adapted to network applicability based on the results discussed in 3.4.3.

As Okabe and Sugihara, 2012, noted, it is desirable for statistical tests that the network links are split into equally spaced segments because then no adaptions have to be made later on to balance these inequalities to prevent bias. So, the `NetworkStatCalculator` offers the possibility to specify the preferred arc length and splits the links accordingly. The functions to realise these implementations are all included in the `spaghetti` package.

TimeSlicer

The `TimeSlicer` class introduces the functionality to slice huge amounts of data into time slices in the size specified by the user. The purpose is that afterwards each time slice is analysed using the Getis-Ord G_i^* statistic such that a time series forms for every link. In each time step the links are labeled with 0 if no hotspot or coldspot was detected by the statistic, labels 1, 2 or 3 stand for the three defined intensities of hotspots (in accordance with significance levels 10%, 5% and 1%) and analogously -1, -2 or -3 for the three types

of coldspots (in accordance with significance levels 10%, 5% and 1%). So, the `TimeSlicer` class opens the pathway to further pattern analysis based on the resulting time series.

PatternFinder

The name `PatternFinder` already suggests that this class contains all pattern definitions as well as logic on time series analysis to arrive at these patterns. The required auxiliary functions were also implemented in this class. As depicted in chapter 4.3.1, there were eight possible patterns defined. Three of those patterns include the application of the Mann-Kendall trend test. For the Mann-Kendall trend test implementation, the existing implementation of Kendall's τ in the library `SciPy` was used as a basis and adapted. As mentioned in 4.1, the trend test was adapted such that existing autocorrelation will be leveled out through an adaption in the calculation of the variance as proposed by Hamed and Rao, 1998.

In the course of the implementation the following `Python` libraries and packages were used: `NumPy`, `SciPy`, `pandas`, `geopandas` and from the `PySAL` library: `libpysal`, `esda`, `spaghetti` and `splot`.

4.4 Evaluation

The implementation of a spatio-temporal hotspot analysis tool that was introduced in this chapter is a first suggestion on how to tackle the vast topic of spatio-temporal analysis not only in the plane but also on network shaped research areas. It exemplifies not only the analysis statistics that could be used but also the definition of patterns that could be detected and of interest following the approach of exploratory data analysis. The application of the proposed analysis workflow will be discussed in more detail in the next chapter (5) that deals with the analysis of taxi entry and exit points within the city of Vienna.

In future enhancements of the tool, it would be desirable to eradicate certain flaws that became apparent within the implementation process. These include long computation time of the functions used from `spaghetti` but especially the issue that when using `spaghetti` all link properties in the data is lost in the reading process and only point positions will be maintained for analysis, should be solved (for reference see that the corresponding ticket that was posted to the `spaghetti` support channel when the tool was first implemented and is still open Gaboardi, 2020). This issue actually prevented the implementation of additional evaluation possibilities concerning attribute values such as speed of the car in

case of the taxi data that was processed in the case study (chapter 5). Furthermore, a comprehensive documentation and workflow guide could be formulated such that the tool increases in practicability and user-friendliness.

The validity of the tool, existing functions that were modified for the purpose as well as newly implemented features were tested in-depth using unit tests. It is available for inspection and usage on GitHub.

Chapter 5

Case Study: Analysing Taxi Movement Data

This chapter elaborates on the case study of taxi movement data in the city of Vienna. The implementation of a `Python` tool for spatio-temporal analysis is demonstrated investigating the point patterns of taxi entry and exit points in the area of the railway stations Wien Westbahnhof and Wien Hauptbahnhof in Vienna. The station Wien Westbahnhof used to be the most important railway junction in Vienna until the new Hauptbahnof was taken into full operation on the 13^{th} of December in 2015. From that moment onward every Railjet, Intercity, Eurocity and Euronight train to and from the west of Austria and Europe, departed and arrived from the Hauptbahnhof instead of the Westbahnhof as before (Schrenk, 2015). Though, the launch of the Hauptbahnhof consisted of multiple stages, the first tracks went into operation in December 2012, due to timetable changes in December 2013 and December 2014 more and more trains were transferred from the Westbahnhof to the Hauptbahnhof ("Eröffnung des Wiener Hauptbahnhofs in Bildern", 2014).

In the scope of this thesis it will be investigated if due to this gradually commissioning, a change in hotspot patterns of taxi stops can be detected. In particular, it will be interesting to see if there are diminishing hotspots around the Westbahnhof while there are emerging hotspots in the area of the new Hauptbahnhof with every launching phase.

5.1 Data and Study Area

The data that is used for the case study was provided by the data owner 'Taxi 31300' and the Austrian Institute of Technology. It comprises taxi movement data within the city

of Vienna from September 2012 until the end of June 2016. This time frame is broadly formulated such that both the time before the first launch stage of the Hauptbahnhof was coming into effect as well as the time after the final launch stage was completed.

The data is provided in form of `.csv` files. For building the network, links are characterised by their ID for OpenStreetMap and their well-known-text, which is a form or representation for vector geometry objects that includes the minimum and maximum longitude and latitude. Further information that is provided includes each link's junctions, length, means of transportation (foot, bicycle, car, etc.) and the information if they are one-way. In case of the taxi data, the provided information includes the trip ID, its starting and ending time as well as their starting and ending points in form of specification of longitude and latitude such that they can be mapped correctly onto the corresponding network.

The study area comprises of the region around the station Wien Westbahnhof and Wien Hauptbahnhof, as can be seen in figure 5.1. This area also includes the important rail station Wien Meidling. Every train starting or ending at the new Hauptbahnhof will also pass through Wien Meidling. Therefore, it presents a notable point of interest as well. While the Hauptbahnhof is connected to a different underground line than the station Wien Meidling, almost as many people will enter and exit trains at this station as they do at the Hauptbahnhof itself.

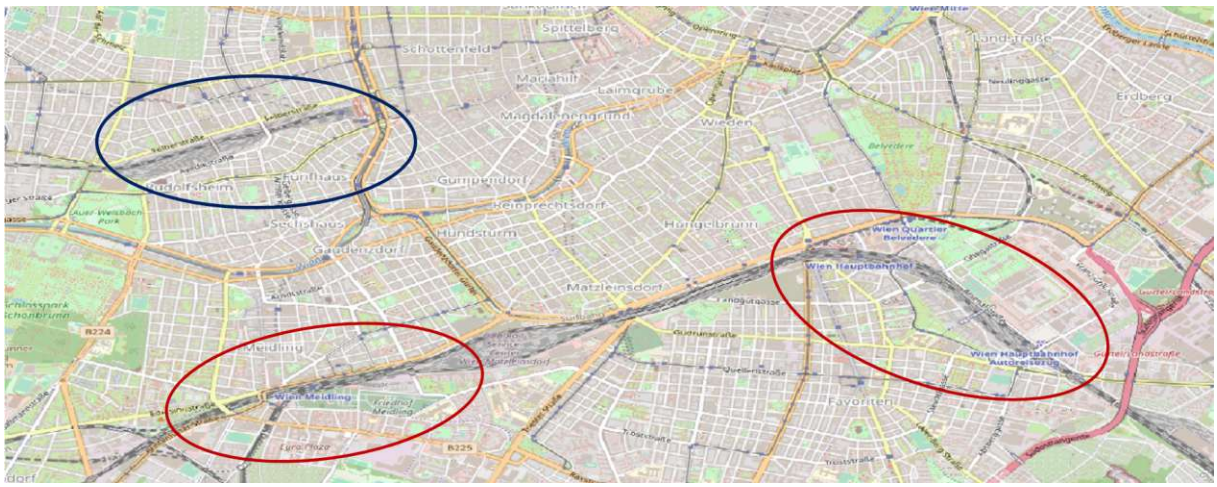

Figure 5.1: This is the research area of this case study, the Westbahnhof is marked in dark blue and the station Meidling (left side) and the Hauptbahnhof (right side) are highlighted in red ("OpenStreetMap", 2022).

Preprocessing of Data

The data for the links illustrate the whole city of Vienna. So, as a first step only the links are extracted that are available for car traffic, then these links have to be reduced to the specified research area (see figure 5.1). This can be done with the help of QGIS, which is a software for geographical information systems ("QGIS", 2022). Here, one can generate a vector layer containing the links of all Vienna. Then one can create a new layer with the area of ones choice, containing only the links within this area. So, for this case study the street network extracted for further analysis resembles the specified area, see figure 5.2. This vector layer that was generated in QGIS and then saved as a shapefile such that it suits the conditions for further processing with the tool.

Figure 5.2: Vienna's street network surrounding the stations Wien Westbahnhof, Wien Meidling and Wien Hauptbahnhof.

Note, that due to the selection of this area, it can happen that there are streets that are not connected to the network. The automatic removal of those unconnected links is implemented in the spatio-temporal analysis tool and therefore can be ignored during the preprocessing.

Just as the links, the provided taxi data is spread over the whole of Vienna as well. Since the data is originally stored in a database, one can already filter all trips that have their starting or ending geometry within the specified research area. The result can then be extracted in form of a `.csv` file. Though, this still leaves us with more than 14 million trips from September 2012 till June 2016. Approximately half of them starting in the specified area and half of them ending within the defined street network. Again, as determined in chapter 4.3.2, the inputs for the spatio-temporal hotspot analysis tool have to be in form of shapefiles. Similar to the links, the data for the point patterns can be handled by

QGIS to meet this condition. Since there are so many trips within the broad time frame chosen for the analysis, the `.cvs` data first has to be split into several parts, for each part a point layer is created in QGIS and afterwards all those layers are merged again. Then, this final layer can be saved as a shapefile that can be processed by the spatio-temporal analysis tool. As one has to specify if the starting geometry or the endpoint geometry has to be used when working with QGIS, the above process has to be done seperately for trips starting or respectively ending in the research area.

