Robust Modeling of Spatio-Temporal Dependencies and Hot Spots

Zur Erlangung des akademischen Grades eines
Doktors der Ingenieurwissenschaften

(Dr.-Ing.)

von der Fakultät für
Wirtschaftswissenschaften
am Karlsruher Institut für Technologie (KIT)

genehmigte

DISSERTATION

von

M.Sc. Inform.-Wirt. Julian Bruns

Tag der mündlichen Prüfung:	08.06.2018
Referent:	Prof. Dr. Thomas Setzer
Korreferent:	Hon.-Professor Dr. Hansjörg Fromm
Korreferent:	Professor Dr. York Sure-Vetter

Karlsruhe, 2018

Contents

List of Figures

List of Tables

Part I

Introduction

Chapter 1

Introduction and Motivation

1.1 Motivation

SPATIAL data is everywhere (Parsons (2017)). In the modern world almost all data has a spatial and a temporal context. This ranges from route-planning, to location based services, environmental processes and almost any human interaction has a place and a time, to name just a few. Smart phones, now an integral part of our lives, were introduced on the premises of their location based services by Steve Jobs of Apple Inc.. The possibilities of Google Glass was shown with its navigation option and spatial context. To use these data sources, appropriate models are needed. The fields of geo-statistic as well as spatial econometrics are today well-established, see e.g. Anselin (2010), and provide new insights to our understanding of the world. But these methods and the underlying paradigm is not yet fully established in the field of information systems (IS). This is surprising as most of the recent trends in the field of IS are based on spatial data:

(1) On of the biggest trends in the field is the rise of *Big Data* and its implications. This is fueled particular by the birth of social media and social networks such as Facebook and Twitter. These produce an immense amount of data. The high percentage of publications in journals such as

the DSS[1] which are focused mainly on these social data show their importance for IS. The context of this data is often dependent on the location; most of the underlying services even demand the geo-location for their use. Smart phones in particular are multi-purpose geo-referenced sensors. Today, we measure the mobility behavior of people based on their smart phones, we can measure the emotional well-being based on the position and even use the sensors directly, e.g. the cameras or additional sensors for noise. In the words of Blaschke et al. (2011) "Location-based services on mobile smart phones are penetrating our daily communication behaviour more and more". Other data sets such as from car-sharing (Wagner et al. (2016)) or taxi data [2] are used to optimize services and (urban) mobility, predict future demand, plan optimal locations and the optimal tours. James Steiner, Vice President of Oracle, emphasized the importance of spatial information quite strongly: 'Location information and geospatial data are at the core of most Big Data use cases" (Steiner (2017)).

(2) Another important trend is the resurgence of *Cyber Physical Systems* (Broy (2010)). Under this term the combination of digital processes and data representation with real world processes is understood. Typical applications are sensor placement in production pipelines and tracking of goods in a supply chain. Standard approaches are based on process mining and know the placement of sensors in the process. But more recent developments, especially in the case of the tracking of goods, are using and needing spatial information. For these new, enhanced applications the exact location of a good or sensor is important for a complete understanding of the processes and potential impact on the overall production and supply chain process. An example for this is presented in the vision of Work and Bayen (2008), where smart phones are proposed as multi-purpose monitoring tool for Cyber Physical Systems, as these provide visual as well as auditory information coupled with the geo-position. Particular in the early

[1] https://www.journals.elsevier.com/decision-support-systems/
[2] http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml

stages of new production lines as well as for general purpose applications the spatial component is essential.

(3) The most obvious trend for the use of spatial data is the *Smart City*. While no clear definition of the term is agreed upon (Resch et al. (2012)), we use the understanding provided by the BISE special issue regarding smart cities[3]: "Using information systems to improve all of the facets of urban life is the core of the Smart City paradigm." A smart city has many components, goals and actors as well as underlying processes and a huge variety of different data sources. And all of these different parts are highly interconnected based on their temporal and spatial closeness. For example, temperature influences human well-being directly through heat-stress and indirectly through its influence on other environmental variables such as air pollution. In addition this in turn influences the energy consumption of a city through heating or cooling devices, dependent on the temperature. Hay et al. (2011) developed a demonstrator to show the benefits of spatial technologies for influencing the energy consumption. While in the past, the examination of this highly complex construct city has had not enough data for meaningful analysis, today the rise of volunteered geographic information (VGI), collective sensing apparatus, IoT devices and the ubiquity of mobile devices will allow a more in-depth analysis and understanding of these processes (Arribas-Bel (2014)). But these new data sources also include new challenges, in particular the heterogeneous quality of the available data, the different purpose for their original gathering and the requisite to include the spatial and temporal dimensions. This leads to the paradigm of *Collective Sensing* (Blaschke et al. (2011)). Furthermore, in the context of a truly smart city, all the stakeholder in a city have to be included. This leads to a stronger focus on the participation of individuals and a focus on the individual level. Zeile (2017) showed a way to do so by combining mobile emotion sensing with urban planning methods to detect and use critical points in a city. Finally the paradigm of a smart city is even

[3] http://www.bise-journal.com/?p=1224

now extended to the vision of a *Live City*, "in which the city is regarded as an actuated near real-time control system creating a feedback loop between the citizens, environmental monitoring systems, the city management and ubiquitous information services" (Resch et al. (2012)).

As these examples show, most real-world processes relevant to the field of IS are dependent on spatial as well as spatio-temporal factors. These factors include often spatial as well as temporal autocorrelations as well as their interaction. While the temporal aspects are well established in the field of IS, e.g. time series forecasting, the spatial aspects and the spatio-temporal aspects are rarely discussed or included in models. In the few instances, e.g. in logistics and decision support systems (DSS), often only abstractions such as graphs are used. The arising problems in the spatial sense are best described by LeSage (1999): "Two problems arise when sample data has a locational component: 1) spatial dependence between the observations and 2) spatial heterogeneity in the relationships we are modeling." Apart from the possible miss-specification of models this also reduces the effective sample size and therefore the validity of the results of statistical tests such as hypothesis testing or structural models. Again, this phenomenon of correlated data is well-known in the field of IS, but often disregarded in the case of spatial data.

The goal of this thesis is to enhance the foundation of the field of IS by developing methods and approaches to tackle challenges of spatio(-temporal) data for the field of information systems. These new methods extend the tool box of researchers as well as practitioners to detect and predict spatio-temporal developments and dependencies. The focus is to provide methods to automatically detect and evaluate points of interest given big data sizes and providing the methods to explain and predict the occurrence of the phenomena, even when there is little data available. This follows from the general motivation in that almost all Big Data is *Spatial* Big Data. This works further aims to also contribute to the final challenge stated in An-

selin (2010): "A final challenge parallels the previous one and pertains to the computational techniques needed to handle the complex space-time interactions in increasingly large data sets. New algorithms will need to be developed and effective use made of the rapidly changing computing technology, such as distributed computing, cloud computing and the use of handheld devices." The methods and approaches presented in this thesis are developed so that they can be easily performed in a distributed, parallel computation and deal with the inherent uncertainty of the data sources. By developing methods for both, detection as well as prediction, we follow the idea presented in Appice and Malerba (2014): "Nowadays ubiquitous sensor stations are deployed worldwide, in order to measure several geophysical variables (e.g. temperature, humidity, light) for a growing number of ecological and industrial processes. Although these variables are, in general, measured over large zones and long (potentially unbounded) periods of time, stations cannot cover any space location. On the other hand, due to their huge volume, data produced cannot be entirely recorded for future analysis. In this scenario, summarization, i.e. the computation of aggregates of data, can be used to reduce the amount of produced data stored on the disk, while interpolation, i.e. the estimation of unknown data in each location of interest, can be used to supplement station records."

While the new approaches and methods proposed in this thesis can be applied in a variety of scenarios and use cases, we focus in the evaluation of our approaches on temperature data. We base this decision on several facts: (1) Temperature is one of the main underlying environmental factors (Oke et al. (2017)). Temperature impacts directly human health, well-being, work performance, energy consumption and many more. A sound understanding of causal dependencies of temperature is mandatory for any information system regarding city planning and operation. The difference of temperature within cities and its surrounding areas is called *Urban Heat Island* (UHI) and its study the focus of a whole field of science. (2) Today,

there exists an increasing importance to detect extreme local temperatures in cites, see e.g. Hansen et al. (2010); Chase et al. (2006); Department of Economic and Social Affairs, Population Division, United Nations (2014). Therefore models and methods to reliably detect and mitigate the effect of temperature extremes in a proactive fashion are needed. (3) Temperature data, as all meteorological data, is inherently complex and good generalizable models are needed. Achieving this feat for temperature indicates that these models can be applied to other fields and applications. (4) Finally, while there still is a sparseness of available meteorological data, temperature data is one of the most active field of VGI (Meier et al. (2017)). But while the amoung of open data is higher then in other domains, the quality of its data is highly heterogeneous. Meier et al. (2017) found that over 60% of VGI temperature measurements in the city of Berlin can not be used as their quality varies strongly and is too low for meteorological models. A selected overview of temperature, urban climate and its impact on cities will be presented in chapter 2.

1.2 Research Questions

This thesis focuses on developing insights in the exploratory data analysis, the examination of causal reasons and the use of these analyses to predict spatio-temporal developments with the requirements of robustness, generalizability and applicability in big data use cases. For this purpose three research question, derived and defined in this section, are addressed. The first two research questions focus on exploratory data analysis and the robust detection of points of interest. The third research question focus on the explanatory data analysis and prediction based on causal factors.

The first question tackles the problem how a stable hot spot analysis can be modeled and created. Hot spot analysis deals with the detection of points of interest with statistical tests and significance and is therefore

the preferred method to detect points of interest automatically without human input. As this is paramount in a big data setting, the new methods are developed for these statistical tests. The overall question is subdivided into three important tasks. First, we have to understand which modeling parameter and effects influence the stability of a hot spot analysis and in what way this happens. The answer to this question allows an analyst to modify the model deliberately and in accordance to his goal. The second question is built upon the results of this examination. How can these results be used to modify existing hot spot analyses to be more stable? If the different influences are known, we can decide how to change the parametrization to create more stable results, which are ideally invariant over all parametrizations. Finally, it is important to know what the optimal parametrization is. As it is most likely impossible to have a perfect, invariant parametrization, it is of interest if there are certain parametrizations which are more stable then others and why. The answer to this question could lead to an easy rule of thumb for practitioners and minimize the effect of suboptimal parametrizations. Therefore, the three research questions (RQ) can be formulated concisely:

RQ 1 **How to create a stable Hot Spot Analysis?**

- **a)** What effects and parameter influence the stability of hot spot analysis?
- **b)** How can existing methods be modified to be more stable?
- **c)** What is the optimal parametrization for an hot spot analysis?

But to quantify the results of the previous research question, a metric is needed to evaluate the stability of the results. In the literature, only visual approaches are used to determine the stability of hot spot analyses. But while the visual approach works for small data sets and a limited number of different parametrizations, it requires a human to decide whether the results are stable or not. For a typical big data task and many different

possible parametrizations this is unsatisfactory. But, as there is no ground truth, a metric or definition of the stability is difficult to determine. As the hot spot analysis is an unsupervised approach, the problem is similar to measuring the quality of a clustering. This leads to the second overall research question:

RQ 2 How can the stability of found Hot Spots be measured?

However, the detection of points of interest is a exploratory analysis and relies on sufficient data. It does not explain *why* the points are different. But an understanding of the causal drivers of the differences between spatial locations can help to influence or mitigate the effects of these differences. The effects themselves can dependent on a manifold of developments and differences, so it is important to first understand what causal drivers exists and which are best used to explain those differences. Temperature is in this case one, if not the, most relevant environmental variable to understand. As stated before, temperature influences most other, environmental or not, factors, particular in and outside of urbanizations. But there does not yet exist a reliable prediction method for temperatures within a city. This is the result of two main problems of this field. The inherent complexity of the underlying meteorological, environmental and physical processes and the sparseness of available meteorological data, e.g. lack of available long-term weather station data. We therefore want to combine the causal drivers in an intelligent way to create a temperature prediction model, which builds upon existing measurement networks and complements these. But any model has to be generalizable to a manifold of different areas, has to be robust and needs to provide a spatio-temporal high resolution to be of use for further analyses. This leads to the formulation of the final overall research question and the subdivided tasks:

RQ 3 **How can temperatures in an intra-urban setting be predicted?**

a) What are causal drivers behind local temperature differences?

b) Given the inherent complexity of the underlying meteorological, environmental and physical processes and the sparseness of available meteorological data, how can those drivers be modeled to produce an accurate and robust prediction?

Based on the derived research questions, this thesis makes contributions to the literature on the robust detection of points of interest, the causal, robust prediction of temperature in a high resolution and to the overall toolbox of IS research.

1.3 Structure of the Thesis

To embed and introduce the new spatio(-temporal) methods and approaches, this thesis is structured into four parts. The first parts focuses on the motivation and the overarching foundations. In the motivation chapter, the need and context of the new spatial approaches for the field of IS are discussed. Based on the this discussion, the prevalent topics are defined with the formulation of the research questions, which guide this thesis. The foundation chapter provides an understanding of the methodical and domain specific background for the methods. A selected overview of spatio(-temporal) approaches and core concepts as well as an overview of the domain of urban climates and the importance of temperature is presented. This domain knowledge is then used throughout the empirical data sets and evaluation.

The second part addresses the challenge how to detect reliable points of interests in space and time. Existing approaches are presented and discussed. Based on the discussion of the most prevalent approach, influences

on the stability of the Getis-Ord (G^*) Statistic are analyzed and the statistic is then modified to be more stable. To evaluate this stability, existing approaches for the measurement of stability from the field of geo-statistics and computer science are compared and used as basis for the development of the quantifiable stability metric for hot spot analysis. Finally, the results are validated on an empirical temperature data set for the city of Karlsruhe, both visually as well as based on the new metric. This allows a fast and reliable detection of intra-urban heat islands for the data set and of points of interest in general.

While the second part is an exploratory approach, the third part deals with the challenges on how to understand the causal reasons behind differences in temperature between different nearby areas as well as how to predict temperature values at unknown areas based on these causal reasons. We propose an analytic model to explain and predict temperature differences within urbanizations. This model, the *Land use-based Temperature Model* (LTM), combines land use information, time series of weather station-based temperature measures and the interactions of both types of predictors to derive a causal prediction model. It solves the two main problems of temperature prediction, as it deals with the sparseness of the data through its simplicity and inherent robustness. We evaluate this model on empirical data of cities in the German federal state of Baden-Württemberg and we can show an increase om the accuracy of mean air temperature predictions up to a MAE of 2°C compared to standard models solely based on temperature and distance data.

The fourth and final part concludes this work and provides an outlook for further extensions and future work.

A graphical overview of this thesis with its four parts is additionally provided in Figure 1.1.

Extracts of this work have already been published, are under review or are working papers. Part II is based on Bruns and Simko (2017) and con-

tains insights, models, evaluations and textual paragraphs from that publication . Part III is based on Bruns and Setzer (2018) and contains insights, models, evaluations and textual paragraphs from that publication. Part I contains paragraphs from all publications. Those works are extended and discussed in more detail.

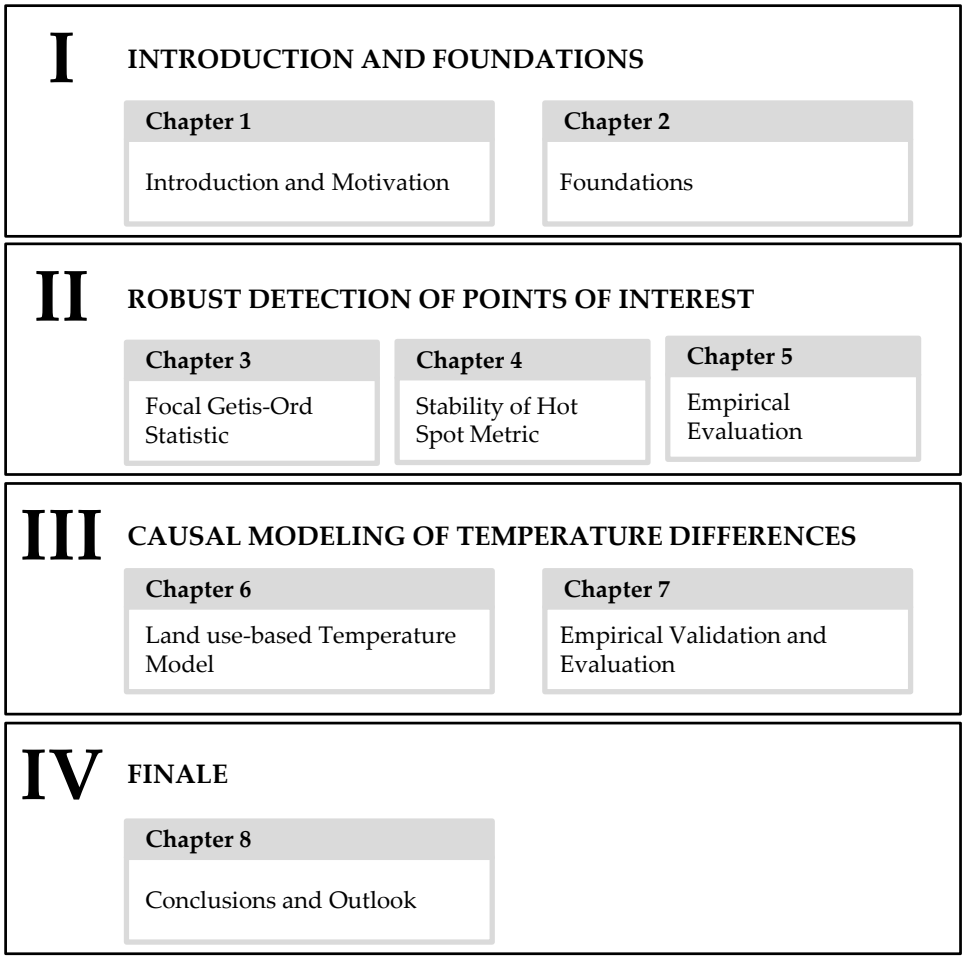


Figure 1.1: The Thesis is structured into four parts. The first part motivates and introduces the overarching topic of the thesis as well as the overarching foundations. The second part addresses the problem how to reliably detect points of interests unsupervised in spatial and spatio-temporal settings. The third part presents a novel modeling approach to detect and use causal temperature differences in a sparse data setting. In the last part the conclusion, future work as well as an outlook for further extension is presented.

Chapter 2

Foundations

In this chapter we present a selective overview of overarching concepts and topics relevant to this thesis. As temperature, in particular in cities, is used to empirically evaluate our models, we discuss the idea and background of the *Urban Heat Island* (UHI), urban climates and the impact of temperature on human health. This provides the background for the use cases and their importance. The other topic is geographical analysis. We introduce briefly key concepts relevant to this work and provide references for further reading.

2.1 Urban Heat Island and Urban Climates

The phenomenon of higher temperature levels in cities as well as its impacts on urban planning und human health, coined Urban Heat Island (UHI), is subject to research for many decades. The term was termed by Oke (1982). One of the earliest known overviews of the scientific literature for city climates is given by Kratzer (1937). At that time already, relations between temperature, humidity, human heat fluxes and air pollution are investigated.

Today, developments like the aging of society, the increasing urbanization and the climate change is making the adaptation to heat stress danger

more and more important. Due to the tendency that a rising number of people is moving into the cities, the urban heat island effect (UHI) is gaining more importance in the future.

Heat is an important factor to human health and comfort. High temperatures cannot only lead to a discomfort, it also has serious negative effects on the health as well as the ability to work. An overview is presented in Basu (2009), where numerous studies are compared and show that high temperature is associated with an increase in mortality as well as morbidity. The most well-known example of this association in recent history is the 2003 heat wave in Europe. Over 19,490 heat related deaths in France alone were registered, an excess mortality of 60% for the whole country. In Paris the excess mortality reached 142% (Robine et al. (2008)). These numbers are based on the difference between the mortality level of 2003, the hottest summer in Europe on record (Chase et al. (2006)), in comparison with the mean of the previous five years. This effect is not equally distributed in the population. Hübler et al. (2007) for example shows that particular groups such as older people or people with health problems are especially vulnerable to the heat. Davis et al. (2010) examine the effect of increased body temperature on patients with multiple sclerosis. This increased temperature can lead to a worsening of their symptoms.

Our overview of the UHI and urban climates presented here focuses on new and selected insights and works of the last 15 years, extending the overview provided in Arnfield (2003). The focus of his work lies on the development in the field of climatology between 1980 and 2003. In comparison to 1982 the understanding of the UHI effect is increased, but, to quote Arnfields conclusion, “simple methods are still required to estimate UHI intensity within urban areas, as a function of time, weather conditions and structural attributes, for practical applications such as road climatology, phenology, energy conservation, and weather forecasting.” He continues in that simulations are one option to deal with the underlying complexit-

ies of city temperature modeling, but their prediction performance remains low.

A strong focus in the last 15 years was to standardize the measurement and modeling methods and improve the quality of the resulting insights.

A general overview of the approaches is presented in Mirzaei and Haghighat (2010). They review the different approaches of UHI studies, observational and simulation methods in particular. The observational approaches are divided into field measurement, thermal remote sensing and small-scale modeling, the simulation approaches into energy balance models, computational fluid dynamics, meso- and micro-scale models and turbulence treatments. They criticize the high computational cost of the state-of-the-art models and propose the integration of models to "take advantage of multi-scale models" (Mirzaei and Haghighat (2010)).

A harsh critique of the quality of existing studies and their methodology is presented in Stewart (2011). It combines a systematic review of 190 UHI studies between 1950 and 2007 of nocturnal air temperature with a scientific critique of the methodology. The main focus lies in the measurement methods and the description of the whole experimental settings of the examined papers. According to his findings, the quality of the UHI literature as well as its implications are unreliable. Ten percent of the studies are regarded as top-tier studies which have an high methodology standard. Those studies are then presented as examples for the highest quality and should be used as reference for other studies on UHI. In addition, guidelines are presented for conducting measurement and reporting results.

In a similar vein were studies from Shao et al. (2011), Mohsin and Gough (2012) and Siu and Hart (2013). They discussed the need to choose representative stations to measure UHI and the existing challenges in the field to determine these representative studies. Their argument are similar to older discussions about the problems behind choosing the correct locations

to perform measurements can be found e.g. in Sundborg (1950). Shao et al. (2011) discuss the new problems of the rapidly changing urban surfaces and increasing size of urban areas. Formerly rural stations may become urban stations and thereby underestimate the UHI intensity.

To solve these problems, Stewart and Oke (2012) propose the use of local climate zones (LCZ) to standardize the methodology and terminology. They present their concept of these LCZ and propose 17 different LCZ, which are based on the underlying land usage. The land usage can range from forests up to heavy industry areas. The LCZ are evaluated on three different mid-latitude cities in 2014 (Stewart et al. (2014)). It is shown that each LCZ has a different climate and delivers a better understanding of the UHI effect based on the land cover usage. This results in a better understanding and differentiation compared to the classical urban/rural differentiation.

