Visual Analytics of Spatio-Temporal Event Predictions: Investigating Causes for Urban Heat Islands

Visuelle Analyse von raumzeitlichen Ereignisvorhersagen: Untersuchung der Ursachen für Urbane Hitzeinseln

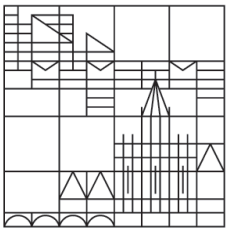
Master Thesis

by

Matthias Miller

at the

Universität Konstanz



to attain the academic degree ***Master of Science***

Group: Data Analysis and Visualization

Department: Computer and Information Science

1. **Reviewer:** Prof. Dr. Daniel A. Keim
2. **Reviewer:** Prof. Dr. Bela Gipp

Konstanz, 2018

Miller, Matthias:
Visual Analytics of Spatio-Temporal Event Predictions: Investigating Causes for Urban Heat Islands
Master Thesis, Department of Computer and Information Science, 2018.

Abstract

Due to ongoing urbanization and industrialization processes, city planners are facing an increasing number of challenges. One such challenge is the phenomenon of locally increased temperature – the urban heat island effect. This effect is a well-known phenomenon in both the domain of city planning and meteorology. The appearance of urban heat islands is not only influenced by many apparent variabl, such as weather, vegetation, and surface characteristics, but also by rather inconspicuous parameters, like industry, transportation infrastructure, population density, energy management, and air pollution. Because of this large number of influencing factors, analyzing the causes for this effect is a complex task. To improve urban climate and energy management, innovative applications that provide support for decision-making tasks are needed. We propose a visual analytics system which enables expert users to explore temperature conditions of a city area. At its core, the system uses a random forest classification model, trained on heterogeneous, nationwide collected data, which facilitates interactive parameter steering of spatial and meteorological features. Users can explore the influence of several variables on urban heat island forecast events. The system provides spatio-temporal event predictions that offer insights about future conditions and the effect of various variables on the formation of urban heat islands. To present connections between diverse urban heat island properties and forecast events, we compose different mature, interactive visualizations. Through several use cases, we demonstrate that our system allows users to focus on relevant features while getting a solid overview of the urban heat island situation in a specific area of interest. The integration of a combination of selected reasonable visualizations with a prediction model based on ensemble learning offers a viable solution for an adequate analysis of the urban heat island effect that can be enhanced by further functionality in the future.

Acknowledgment

I want to thank my advisor Daniel Seebacher for his constant feedback that proved to be very helpful during the conceptualization process and writing of this thesis. Your ideas supported me with essential design decisions. Thank you for your openness and patience during long and frequent discussions. I learned a lot while working with you on this project and you have broadened my horizon in various ways. Manuel Stein, thank you very much for sharing your experience in the field of data visualization and for helping me with any questions that I had during the writing phase of my thesis.

Moreover, I thank Prof. Dr. Daniel A. Keim for his input that pointed me to the right direction, and I am thankful that Prof. Dr. Bela Gipp agreed to act as the second reviewer for my work.

Prof. Dr. Stefan Emeis and Dr. Saskia Buchholz provided valuable domain knowledge which helped me to focus on relevant aspects during the design, implementation, and evaluation process. I am very grateful for your assistance.

I want to thank my siblings Nelli and Daniel Miller as well as my colleagues Fabian Sperrle, Manuel Hotz, Matthias Kraus, Nicolas Siebeck, Leonard Krämer, and Steffen Fissler who were willing to act as proofreaders and provided invaluable feedback to support me getting the best out of this master thesis. I am incredibly grateful to be able to call you all my friends.

Finally, let me express a special thanks to my family constantly supporting me through prayer and encouraging and releasing me when I needed it the most. I am grateful that you always believe in me. Thank you, Céline, for comforting me and for your love you have shown me during the recent year.

Contents

List of Figures

List of Tables

Abbreviations

DWD – Deutscher Wetter Dienst
GMS – German Meteorological Service
IFPP – Input Feature Prediction Panel
J48 – Decision Tree (C4.5 standard)
KDD – Knowledge Discovery in Databases
kNN – k-Nearest-Neighbor
LST – Land Surface Temperature
MLP – Multilayer Perceptron
NB – Naive Bayes
MODIS – Moderate Resolution Imaging Spectroradiometer
OSM – Open Street Map
PCG – Pie Chart Glyph
PCP – Parallel Coordinates Plot
PVA – Predictive Visual Analytics
PWS – Personal Weather Station
RF – Random Forest
SC – Stacking Classifier
TRS – Time Range Selector
UHI – Urban Heat Island
VA – Visual Analytics
WU – Weather Underground
WUSN – Weather Underground Station Network