5.2 Spatio-Temporal Analysis Workflow

In the following section the spatio-temporal analysis workflow will be described in more detail. As there was no graphical user interface developed, the presented screenshots in this section are taken directly from the command lines in a `Jupyter Notebook`.

After the preprocessing procedure, one is left with one shapefile containing the links of the network and two shapefiles containing the information about the point patterns that are to be examined. One for the starting geometry of each trip and one for the endpoint geometry. The following description of the analysis workflow is demonstrated on the example of examining the taxi exit points and therefore the endpoint geometry of the point data. As a first step, this data has to be read into the notebook to be able to process it by the analysis tool, see figure 5.3.

```
In [ ]: file_dir = 'C:/User/Spatio-Temporal_Hotspot_Analysis/shp-files/'
        network_file = os.path.join(file_dir, "bhf.shp")
        pp_file = os.path.join(file_dir, 'ppt_bhf_end.shp')
```

Figure 5.3: First command to read in the data in form of shapefiles for the street network and the point patterns.

For the network data two further steps are necessary (figure 5.4). First, the shapefile has to be processed in form of a `geopandas` DataFrame, this is necessary to be able to plot the street network and visualise the results. Second, the network data has to be processed with the `Python` package `spaghetti` such that a network object is created that can be processed in the course of the point pattern mapping. Since a tool always aims at relieving the user from tedious work, this is done automatically by using the implemented class of the `NetworkStatCalculator` which constructs the required network object (see section 4.3.3).

As can also be seen in figure 5.4, for the point patterns there are also a couple of initial steps to be taken before the tool functions automatically. Similar to the network data, the

```
In [ ]: network_gdf = gpd.read_file(network_file)
        bhf = NetworkStatCalculator(network_file)

In [ ]: pp_stops = gpd.read_file(pp_file)
        pp_stops["end"]=pd.to_datetime(pp_stops.end)
```

Figure 5.4: For both network data as well as point patterns initial steps have to be taken. The data has to be processed as a geodataframe and timestamps have to be transformed to a datetime format.

point patterns have to be processed as a `geopandas` DataFrame. This is necessary such that the point patterns can be mapped onto the provided network links. It is a functionality provided by the package `spaghetti` that is used within the tool implementation. Then, for `Python` to be able to work with timestamps, the corresponding columns of the dataframe have to be transformed to a datetime format. In case of this demonstration, this is done to the column `end` which contains the end time of each trip.

Now, the data is prepared to be processed by the function `get_pattern_as_gdf` which is implemented in the class of the `NetworkStatCalculator`, see figure 5.5. This function maps the specified point patterns onto the network object (using the package `spaghetti`, section 4.2.2), creates time slices of the requested size (using the class `TimeSlicer`, section 4.3.3), applies the Getis-Ord G_i^* statistic (section 2.1.4) and the Mann-Kendall trend test (section 2.2.4) to categorise each link with its corresponding pattern. Furthermore, the defined value corresponds to the value that has to be specified for patterns 1 - consecutive hotspot and 7 - oscillating hotspot (for more details see section 4.3.1).

```
In [ ]: pattern_gdf = bhf.get_pattern_as_gdf(pp_shp=pp_stops, time="end", slice_size='m', value=7)
```

Figure 5.5: The *get_pattern_as_gdf* function is part of the *NetworkStatCalculator* class. Therefore, it is applied to the created network object and takes the analysis parameters as input.

In the above figure 5.5, the network object `bhf` is used and utilised as a basis for the analysis of the previously read-in point patterns `pp_stops`. Furthermore, the time column that is to be used for creating the time slices is the column `end`, the size of the time slices is that of a month and the value is set to seven. The result of this function is a `geopandas` DataFrame that provides information on every link, namely its ID, geometry and the detected pattern. Together with the initially defined `network_gdf` the `pattern_gdf` can be processed by the `plot_pattern` function to visualise the calculated results as can be seen in figure 5.6.

Note, that the classes `TimeSlicer` and `PatternFinder` are implemented and used by the `NetworkStatCalculator` in the background. Therefore, the class `NetworkStatCalculator`

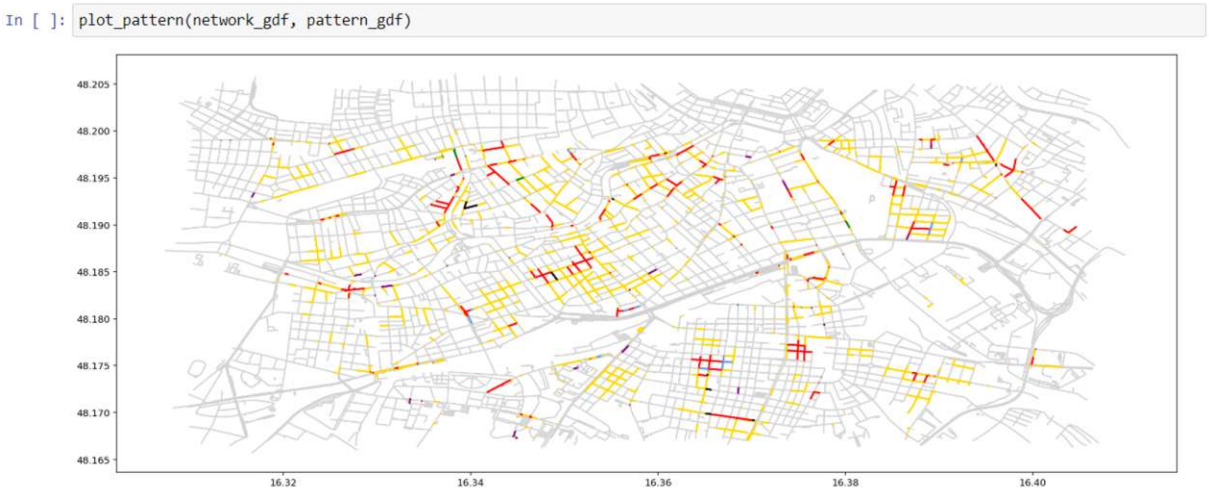

Figure 5.6: The *plot_pattern* function uses the dataframes for network and pattern data as an input and plots the corrsponding pattern visualisation. Note, that this image is only for demonstration purposes, the detailed analysis of the case study will be discussed in the following chapter 5.3.

is the only one the user of the analysis tool has to interact with.

5.3 Results

The provided taxi movement data in the city of Vienna covers the period from September 2012 until June 2016. Every year in the middle of December a new part of the new Hauptbahnhof went into operation. That is why, in the course of the analysis, for once the period as a whole was analysed with help of the spatio-temporal hotspot analysis tool. Secondly, every year the period from September until the end of March in the next year was examined, this provided a time window of approximately the same length before and after the launch date.

Moreover, the taxi stop data comprises both rides with the exit point within the research area and trips with the starting point within the specified region. The number of trips that start or end within the area are about the same. Hence, under the assumption that people tend to drive to a rail station at the start of their journey, when they are for example late, the ride with their *exit point* within the research area are hereafter examined in greater detail.

Before discussing the results in-depth, the following table 5.1 gives a short overview over the colours used within the analysis. As no coldspots were deteced in the total analysis, the colour codes for coldspots aren't stated in the table.

Hotspot type	Colour
No pattern	grey
New hotspot	purple
Consecutive hotspot	orange
Intensifying hotspot	maroon
Persistent hotspot	red
Diminishing hotspot	green
Sporadic hotspot	gold
Oscillating hotspot	cornflowerblue
Historical hotspot	black

Table 5.1: Overview over the colour codes used in the following analysis.

Entire time period

As stated above, the provided data extends from September 2012 to June 2016. This covers 46 months or 201 weeks. Hence, for the analysis the time slices were decided to have the size of a month. Furthermore, the value that is needed for the calculation of pattern of a *consecutive hotspot / coldspot* and the pattern of an *oscillating hotspot / coldspot* is set to five, which is equivalent to a share of almost 11%. For the hotspot definition that means that for a consecutive hotspot at least the last five time steps have to be categorised as a hotspot (or coldspot respectively). For the link to be labeled as an oscillating hotspot there have to be at least five switches from a hotspot to a coldspot in the whole time series.

Figure 5.7 shows the results given by the spatio-temporal analysis tool. Possible drop-off points for taxis in front of the examined rail stations are marked in blue for the Wien Westbahnhof and red for the station Wien Meidling and Hauptbahnhof. As can be seen, when analysing the complete research period, there is still a persistent hotspot detected in front of the Wien Westbahnhof. At the same time, when considering the station Wien Meidling (left red circle), in the area in front of the station there are road sections categorised as persistent or sporadic hotspots and parts of it even as diminishing hotspots. The last point of interest is the area around the Hauptbahnhof (right red circle), here we can see persistent, sporadic and intensifying hotspots as well, but also newly emerging hotspots. Therefore, the hypothesis of a shift from the Wien Westbahnhof to the station Wien Meidling and Hauptbahnhof can't be wholeheartedly supported when analysing the complete time period as a whole.