Apart from the problem of the quality of the measurement, another recent topic is the use of the correct indicator or metric to determine whether an area is a UHI or not. Schwarz et al. (2011) compare 11 different Surface Urban Heat Island (SUHI) indicators on a data set of 263 European cities with monthly mean temperatures. They show that the selection of indicators is important for the detection of UHI due to possible instabilities of each indicator. To follow this up, they compare in Schwarz et al. (2012) different measurement methods for the UHI effect and come to the conclusion that the UHI effect is dependent on the *exact* placement of the rural as well as urban station and it is therefore important to take the effect of the immediate surrounding into consideration when comparing the UHI between different cities. On the one hand, the authors state that the reduction of an UHI to a single value for a whole city is questionable regarding its explanatory power. On the other hand, they conclude that there is currently no other way to quantify the temperature difference of the UHI between different cities.

To solve this problem, Martin et al. (2015) argue to drill down the resolution of the indicator and the subject of study. They define the surface intra urban heat island (SIUHI), which measures the temperature differences *within* a city. This provide a more detailed overview. Accordingly, the results can then be used to detect vulnerable areas in a city and trigger alerts for a much finer spatial granularity. The study was done in the city of Montreal by the use of Landsat satellites data between the years 1984 and 2011. The SIUHI is determined by defining thresholds with respect to spatial reference and compare the absolute deviation from the mean temperature given a survey area.

A new reference work regarding urban climates was published in 2017 with Oke et al. (2017). It provides an in-depth overview of the different concepts, methods and impacts on the urban climates. These range from air flow and heat fluxes to pollution, climate change and climate-sensitive design of cities and buildings. We refer an interested reader to this work for a deepening in the topic of UHI and urban climates.

2.2 Spatial Analysis

We want to focus on two core concepts - the idea of spatial association and how to measure this spatial association. Other, more specific concepts are discussed in their relevant chapter in thesis, but these two concepts are a common foundation for this thesis.

Spatial Association: The core idea of spatial analysis is that near areas influence each other. We use this term interchangable with the term spatial depdence. This idea is, in its most popular form, formulated in Tobler (1970) and is well known as Tobler's First Law (of Geography). It states that "everything is related to everything else, but near things are more related than distant things". This is later extend with Tobler's second law, which states that: "The phenomenon external to an area of interest affects

what goes on inside". LeSage (1999) define this formally as: "Spatial dependence in a collection of sample data means that observations at location i depend on other observations at locations $j \neq i$. Formally, we might state:

$$y_i = f(y_i), i = 1, \ldots, n \quad j \neq i \tag{2.1}$$

"

Measuring Spatial Association: The understanding of the concept of spatial association itself helps analysts to incorporate this in their work. Otherwise, this may impact the validity of the results. However, to truly incorporate this into models and their specification of the parameter, the spatial association has to be known and measured. To cite Getis and Ord (1992): " To geographers, the best-known statistics are Moran's I and, to a lesser extent, Geary's c (Cliff and Ord (1981)). To geologist and remote sensing analysts, the semi-variance is most popular (Davis (1986)). To spatial econometricians, estimating spatial autocorrelation coefficients of regression equations is the usual approach (Anselin (1988))." In this work, we use models, which are based on the Moran's I or the semi-variance. The idea behind Moran's I (Moran (1950)) is to measure the correlation of each value x_i to all other values $x_{j \neq i}$ within a predefined distance d of i. The idea for the semi-variance is the expression of the spatial association through the covariance function and formulating this as a (semi-)variogram. This is formally expressed in Cressie and Wikle (2015) for the stationary variogram: " Let $\left\{ Y(s) : s \in D_s \subset \mathbb{R}^d \right\}$ be a real-valued spatial process defined on a domain D_s of the d-dimensional Euclidean space $\mathbb{R}^d$, and suppose that differences of variables displaces **h**-apart vary in a way that depends only on **h**. Specifically, suppose that

$$var(Y(\mathbf{s} + \mathbf{h}) - Y(\mathbf{s})) = 2\gamma_Y(\mathbf{h}), \quad for\ all\ \mathbf{s}, \mathbf{s} + \mathbf{h} \in D_s. \tag{2.2}$$

The quantity $2\gamma_Y(*)$, which is a function only of the *difference* between

the spatial locations **s** and (**s+h**), is called the stationary *variogram*. ”

Several additional concepts and methods of spatial analysis are not are not discussed in this thesis. While they go beyond the scope of this work, we want to give the interested reader several starting points for further study.

The first such concept is the use of the (semi-)variogram to predict unknown variables: Kriging (Krige (1951), Matheron (1963)). It was developed to predict ore deposits based on few samples, but is considered today a standard method for prediction of a manifold a variables. Built upon this is regression-kriging. Hengl et al. (2007) provide an overview for this concept, both in theory and in practice. Gräler (2014) ”presents a new approach that allows to build vine copulas that are aware of separating distances across space and time.” This provides an alternative to Bayesian Hierarchical Models presented in Cressie and Wikle (2015). The final concept for further reading presented here is the idea of spatial decision support. This topic can be approaches from the geo-statistical perspective, e.g. Jankowski et al. (2014), or from an IS perspective, e.g. Ferretti and Montibeller (2016).

Further general reading can be found in Cressie and Wikle (2015)m which focuses more on the combined spatio-temporal statistics and was awarded the DeGroot Prize 2013 as well as the PROSE Awards in 2011 in the Mathematics category.

Part II

Robust Detection of Points of Interest

Chapter 3

Focal Getis-Ord Statistic

3.1 Introduction

THE identification of points of interest is one of the most fundamental tasks in spatio (-temporal) analysis. It shows the spatial distribution and significant deviations of the phenomena under investigation. This allows for a simple, but in-depth overview and guides further analysis. By knowing where regions with higher or lower values are, an analyst can then further delve into the underlying structures or factors, which are present or absent at those locations. An example for this can be found in Wagner et al. (2016). In their work, the authors identify areas and zones, where there is a high demand for cars in a free-floating car-sharing model. They use self-defined thresholds, to determine whether an area has a critical mass of rentals to be considered a point of interest. This is then used as input for further analysis to predict future demand in different areas. A similar example application within the domain can be found in the GIS-Cup 2016 of the ACM SigSpatial [1]. Here, the New York City Yellow Cab taxi trip record was used and the goal was to identify those locations, where the most people exit from the Yellow Cab taxis, given a spatio-temporal spheroid. In this case no threshold was used but instead a test-statistic, the Getis-

[1] sigspatial2016.sigspatial.org/giscup2016/home

Ord statistic (Ord and Getis (1995)). For urban city planers the detection of areas with high temperatures, so called Intra Urban Heat Islands (IUHI), is of particular interest. High temperatures impact energy consumption (Hassid et al. (2000)) as well as human health (Ye et al. (2012)). The effect that the temperatures between an urban area and its surroundings differ, called the Urban Heat Island effect (Oke (1982)), has long been the subject of research. Other examples include crime detection, disease prevention, economic development to name but a few.

The methods to detect areas and points of interest are summarized under the term *Hot Spot Analysis*. The most well-known method, the Getis-Ord statistic, allows to detect areas where the values are significantly different from the mean value of the study area. This enables the identification of points of interest without the need to pre-process the data or pre-define fixed thresholds.

Although most existing methods are independent of concrete values, their results are highly dependent on the size of the study area and their parametrization such as the weight matrix in the case of the Getis-Ord statistic. This dependency can lead to unstable hot spots, where the identified hot spots only appear in one specific combination of parameter. The generalization of insights gained from unstable hot spots is sub-optimal. A researcher or analyst like a city planer who has to rely on those insights will most likely prioritize the wrong area to invest his limited resources.

The goal of this chapter is to identify the factors which influences the results of hot spot analyses and how they can be modified. In section 3.2 we present an overview of methods for hot spot analysis as well as existing approaches to mitigate or eliminate the instability problem. Following this overview, in section 3.3 we will examine and deconstruct the Getis-Ord statistic G^*, as this is the most often used statistic and the basis for most hot spot analyses. Based on the results we propose in section 3.4 a modification of the well known Getis-Ord statistic (G^*): the *Focal Getis-Ord* statistic

(Focal G^*). Instead of the global mean and variance used by G^*, it only uses the mean and variance of a predefined region around each point. This region is a subset of the whole study area. By doing this, the instability is contained within a smaller region and thereby an increase in stability as well as more fine-grained analysis results are achieved.

3.2 Overview of Methods for Hot Spot Analysis

There are two fundamental approaches to detect points of interest: A semi-supervised approach, where there is a pre-known or pre-defined threshold value for the variable of interest and an unsupervised approach similar to clustering, the *Hot Spot Analysis*, which is based on statistical significance levels and tests.

Examples for the semi-supervised approach include the work of Wagner et al. (2016), which was already discussed and the work of Martin et al. (2015) in the field of urban heat islands and temperature. Martin et al. (2015) introduce in their work the definition of intra urban heat islands (IUHI). By defining thresholds with respect to spatial reference, these enable the detection of hot spots in a city, which they call surface intra-UHI. This boils down to five steps, i.e. essentially a comparison of absolute deviation from the mean temperature given a survey area. The results can then be used to detect areas of interest in a city and potentially trigger alerts for a much finer spatial granularity. This approach is limited to a simple counting of measurements or values. Two other well known method, which allow for more complex computations, are the kernel density estimation (Pulugurtha et al. (2007)) and kriging (Oliver and Webster (1990)). Different then the approach discussed before, they estimate the values for each location based on the rest of the study area and a threshold value (Thakali et al. (2015)). Therefore, results for different areas are not comparable, especially in the case of differing value distributions. Kriging was developed

for the estimation of ore deposit (Krige (1951)), but today, applications for geo-temporal forecasts with this approach can be found, e.g. for the city of Zurich[2].

In the rest of the work, the focus will be only on the unsupervised approach. This is derived from the overall goal of this thesis: To provide and extend the existing foundations for spatio-temporal analysis in the field of information science. Therefore for the explanatory analysis, an automated method is needed. It has to be independent from absolute values or thresholds and to be applicable in any setting or area. If for example the method of using pre-defined thresholds is used, it needs a-priori knowledge of the variable of interests and the study area. In addition, any threshold is only applicable to the specific circumstances for which is was determined. In the case of the car-sharing data of Wagner et al. (2016), the used thresholds can only be used for Berlin and not for any smaller or bigger cities such as Freiburg or New York, as the overall number of people and therefore demand is different. The unsupervised approach provides this independence by being based on significance levels. In practice, the goal is to focus on local hot spots and to measure the significance of those local areas. Spatial associations have to be included, i.e. the (local) neighborhood of each point has an influence and has to be included in its value. The following subsections will present a short overview of the most common approaches for this task.

3.2.1 Hot Spot Analysis

One of the most fundamental approach is Moran´s I (Moran (1950)). It provides a hypothesis test for the existence of spatial dependency. This gives the information on global dependencies in a data set. Upon this hypothesis test several geo-statistical tests are based. The most well known

[2] https://r-video-tutorial.blogspot.de/2015/08/spatio-temporal-kriging-in-r.html

are the Getis-Ord statistic (Ord and Getis (1995)) and LISA (Anselin (1995)). In both cases the more general, the global statistic of Moran´s I is applied in a local context. The goal is to detect not only global values, but instead to focus on local hot spots and to measure the significance of those local areas.

The idea behind the Getis-Ord statistic is to transform the existing values to their spatial z-scores. As a z-score can be then transformed to a p-value, we have the significance level for each location. This can then be used to only select those locations which have a significant deviation from the mean, e.g. a z-score of 1.96 for a p-value equivalent of 0.05.

The local Getis-Ord statistic (Ord and Getis (1995)) is defined as follow:

Def. 1 (Getis-Ord G_i^* statistic)**.** *Assuming a study area with n measurements, let $X = [x_1, \ldots, x_n]$ be all values measured in this area. Let $w_{i,j}$ be a spatial weight between two points i and j for all $i, j \in \{1, \ldots, n\}$. The Getis-Ord G_i^* statistic is given as:*

$$G_i^* = \frac{\sum_{j=1}^{n} w_{i,j} x_j - \bar{X} \sum_{j=1}^{n} w_{i,j}}{S \sqrt{\frac{n \sum_{j=1}^{n} w_{i,j}^2 - (\sum_{j=1}^{n} w_{i,j})^2}{n-1}}} \tag{3.1}$$

where:

- *$\bar{X}$ is the mean of all measurements,*
- *S is the standard deviation of all measurements.*

As can easily be seen, this statistic creates a *spatial* z-score, which denotes the significance of an area in relation to its surrounding areas. The standard z-score is defined as the deviation from the mean, measured in z times the standard deviation. This can be transformed directly to a p-value for the statistical significance. By excluding all instances of the weight matrix W, the original z-score is given. As this approach allows for significance tests it is considered a test-statistic.

LISA is quite similar, as it is the local statistic for Moran's I (Anselin (1995)), but the z-score has a different meaning. Apart from G_i^*, LISA does

not distinguish between cold spots and hot spots as it assigns high z-score to most similar areas.

These methods use weights between pairs of points, usually based on their geographical distance. However, in most real world applications, the points are aggregated into a raster representation and the weights are represented as a weight matrix. This allows for expressing the algorithms in terms of map algebra operations, a term first coined by Tomlin (1990) and computed in a distributed fashion (e.g. using GeoTrellis framework running on Apache Spark Eclipse Foundation et al. (2016)).

For further reading the work of Shekhar et al. (2011) is recommended. They present an extended overview of methods to identify and visualize spatial patterns and areas of interest.

3.2.2 Instability in Hot Spot Analysis

While the standard hot spot analysis approaches allow for a fast and automated exploratory analysis of spatial data sets, they are highly dependent on their parametrizations and underlying data. This is quite similar to the challenges in clustering, in particular for the well-known k-means, first coined in MacQueen et al. (1967), or DBScan (Ester et al. (1996)) algorithm. This leads to an instability of the analysis results, which are then difficult to use with high certainty. Here, a selection of approaches and discussions of the last 15 years are presented how to tackle this instability.

The most general approach is to pre-determine and calculate spatial dependencies, based on the a-priori knowledge of the analyst, the empirical data set or best-practices from previous, similar analyses. A good example for this approach can be found in Suomi et al. (2012). In their work, the authors examine the effects of scale on temperature and in particular urban heat island modeling. The pre-determined range of spatial influences is called *buffer zones*; these buffer zones indicate the range of impact in an inner-city temperature measurement scenario. They found that for their

empirical data set a buffer zone of 1000m provides the best results. This is familiar to the well-known approach to solve the inherent problem of DBScan, the determination of its distance: OPTICS (Ankerst et al. (1999)). The disadvantage of this approach is that it has to be manually pre-determined, which results in similar problems as the semi-supervised methods to detect points of interest. It can not be done automatically without introducing an element of uncertainty, which is opposed to our goal to create a more stable approach.

An automated approach is presented with the *A Multidirectional Optimal Ecotope-Based Algorithm* (AMOEBA) in Aldstadt and Getis (2006). The idea behind this approach is to automatically create the optimal, scale-invariant weight matrix and then use this weight matrix in conjunction with a clustering approach to create a graphical overview map of areas of interest. The term ecotope is used for this areas, which is the technical term from the field of biology for the habitat of species. The result is a consistent identification of spatial clusters on a map. In their work they use the G^* statistic as the underlying statistic. The clustering approach is quite similar to DBScan in its approach of creating ecotopes.

A true modification of the G^* statistic is presented in a later work (Getis and Aldstadt (2010)) of the same authors called the LSM (local statistics model). They base their modification on the Kriging approach and its ability to model the spatial autocorrelation as a function dependent on the distance. The idea is to model the weight matrix W as a function of the spatial autocorrelation, where each entry of the matrix is a value derived from the empirical (semi-)variogram. This leads to continuous values up to the so called *critical distance*, which is "defined as the distance beyond which no discernible increase in clustering exist" (Getis and Aldstadt (2010)). They compare their configuration to other, well-known spatial configuration approaches for the weight matrix W. These are taken from Griffith (1996) and in the words of Getis and Aldstadt (2010): ”Research on W has been re-

viewed by Griffith (1996, p. 80), who concludes that five rules of thumb aid in the specification of weights matrices:

1. "It is better to posit some reasonable geographic weights matrix than to assume independence." This implies that one should search for or theorize about an appropriate W and that better results are obtained when distance is taken into account.

2. "It is best to use surface partitioning that falls somewhere between a regular square and a regular hexagonal tessellation." Griffith suggests that for planar data, a specification between four and six neighbors is better than something either above six or below four. Of course, the configuration of the planar tessellations will play a role here (Boots and Tiefelsdorf (2000)).

3. "A relatively large number of spatial units should be employed, n > 60." Following from the law of large numbers, most spatial research, especially due to unequal size spatial units, would require fairly large samples.

4. "Low-order spatial models should be given preference over higher-order ones." Following from the scientific principle of parsimony, it is always wise to choose less complicated models when the opportunity presents itself.

5. "In general, it is better to apply a somewhat under-specified (fewer neighbors) rather than an over-specified (extra neighbors) weights matrix." Florax and Rey (1995) found this result by identifying the power of tests. Overspecification reduces power. They recognize that "Uncertainty with respect to proper specification has long been recognized as a fundamental problem in applied spatial econometric modeling" (p. 132).

"

Ord and Getis (2001) discuss the question in how to formulate the G^* statistic to focus more on local pattern, while still accounting for the global autocorrelation. They propose the O statistic which uses the (semi-)variogram to subdivide the data set into several " 'relatively homogeneous' subregions" (Ord and Getis (2001)). This allows the identification of smaller, more local hot spots, which can be overshadowed in bigger data sets. Finally, they restrict the general applicability in that the version presented in their work requires spatial stationarity.

Westerholt et al. (2015) present a scale-sensitive version of the local G^* statistic, which they call the *GS statistic*. The motivation for this modification is to account for the differences in the scale (the impact of the area under investigation) of the data set, i.e. whether a data set includes only the inner city or also its surrounding area. The problem lies in the detection and use of the local context of the gathered data. A fixed weight matrix W does not include the difference in context, e.g. in Twitter feeds. Their approach is to redefine the neighborhood of a data point with upper and lower distance thresholds, which are then used in pairwise comparisons. Only sufficiently connected data points within their thresholds are considered to be viable as a hot spot and only those points are used for the global mean and global deviation values. They evaluated their approach on Twitter data of the city of San Francisco, USA and show that this leads to reduction or even negation of cross-scale interference. The main restrictions of this approach lies in its increased computational costs as well as its reliance on continuous distance functions.

For futher literature regarding the creation of optimal weight matrices for spatial associations we refer to the work of Aldstadt and Getis (2006), where they provide an exhaustive overview of the state of the art.

In this work, the Getis-Ord statistic is used as the basis for a more stable hot spot analysis. The reasons for this is the simplicity of its formula and the ease of interpretation of the resulting z-score, while retaining a high

explanatory power. A similar reasoning can be found in the the aforementioned works and is one of the reasons why the G^* statistic remains the most popular method for hot spot analysis in research as well as practical applications. For the analysis of the parameter influences the form presented in definition 1 is used, and in the later formulation both the mathematical as well as focal variant is used and described.

3.3 Analysis of Influences on Stability

Based on the existing work, the goal of this section is to answer RQ1a: "What effects and parameter influence the stability of hot spot analysis?" in an analytic way by dissecting the G^* statistic into its single parts. Recall definition 1:

Def. 2 (Getis-Ord G_i^* statistic)**.** *Assuming a study area with n measurements, let $X = [x_1, \ldots, x_n]$ be all values measured in this area. Let $w_{i,j}$ be a spatial weight between two points i and j for all $i, j \in \{1, \ldots, n\}$. The Getis-Ord G_i^* statistic is given as:*

$$G_i^* = \frac{\sum_{j=1}^{n} w_{i,j}x_j - \bar{X}\sum_{j=1}^{n} w_{i,j}}{S\sqrt{\frac{n\sum_{j=1}^{n} w_{i,j}^2 - (\sum_{j=1}^{n} w_{i,j})^2}{n-1}}} \tag{3.2}$$

where:

- *$\bar{X}$ is the mean of all measurements,*
- *S is the standard deviation of all measurements.*

It can be easily seen that this equation can be divided into three different parts and their separate influences:

1. The variable under observation X and its single elements x_j.
2. The weight matrix W, i.e. the neighborhood, and its elements $w_{i,j}$.

3. The global mean $\bar{X}$ and the global standard deviation S.

The term $\sqrt{\frac{n\sum_{j=1}^{n} w_{i,j}^2 - (\sum_{j=1}^{n} w_{i,j})^2}{n-1}}$ is only dependent on the weight matrix W and represents the standard deviation of the weight matrix. It it therefore already included in the discussion of the weight matrix W.

The first parameter to discuss is X. As the variable under observation, it is the key part of the method. If the spatial influences are negated, the formula would reverse to the computation of a z-score and the parameter X would be the only influence. It is only dependent on the values of the variable under observation. Therefore there do exist three basic possibilities to influence or modify this parameter: (1) The determination in how to measure the data used to compute the G^* values in the real world, or (2) the data pre-processing step, i.e. how to filter low quality measurements, or (3) the decision whether and how to aggregate the measurements before analysis. Each of these possibilities are decisions made before the analysis and are often outside the control of the analyst or highly dependent on the context.

The second parameter to discuss is the weight matrix W. As the name implies, it is the weight for each measurement x_j and models the spatial association. It determines how much each (spatially) neighboring measurement influences the measurement at point i. It is dependent on the overall spatial association of the variable and the context, under which the analysis is performed. The greater the value of $w_{i,j}$, the higher the influence is on the measurement at j. The number of values $\neq 0$ of W is often called the size of the weight matrix. In its most simply implementation, the values are binary and a W of 3x3 means that only the values of direct neighbors in a queen distance have an influence. Overall, W is highly dependent on the context as well as the underlying data. Its determination is one of the main tasks of the analyst before the analysis. Most of the existing work in creating more stable or scale-invariant modifications of hot spot analysis have therefore focused on this parameter.

The third parameters are the global mean as well as global standard de-

viation. As with the standard z-score these parameter are the basis for the normalization. The global mean provides the zero point of the given empirical distribution. By being weighted with the sum of $w_{i,j}$ the spatial global mean is computed, which represents the mean value given the spatial association. The global standard deviation, multiplied with the standard deviation of the weight matrix, represents the spatial normalization factor. The benefit of their use is that they still represent a z-score. As it is well-known, the z-score can be easily transformed to a p-value. The typical p-value for high significance of 0.05 is equivalent to an absolute z-score of 1.96, meaning that the value of the variable at this point is 1.96 times the standard deviation higher (or lower) than the mean value. As W as well as X are already discussed, the only influence on these global values is the size of the study area, which is also called the reference area.

3.4 Focal Getis-Ord

In this section an approach to answer the RQ 1.b: ”How can existing methods be modified to be more stable?” is proposed and discussed. The existing approaches for increasing the stability of hot spot analyses are based on modifications of the weight matrix W. The reasoning is that this parameter is influenced by the spatial association of the empirical data set and can therefore be modeled as a dependent variable on the association. Here, we propose a different approach: To modify the global parameter. The reasoning is similar to Westerholt et al. (2015): "One recurring problem with spatial autocorrelation statistics is their sensitivity to spatial scale effects". The problem with modifying W lies in the influence and meaning behind this parameter. The specification of W can be modified by the analyst based on the goal as well as the context of the analysis. Given the example of an urban planner, who wants to determine the availability of shopping or medical facilities. W in this case has to be chosen, at least partly, based on

laws and regulations to determine if there is a shortage of these facilities for the number of persons in their influence area. Particular for questions in the field of information science, this context has to be included and can often not be automated. The global parameter on the other hand are often chosen without regard to their influence. Therefore their impact on the G^* value is not determined by the intent of the study or regulation and therefore avoidable. In addition, even a small change in the size or exact location of the study area changes these global values and therefore the normalization results. This prevents the comparison of two different study areas with the standard G^* statistic, as their global values will likely differ.