To dig a little deeper, we can have a closer look at the evaluation of road segments in front of the station Wien Westbahnhof and Hauptbahnhof. For each time slice the

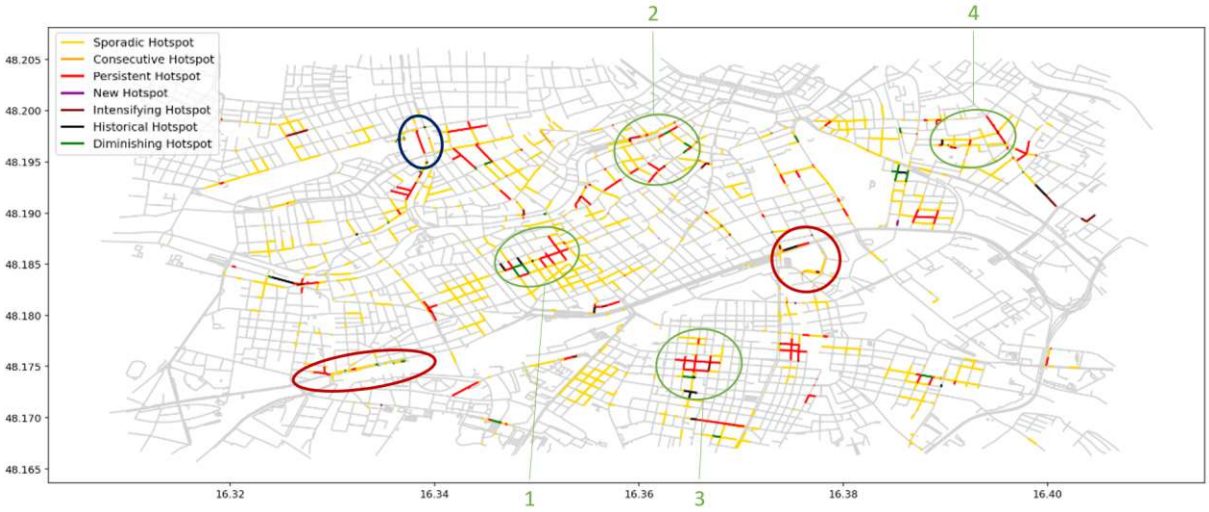

Figure 5.7: Spatio-temporal hotspot analysis for the entire time frame (September 2012 - June 2016) based on monthly time slices.

Getis-Ord statistic is evaluated, the resulting time series is then analysed using the Mann-Kendall trend test. The results of the Getis-Ord statistic are shown in figure 5.8, it consists of two parameters, namely the p-value and the z-value.

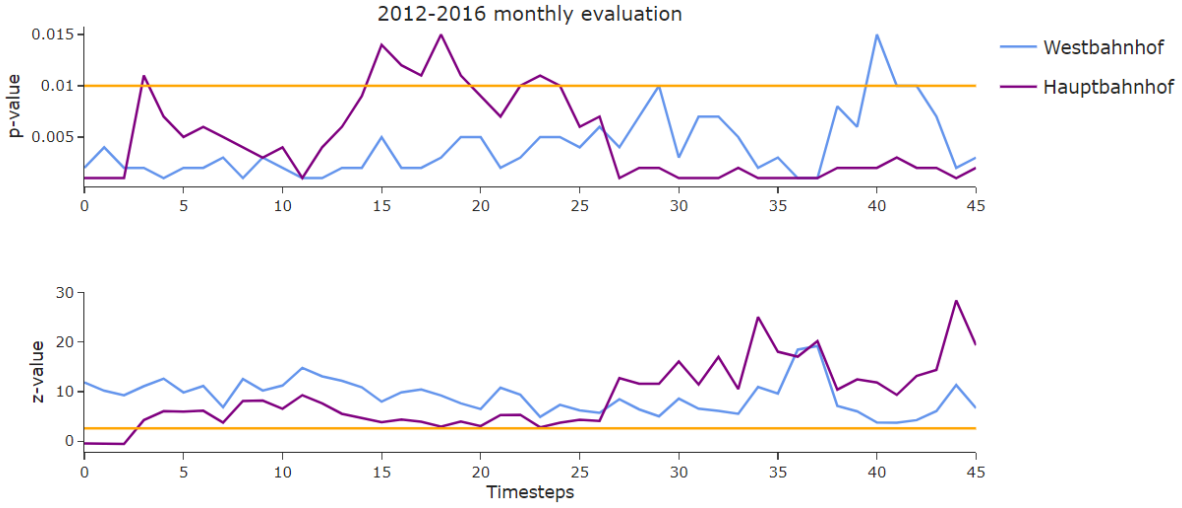

Figure 5.8: Time series of the evaluation of the Getis-Ord statistic at road segments at Wien Westbahnhof and Hauptbahnhof for the entire time frame (September 2012 - June 2016) based on monthly time slices.

To be categorised as a hotspot within a significance level of 1%, the p-value has to be less than 0.01 and the z-value has to be greater than 2.58, these boundaries are also highlighted in the graphic. As can be seen, the evaluation of the Wien Westbahnhof yield

a *persistent hotspot* - both the p-value and the z-value remain within the boundaries for the whole time period, except for the p-value at the 40^{th} timestep. Still, the p-value can be categorised as significant within a significance level of 5% at this point in time. For the station Hauptbahnhof the road segment showing an *intensifying hotspot* was examined in figure 5.8. As can be seen, both p- and z-value exceed the boundary in the first half of the time series but then gain in significance. This results in the mentioned pattern by the Mann-Kendall trend test.

Coming back to figure 5.7, what arouses attention are the areas highlighted in green circles. Examining these regions in more detail, the following points of interest were detected:

(1) Metropol theater

(2) Area around the famous Naschmarkt and Theater an der Wien

(3) -

(4) Hospital Rudolfstiftung

Even though not all of these highlighted areas can be assigned to a point of interest, most of them did. So, the identification of these locations using the spatio-temporal analysis tool indicates that plausible results are achieved.

As the analysis of the entire research period didn't support the initial hypothesis, each launching phase of the Hauptbahnhof will be analysed separately. Every year in December an additional part of the Hauptbahnhof went into operation. Therefore, in the following the time window from September till the end of March will be investigated. As the occasion of interest usually took place in the middle of December, the time window was selected such that the event is situated in the middle of the period.

September 2012 - March 2013

The analysis period from September 2012 till March 2013 covers 30 weeks, therefore a weekly analysis is chosen with the value that is needed for the patterns *consecutive hotspot* and *oscillating hotspot* set to three, which is equvalent to a share of 10%.

As can be seen in figure 5.9, the area in front of the Westbahnhof is, analogous to the previous analysis, categorised as a persistent hotspot. But in this time frame the area in front of the station Wien Meidling is mostly labeled as a persistent hotspot as well, only parts are categorised as sporadic hotspots. For the Hauptbahnhof on the other hand,

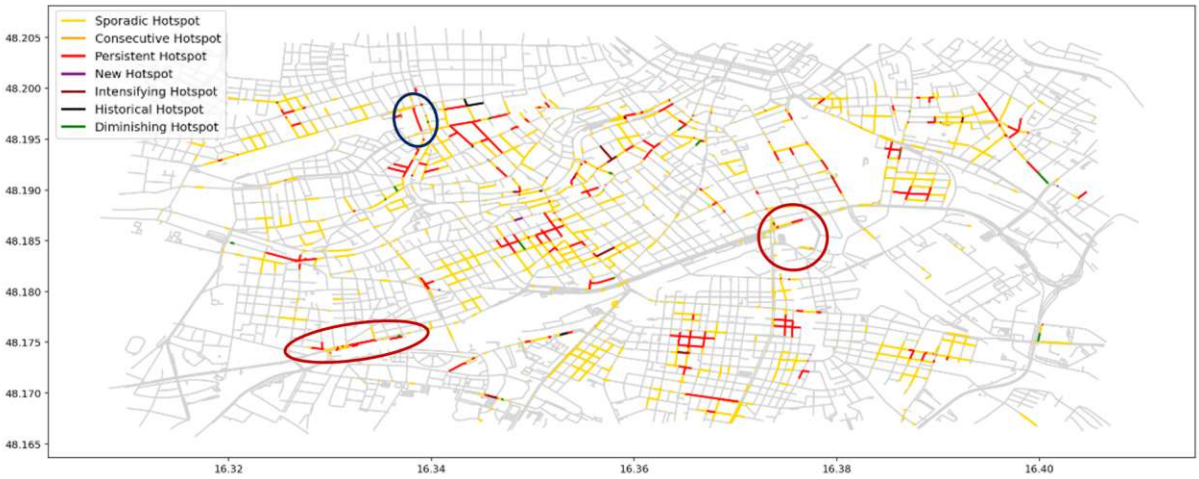

Figure 5.9: Spatio-temporal hotspot analysis for the time frame September 2012 till June 2013 based on weekly time slices.

hardly any street near it is categorised as a hotspot at all, mostly there is no pattern to be detected. This supports the initial hypothesis at least partly, namely that the Hauptbahnhof was no point of interest before its renovation and the timetable shift from Westbahnhof to Hauptbahnhof. Other points of interest to be discovered are similar to the example of the whole research period.

Again, looking at the time series of Getis-Ord G_i^*'s evaluation, one can see that the p- and z-value of the road segment at Wien Westbahnhof remain signifcant in every time step within a significance level of 1%.

For the Hauptbahnhof station on the other hand, the evaluation shows that it could even have been categorized as a coldspot at the start of the time series. Though, over time it gains in significance and to a part can also be categorised as hotspot within a significance level of 1%. This road segment was categorised as a *consecutive hotspot* by the Mann-Kendall trend test.

September 2013 - March 2014

Similar to the year before, the time window from September 2013 till March 2014 consists of 30 weeks. That's why, again the size of each time slice was set to a week and the value necessary for patterns two and seven is set to three, which constitutes a share of 10%.