The mitigate the influence of the study area and create a more stable hot spot analysis we propose the *Focal Getis-Ord* statistic (Focal G^*). The idea behind this modification of the existing G^* statistic is the replacement of the global parameter through *focal* (regional) parameter. These are used instead of the global parameter to normalize the values and thereby creating a *focal* z-score. The mathematical formulation is given in Def. 3:

Def. 3 (Focal Getis-Ord G_i^* statistic)**.** *Assuming a study area with n measurements, let $X = [x_1, \ldots, x_n]$ be all values measured in this area. Let $w_{i,j}$ be a spatial weight between two points i and j for all $i, j \in \{1, \ldots, n\}$ and $w'_{i,k}$ be a spatial weight between two points i and k for all $k \in \{1, \ldots, m_i\}$, where $m_i \leq n$. The Focal Getis-Ord G_i^* statistic is given as:*

$$FocalG_i^* = \frac{\sum_{j=1}^{n} w_{i,j}x_j - (\frac{1}{m_i}\sum_{k=1}^{m_i} w'_{i,k}x_j)\sum_{j=1}^{n} w_{i,j}}{\sqrt{\frac{\sum_{k=1}^{m_i} x_k}{m_i} - \bar{X_{m_i}}}\sqrt{\frac{n\sum_{j=1}^{n} w_{i,j}^2 - (\sum_{j=1}^{n} w_{i,j})^2}{n-1}}} \tag{3.3}$$

The focal values for each point and thereby the normalization are computed for each point separately and are therefore independent of the size of the study area. As can easily be seen, as the value for m_i converges to n, the results of the Focal G^* will converge to the results of G^*. The Focal G^* statistic still results in a z-score for the same reasons as shown in Ord and

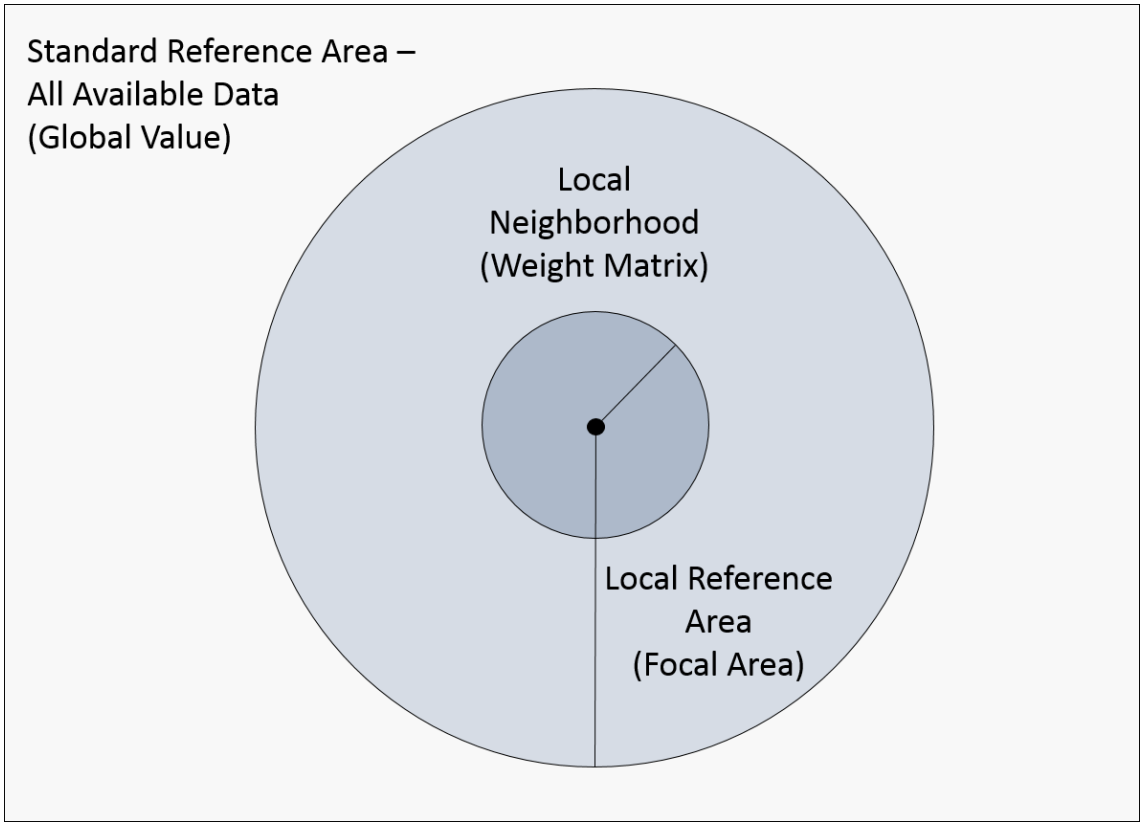


Figure 3.1: Overview of the different influence areas for hot spot analysis. The local neighborhood describes the area of direct influence on each single point. The local reference area is the user-defined comparison area for the Focal G^* statistic and redefines the used reference area to determine the focal mean and variance. The standard reference area is used for the determination of the global mean and variance in the standard G^* statistic. It is given by the study area.

Getis (1995). But the z-value is only applicable in comparison to values of the focal radius defined by m_i.

A graphical overview of the different areas can be found in Figure 3.1.

A more efficent computation can be shown by applying the use of the hot spot analysis to raster data. As stated before, most spatio (-temporal) data sets are available as raster data, e.g. each pixel of a satellite measurement represents a single entry in a raster file. This allows the use of the aforementioned focal operations and enables the reformulation and transformation of the formula for the G^* and Focal G^* statistic into a computationally more efficient form, as presented in Bruns and Simko (2017): " In the following

text, we use the notation $R \overset{\mathtt{op}}{\circ} M$ to denote a focal operation op applied on a raster R with a focal window determined by a matrix M. This is roughly equivalent to a command `focal(x=R, w=M, fun=op)` from package *raster* in the R programming language Hijmans (2016b).

Def. 4 (G^* function on rasters)**.** *The function G^* can be expressed as a raster operation:*

$$G^*(R, W, st) = \frac{R \overset{\mathtt{sum}}{\circ} W - M * \sum_{w \in W} w}{S \sqrt{\frac{N * \sum_{w \in W} w^2 - (\sum_{w \in W} w)^2}{N-1}}}$$

where:

- *R is the input raster.*
- *W is a weight matrix of values between 0 and 1.*
- *$st = (N, M, S)$ is a parametrization specific to a particular version of the G^* function. (Def. 5 and 6).*

Def. 5 (Standard G^* parametrization)**.** *Computes the parametrization st as global statistics for all pixels in the raster R:*

- *N represents the number of all pixels in R.*
- *M represents the global mean of R.*
- *S represents the global standard deviation of all pixels in R.*

Def. 6 (Focal G^* parametrization)**.** *Let F be a boolean matrix such that: $all(dim(F) \geq dim(W))$. This version uses focal operations to compute per-pixel statistics given by the focal neighbourhood F as follows:*

- *N is a raster computed as a focal operation $R \overset{\mathtt{sum}}{\circ} F$. Each pixel represents the number of pixels from R convoluted with the matrix F.*

- *M is a raster computed as a focal mean* $R \overset{\texttt{mean}}{\circ} F$*, thus each pixel represents a mean value of its F-neighbourhood.*
- *S is a raster computed as a focal standard deviation* $R \overset{\texttt{sd}}{\circ} F$*, thus each pixel represents a standard deviation of its F-neighbourhood. "*

Def. 5 and Def. 6 show that for raster files the Focal G^* formulation represents a more general formulation.

3.5 Conclusion Chapter Focal G^*

In this chapter the research question 1.a and 1.b are discussed. Based on the discussion in section 3.3 RQ 1.a is answered in that three parameter and their effects can be found: (1) The variable under investigation and (2) the spatial association and its context and (3) the study area. Each of these parameter has different underlying effects. The first two are dependent on the on the context of the analysis, regulations, available measurements, quality of measurements and other factors which are often outside the influence of the researcher or have to be carefully chosen before the analysis and therefore can be considered to be fix. The third parameter is important for the comparability of the results, but often chosen randomly or by circumstances.

In section 3.4 RQ 1.b is (partly) answered by the development of a new, more stable hot spot analysis based on the G^* statistic: The Focal G^* statistic. It is based on the insights of the existing literature as well as the results of RQ 1.a.

In addition to creating a more stable version of G^* by eliminating the influence of the overall study area, the Focal G^* statistic has an additional benefit: It provides a more detailed view of the results and the identification of more localized hot spots. In the standard statistic, few high (or low) value points or areas can have a high impact on the overall computation,

similar to outliers, which may overshadow other areas of interest. By using the focal values and a suitable focal radius, more local hot spots can be identified reliably and consistently. This in turn allows for the creation of more consistent results, which allow for better decision processes and the better allocation of scarce resources.

Chapter 4

Stability of Hot Spot Analysis

In the previous chapter a new approach to create stable hot spot analysis was presented as well as other existing approaches to do so. However, the evaluation in the existing work was only done on a visual basis and on very specific data sets. A comparison between these different approaches has up to now be done visually and on the same data set. For an automated analysis and comparison this is insufficient and leads to RQ2: "How can the stability of found Hot Spots be measured?".

To answer this question and solve the underlying problem, in this chapter a metric to measure the stability of an hot spot analysis is proposed called the *Stability of Hot spot* (SoH). This metric measures whether a hot spot found for a given parametrization is carried over to the found hot spots with a different parametrization. This enables the quantification of the stability of any hot spot analysis and provides the foundation for automated approaches.

4.1 Motivation

Consider the real-world example depicted in Fig. 4.1. The temperature map of a morning thermal flight data set (a) has been processed using the G^* statistic with an increasing size of a weight matrix (b, c and d).

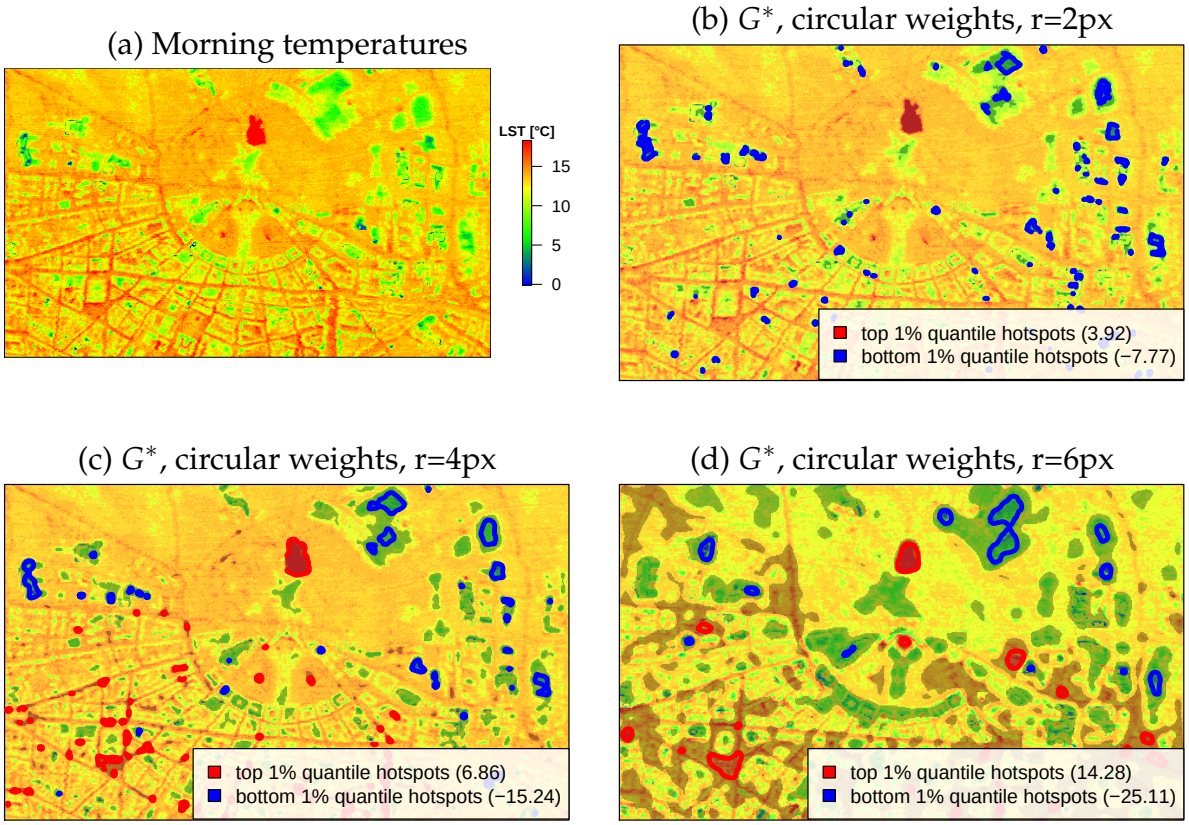


Figure 4.1: Karlsruhe city center. Selected area of $2.4 \times 1.4\,km$. Pixel size $5 \times 5\,m$.

As can be seen, hot spots are oftentimes disappearing or appearing unrelated to previously found hot spots. While these computations indeed show hot spots and the results are correct for their parametrization, they lack stability.

For a data analyst, when exploring the data interactively by choosing different filter sizes (in form of matrices), it is important that the hot spot position and size changes in a predictable manner. This intuition is the basis for the presented stability metric.

As stated before, the goal of hot spot analysis is the detection and identification of interesting areas. It achieves this goal by computing statistically significant deviations from the mean value of a given study area. This allows a decision maker to easily identify those areas of interest and allows

further focus in subsequent data analysis or the decision process. In typical applications, scarce resources are then often applied only in identified hot spots or these are used as the basis for the allocation.

But, similar to a cluster analysis, there does exist a high dependency of the identified hot spots on the detection method and in particular the parametrization of this method. The identified areas as well as their shape can vary highly. This volatility can lead to a decrease in trust in the result and in suboptimal allocations of scarce resources. Therefore it is necessary to measure and evaluate the stability of a hot spot analysis as well as the different parametrizations.

4.2 Overview of Quality Metrics for Unsupervised Learning

There does already exist work regarding the quality of unsupervised learning approach, in particular clustering. This problem of assessing the quality in unsupervised learning is well known, or in the words of Shamir and Tishby (2008): "A major problem [...] is assessing cluster validity." In the most popular case of the k-mean algorithm, the quality of the clustering is mostly dependent on the value of the k and a miss-specification can lead to highly irregular clusters. In a simple 2D clustering, they can be easily recognized by visual analysis, but in higher dimensionality, this is impossible. One method to measure the quality of such a clustering is the compactness of the clusters, see e.g. Song (2010). This enables the automated comparison between different clusters. Another possibility is the Silhouette Coefficient by Rousseeuw and Kaufman (1990). This metric measures the similarity of objects in a cluster in comparison to other clusters. For density based clustering, e.g. for DBSCAN Ester et al. (1996), OPTICS Ankerst et al. (1999) gives a simple method to tune the essential parameter for this clustering. This is only a concise overview of the state of

the art approaches to influence and measure the quality of different clustering methods. But it shows that this problem is not easily solved and dependent on the chosen algorithm.

Shamir and Tishby (2008) discuss the topic of cluster stability in the context of sample size. They motivate their work by stating that according to the literature stability metrics for clustering are lacking as a tool, as with increasing sample size the quality and stability of a clustering converges. But while the quality of clustering may converge, the rate at which this happens deviates between different clustering algorithms. This is similar to the model selection problem and its frameworks, e.g. the VC dimension (Vapnik and Chervonenkis (2015) in a newer translation from the original 1968 publication). They proceed to create a Bayesian framework to demonstrate their idea and state that stability measures for clustering are important in that they help in the decision of which algorithms to use. These measures can be created independent from the sample size. Of further interest for this thesis is the definition of stability, cited from Ben-David et al. (2006):

Def. 7 (Stability of Clustering Algorithm)**.** *"Following [2], we define the stability of a clustering algorithm A on finite samples of size m as:*

$$stab(A; D; m) = \mathbb{E}_{S_1,S_2} d_D(A(S_1), A(S_2)), \tag{4.1}$$

where S_1 and S_2 are samples of size m, drawn i.i.d from D, and d_D is some 'dissimilarity' function between clusterings of X, to be specified later." For this definition is A a clustering algorithm and D is the underlying distribution of the data.

A formalization of clustering quality and the conditions for the metrics are discussed and presented in the work of Ben-David and Ackerman (2009). Their axioms are based on the work of Kleinberg (2003), but instead of a general definition of clustering these axioms are adapted for the definition of clustering quality measures. "A *clustering-quality meas-*

ure is a function that maps pairs of the form (*dataset*, *clustering*) to some ordered set (say, the set of non-negative real numbers), so that these values reflect how 'good' or 'cogent' that clustering is" (Ben-David and Ackerman (2009)). Three main axioms are defined: **Scale Invariance**, **Consistency** and **Richness**. They prove these axioms by present clustering measures which fulfills these axioms, which in Kleinberg (2003) no clustering could achieve. A strong limitation in their work is that clustering cannot be clearly defined, but, in their words, they use common and uncontroversial clustering methods as examples, which are used as relaxed replacements for lack of a formal definition of clustering. Several more clustering quality measures are discussed and compared to show the benefit of their axioms.

Grubesic et al. (2014) present in their work an overall overview of spatial clustering, its techniques as well as its particular challenges. The main goal of their work is to provide the basis for informed decisions by comparing the different approaches and their trade-offs. The identification of true spatial clusters has additional challenges compared to classical clustering which arise from the spatial association. Existing non-spatial clustering approaches have to modified; they describe the impact of particular the distance function in cases of k-means clustering. Here, the squared distance can lead to an overemphasize on outliers. This is similar to the problem of correctly defining the weight matrix W in the previous chapter. In their work, spatial clustering is divided into four broad classes: Non-hierarchical, Scan, Spatial autocorrelation and Hybrid. In their evaluation seven common clustering algorithms are compared. A synthetic data set with pre-defined clusters is used to compare the accuracy of the different methods. These are evaluated by common clustering metrics such as the Rand Index, the F1-Score as well as the log-likelihood ration (Kulldorff (1997)). No overall best score can be achieved, but several trade-offs could be observed. Of interest to the authors is that there is a particular trade-off between the log-likelihood ratio and the spatial accuracy of the clustering

methods.

Lukasczyk et al. (2015) present an alternative approach. In their work, the problem of stability can be circumvented by creating better visualization techniques and approach to compare different parametrizations. They use reeb graphs to select suitable parametrizations and visualizations of hot spots. The focus lies in the temporal changes of hot spots and to show how show how these change over time. This leads in their eyes to a better understanding of the phenomenon under investigation, particular through the inclusion of humans and their visual processing.

4.3 SoH Metric

As seen in the literature there does exist a manifold of metrics for the quality of *clustering*, but apart from visual possibilities, no such metric does exist for the stability of hot spots. Therefore, we propose a simple metric called the *Stability of Hot Spot* (SoH). It measures the deviation from a perfectly stable transformation between different parametrizations. The idea is based on the definition 7 from the clustering domain: Stability is the difference between two clustering results. This achieved by comparing two hot spot results, where one result is defined as *parent* and one is defined as *child*, similar to a tree in the domain of computer science.

A hot spot found in comparably more coarse resolutions is defined as *parent* (larger weight matrix) and in finer resolutions as *child* (smaller weight matrix). To be stable, one assumes that every parent has at least one child and that each child has one parent. For a perfectly stable interaction, it can be easily seen that the connection between parent and child is a injective function and between child and parent a surjective function. The resulting metric consists of two cases and comparisons:

In its downward property (from parent to child, injective) it is defined

as:

$$SoH^{\downarrow} = \frac{ParentsWithChildNodes}{Parents} = \frac{|Parents \cap Children|}{|Parents|} \tag{4.2}$$

And for its upward property (from child to parent, surjective):

$$SoH^{\uparrow} = \frac{ChildrenWithParent}{Children} = 1 - \frac{|Children - Parents|}{|Children|} \tag{4.3}$$

where *ParentsWithChildNodes* is the number of parents that have at least one *child*, *Parents* is the total number of *parent*, *ChildrenWithParent* is the number of children and *Children* as the total number of children. The SoH is defined for a range between 0 and 1, where 1 represents a perfectly stable transformation while 0 would be a transformation with no stability at all.

This simple metric can be computed by a comparison of the differing overlays and its advantages are the results of its simplicity as well as its plausibility. The results can be easily used in an automated comparison of a manifold of parametrizations and be computed in parallel. This solves the time consuming comparison of stability by visual analysis through domain experts.

4.4 Conclusion SoH

In this chapter, a first metric for the stability of hot spots is proposed, the *Stability of Hot Spot*. The metric is derived and inspired by clustering quality measures. This allows for a fast and easy comparison between different hot spot parametrizations in a single number between 0 and 1 and enables a departure from the pure visual analysis of the stability. Based on this number an analyst or algorithm can decide which parametrization to use and researchers can compare the stability of their methods for unsupervised hot spot analyses. The metric is simple by design. As of the time of writing it is the first metric to measure the stability of hot spots. Therefore the answer to RQ 2 is twofold: (1) In the literature visual comparison

is used to evaluate the stability of hot spots. No metric or direct measurement has been developed or used. The stability is, if measured at all, measured afterwards during the spatial clustering with stability measures for the clustering. (2) The proposed SoH metric enables a simple metric to measure the stability of hot spots in a single value. This helps to produce informed decisions.

Chapter 5

Empirical Evaluation for Robust Detection of Points of Interest

AFTER defining a new approach for stable hot spot analysis and creating a metric to measure the stability, in this chapter those results are used in an empirical evaluation to answer RQ 1.c: "What is the optimal parametrization for an hot spot analysis?".

5.1 Empirical Data Set

As stated in chapter 1, temperature data is used in this thesis as empirical data for all evaluations. Temperature itself is one of the most fundamental environmental factor, underlying most processes in nature and particularly in cities. The location of intra urban heat islands is essential for a manifold of tasks a city planner has to do, especially in a so called *Smart City*. In this chapter, the evaluation is build upon two snapshots of temperature data, available in the raster file format.

The two data sets (morning and evening flights) depicted in Fig. 5.3 and Fig. 5.4 were obtained from a thermal flight over the city of Karlsruhe on 26.09.2008 at 6:30–7:45 and 20:00–21:30. The flights were executed by the

Nachbarschaftsverband Karlsruhe[1]. A single pixel in the raster represents an area of approximately $5 \times 5m$. The whole data set of size $35 \times 25\,km$ was cropped into the inner city area of $2.4 \times 1.4\,km$. The temperature in this data set ranges from -1.7°C to 18.3°C. Missing values in the data set are interpolated using a focal median function with a square matrix of 11x11 pixels, mainly for speeding up further computations and to avoid special handling of the missing values.

5.2 Parametrization

In this evaluation two different contributions have to be evaluated:

1. The stability metric SoH.
2. The stability enhancing modification of the Getis-Ord statistic, the Focal G^* statistic.