In this time frame, the station Wien Westbahnhof remains a persistent hotspot as before. As for the station Wien Meidling, here sporadic, persistent and new hotspots are discovered. It seems that this station was a popular station even before the shift from Westbahnhof to Hauptbahnhof. It is interesting though, that in this year, there is the first

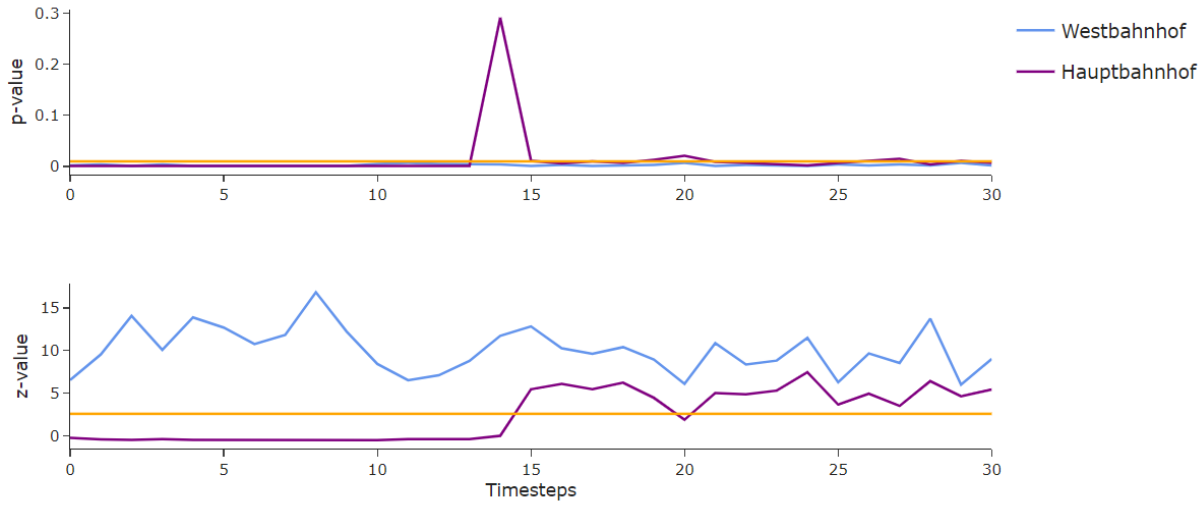

Figure 5.10: Time series of the evaluation of the Getis-Ord statistic at road segments at Wien Westbahnhof and Hauptbahnhof for the time frame September 2012 till March 2013 based on weekly time slices.

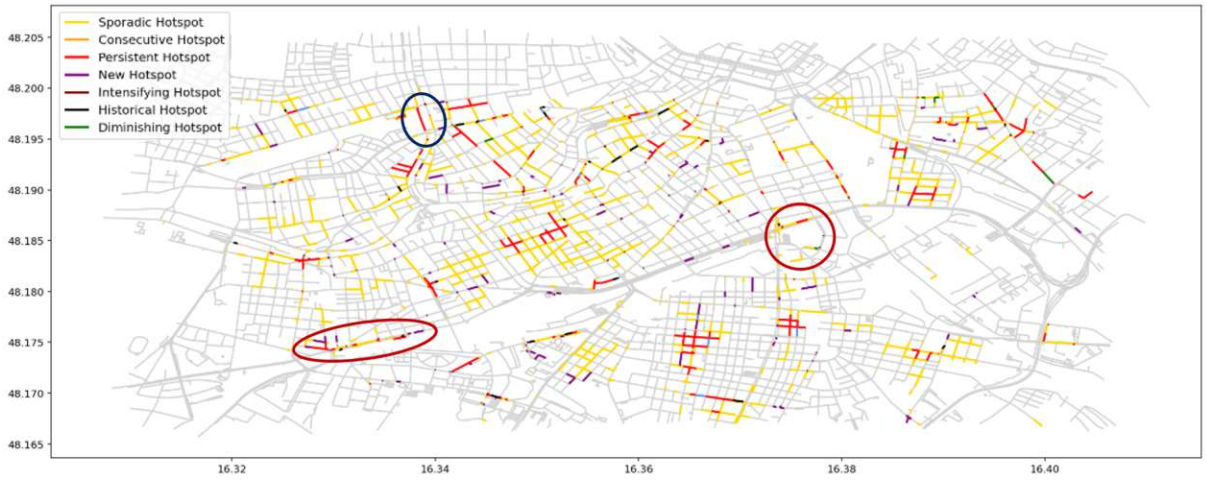

Figure 5.11: Spatio-temporal hotspot analysis for the time frame September 2013 till June 2014 based on weekly time slices.

persistent hotspot around Wien Hauptbahnhof. In December 2013 the first trains were transfered from the Westbahnhof to the Hauptbahnhof, so the first changes in the trains' timetables happened. So, with the knowledge of the analysis of the entire time period and at the current analysis stage, it seems that the initial hypothesis is only valid in terms of the Hauptbahnhof becoming a point of interest due to the switch. Therefore, let's follow the hypothesis that the area around the Hauptbahnhof becomes more and more important with every launching phase. Details on the Getis-Ord evaluation will be omitted at this

point, the evaluation for the time frames September till March for all years 2012-2016 will be presented later, see figures 5.14 and 5.13

September 2014 - March 2015

As the results for the time period from September 2014 until March 2015 resemble the results for the previous year, a detailed analysis is omitted at this point.

September 2015 - March 2016

Again, as in the other time windows from September until March, there are 30 weeks to be analysed. Hence, a weekly analysis with a value set to three, which is equivalent to a share of 10% is chosen.

On December 13^{th}, 2015, the new station Wien Hauptbahnhof was taken into full operation. That's why the investigation of this time frame is the most interesting one. Besides that, it is the last possibility to support or dismiss the hypothesis stated for the analysis. Figure 5.12 shows that for the first time there are diminishing hotspots in front of the Westbahnhof station. So, there has actually been a reduction in taxi rides ending at the Westbahnhof. At the same time there are no notable changes at the station Wien Meidling. While there have been emerging hotspots in the previous years, there are now persistent and sporadic hotspots in front of and around the station. When looking at the Hauptbahnhof station, one can see that there has been an increase in stops at the south entrance of the train station. Now, there is a persistent hotspot in comparison to a sporadic hotspot before. At the same time, at the northern side of the Hauptbahnhof there is a decrease in stops. Here, a historical hotspot as well as a diminishing hotspot was detected.

Taking a closer look at the time series resulting from the Getis-Ord G_i^* statistics' evaluation, we can see that for the selected road segment for the station Wien Westbahnhof, figure 5.13, the p- and z-values remain within the boundary given by the significance level of 1%. There has been only one breach in the 18^{th} time step in 2014 and one in the 19^{th} time step in 2015. At those two points in time, the evaluation is still considered a hotspot but only within a significance level of 5%. The Mann-Kendall trend test evaluated the portraied road segment in every time period as a *persistent hotspot*.
In figure 5.14 the selected road segment for the Wien Hauptbahnhof station is shown. The results for the year 2012 are omitted since because of an outlier the other results couldn't be properly assessed. As the threshold of the 1% signifcance level was passed by the p-value in the time period from September 2013 onwards, also the boundary of the 5% significance level is shown in the figure. One can see, that over the years the p-value

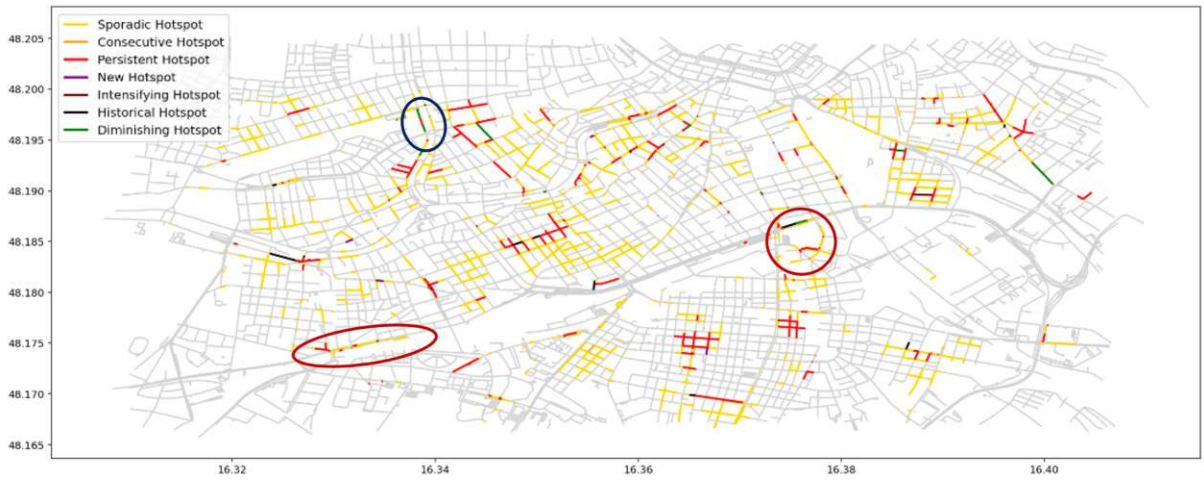

Figure 5.12: Spatio-temporal hotspot analysis for the time frame September 2015 till June 2016 based on weekly time slices.

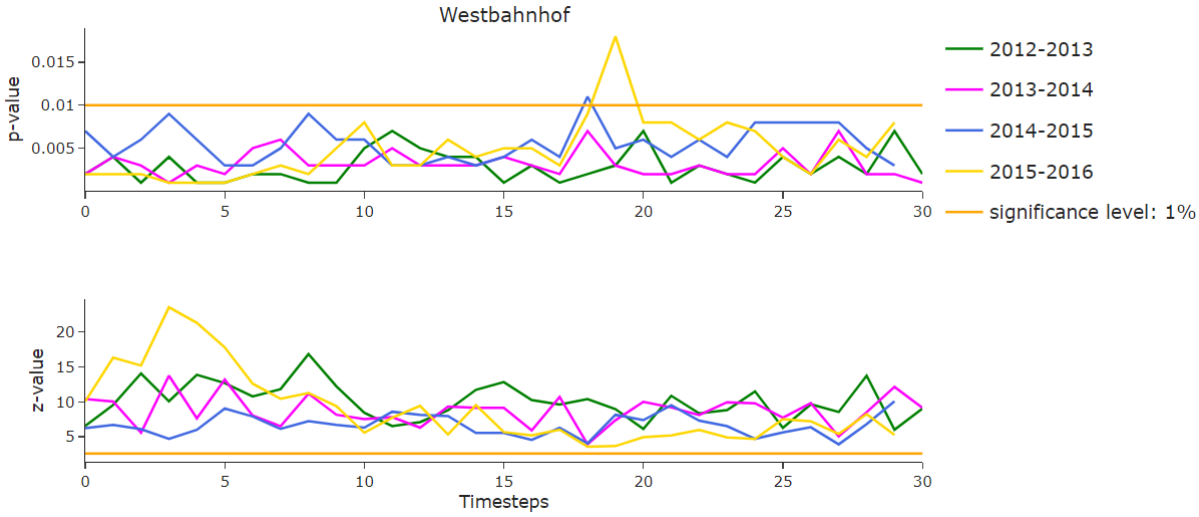

Figure 5.13: Time series of the evaluation of the Getis-Ord statistic at road segments at Wien Westbahnhof for the time frame September till March every year from 2012-2016 based on weekly time slices.

decreases while the z-value increases which means that the selected road segment gains in significance. The categories evaluated by the Mann-Kendall trend test are *diminishing hotspot* in 2013, *intensifying hotspot* in 2014 and *persistent hotspot* in 2015.