The Focal G^* has to be evaluated by the use of the SoH metric, as there does not yet exist another stability metric for the evaluation of the stability of hot spots as of writing this thesis. The evaluation of the SoH has to be done therefore visually. This is achieved by combining the evaluation of the stability of Focal G^* and comparing these results with the visual differences of the found hot spots of both the G^* and Focal G^* statistic. This approach is inspired by the methodology of Visual Analytics (Keim et al. (2008)) in that the human ability to quickly process graphical information is used to confirm the algorithmic results.

To evaluate the stability of the proposed Focal G^*, it is compared to two baselines:

- Standard G^*, which uses the same weight matrix W as the focal version.

[1] http://www.nachbarschaftsverband-karlsruhe.de/

- Standard G^*, which uses square weight matrix with all cells set to 1.

These two weight matrices are chosen to compare both the difference in shape as well as the impact of the focal region to the stability of hot spots. The Focal G^* parametrization uses a round shape for both the focal as well as weight matrix as this guarantees a consistent maximal distance for each raster file. In fig. 5.1 the spatial form of two example matrices is shown.

(a) Weight matrix W (b) Focal matrix F

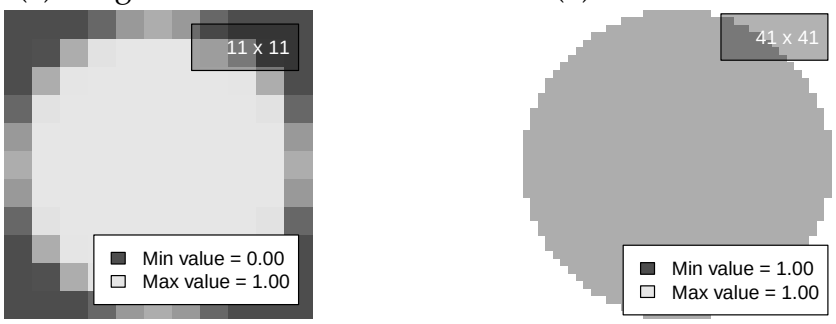


Figure 5.1: Example matrices W and F

The weight matrices W for this empirical evaluation are iterated from a minimal value of 1 to the maximal size of 41. The stability of each iteration is compared to the most similar matrix. Here, only the $SoH^{\uparrow}$ is used and mathematically the following comparisons are performed:

$$SoH^{\uparrow}(G^*(R, W_i, st), G^*(R, W_{i+2}, st))$$

Only $SoH^{\uparrow}$ is used for ease of the visual analysis. The size of the focal matrix is fixed to a square of 41×41.

5.3 Results and Discussion

Fig. 5.3 and Fig. 5.4 show standard and Focal G^* computations for both morning and evening data sets with weight matrix W of size 3, 5, 7, 9, 15 and 31. The evaluation results are plotted in Fig. 5.2, each point in the graph represents the $SoH^{\uparrow}$ metric (Eq.4.3) between two G^* generated using weight matrices of size i and $i+2$.

The results for the hot spot analysis are found in Fig. 5.3 and Fig. 5.4 for a comparison of G^* and the Focal G^* statistic. It can easily be seen that both versions produce similar results, but the focal versions produces a more differentiated picture for larger weight matrices. Small differences on a global scale are more pronounced on a regional scale and result in smaller and finer areas for hot spots. This enables the detection of additional hot spots and interesting areas which are most easily observable for the weight matrix of size 7×7 in the evening (Fig. 5.4). This enables the detection of significant deviations from the surrounding area. In contrast the standard G^* statistic only shows larger areas as important. Therefore, depending on the need of a planner, the Focal G^* statistic is more helpful to identify individual areas of interest whereas the standard G^* statistic gives a more broad overview. For the identification of local phenomenon such as intra urban heat islands this the identification of individual areas is quite important. As a city planer wishes to detect those critical areas, it is important to detect not only general hot areas, but also those points where the most extreme differences in a local context exist. Finding those areas can help identify the underlying reasons or plan individual solutions. This can then be used as the basis for further analysis and its parameter selection.

Based on these images one can also see that the hot spots found by the Focal G^* statistic seem to be more stable. A plot of the results can be found in Fig. 5.2. Fig. 5.2 compares the $SoH^{\uparrow}$ between each increasing size of the weight matrix W. At a first glance, one can see that the typical implementation with a square weight matrix is the most unstable hot spot analysis,

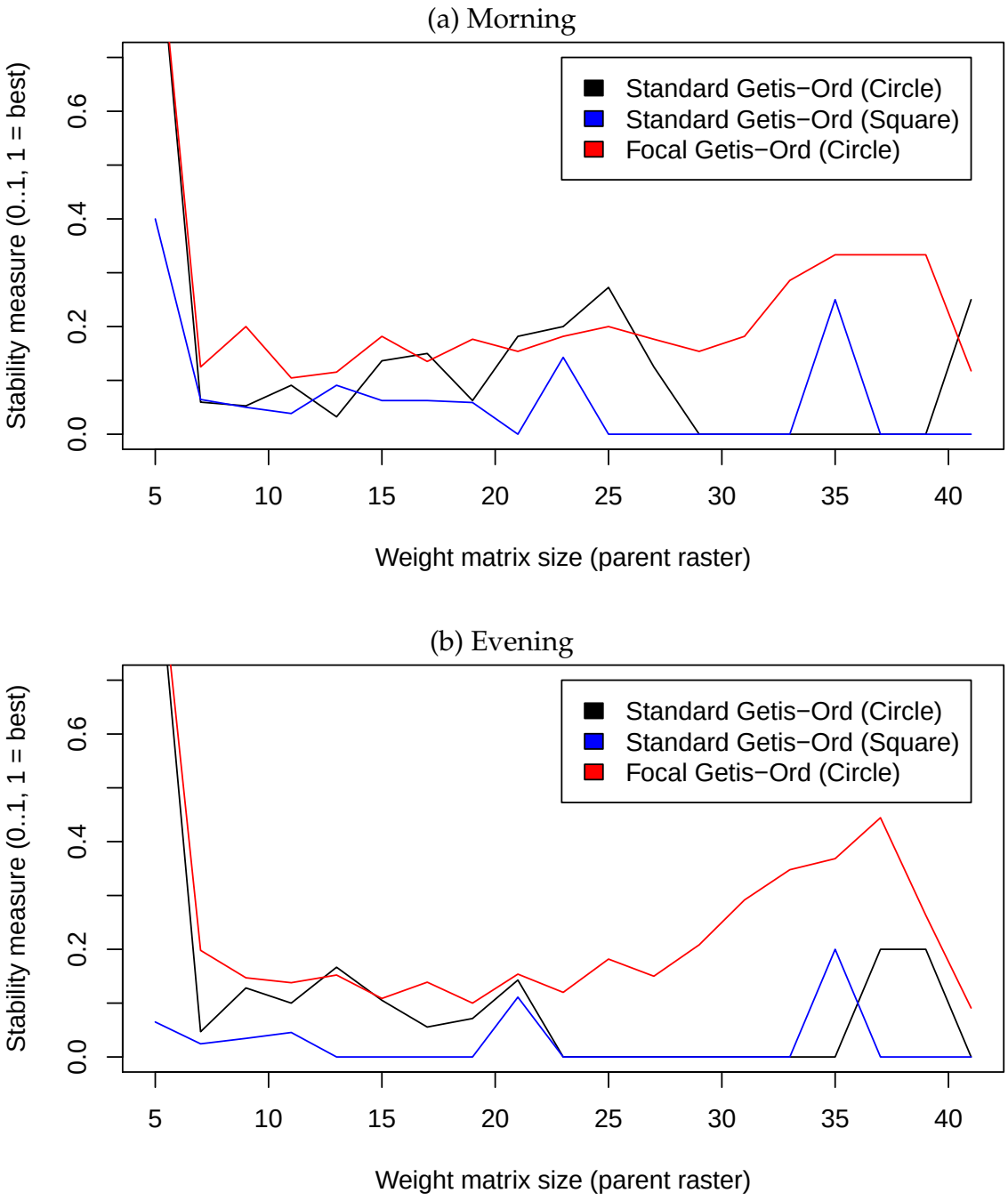


Figure 5.2: Evaluation results - Standard vs Focal G^*

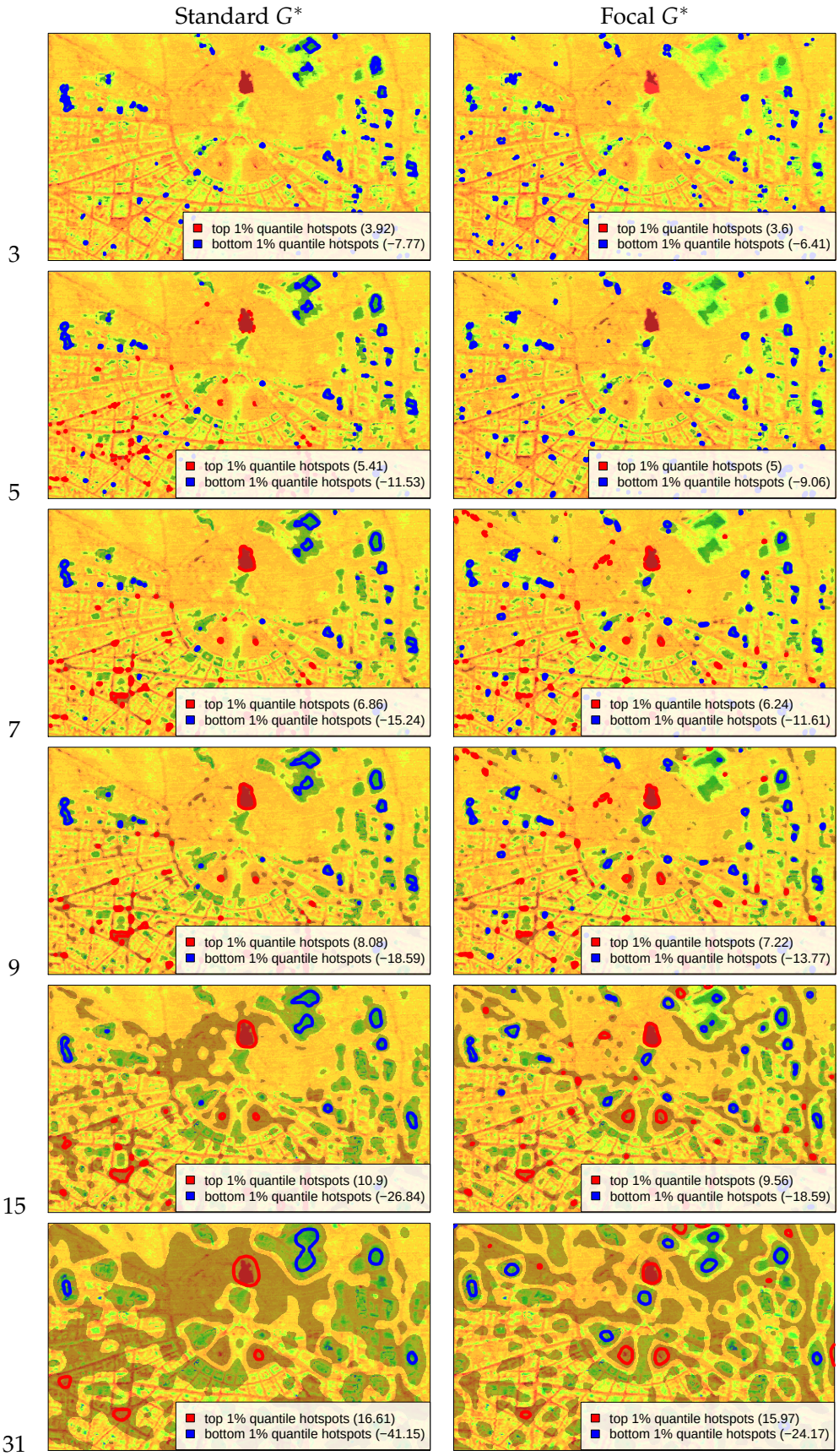


Figure 5.3: Standard and Focal G^* with different weight matrices applied on morning data set.

Figure 5.4: Standard and Focal G^* with different weight matrices applied on evening data set.

regardless of time of day. This is to be expected as the binary weights increase the dependence on the weight matrix. The use of decreasing weight matrices leads to an increase in the stability performance. As the more outlying data points get less weight, this reduces the dependence on the weight matrix and leads therefore to more stable results. Our proposed Focal G^* statistic achieves the most stable results in almost all cases. Only data points in a restricted region around the area of interest may influence the significance result. Through this restriction high values at key points gains more weight regardless of the weight matrix and are therefore more independent of the weight matrix. This increases the stability. The decrease in stability for the largest weight matrices is most likely a result of the parametrization of the focal matrix. With increasing size of the weight matrix in relation to the focal matrix the value of each pixel is approaching to the mean of the area of the weight matrix. As can be easily seen from Def. 4 then the value for every pixel would be zero.

In regard to the research question 1.c, the empirical results are mixed. While the best values for the Focal G^* are concentrated on the biggest weight matrices, the results for the standard G^* are in the range of 20-30 and less stable above that size. This is similar to Florax and Rey (1995) who stated: "In general, it is better to apply a somewhat under-specified (fewer neighbors) rather than an over-specified (extra neighbors) weights matrix." The results indicate that the inclusion of a wider range and therefore increased weight matrix leads to more stable results for Focal G^*, especially as the value of W and F converge. As only a fixed F is chosen, this has to be examined in more detail and no conclusive answer can be given for this question. The results below the value of 7×7 represent almost no spatial association for this data set and can not be considered as most stable. This result is supported by the graphical results.

5.4 Conclusions and Future Work

In this part of this thesis, RQ 1 and RQ 2 were discussed and evaluated. First, the Getis-Ord statistic was analyzed to detect influences on the stability of hot spot analyses. Three main reasons are identified: The weight matrix W, the aggregation level of the variable under investigation X and the size of the study area. This lead to a modification of the existing G^* statistic, the *Focal G^* statistic*. It reduces the impact of the study area used for comparison by replacing the global variables with focal variables and achieves through this modification an increase in stability. This allows for a more local analysis and a reduction of the impact of outliers. Additionally, the analyst can define the focal matric F before the analysis to account for spatial dependencies and context. The computation of this statistic can be easily parallelized by its reformulation as a focal operation. The $G*$ statistic is used for its inherent simplicity as well as overwhelming use in the literature. Therefore, the presented modification can be easily augmented by using existing approaches for the weight matrix W such as Getis and Aldstadt (2010) and Westerholt et al. (2015). These approaches come with a higher computational cost and have therefore to be considered carefully. The presented Focal G^* has already an increased computational cost (see e.g. Gassenschmidt (2017)) and the combination could become prohibitive for standard setups and uses.

To show the improvement in the stability, a quality measure is needed. In the existing literature only visual comparisons such as used in Lukasczyk et al. (2015) existed as of the time of writing. To enable an automated measurement a stability metric called the *Stability of Hot spots* (SoH) was created. It is inspired by quality metrics from the field of clustering. The SoH computes the ratio of dependence of hot spots for different parametrizations. It enables to express the stability between each parametrization using single value restricted between zero and one. Based on this number one can decide which parametrization to use and researchers can compare the sta-

bility of their methods for unsupervised hot spot analyses. In particular, for temperature values one wishes to detect those areas which have high differences regardless of the particular parametrization. If a hot spot only appears for one parametrization, the information gained for general use is quite small and can even lead to an inefficient allocation of resources.

In the empirical evaluation the benefit of the Focal G^* statistic as well as the SoH was shown. The Focal G^* is more stable over the different parametrizations and the visual results provide a better local view of interesting areas. The use of the SoH metric was shown for the $SoH^{\uparrow}$ and the visual results confirm the results of the metric.

This research has several restrictions which have to be taken into account. First, only the $SoH^{\uparrow}$ metric was used. While it can be assumed, based on the graphical analysis, that the $SoH^{\downarrow}$ stability should be similar, there are no quantifiable results. The results themselves are tested on two events in time for a fixed area of the city of Karlsruhe. It is not tested for smaller or larger study areas, but it can be assumed that the stability of the Focal G^* would stay the same whereas the stability of the G^* statistic would increase with a smaller study area and decrease with a larger study area. This follows the reasoning that the impact of a singular point increases with a decrease of the study area. The last restriction is the fixed size in this work of the focal matrix for the Focal G^* approach. Only one focal matrix F was tested, but it is highly probable that the size of the focal matrix has an impact on the stability as could be seen in Fig. 5.2. While an overall trend can be seen in this work when the size of the weight matrix W and the focal matrix F are almost identical, the exact ratio is beyond the scope of this work. The optimal ratio as well as when the stability suffers from a too similar size are interesting question for future work. Further research in this direction was done in Gassenschmidt (2017), where he used the New York Yellow Cap data set as empirical data and compared the stability with the inclusion of the temporal dimension.

Part III

Causal Modeling of Temperature Differences

Chapter 6

Land use-based Temperature Model

THE previous section introduced a new metric as well as a new statistic to reliably detect hot spots. But this can only be done when there is enough fine-grained data. In addition, as discussed before, the exploratory analysis leads a researcher to interesting points, but it does not provide a direct explanation *why* this is interesting. To do so, further and different modeling is needed, first a descriptive model and then a predictive model. The insights gained from the explanatory analysis of this work as well as the existing urban climate literature is therefore used as a basis to answer the RQ 3: *How can temperatures in an intra-urban setting be predicted?* To answer this question, first the RQ 3a is discussed: *What are causal drivers behind local temperature differences?* to identify the causal influences on temperature within cities and this is then used in RQ 3b: *Given the inherent complexity of the underlying meteorological, environmental and physical processes and the sparseness of available meteorological data, how can those drivers be modeled to produce an accurate and robust prediction?* to utilize those insights to produce a robust and fine-grained temperature prediction.

6.1 Introduction

In this part a causal model is proposed to explain and predict intra-urban temperature differences and their developments over time. While it is regularly observed that large temperature differences exist not only between cities and rural areas surrounding them but also between nearby locations within urbanizations, historically research almost exclusively focused on the first phenomenon, referred to as urban heat islands (UHI). This term also describes the research regarding the underlying reasons as well as the implications of increased temperature within cities. But the second phenomenon, small geographic regions within a city with extreme temperatures peaks, received scant attention in the literature so far. While different underlying causal reasons for temperature development, such as wind direction and speed, different land use, green areas or the reflectence of light are generally known, no model exists that makes use of these data sources and their interaction to predict intra urban temperature differences and their developments, e.g. between close-by neighborhoods. In particular, no models are available to determine or anticipate intra urban heat islands (IUHI).

We argue that such models are necessary for the understanding and modeling of urban temperatures and are key to smart city planning and operation as well as urban information systems. Temperature and specifically heat-stress heavily impacts human health, well-being and work performance as well as other city variables such as pollution and energy consumption. Examples for the use of temperature can already be found in the literature. For instance, one can exploit temperature measurements to minimize the impact of heat stress on humans regarding their daily routines and walking paths (Rußig and Bruns (2017)). A graphical example from the paper can be found in Fig. 6.1. In case of load planing for electric grids, the impact of air conditioning load on a grid can be forecasted depending on the temperature distribution in a fine grained manner (Hassid et al.

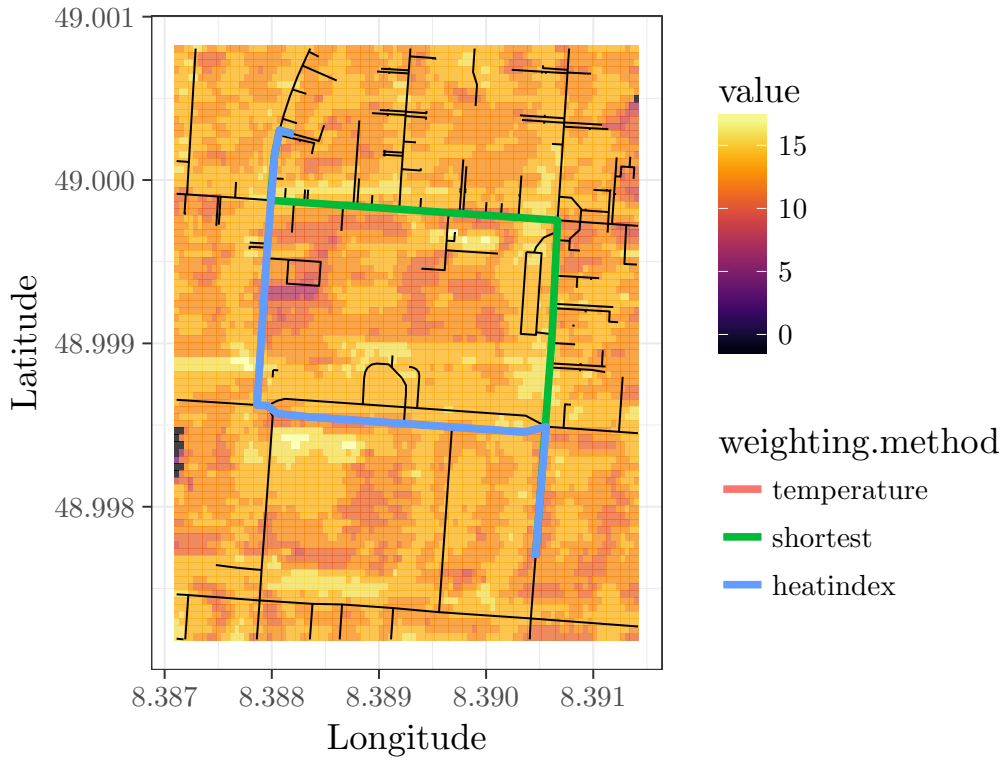


Figure 6.1: Example for heat stress based routing in comparison to shortest route. The benefit is visible even for small trips.

(2000)). Other applications include the use of groundwater for heat saving and to optimize the energy consumption (Benz et al. (2015)). In addition, a better understanding of causal reasons for temperature differences in a city can help shaping decision for city planers regarding city development plans. This part aims to create the foundation for such information systems and decision problems by supplying an accurate temperature data basis.

As of today, temperature models are either too coarse-grained – modelling differences between larger regions such as between urban and rural areas – or very detailed – modelling single buildings. This can be seen in particular for meteorological models or those related to UHI (Oke (1982)). Models proposed from high spatial resolution are typically based on snapshots, e.g. based on one-time thermal flights (see, for instance, the works described in Schwarz et al. (2012) or Cai et al. (2017)). There are four

primary reasons for the lack of models between high temporal or high spatial resolution:

(1) The required data to learn such models was not available in the past. While there have been advancements in temperature measurements, from cars, e.g. Kratzer (1937), to thermal flights and satellites (see Schwarz et al. (2012) or Bhattacharjee et al. (2016)), those measurements are fine grained on a spatial *or* temporal level, e.g. monthly snapshots for the temperature of the overall city areas or hourly measurements at few spatial points. Today, the available data to train and predict temperature on a fine scale is still relatively sparse.

(2) The computational effort to derive respective models is high. Given spatial and temporal references and a combination of many causal predictors dependent of their position in space and time, learning models that make use of such data is computationally expensive. Existing approaches therefore either aggregate high-resolution data to more coarse data, as for meteorological models, or reduce data by focusing on small areas such as single buildings. Meteorological models in particular are based on thermal conditions of higher air layers. The generalizability of such models is not yet given.