In conclusion, the initially formulated hypothesis, namely that there is a hotspot shift from Wien Westbahnhof to the stations Wien Meidling and Wien Hauptbahnhof, can only be approved to some extent. While the decrease of stops at the Westbahnhof became partly apparent in the last observation window, when the last launching phase of the new

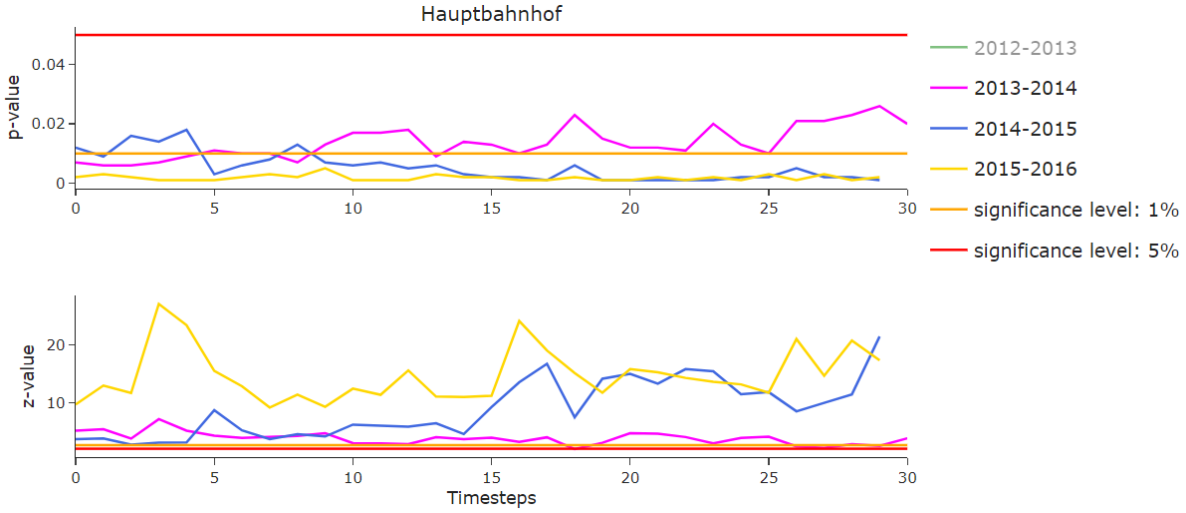

Figure 5.14: Time series of the evaluation of the Getis-Ord statistic at road segments at Wien Westbahnhof and Hauptbahnhof for the time frame September 2012 till March 2013 based on weekly time slices.

main station was finished, the increase in stops at the station Hauptbahnhof can primarily be seen when comparing the categorisations around the station over the years.

Encountered Problems

During the evaluation phase it was detected that using the `Python` package `spaghetti` causes that important information gets lost during the process, this issue was also reported to the developers (Gaboardi, 2020). Unfortunately, not only attribute values get lost, but also the original ID for every line segment is lost and replaced by an ID created by the package. Therefore, there is no possibility to extract the evaluation of a road segment that is chosen by, for example, its address - as would have been useful for the above evaluation. The only information that would have been possible to use as a filtering parameter is the specification of the geometry. Though, by reading in the data by `spaghetti` through a Shapefile, the geometry data was rounded from six to five decimal digits and stored as an geometry object. Therefore it couldn't be compared automated to the data given by reading in the original data from a `.csv` file. This issue resulted in tedious work for finding the ID of a road segment manually in the areas of the stations Wien Westbahnhof and Hauptbahnhof.

5.4 Conclusion

As became apparent in the results of the case study, the spatio-temporal hotspot analysis tool can only assist and support the user by providing a framework for analysis. At the same time, the definition of the time slices' size as well as the value which is critical for the patterns *consecutive hotspot* and *oscillating hotspot* have to be specified by the user. Therefore, the proposed tool is succumbed to subjectivity to a certain degree. The results for the entire observation period and a monthly resolution didn't support the hypothesis as expected. Though, when examining the analysis results for every launching phase in particular with a higher granularity, certain supporting factors for the hypothesis could be discovered. The fact that there was indeed a reduction in taxi stops didn't show when considering the entire time frame and a monthly resolution. This was only discovered when regarding solely the time window around the point in time when the new main station went into full operation.

These subjectively defined parameters can influence and distort the results of the analysis dramatically. That's why it'd be important to answer the question on how to define those parameters ideally or rather to provide some guidelines or approach on how to define a good approximation of an ideal solution. Connected to these considerations, it would also be useful to have some guidelines on how big to choose the time frame of the observation. That means, additionally to the granularity and the fixed value, one has to determine a time window of observation. In the presented case study, the time window for the process of the first launching phase until full operation was almost four years. When investigating every launching phase of its own, the chosen time window covered three and a half months before the selected event and three and a half months after it. But these time windows were subjectively chosen as well.

So, in conclusion it can be said that the spatio-temporal hotspot analysis tool detected various points of interest (apart from the hypothesis testing) that mostly turned out to be points of interest in the real world. At the same time, testing a particular hypothesis is influenced by subjectively chosen parameters which could result in contradictory support or dismissal statements for the hypothesis.

Chapter 6

Conclusion

The analysis of spatio-temporal hotspots is of great interest in various research fields. However, the detection of spatio-temporal patterns is limited by lack of appropriate methods that are also applicable to networks. This thesis contributes to fill this gap by presenting a two step approach to the spatio-temporal pattern analysis and furthermore presents the developed tool to foster the applicability of the proposed procedure.

6.1 Revisiting Research Questions

The main objective of this thesis was to develop a method for spatio-temporal hotspot analysis that meets the following requirements: it integrates the spatial and temporal aspects of the data, is applicable to data collected on a network and is based on methods implemented in the well-established open-source programming language `Python`. The proposed approach models the spatio-temporal analysis process in two stages. The first stage comprises of the slicing of the data into various time steps whose size can be individually specified by the researcher. For each of these time slices a spatial analysis is performed, the resulting time series is then subjected to a trend analysis. This process ends in the assignment of eight different patterns to each link of the research area. That way, both spatial and temporal dimension is integrated in the analysis process. Furthermore, the applicability of network based spatial methods was studied extensively, confirming the applicability of the chosen method to network shaped research areas. The approach has been implemented using the programming language `Python` utilising the various available libraries and packages. With the developed tool the practicability for data collected on networks was further demonstrated in a case study on taxi data in Vienna.

In a conceptual evaluation, the proposed approach was compared to the previous state-of-the-art approaches. Starting from a geographers point of view, the state-of-the-art approaches for spatial analyses are mainly focused on planar analysis utilising Moran's I, Getis-Ord G_i^* statistics and Geary's c. Concerning their applicability to network restrained data, it can be achieved by a few adaptions to the process, mainly by adapting the spatial weights matrix by using the shortest path distance instead of the Euclidean distance. As the thesis is focused on spatio-temporal analysis, these state-of-the-art approaches were introduced as well. For once, the spatio-temporal Moran's I suggested the incorporation of the temporal dimension into the spatial weights matrix. Though, as discussed, the author itself stated that this was only a first suggestion and still harbouring some issues concerning the interoperability of space and time. Another approach, the space-time scan statistic, is based on circular windows or rather cylinders when integrating the temporal component. This approach is problematic when applying it to networks, as the usage of Euclidean distance leads to misconceptions. The last introduced state-of-the-art approach, the space-time kernel density estimation, seems to be a very promising approach. Unfortunately, it is highly exposed to subjectivity due to the bandwidth definition that has to be set by the researcher. Due to this, the resulting hotspots cannot be categorised as statistically significant and therefore an additional analysis is required. In conclusion, a two stage approach was chosen in combining the network Getis-Ord G_i^* statistic and the Mann-Kendall trend test.

Then, the proposed approach was implemented using the programming language `Python`. Utilising various libraries that are available as open-source packages, three main classes were defined to facilitate a straightforward analysis workflow for the user. The implementation allows for simple data input in form of shapefiles and outputs the resulting patterns in form of GeoDataFrame, i.e. every individual link is listed with its corresponding pattern. In addition, the tool offers a function to easily plot the results for visual assessment. The visualisation not only presents the network of the research area but also colours the links according to their resulting patterns. This way, the spatio-temporal analysis tool is also applicable by users with limited expertise in `Python`.

The practicability of the spatio-temporal hotspot analysis tool was demonstrated in the case study of taxi movement data in Vienna. Furthermore, the applicability not only to network shaped research areas but also to planar research space was tested. By a slight modification of the `NetworkStatCalculator` the input of planar data is enabled. By adding a parameter to query if the input is a network or a plane, the internal setup is modified to be able to process the planar area and fall back to using the Euclidean distance

in the course of analysis. For the user the only modification is that a single parameter has to be specified, tagging the research as planar or network spatio-temporal analysis. The application was tested using aggregated Covid-19 data for the individual districts of Austria.

6.2 Discussion and Outlook

The case study has given some insights of what is possible with the proposed approach of spatio-temporal hotspot analysis on networks: constructing a network using link specifications in shapefiles, processing point patterns in form of shapefiles and mapping the points onto the nearest link, splitting links into equally distanced segments to avoid bias in the computation, applying the Getis-Ord G_i^* statistic on the network using the number of points on each link (link segment) as attribute values, specifying the time resolution of computation and therefore determining the length of the resulting time series, computing eight different patterns using logic inference and the Mann-Kendall trend test.

Still, the implementation also unveiled some flaws in existing `Python` packages and raises the necessity to think about multiple refinements before additional features and enhancements should be implemented. These include besides extensive computation time of `spaghetti` functions and input problems for higher amounts of data leading to revealing the underlying complexity of the simple input function provided by the tool implementation. A more sever problem is that when reading in data with `spaghetti`, the attribute values apart from the geolocation get lost in the process. This prevents the implementation of additional features such as adding attribute values such as the speed of taxis, speaking in terms of the presented case study. It can be highlighted though, that this issue was reported to the developers of the `spaghetti` package and therefore hopefully be resolved in the future (Gaboardi, 2020). This would open the possibility for multiple extensions of the tool implementation.

Another weakness is posed by the realisation of the two stage approach for the analysis. In case of using this procedure as a surveillance system for disease or crime monitoring, the emerging hotspots become apparent with a time delay. That means that hotspots are most likely not detected quick enough to employ effective countermeasures. Furthermore, if a continuous streak of hotspots is detected in the last couple of time steps the results can be diluted by the possibility of random fluctuations in earlier time steps (Kulldorff, 2001). This offers the possibility for reconsidering the pattern definitions and either add

additional patterns or implement conditions such that fluctuations at an early stage don't distort the results.

The `Python` implementation of a spatio-temporal hotspot analysis tool for networks presented in this thesis is a first prototype that is aimed at the analysis of a specific data source. Additional work is necessary to enhance this prototype to a more sophisticated and user-friendly implementation that can be made publicly available and be effectively used for spatio-temporal hotspot analysis on networks.

Bibliography

An, L. et al. (2015). “Space–Time Analysis: Concepts, Quantitative Methods, and Future Directions”. In: *Annals of the Association of American Geographers* 105(5), pp. 891–914.

Anselin, L. (1980). *Estimation Methods for Spatial Autoregressive Structures: A Study in Spatial Econometrics*. Vol. 8. Program in Urban and Regional Studies, Cornell University.

Anselin, L. (1995). “Local Indicators of Spatial Association - LISA”. In: *Geographical Analysis* 27(2), pp. 93–115.

Anselin, L. (1996). “The Moran Scatterplot as an ESDA Tool to Assess Local Instability in Spatial Association”. In: *Spatial Analytical* 4, p. 121.

Anselin, L. (2019). “A Local Indicator of Multivariate Spatial Association: Extending Geary’s c”. In: *Geographical Analysis* 51(2), pp. 133–150.

Anselin, L. and A. Getis (2010). “Spatial Statistical Analysis and Geographic Information Systems”. In: Perspectives on Spatial Data Analysis. Springer-Verlag Berlin Heidelberg, pp. 35–47.

Anselin, L. and X. Li (2022). *GeoDa*. URL: `https://geodacenter.github.io/`. (Accessed: 17.07.2022).

Arribas-Bel, D. and E. Tranos (2018). “Characterizing the Spatial Structure(s) of Cities "on the fly": the Space-Time Calendar”. In: *Geographical Analysis* 50(2), pp. 162–181.

Ashraf, M.S. et al. (2021). “Streamflow Variations in Monthly, Seasonal, Annual and Extreme Values Using Mann-Kendall, Spearmen’s Rho and Innovative Trend Analysis”. In: *Water Resources Management* 35(1), pp. 243–261.

Babu, S.K.K. et al. (2011). “Prediction of rainfall flow time series using autoregressive models”. In: *Advances in Applied Science Research* 2(2), pp. 128–133.

Barabási, Albert-László and Márton Pósfai (2016). *Network science*. Cambridge University Press.

Barak, S. and S.S. Sadegh (2016). “Forecasting Energy Consumption Using Ensemble ARIMA–ANFIS Hybrid Algorithm”. In: *International Journal of Electrical Power & Energy Systems* 82, pp. 92–104.

Barrett, A. (2021). *Forecasting the Prices of Cryptocurrencies using a Novel Parameter Optimization of VARIMA Models*.

Besag, J. and J. Newell (1991). "The Detection of Clusters in Rare Diseases". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 154(1), pp. 143–155.

Black, W.R. (1991). "Highway Accidents: a Spatial and Temporal Analysis". In: *Transportation Research Record* 1318, pp. 75–82.

Black, W.R. (1992). "Network Autocorrelation in Transport Network and Flow Systems". In: *Geographical Analysis* 24(3), pp. 207–222.

Bouzerdoum, M., A. Mellit, and A.M. Pavan (2013). "A Hybrid Model (SARIMA–SVM) for Short-Term Power Forecasting of a Small-Scale Grid-Connected Photovoltaic Plant". In: *Solar Energy* 98, pp. 226–235.

Box, G.E.P., G.M. Jenkins, and G.C. Reinsel (2008). *Time Series Analysis: Forecasting and Control*. John Wiley & Sons.

Braithwaite, A. and Q. Li (2007). "Transnational Terrorism Hot Spots: Identification and Impact Evaluation". In: *Conflict Management and Peace Science* 24(4), pp. 281–296.

Briz-Redón, Á. (2021). *Package 'DRHotNet'*. URL: `https://cran.r-project.org/web/packages/DRHotNet/DRHotNet.pdf`. (Accessed: 14.12.2021).

Brockwell, P.J. and R.A. Davis (2016). *Introduction to Time Series and Forecasting*. Springer.

Brunsdon, C., J. Corcoran, and G. Higgs (2005). "Visualising Space and Time in Crime Patterns: A Comparison of Methods". In: *Computers, Environment and Urban Systems* 31(1), pp. 52–75.

Bucher, D. et al. (2021). "Estimation of Moran's I in the Context of Uncertain Mobile Sensor Measurements". In: *11th International Conference on Geographic Information Science (GIScience 2021)-Part I*.

Carlson, R.F., A.J.A. MacCormick, and D.G. Watts (1970). "Application of Linear Random Models to Four Annual Streamflow Series". In: *Water Resources Research* 6(4), pp. 1070–1078.

Chainey, S. and J. Ratcliffe (2005). *GIS and Crime Mapping*. John Wiley & Sons.

Chainey, S., S. Reid, and N. Stuart (2002). "When Is a Hotspot a Hotspot? A Procedure for Creating Statistically Robust Hotspot Maps of Crime". In: *Innovations in GIS* 9, pp. 21–36.

Cheng, T. and M. Adepeju (2014). "Modifiable Temporal Unit Problem (MTUP) and Its Effect on Space-Time Cluster Detection". In: *PloS one* 9(6), e100465.

Cheng, T. and T. Wicks (2014). "Event Detection Using Twitter: A Spatio-Temporal Approach". In: *PloS one* 9(6), e97807.

Clark, P.J. and F.C. Evans (1954). "Distance to Nearest Neighbor as a Measure of Spatial Relationships in Populations". In: *Ecology* 35(4), pp. 445–453.

Clauset, A., C. Moore, and M.E.J. Newman (2008). "Hierarchical Structure and the Prediction of Missing Links in Networks". In: *Nature* 453(7191), pp. 98–101.

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

Cliff, A.D. and J.K. Ord (1981). *Spatial Processes - Models & Applications*. London: Pion.

Cottam, G. and J.T. Curtis (1949). "A Method for Making Rapid Surveys of Woodlands by Means of Pairs of Randomly Selected Trees". In: *Ecology* 30(1), pp. 101–104.

Cressie, N. (1993). *Statistics for Spatial Data*. John Wiley & Sons.

Dahly, D.L. et al. (2013). "The Spatial Distribution of Overweight and Obesity Among a Birth Cohort of Young Adult Filipinos (Cebu Philippines, 2005): an Application of the Kulldorff Spatial Scan Statistic". In: *Nutrition & Diabetes* 3(7), e80–e80.

Delmelle, E. et al. (2014). "Visualizing the Impact of Space-Time Uncertainties on Dengue Fever Patterns". In: *International Journal of Geographical Information Science* 28(5), pp. 1107–1127.

Dezman, Z. et al. (2016). "Hotspots and Causes of Motor Vehicle Crashes in Baltimore, Maryland: a Geospatial Analysis of Five Years of Police Crash and Census Data". In: *Injury* 47(11), pp. 2450–2458.

Dijkstra, E.W. (1959). "A Note on Two Problems in Connexion with Graphs". In: *Numerische Mathematik* 1(1), pp. 269–271.

Dodge, S. et al. (2016). "Analysis of Movement Data". In: *International Journal of Geographical Information Science* 30(5), pp. 825–834.

Douglas, E.M., R.M. Vogel, and C.N. Kroll (2000). "Trends in Floods and Low Flows in the United States: Impact of Spatial Correlation". In: *Journal of Hydrology* 240(1), pp. 90–105.

Dubé, J. and D. Legros (2013). "A Spatio-Temporal Measure of Spatial Dependence: An Example Using Real Estate Data". In: *Papers in Regional Science* 92(1), pp. 19–30.

Eckstein, P.P. (2016). *Statistik für Wirtschaftswissenschaftler - Eine realdatenbasierte Einführung mit SPSS*. Springer.

Encyclopedia of Mathematics (2014). *Greedy Algorithm*. URL: `http://encyclopediaofmath.org/index.php?title=Greedy_algorithm&oldid=34629`. (Accessed: 12.04.2022).

Eröffnung des Wiener Hauptbahnhofs in Bildern (2014). URL: `https://www.derstandard.at/story/2000006679434/shoppen-und-reisen-der-neue-wiener-hauptbahnhof-wurde-offiziell-eroeffnet`. (Accessed: 16.07.2022).