(3) The area of interest, the city itself, adds additional, unique challenges as a city is a diverse environment. Parks can be next to an industrial zone and building materials can vary broadly between cities, questioning the value of closeness of data measurement and target location for the prediction. The number of different combinations leads to the well known curse of dimensionality and is worsened by the sparseness of the data.

(4) Finally, the quality and technique of the measurements can vary. There does exist a great heterogeneity of measurement techniques, ranging from thermal flights, satellites as well ground-based weather stations, to name but a few. These techniques measure different temperatures, such as the land surface or air temperature and are often calibrated for different

environments. Additionally, the quality of each sensor can vary greatly. In the case of a satellite a single cloud leads to a non-usable measurement.

Thanks to recent advances in data availability within cities, together with more efficient algorithms and Geographic Information systems (GIS), more detailed urban temperature modeling is increasingly possible. In particular, historical time-series of hourly temperature measurements within cities are now openly available. Advances in computational capacity and modeling techniques, such as partial pooling, enable the maximization of existing information and robust modeling of their inaction so that the unique challenges inherent in fine grained, *urban* temperature explanation and prediction can be better tackled now and increasingly in the future. The vision is a better understanding of the processes how temperature is influenced in a city at any given point and time, improved prediction of temperatures and detailed simulations of temperature effects new constructions will have.

In this chapter, the *Land use-based Temperature Model (LTM)* is developed to provide a novel explanatory data combination model. LTM is not only based on Tobler's first law[1] by using local temperatures but also on land use information known to impact temperature. In the evaluation it can be shown that modeling the interaction of these land use-based causal factors with temporal temperature features is essential for accurate air temperature prediction with sparse data. LTM is then used as a predictor basis for more advanced modeling approaches and techniques.

LTM combines land use information, time series of weather station-based temperature measures and the interactions of both types of predictors to derive a causal prediction model. More precisely, LTM models temperature differences to the closest reference measurement point to predict the temperature at any nearby location depending on their land use types. To maximize the inherent information and deal with the sparseness of the data, simple Bayesian Hierarchical Models (BHM) are used as an example

[1] "everything is related to everything else, but near things are more related than distant things"

for a more complex modeling approach. They are chosen for their ability of partial pooling of information. This provides the benefit of being able to determine the locations and land-usage types with high predictive uncertainty and to use this to determine where additional sensor should be located within the city to gain maximal additional information. In the empirical evaluation of this section, LTM is applied to estimate temperature within several cities for the state of Baden-Württemberg in Germany. Results show that LTM increases accuracy of temperature predictions up to 60 % compared to standard models solely based on temperature and distance data and reveals comprehensible relationships between land use and temperature that can be operationalized by smart city planning.

6.2 Related Work on Urban Temperature Prediction

The phenomenon of higher temperature levels in cities as well as their impacts on urban planning und human health, coined Urban Heat Island (UHI), is subject to research since decades. Here, we present a focused overview of the research, with focus on insights and methodology, of the last 30 years of temperature prediction and the UHI research, which can be used as basis for the LTM. This extends the foundation chapter.

A recent overview is provided in Arnfield (2003). The focus here lies on the development in the field of climatology between 1980 and 2003. In comparison to 1982 the understanding of the UHI effect is increased, but, to quote Arnfields conclusion, "simple methods are still required to estimate UHI intensity within urban areas, as a function of time, weather conditions and structural attributes, for practical applications such as road climatology, phenology, energy conservation, and weather forecasting." He continues in that simulations are one option to deal with the underlying complexities of city temperature modeling, but their prediction perform-

ance remains low.

Schwarz et al. (2012) compare different measurement methods for the UHI effect and come to the conclusion that the UHI effect is dependent on the exact placement of the rural as well as urban station and it is therefore important to take the effect of the immediate surrounding into consideration when comparing the UHI between different cities. On the one hand, the authors state that the reduction of an UHI to a single value for a whole city is questionable regarding its explanatory power. On the other hand, they conclude that there is currently no other way to quantify the temperature difference of the UHI between different cities.

Difficulties when comparing UHI values amongst cities are also considered in Stewart and Oke (2012). As the measurement of weather stations are influenced by their surrounding areas, a way to classify the surrounding area to properly describe the UHI effect is needed. The authors propose the use of local climate zones (LCZ) to classify those areas by a standardized methodology and terminology. In total, 17 different LCZ are defined in their work based on the land use. The land usage can range from forests up to heavy industry areas. These are evaluated on three different mid-latitude cities in 2014 (Stewart et al. (2014)). It is shown that each LCZ has a different climate and delivers a better understanding of the UHI effect based on their underlying land use. This results in a better understanding and differentiation compared to the classical coarse-grained urban/rural differentiation. But the effects of those LCZ are only compared for their yearly mean temperatures, without regard to potential interactions of seasons or day cycles.

In a similar vein is the use of so called ”green lungs” for cities. It is well-known that large areas with vegetation reduce the yearly mean temperature and are an often sought out solution for city planners. Those areas supply an effective recreational area, but only affect their immediate surroundings (Gill et al. (2007)). Chen et al. (2014) found out that this temper-

ature reduction effect can vary between different months and is dependent on the size and shape of the green area. They showed this for the city of Bejing, China with five Landsat ETM+ images, where they measured the daily mean temperature at five different days in five different months.

Cai et al. (2017) uses the concept of the LCZ to measure their influence on local climate and how they represent the urban area. They use satellite measurements to measure the land surface temperature in the Yangtze River Delta region with ten Landsat images for a time period of 2 years. The authors showed that the LCZ concept applies to both air as well as land surface temperature. For the future they propose to examine seasonal differences, presuming this might lead to a better understanding as their results indicate interactions between LCZ and the season. However, insufficient data was available for further analysis.

Bhattacharjee et al. (2016) propose a semantic kriging approach, where a high-resolution satellite snapshot is used to quantify the effect of the difference between different locations as well as the interaction of the land uses between those points. The different land use classes are learned in a semantic hierarchical network.

Hengl et al. (2012) also uses a kriging approach to predict temperatures. They include a temporal component to predict the daily mean temperature in Croatia for area of $1km^2$ with an accuracy of 2.4°Kelvin (K) by combining Modis satellite images with 57,282 ground measures of daily temperatures in 2008.

In summary, several approaches exist to explain and predict temperature differences and local peaks. Those models explain the effect of isolated parameter in temperature (difference) prediction. The models are built on data sets either from a coarse-grained temporal perspective, e.g. only using few satellite snapshots or daily aggregate temperatures, or from few point measurements in the case of air temperature.

The approach in this thesis combines the fine grained land use informa-

tion with hourly measurements at diverse locations to predict the air temperature within a city for any point in space and time. Motivated by the prior work it makes use of the additional data sources and combines land use and multivariate temperature data by considering their interaction instead of simple additive combinations. To our knowledge, this has not been done before and promises more fine grained and differentiated spatio-temporal modeling.

6.3 Intra-Urban Temperature Modeling

In this section we will discuss several key insights from the existing UHI literature regarding the temperature in cities. The goal is to use data which is easily available at any location in order to make LTM broadly applicable, namely the general temperature, the time and location of the measurement, and the land use of locations. This provides the generalizability of this approach for any area. Consequently, the focus is on insights that can be used for model development and testing based on these types of data, excluding insights for instance related to wind speed and direction at a target location, which is typically not easily available in most real world settings.

The key foundation is based on the insights presented in Arnfield (2003) regarding the heat in cities in comparison to the surrounding area. These insights are the result of an extensive literature review and summarize the findings in the field of UHI. According to his paper, there are eight key insights, of which only insights five, six, seven and eight will be considered in this work as the aim is to consider only data that is broadly available.

1. UHI intensity decreases with increasing wind speed.

2. UHI intensity decreases with increasing cloud cover.

3. UHI intensity is greatest during anticyclonic conditions.

4. UHI intensity is best developed in the summer or warm half of the year.

5. UHI intensity tends to increase with increasing city size and/or population.

6. UHI intensity is greatest at night.

7. UHI may disappear by day or the city may be cooler than the rural area.

8. Rates of heating and cooling are greater in the countryside than the city.

Those insights are used to derive modified as well as new hypotheses, which will then be used to motivate and build the basic LTM model. This model in turn will be used to pre-select the parameter used in the Bayesian LTM.

The other insight, upon which the hypotheses are derived, is from Stewart and Oke (2012) and Stewart et al. (2014): The local climate zone (LCZ). In their work they define those LCZ based on the combination of the underlying land use and show that LCZ has an high impact on the air temperature. This can be seen as an extension of insight eight and is most likely the underlying reason for insight five, as population density is correlated with particular LCZs. LCZ represent a manifold of different causal reasons for the temperature difference in one standardized classification. This is similar to the work of Kalnay and Cai (2003), which also discussed the impact of land use on the climate, in particular the difference in temperature over yearly means in comparison to changing land uses.

6.3.1 Causal Predictors for Urban Temperature

Hypothesis 1

The land use is the primary driving factor for the temperature difference between nearby locations. This is based on Stewart and Oke (2012) and insight five from Arnfield (2003). We use the land use to substitute the different LCZ. Land use classification is most often used by government agencies and are openly available. These are regularly updated by those agencies and do not change as often as other measurements for the LCZ. In today's city planning green areas, so called "green lungs" are one of the most well-known LCZ. They summarize and pre-classify, similar to LCZ, a manifold of different underlying causal influences on the temperature differences.

Hypothesis 2

Temperature follows a typical movement pattern over the day. The hypothesis is based on insight four and five derived from Arnfield (2003). However, a clear definition of day and night is difficult to determine. Depending on the time of year as well as geographical location, simple definitions, such as the setting of the sun, can vary. Instead of a day – night difference we model the temperature development over the day. This effect can easily be observed outside. The temperature has its lowest point in the early hours of a day and increases up to a high in the later half of the day and then decreases again to the minimum.

Hypothesis 3

Temperature follows a typical movement pattern over the different months. The third hypothesis is in regard to the monthly temperatures. This is based on insight four from Arnfield (2003). In the literature, a connection between temperature differences and seasons has been discovered. In this part we further drill-down to the level of months. Similar to the daily pattern, this

circumvents the problems of an exact starting point for seasons. It also allows to model the development of the temperature in more fine detail.

Hypothesis 4

Given the land use and the month, there does exist a unique temperature pattern over the day. This hypothesis is based on the first three hypotheses. The interaction of the effects of hypothesis one to three will result in a much higher increase in the explanatory power than their additive combination.

This hypothesis can be subdivided into two steps. (1) Hypothesis one and two are combined. It can therefore be stated that given the land use, we can observe a different temperature cycle over the day. The impact of land use on temperature can be explained by the ability to create and save thermal energy. For example, a street can not produce heat, but does reflect it quite strongly. In contrast, an area of water saves more energy and regulates or stabilizes temperature. Given the time of day, the accumulative effect of the sun increases the temperature of a street highly, but over night it decreases rapidly. The water area has the opposite effect and decreases the temperature while there is daylight and increases the temperature during night hours. This effect can not be seen in isolation, only with the interaction. By isolating the effects, the warming effects of the water area over night would be overshadowed. (2) By building upon this logic, we add the interaction with the monthly temperatures. Using the example of the water area, one can observe a cooling effect at any hour given the day for a month such as July and a warming effect at any hour for a month such as January relative to the overall temperature. Considering each data set in isolation, or by simple linear modeling, these relationships are not detectable.

To our knowledge there is not yet a model using the combination of those effects, nor is there a model for their interactions to derive a causal prediction of temperature difference. In the existing literature, the different influences were examined in isolation. The previously existing data and

methodology did not allow for an in-depth examination of the parameters on such a fine spatio-temporal scale in combination with the land use. A good example for this is Chen et al. (2014). They could observe an interaction effect of vegetation and month, but they only used the daily mean temperature based on four snapshots at differing months for a broad spatial area.

We use these hypotheses to derive the *Land use-based Temperature Model* (LTM) model, which incorporates each of these interactions to represent most of their inherent variance. This is then used to parameterize a Bayesian Hierarchical Model (BHM) with those interactions to include additional information and reflect locally differing uncertainties in the temperature estimates by partial pooling of the information.

6.3.2 Land use-based Temperature Model

The systematic, causal differences derived from hypothesis 4 are used to predict the temperature at other stations given the differing land uses and points in time. This approach is called the *Land use-based Temperature Model* (*LTM*).

In its basic form, the LTM is an explanatory model. The idea behind this model is to extract the temperature difference at a location from the mean temperature given the interaction of land use and temporal dependencies.

To do so, LTM computes the mean difference of temperature given hour of day, month and land use for all stations. This difference is then added to the measured temperature of a reference temperature (e.g., global mean temperature or nearby reference station) to predict the temperature at any point given the aforementioned parameters; the mean for each category.

Figure 6.2 visualizes the idea and computation using the example of average daily temperature movement of two stations for the July of 2013. In this example one can learn that the temperature difference for instance at hour 16 between land use *Forest* and land use *Industry* is 4.9°Celsius. When

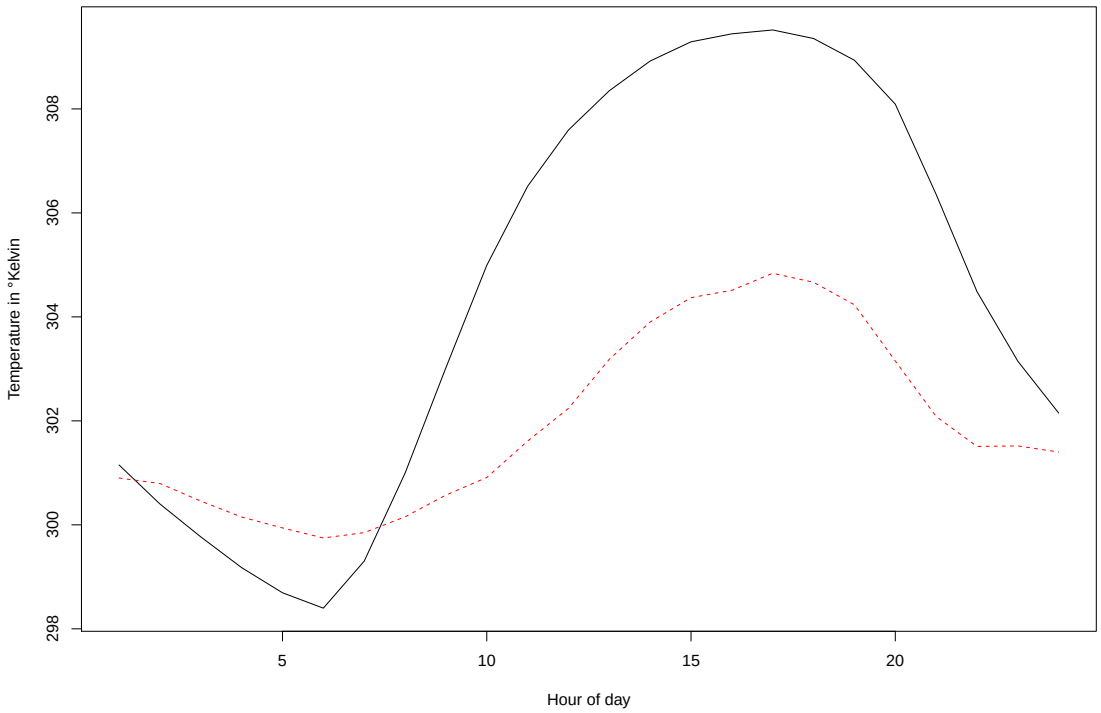


Figure 6.2: Difference of the temperature over the hourly mean temperatures for July 2013 for land use *Forest* (red, dotted) and land- use *Industry* (black). The temperature is in degree Kelvin.

predicting a new station with the land use *Industry* having a nearby station with land use *Forest* one can add this value to the temperature value of the reference station given it is July and the hour is 16. By assuming a temperature of 20°Celsius at the *Forest* station, one would predict a temperature of 24.9°Celsius at the *Industry* station. In doing this for a whole year and for any land use pair, the basic *LTM* enables a simple and fast extraction of the temperature difference of any point. By applying this extracted values on new stations, a fast prediction is possible. This allows for a fine grained prediction, especially for urban areas.

With temperature $T_{l,t}$ as dependent variable, denoting the temperature T at a position or location l at time t, the covariates $hour_t$, $month_t$, $land\ use_l$, and with T_t^R as the temperature at a reference station at time t we can for-

mulate a basic model, derived from a linear model, as shown in (6.1).

$$T_{l,t} = (\alpha | hour_t, month_t, land\ use_l) + T_t^R + \epsilon \tag{6.1}$$

The idea is that by knowing the land use at the station $T_{l,t}$ and the time of the measurement a typical dependent bias α can be computed. This is the deviation from the reference station T_t^R, e.g. an official, standardized measurement from a nearby weather station or a regional mean temperature, for this land use at this time. This allows for a fast and simple prediction at any spatial point by then adding this α to the reference station.

For each interaction, an α is estimated. For each unique land use, given only the months of the year and the hour of the day, 288 combinations are available for one year. Unfortunately, multiplying 288 by the number of unique land use classes, a triple or quadruple digit number of parameters needs to be estimated, which would result in unreliable model outcome and high levels of overfitting.

The LTM therefore computes only the mean for each parameter combination. This allows for greatly reducing the computational effort and the model's variance of outcome depending on randomness in the training data used to fit the model. As a result, each parameter is learned independently by at least 28 training samples, given the minimal number of days in a month.

The main benefit of this approach is its simplicity and therefore easy generalizability. It only needs a reference station and the dependent parameter. This allows for an in-depth explanatory analysis, as it retains most of the variance and the important interaction. As will be seen later in the evaluation sections, it can even be used for predicting temperatures out-of-sample, given that the overall climatic conditions do not change. In accordance to the well-known bias–variance trade-off, this allows it to represent a high degree of variance for the proposed interactions: The interaction of the unique parameter combination.

Although possible, the basic LTM model introduced is of explanatory nature and not particularly well-suited for out-of sample predictions. That is because there is no regulation applied to prevent overfitting and it is not robust to changes in the climatic conditions.

For the purpose of prediction, the LTM is implemented as a Bayesian Hierarchical Model (BHM). This model learns the underlying causal reasons and is capable to handle the varying quality as well as the uncertainty in the data and strongly reduces the impact of outliers in the training data and maximizes therefore the available information in the sparse data. This uses the same reasoning as in the basic LTM but extend it with the increased modelling capabiltites of BHM. this extension is called the Bayesian LTM, which presents two distinct advantages for the given problem setting (see e.g. Gelman et al. (2014), Cressie and Wikle (2015)):

1. Each data point as well as parameter is modeled as a distribution. This allows to include the uncertainty at each data processing step.

2. It allows for partial pooling. This allows to model the interactions found in the causal land usage relationships in an efficient way while avoiding overfitting.

Additionally, strong priors (Gelman et al. (2014)) can be used, e.g. based on physical models, to increase the robustness of the Bayesian LTM.

Overall, while the Bayesian LTM allows for capturing the LTM idea of modeling the fixed interactions – the differing intercepts – it also learns estimates for their slopes. To learn the model, it is assumed that there does exist a global value for the temperature where the local differences are only dependent on the underlying causal reasons. But such a global value is not known in the real world. A reference value derived from other measurements has to be used, adding uncertainty as well as additional dependencies to be considered. This uncertainty can impact the temperature in such a way that apart from the absolute difference between different land uses

there also can exist a difference in the slope of the temperature. By being able to model both the intercept as well as the slope by a multi-level approach, the uncertainty induced by the use of real world reference temperatures can be reduced. In this work, only a general, multipurpose model is used, with only weak priors and robust, non-specialized distributions for each parameter. While this approach does not fully utilize the power and possibilities of BHM, it is more robust and emphasizes more strongly the benefit of the LTM.

To summarize, the LTM enables an analyst to model the effects of the interactions of the land use with temporal parameter on the local temperature within an area, in particular an urban area. The basic LTM provides a strong explanatory model which extracts the impact in an efficient way and provides an easy comparison of those impacts. While it has not the properties required for a predictive model, it is extended with a BHM to the Bayesian LTM, which increases robustness and reduces its ability to blindly reproduce the complete variance inherent in the data.

Chapter 7

Empirical Validation and Evaluation of LTM

In this chapter, the insights found in the literature will be examined in the light of empirical data and compared to the formulated hypotheses, namely *Given the land use and the month, there does exist a unique diurnal temperature pattern* and *Using the interaction of different parameters increases the predictability of temperature at a fine grained geographic scale*. First, the data used for this empirical case study is described and a well-established benchmark prediction model is presented. Second, the different model parametrizations together with the quality metrics are described. Third, the in-sample and out-of-sample predictive results on empirical data of the two models introduced, LTM and Bayesian LTM, are presented and discussed.

7.1 Empirical Data and Benchmark Model

For the empirical case study of this part of this thesis, three different data sets are used:

1. Air temperature from the German meteorological service (DWD)[1].

[1] https://www.dwd.de/

2. Air temperature from the LUBW State Institute for Environment, Measurements and Nature Conservation Baden-Württemberg (LUBW)[2].

3. Land use data from the ATKIS data set[3] of the State Agency for Spatial Information and Rural Development Baden-Württemberg.

The study area consist of all measurements in the federal state of Baden-Württemberg, Germany in the years of 2013 and 2015. We chose those two time periods as they have the most complete time series for the air temperature. The air temperature of the DWD stations consists of 25 ground based weather stations, which measure temperature every hour. The temperature ranges from -16.5°C to 38.1°C for 2013 with a mean temperature of 9.83°C and for 2015 from -11.7°C to 39.4°C with a mean of 10.43°C. The temperature is measured at 2 m height above ground. The time series represent standardized measurements at open places without any interference of their surrounding area. One can therefore ignore the underlying land use class. These measurement present data of the highest quality, which are used in meteorological forecasts.

The air temperature of the LUBW stations consists of 24 station with complete measurements for overall nine different land use classes and one incomplete measurement for an additional land use class (train tracks) for 2013, which measure the air temperature at every hour at 2 m height. For 2015, only 17 stations have complete measurements. Those include 8 different land use classes and are a subset of the 2013 LUBW stations. The temperature ranges from -16.0°C to 38.05°C for 2013 with a mean temperature of 9.93°C and for 2015 from -12.6°C to 39.1°C with a mean of 11.28°C. The stations of the LUBW are used primarily to measure environmental factors at diverse locations. They are therefore used to represent the intra-urban temperatures and provide the generalizability of our results. In contrast to

[2] https://www.lubw.baden-wuerttemberg.de/
[3] www.geoportal-bw.de

the DWD stations, they can be used to model the impact of the land use on the temperature.

The temperature between the DWD and LUBW stations is similar, but has differences which are most likely a result of the underlying land use. The overall temperature between 2013 and 2015 is similar as well, with a small temperature increase in 2015. Each station has 8760 temperature measurements over a year. Therefore, for 2013 there are over 200,000 measurements available and for 2015 we have over 140,000 measurements.