ESRI (2021a). *How Cluster and Outlier Analysis (Anselin Local Moran's I) works*. URL: `https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-`

statistics/h-how-cluster-and-outlier-analysis-anselin-local-m.htm. (Accessed: 02.11.2021).

ESRI (2021b). *How Emerging Hot Spot Analysis works*. URL: https://desktop.arcgis.com/en/arcmap/10.3/tools/space-time-pattern-mining-toolbox/learnmoreemerging.htm. (Accessed: 30.08.2021).

ESRI (2021c). *How Hot Spot Analysis (Getis-Ord Gi*) works*. URL: https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/h-how-hot-spot-analysis-getis-ord-gi-spatial-stati.htm. (Accessed: 02.11.2021).

ESRI (2021d). *What is a shapefile?* URL: https://desktop.arcgis.com/en/arcmap/latest/manage-data/shapefiles/what-is-a-shapefile.htm. (Accessed: 15.07.2022).

Gaboardi, J. (2020). *Retain line segment attributes on network links*. URL: https://github.com/pysal/spaghetti/issues/530. (Accessed: 16.07.2022).

Gaboardi, J.D., S. Rey, and S. Lumnitz (2021). "spaghetti: spatial network analysis in PySAL". In: *Journal of Open Source Software* 6(62), p. 2826. URL: https://doi.org/10.21105/joss.02826.

Gaboardi, J.D. et al. (Oct. 2018). *pysal/spaghetti*. URL: https://github.com/pysal/spaghetti.

Geary, R.C. (1954). "The Contiguity Ratio and Statistical Mapping". In: *The Incorporated Statistician* 5(3), pp. 115–146.

Gebru, G. et al. (2019). "Risk Factors and Spatio-Temporal Patterns of Human Rabies Exposure in Northwestern Tigray, Ethiopia". In: *Annals of Global Health* 85(1).

Getis, A. (2009). "Spatial Weights Matrices". In: *Geographical Analysis* 41(4), pp. 404–410.

Getis, A. and J.K. Ord (1992). "The Analysis of Spatial Association by Use of Distance Statistics". In: *Geographical Analysis* 24(3), pp. 189–206.

Getis, A. et al. (2003). "Characteristics of the Spatial Pattern of the Dengue Vector, Aedes aegypti, in Iquitos, Peru". In: *The American journal of Tropical Medicine and Hygiene* 69(5), pp. 494–505.

Gocic, M. and S. Trajkovic (2013). "Analysis of Changes in Meteorological Variables Using Mann-Kendall and Sen's Slope Estimator Statistical Tests in Serbia". In: *Global and Planetary Change* 100, pp. 172–182.

Goodchild, M.F. (1987). "A Spatial Analytical Perspective on Geographical Information Systems". In: *International Journal of Geographical Information System* 1(4), pp. 327–334.

Griffith, D.A. (2005). "Effective Geographic Sample Size in the Presence of Spatial Autocorrelation". In: *Annals of the Association of American Geographers* 95(4), pp. 740–760.

Hamed, K.H. (2008). "Trend Detection in Hydrologic Data: the Mann–Kendall Trend Test under the Scaling Hypothesis". In: *Journal of Hydrology* 349(3-4), pp. 350–363.

Hamed, K.H. (2009). "Exact distribution of the Mann–Kendall trend test statistic for persistent data". In: *Journal of Hydrology* 365(1), pp. 86–94.

Hamed, K.H. and A.R. Rao (1998). "A Modified Mann-Kendall Trend Test for Autocorrelated Data". In: *Journal of Hydrology* 204(1), pp. 182–196.

Hamilton, J.D. (2020). *Time Series Analysis*. Princeton University Press.

Hamilton, J.P., G.S. Whitelaw, and A. Fenech (2001). "Mean Annual Temperature and Total Aannual Precipitation Trends at Canadian Biosphere Reserves". In: *Environmental Monitoring and Assessment* 67(1), pp. 239–275.

Hart, T. and P. Zandbergen (2014). "Kernel Density Estimation and Hotspot Mapping: Examining the Influence of Interpolation Method, Grid Cell Size, and Bandwidth on Crime Forecasting". In: *Policing: An International Journal of Police Strategies & Management*.

Hinman, S.E., J.K. Blackburn, and A. Curtis (2006). "Spatial and Temporal Structure of Typhoid Outbreaks in Washington, D.C., 1906-1909: Evaluating Local Clustering with the Gi* Statistic". In: *International Journal of Health Geographics* 5(1), paper 13.

Hirsch, R.M. and J.R. Slack (1984). "A Nonparametric Trend Test for Seasonal Data with Serial Dependence". In: *Water Resources Research* 20(6), pp. 727–732.

Hu, Y. et al. (2018). "A Spatio-Temporal Kernel Density Estimation Framework for Predictive Crime Hotspot Mapping and Evaluation". In: *Applied Geography* 99, pp. 89–97.

Huang, B., B. Wu, and M. Barry (2010). "Geographically and Temporally Weighted Regression for Modeling Spatio-Temporal Variation in House prices". In: *International Journal of Geographical Information Science* 24(3), pp. 383–401.

Hudjimartsu, S.A., T. Djatna, A. Ambarwari, et al. (2017). "Spatial Temporal Clustering for Hotspot Using Kulldorff Scan Statistic Method (KSS): A Case in Riau Province". In: *IOP Conference Series: Earth and Environmental Science*. Vol. 54. 1. IOP Publishing, p. 012056.

Hyndman, R.J. and G. Athanasopoulos (2018). *Forecasting: Principles and Practice*. OTexts.

Jeffers, J.N.R. (1973). "A Basic Subroutine for Geary's Contiguity Ratio". In: *Journal of the Royal Statistical Society* 22(4), pp. 299–302.

Jin, W. et al. (2006). "Mining Outliers in Spatial Networks". In: *International Conference on Database Systems for Advanced Applications*. Springer, pp. 156–170.

Kalinic, M. and J.M. Krisp (2018). "Kernel Density Estimation (KDE) vs. Hot-Spot Analysis–Detecting Criminal Hot Spots in the City of San Francisco". In: *Proceeding of the 21st Conference on Geo-Information Science*.

Kang, H. (2010). "Detecting Agglomeration Processes Using Space–Time Clustering Analyses". In: *The Annals of Regional Science* 45(2), pp. 291–311.

Kendall, M.G. (1938). "A New Measure of Rank Correlation". In: *Biometrika* 30(1/2), pp. 81–93.

Kendall, M.G. (1962). *Rank Correlation Methods*. Griffin.

Khan Niazi, M.K. et al. (2014). "Perceptual Clustering for Automatic Hotspot Detection from Ki-67-Stained Neuroendocrine Tumour Images". In: *Journal of Microscopy* 256(3), pp. 213–225.

Kiani, B. et al. (2021). "Spatio-Temporal Epidemiology of the Tuberculosis Incidence Rate in Iran 2008 to 2018". In: *BMC Public Health* 21(1), pp. 1–20.

Kulldorff, M. (1997). "A Spatial Scan Statistic". In: *Communications in Statistics-Theory and Methods* 26(6), pp. 1481–1496.

Kulldorff, M. (2001). "Prospective Time Periodic Geographical Disease Surveillance Using a Scan Statistic". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 164(1), pp. 61–72.

Kulldorff, M., L. Huang, and K. Konty (2009). "A Scan Statistic for Continuous Data Based on the Normal Probability Model". In: *International Journal of Health Geographics* 8(1), pp. 1–9.

Kulldorff, M. et al. (2005). "A space–Time Permutation Scan Statistic for Disease Outbreak Detection". In: *PLoS medicine* 2(3), e59.

Kulldorff, M. et al. (2006). "An Elliptic Spatial Scan Statistic". In: *Statistics in Medicine* 25(22), pp. 3929–3943.

Kwan, M.P. (2013). "Beyond Space (as we knew it): Toward Temporally Integrated Geographies of Segregation, Health, and Accessibility: Space–Time Integration in Geography and GIScience". In: *Annals of the Association of American Geographers* 103(5), pp. 1078–1086.

Law, J., M. Quick, and P. Chan (2014). "Bayesian Spatio-Temporal Modeling for Analysing Local Patterns of Crime over Time at the Small-Area Level". In: *Journal of Quantitative Criminology* 30(1), pp. 57–78.

Lee, J. and S. Li (2017). "Extending Moran's Index for Measuring Spatiotemporal Clustering of Geographic Events". In: *Geographical Analysis* 49(1), pp. 36–57.

Lee, J. and S. Li (2018). "Reply to Comment by Dr. Daniel Griffith on J. Lee and S. Li (2017)."Extending Moran's Index for Measuring Spatiotemporal Clustering of

Geographic Events." Geographical Analysis, 49, 36–57". In: *Geographical Analysis* 50(4), pp. 479–480.

Lettenmaier, D.P. (1976). "Detection of Trends in Water Quality Data from Records with Dependent Observations". In: *Water Resources Research* 12(5), pp. 1037–1046.

Li, Q. and J.S. Racine (2007). *Nonparametric Econometrics: Theory and Practice*. Princeton University Press.

Li, Y. et al. (2020). "Analyzing Traffic Violation Behavior at Urban Intersections: A Spatio-Temporal Kernel Density Estimation Approach Using Automated Enforcement System Data". In: *Accident Analysis & Prevention* 141, p. 105509.

Liu, Hui, Hong-qi Tian, and Yan-fei Li (2012). "Comparison of Two New ARIMA-ANN and ARIMA-Kalman Hybrid Methods for Wind Speed Prediction". In: *Applied Energy* 98, pp. 415–424.

Lu, Yongmei and Xuwei Chen (2007). "On the False Alarm of Planar K-Function When Analyzing Urban Crime Distributed Along Streets". In: *Social Science Research* 36(2), pp. 611–632.

Lütkepohl, Helmut, Markus Krätzig, and P.B.C. Phillips (2004). *Applied Time Series Econometrics*. Cambridge University Press.

Magnussen, S. and L. Fehrmann (2019). "In Search of a Variance Estimator for Systematic Sampling". In: *Scandinavian Journal of Forest Research* 34(4), pp. 300–312.

Magnussen, S. et al. (2020). "Comparison of Estimators of Variance for Forest Inventories with Systematic Sampling-Results from Artificial Populations". In: *Forest Ecosystems* 7, pp. 1–19.

Mann, H.B. (1945). "Nonparametric tests against trend". In: *Econometrica*, pp. 245–259.

matplotlib (2022). URL: `https://matplotlib.org/`. (Accessed: 13.07.2022).

McIntire, P.J. et al. (2019). "Hotspot Enumeration of CD8+ Tumor-Infiltrating Lymphocytes Using Digital Image Analysis in Triple-Negative Breast Cancer Yields Consistent Results". In: *Human Pathology* 85, pp. 27–32.

Milic, N. et al. (2019). "The Influence of Data Classification Methods on Predictive Accuracy of Kernel Density Estimation Hotspot Maps." In: *The International Arab Journal of Information Technology* 16(6), pp. 1053–1062.

Minn, M. (2000–2021). *Basic Spatial Point Pattern Analysis in R*. URL: `http://michaelminn.net/tutorials/r-point-analysis/`. (Accessed: 14.12.2021).

Moran, P.A.P. (1950). "Notes on Continuous Stochastic Phenomena". In: *Biometrika* 37(1), pp. 17–23.

Nakaya, T. and K. Yano (2010). "Visualising Crime Clusters in a Space-Time Cube: An Exploratory Data-Analysis Approach Using Space-Time Kernel Density Estimation and Scan Statistics". In: *Transactions in GIS* 14(3), pp. 223–239.

Naus, J. I (1965). "The Distribution of the Size of the Maximum Cluster of Points on a Line". In: *Journal of the American Statistical Association* 60(310), pp. 532–538.

NetworkX (2022). URL: `https://networkx.org/`. (Accessed: 13.07.2022).

NumPy (2022). URL: `https://numpy.org/`. (Accessed: 13.07.2022).

O'Loughlin, J. (2002). "The Electoral Geography of Weimar Germany: Exploratory Spatial Data Analyses (ESDA) of Protestant Support for the Nazi Party". In: *Political Analysis* 10(3), pp. 217–243.

Offset Aliases (2022). URL: `https://pandas.pydata.org/pandas-docs/stable/user_guide/timeseries.html#offset-aliases`. (Accessed: 15.07.2022).

Okabe, A. and K. Sugihara (2012). *Spatial Analysis Along Networks: Statistical and Computational Methods*. John Wiley & Sons.

OpenStreetMap (2022). URL: `https://www.openstreetmap.org/`. (Accessed: 14.07.2022).

Ord, J.K. and A. Getis (1995). "Local Spatial Autocorrelation Statistics: Distributional Issues and an Application". In: *Geographical Analysis* 27(4), pp. 286–306.

Partal, T. and E. Kahya (2006). "Trend Analysis in Turkish Precipitation Data". In: *Hydrological Processes* 20(9), pp. 2011–2026.

Parzen, E. (1962). "On Estimation of a Probability Density Function and Mode". In: *The Annals of Mathematical Statistics* 33(3), pp. 1065–1076.

Pielou, E.C. (1959). "The Use of Point-to-Plant Distances in the Study of the Pattern of Plant Populations". In: *The Journal of Ecology*, pp. 607–613.

Prasannakumar, V. et al. (2011). "Spatio-Temporal Clustering of Road Accidents: GIS Based Analysis and Assessment". In: *Procedia-Social and Behavioral Sciences* 21, pp. 317–325.

QGIS (2022). URL: `https://www.qgis.org/en/site/`. (Accessed: 17.07.2022).

Qi, Y., Y. Yang, and F. Jin (2013). "China's Economic Development Stage and Its Spatio-Temporal Evolution: A Prefectural-Level Analysis". In: *Journal of Geographical Sciences* 23(2), pp. 297–314.

Rao, H., X. Shi, and X. Zhang (2017). "Using the Kulldorff's Scan Statistical Analysis to Detect Spatio-Temporal Clusters of Tuberculosis in Qinghai Province, China, 2009–2016". In: *BMC Infectious Diseases* 17(1), pp. 1–11.

Rey, S.J. and L. Anselin (2010). "PySAL: A Python Library of Spatial Analytical Methods". In: *Handbook of Applied Spatial Analysis*. Springer, pp. 175–193.

Rey, S.J. et al. (2015). "Open Geospatial Analytics with PySAL". In: *ISPRS International Journal of Geo-Information* 4(2), pp. 815–836.

Rey, S.J. et al. (2021). "The PySAL Ecosystem: Philosophy and Implementation". In: *Geographical Analysis*.

Richardson, D.B. (2013). "Real-Time Space–Time Integration in GIScience and Geography: Space–Time Integration in Geography and GIScience". In: *Annals of the Association of American Geographers* 103(5), pp. 1062–1071.

Rosenblatt, M. (1956). "Remarks on Some Nonparametric Estimates of a Density Function". In: *The Annals of Mathematical Statistics* 27(3), pp. 832–837.

Schrenk, J. (2015). *Der Westbahnhof ist nicht mehr Haupt-Bahnhof*. URL: `https://kurier.at/chronik/wien/wien-der-westbahnhof-ist-nicht-mehr-haupt-bahnhof/168.688.012`. (Accessed: 16.07.2022).

SciPy (2022). URL: `https://scipy.org/`. (Accessed: 13.07.2022).

Sierksma, G. and D. Ghosh (2010). "Networks in Action." In: *International Series in Operations Research and Management Science*.

Silverman, B.W. (1986). *Density Estimation For Statistics And Data Analysis*. Chapman and Hall.

Smith, T.E. (2008). "Spatial Weight Matrices". In: *Encyclopedia of GIS*.

Sokal, R.R., N.L. Oden, and B.A. Thomson (1998). "Local spatial autocorrelation in biological variables". In: *Biological Journal of the Linnean Society* 65(1), pp. 41–62.

Stamp, L.D. (1964). *The Geography of Life and Death*. Fontana.

Tang, J.-H., T.-J. Tseng, and T.-C. Chan (2019). "Detecting Spatio-Temporal Hotspots of Scarlet Fever in Taiwan with Spatio-Temporal Gi* Statistic". In: *PloS one* 14(4), e0215434.

Tenkanen, H. (2017). *Pandas and Geopandas -modules*. URL: `https://automating-gis-processes.github.io/2016/Lesson2-overview-pandas-geopandas.html`. (Accessed: 13.07.2022).

Tobler, W. (1970). "A Computer Movie Simulating Urban Growth in the Detroit Region". In: *Economic Geography* 46, pp. 234–240.

Tukey, J.W. (1977). *Exploratory Data Analysis*. Vol. 2. Reading, Mass.

Valipour, M. (2015). "Long-term Runoff Study Using SARIMA and ARIMA Models in the United States". In: *Meteorological Applications* 22(3), pp. 592–598.

Von Storch, H. (1999). "Misuses of Statistical Analysis in Climate Research". In: *Analysis of Climate Variability*. Springer, pp. 11–26.

Wang, Z. and N.S.N. Lam (2020). "Extending Getis–Ord Statistics to Account for Local Space–Time Autocorrelation in Spatial Panel Data". In: *The Professional Geographer* 72(3), pp. 411–420.

Wesolowska, W.A. (2018). *Short Term Electricity Forecasting Using Smart Meter Data*.

White, D.R., M.L. Burton, and M.M. Dow (1981). "Sexual Division of Labor in African Agriculture: a Network Autocorrelation Analysis". In: *American Anthropologist* 83(4), pp. 824–849.

Whittle, P. (1951). *Hypothesis Testing in Time Series Analysis*. Vol. 4. Almqvist & Wiksells boktr.

Wu, B., R. Li, and B. Huang (2014). "A Geographically and Temporally Weighted Autoregressive Model with Application to Housing Prices". In: *International Journal of Geographical Information Science* 28(5), pp. 1186–1204.

Yamada, I. and J.C. Thill (2004). "Comparison of Planar and Network K-Functions in Traffic Accident Analysis". In: *Journal of Transport Geography* 12(2), pp. 149–158.

Yiu, M.L. and N. Mamoulis (2004). "Clustering Objects on a Spatial Network". In: *Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data*, pp. 443–454.

Yu, D. (2014). "Understanding Regional Development Mechanisms in Greater Beijing Area, China, 1995–2001, from a Spatial–Temporal Perspective". In: *GeoJournal* 79(2), pp. 195–207.

Yuan, C., S. Liu, and Z. Fang (2016). "Comparison of China's Primary Energy Consumption Forecasting by Using ARIMA (the Autoregressive Integrated Moving Average) Model and GM (1, 1) Model". In: *Energy* 100, pp. 384–390.

Yue, S. and C. Wang (2004). "The Mann-Kendall Test Modified by Effective Sample Size to Detect Trend in Serially Correlated Hydrological Series". In: *Water resources management* 18(3), pp. 201–218.

Yue, S. et al. (2002). "The Influence of Autocorrelation on the Ability to Detect Trend in Hydrological Series". In: *Hydrological Processes* 16(9), pp. 1807–1829.

Zetterqvist, L. (1991). "Statistical Estimation and Interpretation of Trends in Water Quality Time Series". In: *Water Resources Research* 27(7), pp. 1637–1648.

Zhai, Xiaoyan, Jun Xia, and Yongyong Zhang (2014). "Water Quality Variation in the Highly Disturbed Huai River Basin, China from 1994 to 2005 by Multi-Statistical Analyses". In: *Science of the Total Environment* 496, pp. 594–606.