The land use is extracted from the ATKIS data set. It has a spatial resolution of 1 m^2 and is updated in regular intervals every few years. The data set is the official classification of the land use from the federal government. These represent 25 different land uses. Here, the official classification of land use is used instead of the LCZ concept from Stewart et al. (2014) as this classification is already in use by government agencies and therefore easily available. They are comprised of similar classes as the LCZ in the original work and eliminate the need to classify the area and thereby induce additional uncertainty. The land use classes given in this empirical data set are show in Table 7.1

As a benchmark the temperature of the nearest station is used. This benchmark is often used in geographical analysis and is based on Tobler's first law (Tobler (1970)). Advanced geo-temporal prediction methods are based upon this idea. It is similar to the well-known random walk without drift from time series analysis. By using the nearest station, most of the regional climates and regional effects are excluded. This enables a simple, but unbiased prediction of the surrounding area of a station. This benchmark is also the result of the meteorological forecasts of temperatures. Weather forecasts are done for and based upon those stations and then extended to the surrounding area as the true value. By using this benchmark one can also compare the additional benefit our prediction generates by focusing on fine grained areas in contrast to state-of-the-art, coarse-grained predic-

Land Use Number	Class Name
1	Industry
2	Sport and Leisure Activity
3	Areas of Special Functional Character
4	Residential Building Area
5	Shipping Lanes
6	Groove
7	Agriculture
8	Combined Use Area
9	Rail road
10	Forest

Table 7.1: Overview of used Land Use classes. Class definitions from http://www.ioer-monitor.de/en/home/

tions. This benchmark also includes by design the interaction of the daily temperature cycle and the months of the year.

The mean distance to the nearest LUBW station for another LUBW stations is 26 kilometers, the distance ranges from 7.6 kilometers to 58 kilometers. The mean distance between a LUBW station and the nearest DWD station is 12.6 kilometers and the distance ranges from 0.48 kilometers to 25 kilometers. Given this empirical data set, the reference stations are geographically near and should represent the regional climate quite well. The overall size of the federal state is $35,751 km^2$, which, in accordance to Hengl et al. (2012), allows us to assume an overall global temperature. In the following section, the data of 2013 is used as training data and the results are evaluated on the data of 2015.

7.2 Empirical Validation of Hypothesis

7.2.1 Insights from Literature

The selected insights from the existing literature are modified to the following hypotheses in the last chapter:

1. The land use is the main driving factors for the temperature difference between different but close-by locations.

2. Temperature follows a typical movement pattern over the day.

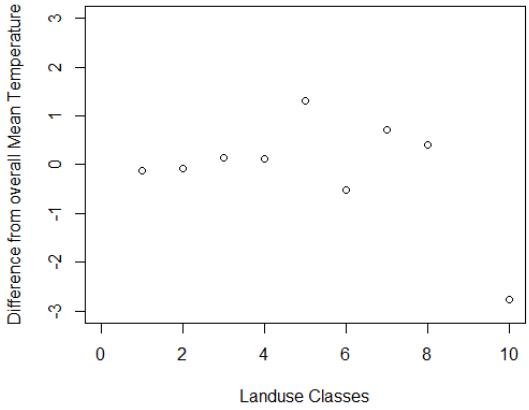
3. Temperature follows a typical movement pattern over the different months.

While those hypotheses are sound in theory, one has to explore how they are represented in our empirical data set. Therefore we show their impact on the training data, and validate the hypotheses according to results of the descriptive analysis. The results are first validated using a visual analysis and then the results are compared using statistical methods.

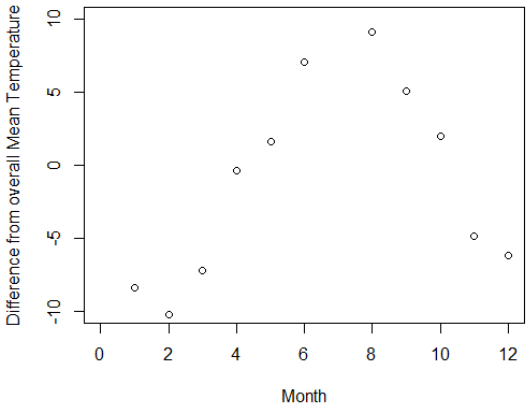
The overall graphical validation can be seen in Figure 7.1. The mean temperature is compared for each of those parameter and it is shown how the different hypotheses differ in relation to the global mean temperature.

One can see and impact of the land use in Figure 7.1(a). Nine different land use classes can be compared (land use class nine has insufficient data points for the whole year). Interestingly, between land use class one to four we see only small differences in temperature even though those are the most developed and urban areas. An increase in temperature can be seen for the land use classes five, seven and eight, which are more open areas and a small temperature decrease for land use class six, the grooves. Land use class ten, forest, shows the most pronounced difference from the global mean temperature, underpinning a strong relation of the land use and the local temperature in our data set. This provides first evidence of for the

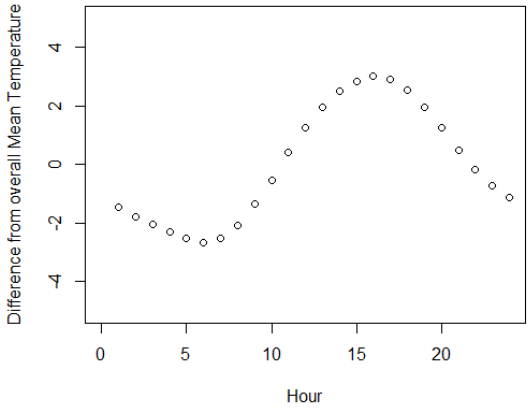
(a) Impact land use



(b) Impact month of year



(c) Impact our of day



(d) Mean growth rates over day

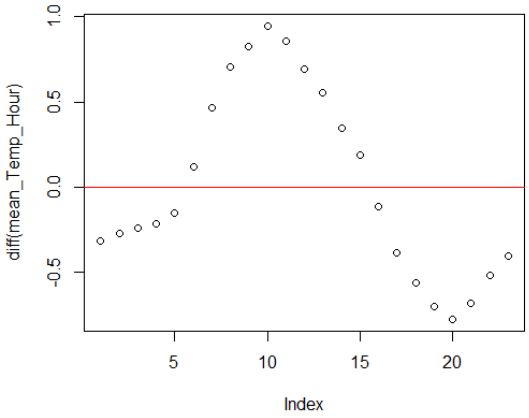


Figure 7.1: Figure (a) - (c) show the absolute temperature deviation from the global mean for each parameter. Figure (d) shows the daily temperature growth rates between the hours, with the red line indicating no temperature growth.

hypothesis by a first visual inspection. In green areas we see a decrease of temperature over the whole year, in urban areas almost no effect and for open areas an temperature increase.

The existence of a typical movement pattern over the day can be seen in Figure 7.1(c). One can clearly see a systematic pattern of temperature over the day with the highest temperature at about 16:00 and the lowest temperature at about 06:00. The temperature seems to move between -3°K and $+3$°K over the day cycle. This validates this hypothesis on the empirical data. For the daily cycle we are also interested how strong the difference between the different hours is. The growth rates are plotted between the different hours in Figure 7.1(d). Here one can see that there is a clear difference between each hour, but also that a division into growth segments could be feasible. Starting at 12:00, the temperature growth rate is reduced to every hour till 20:00 at which it is growing again. By using a more time-series focused approach, e.g. an ARIMA model, this could allow an alternative parametrization.

The final hypothesis from the literature is the impact of the month of year. The results can be seen in Figure 7.1(b). Here one sees a strong overall temperature difference for each month, with a maximal span of 20° K for the most extreme months, February and July. This also strongly supports the hypothesis on the used empirical data. Interestingly, this graphical result indicates also that a categorization into seasons only is too coarse-grained and monthly values are more appropriate as of the somewhat continuous developments over the months of a year.

These results are also reflected using a dummy regression for each hypothesis, where each parameter is a factor. The results for the land use hypothesis are shown in Table 7.2. One can see a high significance of each parameter and a clear direction of the influence for each land use type. For the day cycle, the results can be seen in Table 7.4. Again, One sees a very high significance for every parameter, but a quite low explanatory value

with an R^2 of 0.055. For the monthly cycle the results can be seen in Table 7.3. All parameter are highly significant and like the visual analysis quite pronounced. The explanatory value is relatively high with an R^2 of 0.687.

	Dependent variable:
	Temperature Difference
Industry	-0.102^{***}
Sport and Leisure Activity	0.131^{***}
Areas of Special Functional Character	0.184^{***}
Residential Building Area	0.154^{***}
Shipping Lanes	1.284^{***}
Groove	-0.475^{***}
Agriculture	0.753^{***}
Combined Use Area	0.404^{***}
Rail road	-0.709^{***}
Forest	-2.716^{***}
R^2	0.007
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 7.2: Summary of the impact each land use class has as a predictor on the temperature and the significance levels based on a dummy linear regression. While the explained variance is low, the impact each predictor has is quite strong.

This leads to the first part of the answer of RQ 3a: *Hour of day*, *Month of Year* and the *land use* can be considered as causal drivers for temperature differences. One can argue that the land use itself is not the cause, but instead only an umbrella classification of several causal factors. While this is certainly true, exactly this nature as an umbrella classification provides several benefits for its use in the real world. It includes many factors such as the Normalized Difference Vegetation Index (NDVI), the albedo, human use and many more. As this data is often not available, the use of the land use allows to include their impact indirectly, even if specific data is not available. In addition, its use allows for a more robust prediction, as the number of parameter for a model is reduced.

	Dependent variable:
	Temperature Difference
MONTH 1	-8.33^{***}
MONTH 2	-10.12^{***}
MONTH 3	-7.15^{***}
MONTH 4	-0.19^{***}
MONTH 5	1.93^{***}
MONTH 6	7.17^{***}
MONTH 7	11.32^{***}
MONTH 8	8.91^{***}
MONTH 9	4.92^{***}
MONTH 10	1.68^{***}
MONTH 11	-4.90^{***}
MONTH 12	-6.52^{***}
R^2	0.687
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 7.3: Summary of the impact each month has as a predictor on the temperature and the significance levels based on a dummy linear regression. One sees an overall strong difference for each month an high explanation of the variance.

	Dependent variable:		*Dependent variable:*
	Temperature Difference		Temperature Difference
HOUR 1	-1.46^{***}	HOUR 13	1.96^{***}
HOUR 2	-1.77^{**}	HOUR 14	2.51^{***}
HOUR 3	-2.05^{***}	HOUR 15	2.85^{***}
HOUR 4	-2.29^{***}	HOUR 16	3.04^{***}
HOUR 5	-2.50^{***}	HOUR 17	2.93^{***}
HOUR 6	-2.66^{***}	HOUR 18	2.54^{***}
HOUR 7	-2.53^{***}	HOUR 19	1.98^{***}
HOUR 8	-2.07^{***}	HOUR 20	1.28^{***}
HOUR 9	-1.36	HOUR 21	0.50^{***}
HOUR 10	-0.53^{***}	HOUR 22	-0.18^{***}
HOUR 11	0.41^{***}	HOUR 23	-0.70^{***}
HOUR 12	1.27^{***}	HOUR 24	-1.11^{***}
R^2	0.055		
Note:	*p<0.1; **p<0.05; ***p<0.01		

Table 7.4: Summary of the impact each hour of the day has as a predictor on the temperature and the significance levels based on a dummy linear regression. One sees a low significance level for HOUR 9 as the temperature difference is quite similar to HOUR 1.

7.2.2 Interaction of Parameter

To examine the final hypothesis, *Given the land use and the month, there does exist a unique temperature pattern over the day* and that the *interaction result in increased explanatory power than their simple additive combination*, the number of interactions has to be taken into account. The number of interactions is the number of interactive time intervals, here 288, times the number of land use class, which for our empirical data is nine. We therefore have over 2000 distinct interactions to compare. Given possible similarities between land uses and potential similarities between day cycles, the risk for multicollinearity is high. A comparison with regression models is therefore not feasible. For reasons of brevity, for a graphical analysis we compare only three different land use classes, *industry*, *forests* and *residential building area*. We chose those three, as they represent the most typical built up areas and the strongest green area. The results for the different daily cycles given the month is shown in Figure 7.2.

The comparison of three land uses allows for not only inspecting the difference in temperature movement based on the interaction, but also the influence of the distinct causal reasons. The systematic differences over the months can be clearly seen. But, of more interest, is the influence on the daily cycle pattern and the differing systematic difference for the land uses. We can see very similar movements between the more built up areas in comparison to the forest. But even there, a clear temperature difference can be seen in most of the early hours. This is most pronounced in month twelve, December. The *residential building area* is overall warmer and has a less pronounced temperature curve. We assume that this is the result of a slightly stronger regulating impact of this land use in comparison to *industry*. Factors such as gardens, less isolated buildings and human heat sources such as heating over the night could lead to the smoothing of the curve. Less surprising is the strong difference between those two land uses and *forest*. This difference is strong in any month, as the expectation of the

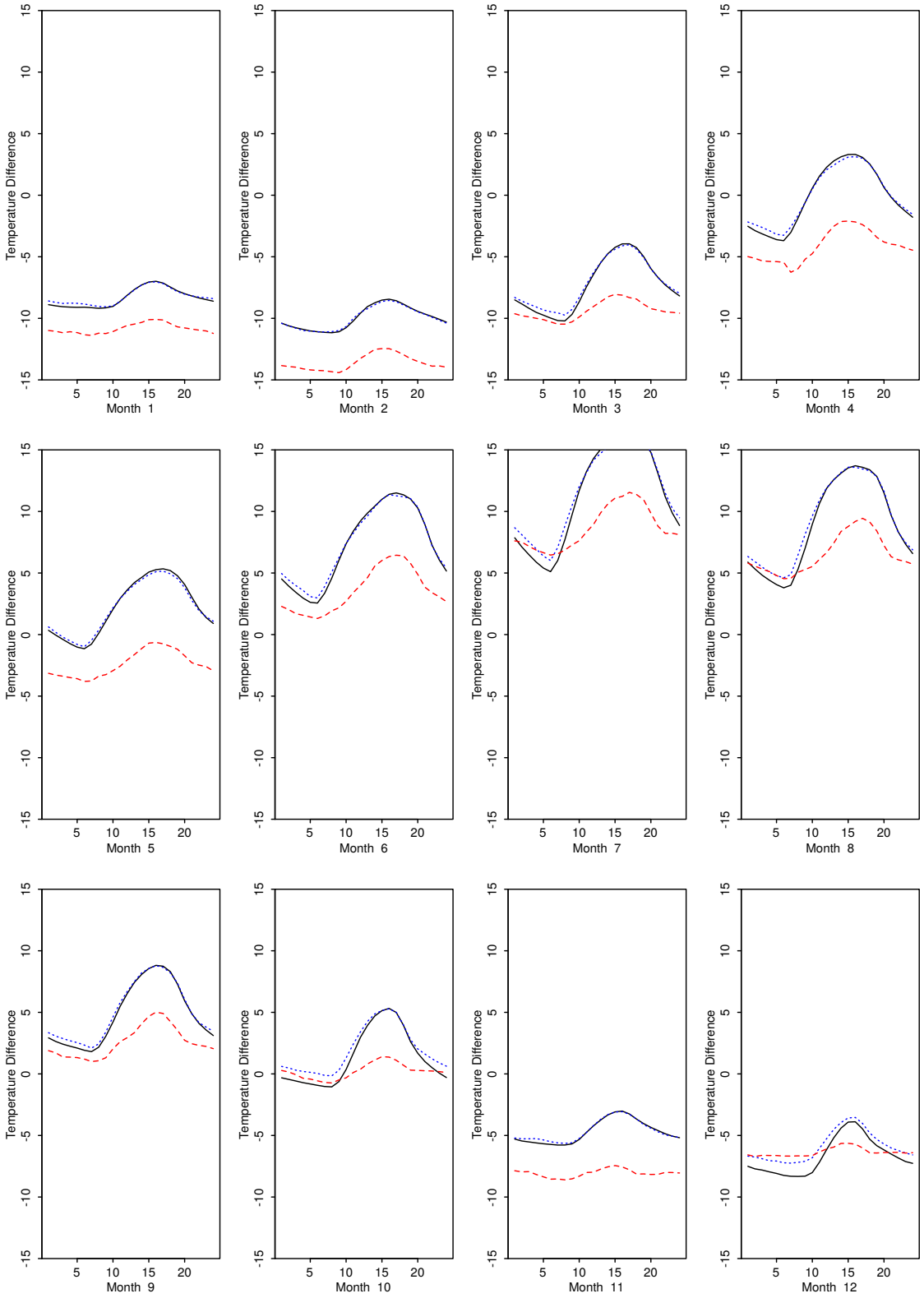


Figure 7.2: Here we see the Temperature difference from the global mean temperature over the daily cycle for each month and three different land uses. In black are the differences for *Industry*, in red (dotted line) the differences for *forests* and in blue (small dotted line) the differences for *residential building area*. The x-lab shows the hour of day.

UHI effect suggests. But we can see several interesting deviations in month seven, July, and month twelve, December. In July the early temperatures of *forest* is higher than the temperature of the built up areas. The UHI effect would suggest that the temperature should be lower overall for the *forest*. In December the reverse effect can be seen and is even stronger. The temperature in *forest* is higher for almost the whole day. In addition it stays relatively constant and has almost no discernible movement over the day. The overall mean temperature for this month is also higher for *forest*. We believe that this is the result of an inherent regulating process in areas with more and denser vegetation.

When we compare Figure 7.2 to Figure 7.1, we see the impact of our inclusion if the interactions. The yearly mean temperature for each land use remains the same, but there are clear daily and monthly pattern observable, which could not be discerned before. The overall averages, which are be observed for the isolated parameters, can still be seen in the overall comparison, but we can better observe their deviation. While only the comparison of three land uses has been shown, other land use pairs show similar relationships and we argue that our hypotheses are strongly supported by our empirical data, namely that the interaction of the causal reasons is beneficial when the aim is to describe or predict temperature. This is further cemented by the improved R^2 of 0.7683 of the regression results.

This allows a full answer to RQ 3a given our empirical data set: The overall causal drivers behind local temperature are in their interaction: *Hour of day*, *Month of Year* and *land use*. While in isolation these factors can explain the temperature difference inherent in a data set, only their interaction can sufficiently explain the difference between different local measurements.

7.3 Evaluation Design

By using the insights from RQ 3b, the RQ 3c can be answered in the rest of this chapter. The LTM approach is used as a basis for the model building to produce accurate and robust predictions. To evaluate the performance of the our proposed approaches, for the basic LTM model and the Bayesian LTM, we use the mean absolute error (MAE) in Kelvin to measure the accuracy of the predictions as well as the percentage difference in accuracy between the benchmark and the different parametrizations. In the UHI as well as temperature prediction literature the MAE is a standard metric; alternatives include the R^2, e.g. in Suomi and Käyhkö (2011). As we want to provide the foundation for information systems, a low MAE is of greater importance as compared to the explained variance and is therefore used. Our treatment structure is shown in Table 7.5:

Model	Interaction
Nearest Station	No learning
Basic LTM	Systematic Differences from mean temperature
Bayesian LTM 1	Interaction for intercepts and the slope
Bayesian LTM 2	Interaction for intercepts and global slope

Table 7.5: This table shows the different model treatments applied on the data sets. Nearest station uses only Tobler's first law, Basic LTM applies the temperature differences of the LTM model. The Bayesian LTM 1 and 2 use the predictors of the Basic LTM as the basis for the BHM.

We use two different reference temperatures for our models. First, the data of the nearest DWD stations are used to predict the temperature at the each LUBW station. This is done as the DWD stations represent the standardized measurements, which should not be influenced by their land use. Thereby the local climate is taken into account and only the impact

of the differences in land use and the interaction with the time parameters are compared. This represents the standard real-world situation, where there does exist an official weather station or official weather forecast in a geographically close location. Its generalizability follows as this setup is valid for most developed countries. Second, a global temperature from the LUBW stations can be derived for the whole area of Baden-Württemberg. This temperature allows to analyze not only the difference to a nearby station, but to compute the impact the land use and its interactions have on difference from the overall regional climate. This would allow to use broad spatial predictions and measurements and reduce their spatial resolution. Here, we use the combined differences for different land uses as motivated in Section 6.3.2. We use the global mean temperature of the LUBW stations to predict the temperature at each LUBW station. This represents the case when there are no official, high-quality measurement networks, but a reference temperature can be derived from other sources such as low quality measurements, satellite images or other aggregates.

7.4 Empirical Results

The results for the presented models in-sample for the year of 2013 can be found in Table 7.6.

We see that both proposed models, regardless of their parametrization, show a great improvement over the established benchmark. The basic LTM model performs better in both instances, but the margin of improvement is reduced for the DWD reference temperature. For the Bayesian LTM model we can see that the use of the interaction term for the slope as well as the intercepts leads to slightly better results. We also see better results if the DWD stations are used as reference station for the benchmark. This is a likely results of the lower average distance to those stations. Alternatively, the standardized placement of the DWD stations leads to a better general-

Model	LUBW	DWD
Nearest Station	1.741	1.188
Basic LTM	1.083 (37.8 %)	0.915 (23.1 %)
Bayesian LTM 1	1.295 (25.6 %)	0.969 (18.4 %)
Bayesian LTM 2	1.324 (23.9 %)	1.010 (15.0 %)

Table 7.6: This table shows the MAE Results in-sample and in brackets the percentage improvement to the baseline of *Nearest Station*. The percentage improvement is strongest for the LUBW data set, the MAE results are better for the DWD data set.

ization power of their measurements and therefore to better results. This could also explain the reduced improvement rate in-sample. The in-sample results represent the power of the models to capture the variance inherent in the data without regard to their generalization power.

Out-of-sample prediction results with the models on the data data of the year 2015 are shown in Table 7.7.

Model	LUBW	DWD
Nearest Station	3.404	2.498
Basic LTM	1.916 (43.7 %)	2.445 (2.1 %)
Bayesian LTM 1	1.362 (60.0 %)	2.339 (6.4 %)
Bayesian LTM 2	1.392 (59.0 %)	2.317 (7.3 %)

Table 7.7: This table shows the MAE Results out-of-sample and in brackets the percentage improvement to the baseline of *Nearest Station*. Out-of-sample, the DWD results are worse then the in-sample results. For the LUBW data set, the percentage improvement is stronger then in the in-sample data set. Overall, the Bayesian LTM models produce the most accurate predictions.

For the out-of-sample forecast we can see a change in the prediction accuracy for the models. The use of the DWD stations as reference temperature produces less accurate forecasts than using LUBW reference stations.

Given the benchmark, we see a decrease of 95.5 % for the LUBW reference stations and of 110.3 % for the DWD reference stations for the MAE. If we compare this with the general information for our data sets from Section 7.1, the high increase of the MAE for the benchmark model is unexpected. The overall difference in temperature between 2013 and 2015 is almost identically for both the LUBW as well as the DWD data set.

Overall, the proposed models including the land use data outperform the benchmark model. For the basic LTM we see a sharp decrease in absolute improvement over the benchmark for both the LUBW as well as DWD reference stations. This indicates that the high improvement of the basic LTM model in the training data is achieved by a decrease in the ability to generalize. The strong increase in the error for the benchmark given DWD reference stations supports this. It is our belief that this is the result of a stronger variability in the local temperatures between 2013 and 2015. As the standardized measurements of the DWD stations do not reflect this impact, the MAE is increasing. This change in variability cannot be reflected by the basic LTM model and therefore the improvement and quality of the prediction is reduced.

In general, the Bayesian LTM models outperform the benchmark for both reference stations and the basic LTM model. We can see this in Table 7.8, where we compare the difference in a paired t-test for both reference temperatures as well as in-sample and out-of-sample. The Bayesian LTM 2 model, where the interaction of the causal reasons is only modeled with fixed intercepts, is slightly better than the Bayesian LTM 1 model, where both the intercepts as well as the slope is modeled to include the interaction, in contrast to the in-sample results. For the LUBW reference temperature, both Bayesian LTM models perform significantly better than the benchmark as well as the basic LTM model. Noteworthy, the relative improvement is also higher than in-sample. For the LUBW reference temperature the Bayesian LTM 1 model performs slightly better than the Bayesian

LTM 2 model. These results show that the improved ability of the Bayesian LTM to generalize in comparison to the basic LTM models, while still representing more of the variance than the benchmark model. In case of the DWD stations, a more robust, bias-focused approach performs better as the changing variability between both years is more pronounced at those stations in relation to the predicted stations.

Interestingly, out-of-sample the models of the LUBW reference temperature outperform the DWD reference temperature in contrast to the in-sample results. The Bayes LTM models remain constant in their MAE. This further supports our assumption of a strong variability in the temperatures between 2013 and 2015.

Comparison	Test	Results
LUBW In Sample	Paired t-test:	t = 100.83, p = $<$ 2.2e-16***, df = 197530
LUBW Out of Sample	Paired t-test:	t = 251.93, p = $<$ 2.2e-16***, df = 147020
DWD In Sample	Paired t-test:	t = 117.42, p = $<$ 2.2e-16***, df = 205650
DWD Out of Sample	Paired t-test:	t = 62.947, p = $<$ 2.2e-16***, df = 147670
Note:	*p$<$0.1;	**p$<$0.05; ***p$<$0.01

Table 7.8: Here we show the results of the paired t-test for the Bayesian LTM 1 model and the benchmark model. The results confirm the clear improvement of the Bayesian LTM model over the benchmark even for the small percentage improvement out of sample for the DWD comparison.

These results allow us to answer RQ 3b: Using the interaction of different causal drivers, we can reduce the complexity of a prediction model to a manageable level and produce accurate and robust predictions. This is done by using the LTM as a basis for more complex statistical models. The results shown in this evaluation show that this increase is especially strong if using an overall regional climate, especially out of sample.

7.5 Error Analysis

As the previous section showed, the different models differ highly in regard to their in- and out-of-sample error as well as the used reference temperatures. The unequal distribution of the change in prediction accuracy indicates the existence of a systematic cause. We are therefore interested how exactly the error is distributed in regard to underlying causal factors, in particular the land use and the reference station. For diagnostic purposes, we conduct a regression analysis for each model type and reference temperature with the land use as dummy variable. The mean error (ME) is used instead of the MAE, as the direction of the error offers additional insights. The results are shown in Table 7.9 .

	Basic LTM		*Bayesian LTM*		Sample
	DWD	LUBW	DWD	LUBW	Size
Land use 1	0.304^{***}	-0.660^{***}	0.218^{***}	-0.187^{***}	35057
(Industry)	(0.014)	(0.031)	(0.014)	(0.009)	
Land use 2	-0.056^{***}	-0.234^{***}	-0.064^{***}	0.241^{***}	35057
(Sport/Leisure)	(0.020)	(0.031)	(0.019)	(0.013)	
Land use 3	-0.077^{***}	-1.039^{***}	0.104^{***}	-0.409^{***}	17537
(Special)	(0.025)	(0.032)	(0.023)	(0.016)	
Land use 4	-0.725^{***}	-0.683^{***}	-0.635^{***}	0.047^{***}	17537
(Residential)	(0.025)	(0.032)	(0.023)	(0.016)	
Land use 5	-3.693^{***}	-1.088^{***}	-3.459^{***}	-0.124^{***}	8777
(Shipping Lanes)	(0.032)	(0.033)	(0.030)	(0.021)	
Land use 7	-3.415^{***}	-2.806^{***}	-3.240^{***}	-0.108^{***}	8777
(Agriculture)	(0.033)	(0.033)	(0.031)	(0.022)	
Land use 8	-2.45^{***}	-1.563^{***}	-2.668^{***}	-0.505^{***}	8777
(Combined Use)	(0.032)	(0.033)	(0.030)	(0.021)	
Land use 10	-2.441^{***}	3.817^{***}	-2.403^{***}	-0.480^{***}	8777
(Forest)	(0.032)	(0.030)	(0.030	(0.021))	
Note:	*p<0.1;	**p<0.05;	***p<0.01		

Table 7.9: Mean error of the predictions for each reference temperature, model and land use, the significance level and the standard error. Only the Bayesian LTM 1 model is used as the prediction accuracy is similar.

Given the results, we detect three interesting aspects: First, as expected, overall ME is low for the Bayesian LTM model with the LUBW reference

temperature but the lack of impact of land use on ME is surprising. For the other models, we see a strong influence of the land use. We argue that the reason for this lies in the LUBW reference temperature and how the model learns the values for the different parameter. As hourly mean temperatures for the study area are used to learn the influence of the combinations, the impact of changes at individual stations is reduced. The abstraction level allows the model to learn the overall underlying temperature difference for any land use instead of the relation between the nearest station and the land use. The results for the basic LTM and the Bayesian LTM for the DWD reference temperature seem to be negatively impacted by their method to learn the underlying relation and therefore their parameters. These models learn from the temperature value difference between the temperature measurements of two stations. As all stations have, in contrast to the global LUBW temperature, a land use, even the standardized DWD measurements, their additional interaction may influence the accuracy of the prediction.

Second is the mentioned impact of the land use class on the error. We see for the other three models that the highest errors are for the land use classes 5, 7, 8 and 10. These represent *Shipping Lanes*, *Agriculture*, *Combined Use Area* and *Forest*. They represent mostly vegetation areas associated with lower temperatures, which are regulated by their underlying land use. We argue that the reasons for this increase in the error is a change in the overall climatic conditions between 2013 and 2015. This could lead to cumulative effects for the different land use classes or intensifying effects through surrounding neighborhood. For example, an *Agriculture* area could dry out over a long heat period, leading to a reduction or reversion of the cooling effect. Another example would be the increase of the temperature in a city park by its surrounding built-up areas.

The third interesting aspect lies in the direction of the error between the Bayesian LTM and basic LTM models. We see for almost all models a negative ME. This means that our models systematically predicts the temperat-

ure too low for 2015. As the basic LTM is a explanatory model, it focuses on representing the given variance. This allows us to specify the change in the climatic conditions between 2013 and 2015. It underestimates the temperature and therefore an overall increase in the temperature can be derived. This supports our assumption from the second aspect, in that the overall cooling effect of those land use classes is diminished in 2015 in relation to 2013. We see the same results for the Bayesian LTM with a DWD reference temperature. This result is most likely derived from reasons presented in the first two aspects: The dependence of the prediction upon the climate development of the target as well as reference station.

7.6 Discussion and Conclusion

Motivated by the increasing availability of time-series data for the urban environment and the need for fine grained temperature predictions, we proposed the combination of land use information and time series of weather based temperature measurement to address this need. We argued both from a theoretical point of view as well as with our empirical data that it is essential to model the interaction of the land use with temporal aspects.

We proposed the Land use-based Temperature Model (LTM), an explanatory model, which extracts the interactions as dependent temperature differences. The results of this explanatory model are used in the parametrization of a Bayesian Hierarchical Model (BHM) that we refer to as Bayesian LTM. Applying the models to empirical data showed a clear improvement of the predictive results for the air temperature even with the sparseness of the available data. Our results indicate that indeed the interaction of the land use information with time-dependent factors is essential to model fine grained temperature predictions. It was also shown that this interaction is dependent on additional climatic conditions. By examining the results and

error distribution for the basic LTM in 2015 it gets clear that a change in the relationship between different measurement stations and their respective land uses occurred.

The benefit for explanatory data analysis is in the detection of where the error increases. As the overall prediction accuracy in-sample is high, one can focus on those land uses as well as points in time where the error increases systematically. This helps in understanding and detecting additional underlying reasons for the development of temperature predictions. As the negative impact of heat on human health and energy consumption is the strongest at high temperatures and so called "greening" is one of the most proposed method to reduce this impact, the implications of our study questions the effectiveness of that approach. As described in Section 6.3, such land use reduces the yearly mean temperature. But we also saw in our empirical results that this effect is dependent on the interaction of several parameters and can be reduced by changing climatic conditions. Especially when the temperature is at its highest, the impact of those green areas is the lowest. Additional effects have to be considered in urban planning to mitigate the risks inherent in e.g. heat stress and the volatile energy consumption.

Our prediction is based on the explanatory results of the LTM and then used in a BHM. The results for the BHM are more stable and show in particular the impact of the reference temperature on the prediction quality. But, most importantly, it is generalizable to almost any region. Also, it allows us to incorporate additional data sources in the future to improve our understanding and the prediction quality of intra-urban air temperature.

The insights gained from our predictions are already used in the future placement of additional weather stations for the city of Karlsruhe in Baden-Württemberg, Germany. In cooperation with the environmental monitoring agency and based on the uncertainty results a new, high quality weather station was already placed inside the city. Additionally weather

stations of the type sensebox[4] are placed according to these results to complement the LUBW measurements. Temperature based routing approaches are used to mitigate heat stress in every day situations for vulnerable and elderly people. Apart from few practical applications, today there are not many information systems that utilize fine-grained temperature distributions as these were not easily available. This thesis provides the foundations for such systems. These can range from the placement of temperature influenced shopping locations, such as ice vendors, to the management of regional health plans.

Overall, this allows us to answer the overall RQ 3 of this thesis: Temperatures in an intra-urban setting can be predicted, with regard to generalizability, robustness and accuracy, by using the *interaction* of causal drivers to produce fine-grained interpolations. This allows the use of other, more complex models such as from the field of meteorology as a basis for the here presented approach. Using these existing models and their mostly spatially coarse-grained predictions, we can then interpolate their results to produce our fine-grained maps. This has been achieved by the introduction of the LTM. It also allows an explanation of the local differences in an easy and understandable way. It therefore can be combined with with the contributions of the previous part, the stable detection of points of interest. It was stated in that part that an hot spot analysis only provides points of interest, not explanations. The LTM allows this explanation and therefore not only provides sufficient data points for an hot spot analysis, but also can also use the results of that exploratory analysis for its own parametrization. This allows for a reinforcement cycle of different analysis approaches, into which also optimization as well as visual analytic approaches can be integrated. This overall combination was presented as the overall goal for any Big Data GIS in Wiener et al. (2016) and this thesis provides a substantial addition to this vision.

The primary limitation of this part lies in the sparse empirical data

[4] https://sensebox.de/

sources. While there will be new measurements in the future, up to now there exists only few weather stations with a fine temporal scale and long historical time-series. We used the two years with the most available measurements and to include as many land use classes as possible. But even so, only nine land use classes could be compared for 2013 and only eight for 2015. We also saw the impact of the yearly climate on the interaction results, which could only be compared between two years. More measurement stations and an increase of the study area could lead to additional insights, but the comparability between different measurements methods and classifications has to be considered. This could lead to the possibility of modeling different lead and lag effects for the interactions and the improvement of the prediction in irregular events such as heat waves. In the future, we will extend this research from considering only directly surrounding neighborhood for the land use class to an increased neighborhood and their land use classes. The problem herein lies in the complexity of the increased combinatoric which leads to the curse of dimensionality.

A different approach to increase the number of observation could also lie in the inclusion of more measurements with highly heterogeneous quality: Volunteered Geographic Information. While work such as in Meier et al. (2017) propose a rigorous filtering approach, a more big data focus approach is presented in Bruns et al. (2018). There, an evolutionary algorithm is combined with a kriging algorithm to first learn the quality of measurements and then interpolate their measurements. Such an approach could be combined with the LTM to provide a more in-depth explanation of the causes, an improved prediction and also to reduce the risk of overfitting even further.

Based on the positioning of our data sources, the impact of the distance could not be incorporated into the our reasoning. Modelling extensions and approaches such as the semantic krigingBhattacharjee et al. (2016) or a regression-kriging approach could only be learned on the impact of several

kilometer, while for the urban prediction the spatial accuracy has to be reduced to several meters. While we included the distance in several models, the results out-of-sample were of very low quality - This was most likely the result of overfitting. With additional stations on a small geographical area with distances between those stations in meters instead of kilometers such methods could be applied.

Part IV

Finale

Chapter 8

Conclusions and Outlook

TODAY, spatio-temporal data is the basis for almost any IS big data real-world process. However, the inherent challenges and properties of these data sources are often neglected. Whereas temporal dependencies and their auto-correlations are often accounted for, the spatial aspects as well the interaction of these dependencies are not. This neglect can lead to unreliable or even false insights. Therefore, a stronger emphasis on methods and approaches for spatio-temporal analyses is needed.

In this work, new, robust methods for spatio-temporal data analysis were developed, discussed and evaluated. First, an exploratory approach on the basis of hot spot analysis was presented. It modified the well-known Getis-Ord statistic to be more robust by considering the impact of the study area on the stability of the resulting analysis. To enable an easy, quantifiable comparison between different hot spot detection algorithms a metric for the stability was introduced, the *SoH*. To provide more in-depth insights, an explanatory data analysis as well as prediction approach based on this analysis was presented, the *LTM*. It uses the openly and freely available data to explain and predict the temperature differences at different locations while being robust, computationally efficient as well as generalizable to most areas in middle Europe.

The development of these methods and approaches provide further

methodical foundations for future information systems. Spatial processes are key requirements for future development and the usage of these systems to further the in-depth understanding and analysis of real world processes.

8.1 Contribution

This work aimed to develop new robust methods and models for the understanding of spatio-temporal phenomena. This was achieved for exploratory data analysis with the robust detection of points of interest and for explanatory data analysis and prediction it was achieved with the new LTM. Existing approaches and insights were explored, discussed and built upon to develop robust models to detect and predict spatio-temporal dependencies and hot spots.

This section summarizes the results for the research questions formulated in chapter 1.2 and, based on these, outlines the contribution as well as limitations of each result for research and practice. In addition, the overall contribution of the combination of these questions will be discussed.

How to create a stable Hot Spot Analysis: To provide reliable information and overview of existing spatio-temporal patterns, exploratory analysis is the first step in most analyses. The detection of points of interest helps to focus the analysis and to provide a quick and easy to understand visualization for a practitioner as well as researcher of any spatial as well as spatio-temporal phenomenon. But to be used on big data sets, the methods for the analysis has to be done independently of the chosen area and reliable regardless of the initial parametrization – it has to be *stable*. In this work we used the Getis-Ord statistic as a basis for our stability enhancing modification, as this is the most well-known hot spot analysis to detect points of interest and it achieves this by transforming a given data set to the equivalent z-score of each area.

(a) What effects and parameter influence the stability of hot spot analysis: To produce a reliable method for hot spot analysis, the different influences on the stability have to be analyzed and fully understood. On the basis of the G^* statistic, three parameter and their influences on the stability are identified: (1) The variable under observation X and its single elements x_j, (2) the weight matrix W, i.e. the neighborhood, and its elements $w_{i,j}$ and (3) the global mean $\bar{X}$ and the global standard deviation S. Each of these parameter influences the resulting z-score and therefore the stability differently. The influence of the variable under observation is (mostly) independent of spatial effects; the influence in its most simple form can be observed in the formula as the standard z-score which is well-known. The weight matrix represents the spatial (local) influences on each variable under observation and are dependent on the underlying data and its spatio-temporal autocorrelation. The global variables function as the standardization to transform the original variables to their z-scores.

(b) How can existing methods be modified to be more stable: Building upon existing approaches for hot spot analysis provides many advantages. We can use existing insights as well as proven properties. In particular the G^* statistic is well grounded in the theory of field of statistics. In doing so, we found that while the problem of the stability is often discussed and there do exist a a substantial amount of scientific work to solve this problem, the approach is almost exclusively to modify the weight matrix by automating its computation. And while this incorporates the spatial dependency, it does not allow an easy comparison between different reference areas. Built upon the insights of the previous question, our goal was to modify the G^* statistic to be more independent of the reference area. Therefore we introduced an additional comparison area, the *focal area*, which substitutes the global mean and standard deviation. This allows the values to be more stable to a change in the reference area. This approach, to focus solely on the reference area, is unique to our knowledge. The results are quite prom-

ising in regard to the improved stability and, in addition, this approach allows a more differentiated analysis of points of interest as the impact of outliers are much reduced and more local hot spots can be detected.

(c) What is the optimal parametrization for an hot spot analysis: For practical use, the results of a hot spot analysis have to be stable as unstable results lead to a miss-allocation of resources. And while the new *Focal $G*$ statistic* provides more stable results, there exists a manifold of potential parametrizations. Regarding the influences of the parameter, the choice is often dependent on the goals of the study itself as well as a-priori domain specific knowledge and the availability of the data. But there do exist degrees of freedom for a researcher as well as a practitioner. As the number of potential parametrizations increase, the potential for a suboptimal parametrization increases as well and therefore a "rule of thumb" is needed for most practical purposes. In our empirical evaluation it is strongly indicated that a broader weight matrix improves the stability overall. A circular weight matrix is also more stable then a square weight matrix. But, as the empirical evaluation was limited to two data sets and a step-wise comparison of the results, the overall answer remains inconclusive.

Overall, the existing hot spot analyses can be made more stable by mitigating the effect of changing reference areas. We showed a simple, but effective implementation of this approach with the new Focal $G*$ statistic. Our implementation is written in way to allow for parallel computation of the results and easy modification with other stability enhancing approaches and modifications. This provides methods to allow researchers of all fields to detect points of interest in an automated and robust way. Furthermore, this enables the detection of more local hot spots and therefore a more fine-grained overview of areas of interests. The approach presented here is designed for the use in big data use cases and was implemented using the Apache Spark and GeoTrellis framework.

How can the stability of found Hot Spots be measured: Existing meth-

ods to measure the stability of hot spots rely on visual analysis and thereby on the intuition of the human analyst. This assumes a relatively low number of comparisons to be sufficient. In turn, only few parametrizations are evaluated and therefore have to based on a-priori knowledge of the analyst. If such knowledge is not available, best practices from different analyses are often used. However, this evaluation approach is expensive, reliant on pre-existing knowledge and work and by its nature subjective. A comparison between different analyses or even algorithms is difficult with such a method. In today's world of big data and the potentially huge number of different parametrizations and results, the resulting difficulties and challenges are increased. To solve this problem, we proposed a definition of the stability, derived from the field of clustering, and proposed a metric, the *SoH*, to automatically quantify the stability of a hot spot analysis result in comparison to different parametrizations or approaches. The proposed metric measures the difference in found hot spots between different parametrizations. This allows for an comparison of the stability by machines. An analyst does not have to manually compare the different results. For the purpose of productive use of information systems, a user can be provided with a reduced selection of the most stable parametrizations. These can then be manually compared in case of an experienced user or directly applied in case of an inexperienced user. By pre-determining the most robust results, the risk and therefore costs of unstable parametrizations and suboptimal results is minimized.

How can temperatures in an intra-urban setting be predicted: But to mitigate and act upon found points of interest, an understanding of why the these points differ is needed. Often there does not exist sufficient data points to do an exploratory analysis. Therefore, explanatory methods and predictions are needed to provide this data and insight for use in advanced information systems. We used temperature prediction as our use case for this, as temperature has a profound impact on most spatio-temporal pro-

cesses and applications.

(a) What are causal drivers behind local temperature differences: To build robust statistical models, a thorough understanding of the subject matter is needed. We used the urban heat island literature as our basis, as this field examined different drivers for local temperature differences over many decades. Additionally, we restricted and pre-selected potential drivers in regard to their generalizability as well as available data sources. This was done to ensure that the resulting new insights and predictions can be applied to all real world scenarios. We identified three main drivers: The land use, the month and the hour of the day. We showed the explanatory power of each driver with statistical as well as visual tests. Furthermore, we derived and empirically evaluated our hypothesis that the interaction of these parameter would produce a better explanation while still retaining the robustness and generalizability.

(b) Given the inherent complexity of the underlying meteorological, environmental and physical processes and the sparseness of available meteorological data, how can those drivers be modeled to produce an accurate and robust prediction: Based on the results of the previous question, we argued both from a theoretical point of view as well as with our empirical data that it is essential to model the interaction of land use with temporal aspects. This focus on the interaction allowed us to build the *Land use-based Temperature Model (LTM)*. The LTM in its basic form can be used to explain the local temperature differences and improves in-sample predictions by a substantial amount compared to the baseline model. By using the parameter and their interaction in more advanced models, in our case Bayesian Hierarchical Modeling to incorporate semi-pooling, we achieved an even stronger improvement out of sample. Our empirical results indicate that indeed the interaction of the land use information with time-dependent factors is essential to model fine grained temperature predictions. The LTM provides simple, stable and general prediction with good results. This allows the

prediction of fine-grained temperature over a broad area and may lead to highly accurate temperature maps which can then be used in advanced IS applications such as heat based routing algorithms or smart city planning. Interestingly, our empirical results also suggests that "greening" has to be regarded more carefully. We found a high deviation in the moderating effect of green areas between the years 2013 and 2015 in our data set. Whereas the existing literature often argues that greening of urban places is always beneficial (e.g. Gill et al. (2007)), our results indicate that this moderating effect weakens in case of long heat periods.

As of the nature of the research questions, they were divided into two parts during this thesis: The explanatory and the explanatory approach. The contribution of each of these parts was discussed separately, but these parts are deeply interconnected (see e.g. Appice and Malerba (2014)). To truly understand a process, the data has to be explored, its structures analyses and explained, unknown areas or combinations predicted and then the so generated new data set explored in more detail. This process leads to a more thorough understanding and self-reinforcing learning mechanism. The exploration phase can also be used to reduce the amount of data to focus on the most interesting or relevant data points or areas, which then guides the in-depth explanatory analysis and in turn the prediction. This can then again be generalized. The research questions and the contribution of this thesis are built upon this idea. The robust identification of points of interests allows for a fast identification of relevant study areas. The focal G^* statistic provides the means to produce this overview; the definition of the focal radius allows the researcher to pre-determine the resolution level and comparison area to provide the needed level of detail. The *SoH* metric guarantees the robustness of the result and minimizes the risk of using unstable hot spots and parametrizations. The results of such an exploratory analysis can then be used with high confidence for the more in-depth exploratory analysis and prediction. We provided such an analysis

our causal modeling of temperature differences and the resulting *LTM*. It provides an understanding of the local temperature differences and can be used in advanced prediction models. We showed this with a Bayesian LTM and proved the benefit in our empirical evaluation. These predictions can then be used to generate a temperature map for a greater area, upon which a new hot spot analysis can be performed to use limited resources more efficiently.

A vision for this interconnected approach can be found in Wiener et al. (2016). The contributions of this work can be used in the model as well as visualization part of this BigGIS approach. They are embedded in this pipeline of continuous refinement and provide users and researchers with the tools and options to generate meaningful and robust insights even in high uncertainty environments.

Apart from their contribution to the scientific field, in particular smart city research, the methods and insights of this thesis are used in the development of commercial products as part of the BigGIS project[1].

Overall, this thesis contributes to the field of IS with the creation of robust and efficient models and algorithm for exploratory and explanatory data analysis in the field of spatial and spatio-temporal data. The methods provide an increased data foundation for new, innovative application. They also provide a high degree of generalization and can be applied to many different domains and use cases in the field of IS and other fields of research. In the domain of temperature this thesis provides additional contributions. First, it enhances the understanding of UHI by providing a reliable detection method with the Focal G^* statistic as well as an understanding of why the temperature differences exist with the *LTM*. Second, it emphasizes the importance of modeling the interaction of different causal and temporal drivers of temperature for more robust and accurate temperature prediction in case of limited data availability, both in scope as well as in diversity. Finally, this provides new insight regard the greening of

[1] biggis-project.eu

areas. Whereas before, the focus was often on low resolution aggregates, both in time and space, a more fine-grained examination leads to a more differentiated view of the impact of green areas on temperature.

8.2 Future Work and Restrictions

We have shown the importance of developing new models and methods to deal with the challenges of spatial and spatio-temporal in this thesis. The following proposed directions are built upon the insights of this work and show extension as well as alternative approaches to deal with these challenges and further extend the toolbox of IS research. We present three different areas for future work:

(1) Exploratory data analysis: In this work, the stability of a hot spot analysis was improved by modifying the computation of the reference value for the global mean and standard deviation. While the benefits of this approach were shown and discussed, the modification of the weight matrix is also feasible, as the existing literature shows. A combination of the focal G^* statistic with a modification of the weight matrix could lead to more robust results. It would be quite interesting to combine this e.g. with the work of Westerholt et al. (2015) to combine the strength of these different approaches. However, both approaches increase the computational cost. The challenge of this direction lies in the trade-off between the increased robustness and the increased cost. While the formulation of Focal G^* allows for these modifications in its computationally efficient form, other modifications have to be re-written to allow their efficient use.

Another extension of the Focal G^* statistic would be the use of different spatial data, such as social media data, "data in motion" (e.g. data from moving vehicles) or continuous measurements. In this thesis, the conception as well as empirical evaluation was performed on data in the raster format with discrete values. However, by changing the nature of the

variable under observation, additional challenges have to be addressed. Whereas here the weight matrix has several fixed values, now a function dependent on the temporal or spatial distance has to be considered. This incurs additional computational costs and increases the complexity of the method. The inherent problem is increased with moving measurement devices. In this case a trade-off between aggregates and the detail level of the found points of interest has to be explored.

The existing SoH metric allows a quantifiable, robust and understandable way to measure the stability of hot spots. But there do exist several limitations at the current state. As of the nature of hot spots, the SoH is only applicable in a singular direction. A comparison between the $SoH \uparrow$ and $SoH \downarrow$ is not possible. It is not yet bijective. Another limitation is that it only measures the distance between two different parametrizations. A comparison of the stability between all possible parametrizations incurs a high computational costs. Therefore, modifications and extensions to this metric which reduce these problems could increase the value of this metric even further. In his master thesis, Gassenschmidt (2017) shows several potential extensions for the SoH metric, inspired by the error metrics of classification algorithms. These provide a good starting point for more theory based research for the extension of the SoH.

(2) Prediction methods: The presented approach of the LTM and the Bayesian LTM show the benefit of the interaction modeling. It provides a robust and generalizable prediction approach. However, in the presented evaluation, the uncertainty of the predictions has not been utilized. Instead, only the mean prediction values are used. The use of uncertainty intervals as well as the overall information of the distributions can provide additional information for the analyst as well as for models which use the prediction results. An example for this idea was presented in Bruns et al. (2017), where uncertainty intervals of the prediction are visualized. This is highly dependent on the study as well as the use case. The challenge lies in

the efficient utilization of such intervals and in particular in the visualization for a user, e.g. embedded in an information system. The field of Visual Analytics would be particularly well-suited for this challenge.

Another approach of interest would be the use of additional data such as VGI measurements. However these are often of varying quality and only insufficient information about their placement and sensors is available. New methods to account for this heterogeneous information is needed to fully utilize the potential wealth of information. In particular an automated pre-processing is essential, as these data sets are big data sets also in volume and veracity. A potential solution to this problem is presented in Bruns et al. (2018), where spatial statistical models in the form of a modified kriging are combined with a simple genetic learning algorithm to learn the quality of each sensor.

Finally, an extension of the LTM could be found with the use of regression-kriging. In the available empirical data set, the data was insufficient to built such a model. The LTM represents the regression part of this modeling approach and the incorporation of additional information in form of the distance could improve the results. However, as of writing this thesis, the necessary data set was not openly available to the author to implement and evaluate such a combined model.

(3) Application to new data sets: A major limitation in this work was the available data sources for the empirical evaluation. Therefore, additional and new data sets as well as use cases could provide additional insights as well as new extensions to the toolbox of researchers and practitioners. For example, Bernsdorf and Bruns (2016) discuss the potential use of satellite based measurements for the urban environment, e.g. for temperature prediction. While their use for prediction models and exploratory analyses is well-discussed in the field of environmental sciences, they are rarely used in IS research. Other interesting data sets include pollution, car sharing or socio-economic interaction in urbanizations.

In summary, various directions and areas for future research for the use of spatial and spatio-temporal data sets in the field of IS exist, in particular for exploratory and explanatory data analysis and prediction methods. The development of new methods and their inclusion in the IS toolbox promises to provide new insights and improve our understanding of complex, real-world processes. This thesis provided a valuable extension to the toolbox of researchers and practitioners for robust spatial data analysis, which should help in the task of understanding these processes.

Part V

Appendix

Software Used

To promote credit for software that powers science, here is a list of software packages employed in our data analysis: Hlavac (2015), Bivand et al. (2013), McElreath (2016), Stan Development Team (2016), R Core Team (2016), Hijmans (2016b), Xie (2014), Zeileis and Grothendieck (2005), Hijmans (2016a), Gräler et al. (2016), Bivand et al. (2016), Wickham (2007).

Bibliography

Aldstadt, J. and A. Getis (2006). Using amoeba to create a spatial weights matrix and identify spatial clusters. *Geographical Analysis 38*(4), 327–343.

Ankerst, M., M. M. Breunig, H.-P. Kriegel, and J. Sander (1999). Optics: ordering points to identify the clustering structure. In *ACM Sigmod record*, Volume 28, pp. 49–60. ACM.

Anselin, L. (1988). *Spatial Econometrics: Methods and Models*, Volume 4. Springer Science & Business Media.

Anselin, L. (1995). Local indicators of spatial association - lisa. *Geographical Analysis 27*(2), 93–115.

Anselin, L. (2010). Thirty years of spatial econometrics. *Papers in Regional Science 89*(1), 3–25.

Appice, A. and D. Malerba (2014). Leveraging the power of local spatial autocorrelation in geophysical interpolative clustering. *Data Mining and Knowledge Discovery 28*(5-6), 1266–1313.

Arnfield, A. J. (2003, jan). Two decades of urban climate research: a review of turbulence, exchanges of energy and water, and the urban heat island. *International Journal of Climatology 23*(1), 1–26.

Arribas-Bel, D. (2014). Accidental, open and everywhere: Emerging data sources for the understanding of cities. *Applied Geography 49*, 45–53.

Basu, R. (2009). High ambient temperature and mortality: a review of epidemiologic studies from 2001 to 2008. *Environmental Health 8*, 40.

Ben-David, S. and M. Ackerman (2009). Measures of clustering quality: A working set of axioms for clustering. In *Advances in neural information processing systems*, pp. 121–128.

Ben-David, S., U. Von Luxburg, and D. Pál (2006). A sober look at clustering stability. In *International Conference on Computational Learning Theory*, pp. 5–19. Springer.

Benz, S. A., P. Bayer, K. Menberg, S. Jung, and P. Blum (2015). Spatial resolution of anthropogenic heat fluxes into urban aquifers. *Science of The Total Environment 524*, 427 – 439.

Bernsdorf, B. and J. Bruns (2016). Big-data und data-mining im umfeld der städtischen nutzungskartierung. In *8. Dresdner Flächennutzungssymposium (DFNS)*.

Bhattacharjee, S., M. Das, S. K. Ghosh, and S. Shekhar (2016). Prediction of meteorological parameters: An a-posteriori probabilistic semantic kriging approach. In *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '16, New York, NY, USA, pp. 38:1–38:10. ACM.

Bivand, R., T. Keitt, and B. Rowlingson (2016). *rgdal: Bindings for the Geospatial Data Abstraction Library*. R package version 1.1-10.

Bivand, R. S., E. Pebesma, and V. Gomez-Rubio (2013). *Applied spatial data analysis with R, Second edition*. Springer, NY.

Blaschke, T., G. J. Hay, Q. Weng, and B. Resch (2011). Collective sensing: Integrating geospatial technologies to understand urban systems—an overview. *Remote Sensing 3*(8), 1743–1776.

Boots, B. and M. Tiefelsdorf (2000). Global and local spatial autocorrelation in bounded regular tessellations. *Journal of Geographical Systems 2*(4), 319–348.

Broy, M. (2010). Cyber-physical systems—wissenschaftliche herausforderungen bei der entwicklung. In *Cyber-Physical Systems*, pp. 17–31. Springer.

Bruns, J., J. Riesterer, B. Wang, T. Riedel, and M. Beigl (2018). Automated quality assessment of (citizen) weather stations. *Journal GI_Forum 2018* (Accepted).

Bruns, J., D. Seebacher, and M. Stein (2017). Predicting habitable zones of invasive species based on temperature. OR2017.

Bruns, J. and T. Setzer (2018). Causal modelling of spatial temperature differences by combining temperature and land use data. *Urban Climate (under Review)*.

Bruns, J. and V. Simko (2017). Stable hotspot analysis for intra-urban heat islands. *GI_Forum 1*, 79–92.

Cai, M., C. Ren, Y. Xu, K. K.-L. Lau, and R. Wang (2017, jun). Investigating the relationship between local climate zone and land surface temperature using an improved WUDAPT methodology – a case study of yangtze river delta, china. *Urban Climate*.

Chase, T. N., K. Wolter, R. A. Pielke, and I. Rasool (2006). Was the 2003 european summer heat wave unusual in a global context? *Geophysical Research Letters 33*(23), n/a–n/a. L23709.

Chen, A., X. A. Yao, R. Sun, and L. Chen (2014). Effect of urban green patterns on surface urban cool islands and its seasonal variations. *Urban forestry & urban greening 13*(4), 646–654.

Cliff, A. D. and J. K. Ord (1981). *Spatial processes: models & applications*. Taylor & Francis.

Cressie, N. and C. K. Wikle (2015). *Statistics for spatio-temporal data*. John Wiley & Sons.

Davis, J. C. (1986). *Statistical and data analysis in geology*. J. Wiley.

Davis, S. L., T. E. Wilson, A. T. White, and E. M. Frohman (2010). Thermoregulation in multiple sclerosis. *Journal of Applied Physiology 109*(5), 1531–1537.

Department of Economic and Social Affairs, Population Division, United Nations (2014). World urbanization prospects: The 2014 revision, highlights (st/esa/ser. a/352).

Eclipse Foundation, Azavea, and contributors (2016). GeoTrellis, Apache 2.0 License. https://github.com/locationtech/geotrellis.

Ester, M., H.-P. Kriegel, J. Sander, X. Xu, et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, Volume 96, pp. 226–231.

Ferretti, V. and G. Montibeller (2016). Key challenges and meta-choices in designing and applying multi-criteria spatial decision support systems. *Decision Support Systems 84*, 41–52.

Florax, R. J. and S. Rey (1995). The impacts of misspecified spatial interaction in linear regression models. In *New directions in spatial econometrics*, pp. 111–135. Springer.

Gassenschmidt, M. (2017). Stabile geotemporale hotspot analyse. Master's thesis, Karlsruhe Institute of Technology.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin (2014). *Bayesian data analysis*, Volume 2. CRC press Boca Raton, FL.

Getis, A. and J. Aldstadt (2010). Constructing the spatial weights matrix using a local statistic. In *Perspectives on spatial data analysis*, pp. 147–163. Springer.

Getis, A. and J. K. Ord (1992). The analysis of spatial association by use of distance statistics. *Geographical analysis 24*(3), 189–206.

Gill, S. E., J. F. Handley, A. R. Ennos, and S. Pauleit (2007). Adapting cities for climate change: the role of the green infrastructure. *Built environment 33*(1), 115–133.

Gräler, B. (2014). *Developing Spatio-temporal Copulas*. Ph. D. thesis.

Gräler, B., E. Pebesma, and G. Heuvelink (2016). Spatio-temporal interpolation using gstat. *R Journal 8*(1), 204–218.

Griffith, D. A. (1996). Some guideline for specifying the geographic weights matrix contained in spatial statistical models. *Practical handbook of spatial statistics*, 65–82.

Grubesic, T. H., R. Wei, and A. T. Murray (2014). Spatial clustering overview and comparison: Accuracy, sensitivity, and computational expense. *Annals of the Association of American Geographers 104*(6), 1134–1156.

Hansen, J., R. Ruedy, M. Sato, and K. Lo (2010). Global surface temperature change. *Reviews of Geophysics 48*(4).

Hassid, S., M. Santamouris, N. Papanikolaou, A. Linardi, N. Klitsikas, C. Georgakis, and D. Assimakopoulos (2000). The effect of the athens heat island on air conditioning load. *Energy and Buildings 32*(2), 131 – 141.

Hay, G. J., C. Kyle, B. Hemachandran, G. Chen, M. M. Rahman, T. S. Fung, and J. L. Arvai (2011). Geospatial technologies to improve urban energy efficiency. *Remote Sensing 3*(7), 1380–1405.

Hübler, M., G. Klepper, and S. Peterson (2007). Costs of climate change: The effects of rising temperatures on health and productivity in germany. Kiel Working Paper 1321, Kiel Institute for the World Economy, Kiel.

Hengl, T., G. B. Heuvelink, and D. G. Rossiter (2007). About regression-kriging: from equations to case studies. *Computers & geosciences 33*(10), 1301–1315.

Hengl, T., G. B. M. Heuvelink, M. Perčec Tadić, and E. J. Pebesma (2012). Spatio-temporal prediction of daily temperatures using time-series of modis lst images. *Theoretical and Applied Climatology 107*(1), 265–277.

Hijmans, R. J. (2016a). *geosphere: Spherical Trigonometry*. R package version 1.5-5.

Hijmans, R. J. (2016b). *raster: Geographic Data Analysis and Modeling*. R package version 2.5-8.

Hlavac, M. (2015). *stargazer: Well-Formatted Regression and Summary Statistics Tables*. Cambridge, USA: Harvard University. R package version 5.2.

Jankowski, P., G. Fraley, and E. Pebesma (2014). An exploratory approach to spatial decision support. *Computers, Environment and Urban Systems 45*, 101–113.

Kalnay, E. and M. Cai (2003). Impact of urbanization and land-use change on climate. *Nature 423*(6939), 528.

Keim, D., G. Andrienko, J.-D. Fekete, C. Görg, J. Kohlhammer, and G. Melançon (2008). Visual analytics: Definition, process, and challenges. In *Information visualization*, pp. 154–175. Springer.

Kleinberg, J. M. (2003). An impossibility theorem for clustering. In *Advances in neural information processing systems*, pp. 463–470.

Kratzer, A. (1937). Das Stadtklima. Die Wissenschaft, Einzeldarstellungen aus der Naturwissenschaft und der Technik, Band 90.

Krige, D. G. (1951). A statistical approach to some basic mine valuation problems on the witwatersrand. *Journal of the Southern African Institute of Mining and Metallurgy 52*(6), 119–139.

Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics-Theory and methods 26*(6), 1481–1496.

LeSage, J. P. (1999). The theory and practice of spatial econometrics. *University of Toledo. Toledo, Ohio 28*, 33.

Lukasczyk, J., R. Maciejewski, C. Garth, and H. Hagen (2015). Understanding hotspots: A topological visual analytics approach. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 36. ACM.

MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Volume 1, pp. 281–297. Oakland, CA, USA.

Martin, P., Y. Baudouin, and P. Gachon (2015). An alternative method to characterize the surface urban heat island. *International Journal of Biometeorology 59*(7), 849–861.

Matheron, G. (1963). Principles of geostatistics. *Economic geology 58*(8), 1246–1266.

McElreath, R. (2016). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. CRC Press.

Meier, F., D. Fenner, T. Grassmann, M. Otto, and D. Scherer (2017). Crowdsourcing air temperature from citizen weather stations for urban climate research. *Urban Climate 19*, 170 – 191.

Mirzaei, P. A. and F. Haghighat (2010). Approaches to study urban heat island – abilities and limitations. *Building and Environment 45*(10), 2192–2201.

Mohsin, T. and W. A. Gough (2012). Characterization and estimation of urban heat island at toronto: Impact of the choice of rural sites. *Theoretical and Applied Climatology 108*(1-2), 105–117.

Moran, P. A. (1950). Notes on continuous stochastic phenomena. *Biometrika 37*(1/2), 17–23.

Oke, T. R. (1982). The energetic basis of the urban heat island. *Quarterly Journal of the Royal Meteorological Society 108*(455), 1–24.

Oke, T. R., G. Mills, A. Christen, and J. A. Voogt (2017). *Urban climates*. Cambridge University Press.

Oliver, M. A. and R. Webster (1990). Kriging: a method of interpolation for geographical information systems. *International Journal of Geographical Information Systems 4*(3), 313–332.

Ord, J. K. and A. Getis (1995). Local spatial autocorrelation statistics: Distributional issues and an application. *Geographical Analysis 27*.

Ord, J. K. and A. Getis (2001). Testing for local spatial autocorrelation in the presence of global autocorrelation. *Journal of Regional Science 41*(3), 411–432.

Parsons, E. (2017). Here come the machines, the potential of machine learning for the geospatial industry. Keynote GI Forum 2017.

Pulugurtha, S. S., V. K. Krishnakumar, and S. S. Nambisan (2007). New methods to identify and rank high pedestrian crash zones: An illustration. *Accident Analysis & Prevention 39*(4), 800–811.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Resch, B., A. Zipf, P. Breuss-Schneeweis, E. Beinat, and M. Boher (2012). Towards the live city–paving the way to real-time urbanism. *International Journal on Advances in Intelligent Systems 5*(3), 470–482.

Robine, J.-M., S. L. K. Cheung, S. Le Roy, H. Van Oyen, C. Griffiths, J.-P. Michel, and F. R. Herrmann (2008). Death toll exceeded 70,000 in europe during the summer of 2003. *Comptes rendus biologies 331*(2), 171–178.

Rousseeuw, P. J. and L. Kaufman (1990). *Finding Groups in Data*. Wiley Online Library.

Rußig, J. and J. Bruns (2017). Reducing individual heat stress through path planning. *GI_Forum 1*, 327–340.

Schwarz, N., S. Lautenbach, and R. Seppelt (2011). Exploring indicators for quantifying surface urban heat islands of european cities with modis land surface temperatures. *Remote Sensing of Environment 115*(12), 3175–3186.

Schwarz, N., U. Schlink, U. Franck, and K. Großmann (2012). Relationship of land surface and air temperatures and its implications for quantifying urban heat island indicators—an application for the city of leipzig (germany). *Ecological Indicators 18*, 693–704.

Shamir, O. and N. Tishby (2008). Cluster stability for finite samples. In *Advances in neural information processing systems*, pp. 1297–1304.

Shao, Q., C. Sun, J. Liu, J. He, W. Kuang, and F. Tao (2011). Impact of urban expansion on meteorological observation data and overestimation to regional air temperature in china. *Journal of Geographical Sciences 21*(6), 994–1006.

Shekhar, S., M. R. Evans, J. M. Kang, and P. Mohan (2011). Identifying patterns in spatial information: A survey of methods. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 1*(3), 193–214.

Siu, L. W. and M. A. Hart (2013). Quantifying urban heat island intensity in hong kong sar, china. *Environmental monitoring and assessment 185*(5), 4383–4398.

Song, Y. (2010, Aug). Class compactness for data clustering. In *2010 IEEE International Conference on Information Reuse Integration*, pp. 86–91.

Stan Development Team (2016). RStan: the R interface to Stan. R package version 2.14.1.

Steiner, J. (2017). I know gis ... should i become a data scientist? Keynote GI Forum 2017.

Stewart, I. D. (2011). A systematic review and scientific critique of methodology in modern urban heat island literature. *International Journal of Climatology 31*(2), 200–217.

Stewart, I. D. and T. R. Oke (2012, dec). Local climate zones for urban temperature studies. *Bulletin of the American Meteorological Society 93*(12), 1879–1900.

Stewart, I. D., T. R. Oke, and E. S. Krayenhoff (2014). Evaluation of the 'local climate zone' scheme using temperature observations and model simulations. *International Journal of Climatology 34*(4), 1062–1080.

Sundborg, A. (1950). Local climatological studies of the temperature conditions in an urban area. *Tellus* (2.3), 222–232.

Suomi, J., J. Hjort, and J. Käyhkö (2012). Effects of scale on modelling the urban heat island in turku, sw finland. *Climate Research 55*(2), 105–118.

Suomi, J. and J. Käyhkö (2011, jan). The impact of environmental factors on urban temperature variability in the coastal city of turku, SW finland. *International Journal of Climatology 32*(3), 451–463.

Thakali, L., T. J. Kwon, and L. Fu (2015, jun). Identification of crash hotspots using kernel density estimation and kriging methods: a comparison. *Journal of Modern Transportation 23*(2), 93–106.

Tobler, W. R. (1970). A computer movie simulating urban growth in the detroit region. *Economic geography 46*(sup1), 234–240.

Tomlin, C. (1990). *Geographic information systems and cartographic modeling*. Prentice Hall series in geographic information science. Prentice Hall.

Vapnik, V. N. and A. Y. Chervonenkis (2015). On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity*, pp. 11–30. Springer.

Wagner, S., T. Brandt, and D. Neumann (2016). In free float: Developing business analytics support for carsharing providers. *Omega 59*, 4 – 14. Business Analytics.

Westerholt, R., B. Resch, and A. Zipf (2015). A local scale-sensitive indicator of spatial autocorrelation for assessing high-and low-value clusters in multiscale datasets. *International Journal of Geographical Information Science 29*(5), 868–887.

Wickham, H. (2007). Reshaping data with the reshape package. *Journal of Statistical Software 21*(12), 1–20.

Wiener, P., M. Stein, D. Seebacher, J. Bruns, M. Frank, V. Simko, S. Zander, and J. Nimis (2016). Biggis: a continuous refinement approach to master heterogeneity and uncertainty in spatio-temporal big data (vision paper). In *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 8. ACM.

Work, D. B. and A. M. Bayen (2008). Impacts of the mobile internet on transportation cyberphysical systems: traffic monitoring using smartphones. In *National Workshop for Research on High-Confidence Transportation Cyber-Physical Systems: Automotive, Aviation, & Rail*, pp. 18–20.

Xie, Y. (2014). *knitr: A Comprehensive Tool for Reproducible Research in R*. Chapman and Hall/CRC. ISBN 978-1466561595.

Ye, X., R. Wolff, W. Yu, P. Vaneckova, X. Pan, and S. Tong (2012). Ambient temperature and morbidity: a review of epidemiological evidence. *Environmental health perspectives 120*(1), 19–28.

Zeile, P. (2017). Urban emotions and realtime planning methods. In *REAL CORP 2017–PANTA RHEI–A World in Constant Motion. Proceedings of 22nd International Conference on Urban Planning, Regional Development and Information Society*, pp. 617–624.

Zeileis, A. and G. Grothendieck (2005). zoo: S3 infrastructure for regular and irregular time series. *Journal of Statistical Software 14*(6), 1–27.