Chapter 1

Introduction

An ever increasing amount of people tend to move to urban areas. There are many reasons for this including the expectation to live a preferable life due to better career and shopping possibilities, and improved mobility that offers higher flexibility. Even old people tend to move into suburban or even urban areas since many clinics are situated in cities. Especially if they are not able to drive a car on their own and are in need of outside help, it is advantageous to reside in the vicinity of health facilities. Additionally, the high population density offers anonymity that promises individual freedom. In 2013, about 54 percent of the earth's population was living in urban regions [22]. Urbanization experts predict that this number will increase up to 70 percent within the next 30 years [79]. Hence, the world's city population will grow immensely. This acceleration of the world's urbanization and industrialization process entails several human modifications, for example, the decrease of green spaces by replacing natural vegetation such as forests with buildings and roads [98]. This is a major alteration of a city's morphology and limits evaporative cooling by reduced plant transpiration [67]. Hence, within areas with a higher vegetation intensity, the temperature is lower than in areas that are paved or have surfaces like concrete, asphalt or other material from anthropogenic sources [85]. Compared to rural regions, the surface characteristics of urban areas are modified by multiple human activities (e.g., urban construction). This has several negative consequences on human health that include, but are not limited to, increased heat stress for urban citizens, raised noise exposure of increased transportation load and heightened air pollution due to the industrial sector [85]. Large buildings that often act as wind breaks are present which reduces the wind speed, especially in the center of metropolises with a high residential density [100]. Besides air and surface temperature, wind speed is one of the most important meteorological variables that is directly connected to the temperature level [68]. Lower wind speeds have the tendency to intensify the heat storage because less heat is convected from surface to air [33]. In cities, anthropogenic heat generation is higher due to energy use coming from different sources like industrial processes, building air conditioning and transportation [39]. The growth of urbanization leads to more acute public traffic levels. Urban transport produces emissions that add air pollutants, contributing to the concentration of increased temperature levels [37].

1.1 Urban Heat Island Effect

The temperature level increase, which is illustrated in Figure 1.1, that can be observed within large cities and metropolitan zones, is a well-known climatic phenomenon [100]: the **urban heat island (UHI)** effect. This phenomenon has already been subject to many studies in the last three decades [75, 74, 19, 118]. The UHI effect describes that, in urban areas, surface and atmospheric temperatures are higher than in surrounding rural regions, with the highest UHI intensity (temperature difference) 3-5 hours after sunset [73]. The UHI effect is considered to be one of the best-documented aspects within the domain of city climatology. It is caused by different human activities such as urban construction and anthropogenic climate modification [118]. Moreover, it is a consequence of industrialization as a byproduct of civilization [63]. Nonetheless, domain experts could profit from effective applications with visual analytics approaches. Ideally, they facilitate the analysis by providing both meaningful visualizations and methods which improve the classification of interesting phenomena and allow manipulation of influencing parameters.

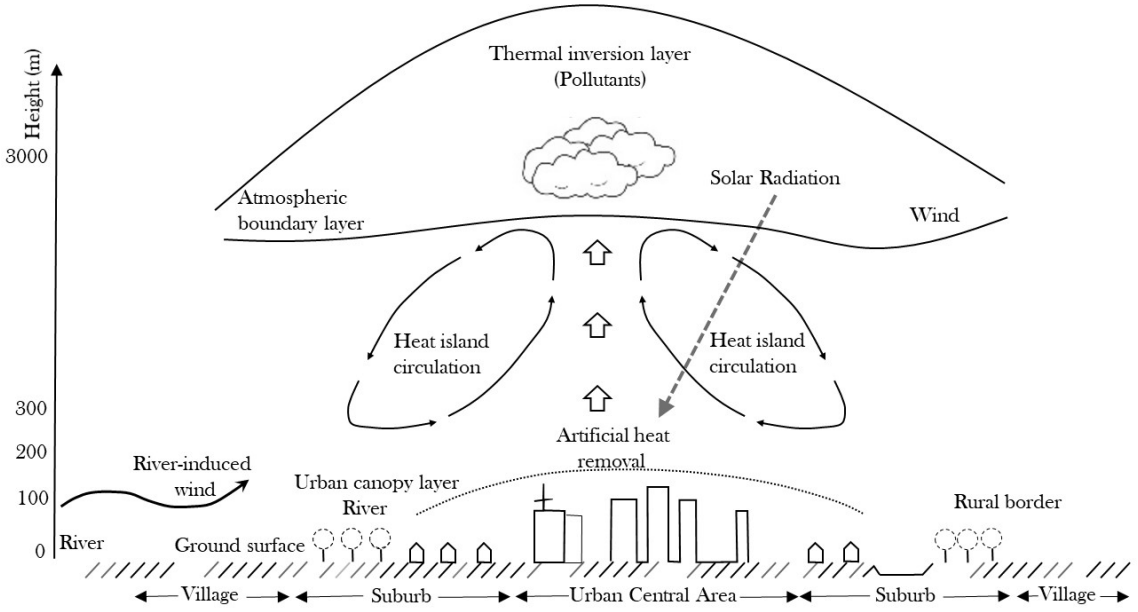


Figure 1.1: The urban heat island (UHI) effect [118].

Air temperatures crucially influence the temperature magnitude of a city's urban heat island. This temperature magnitude varies significantly between different cities. The exact characteristics of urban heat islands vary considerably from city to city. Among other reasons, this can be explained by the differences of the cities' structures, the amount of traffic [37], and the climatic type. Hence, the climatic situation is unique for each city and depends on a complex interplay of several circumstances influencing the intensity of its UHI effect.

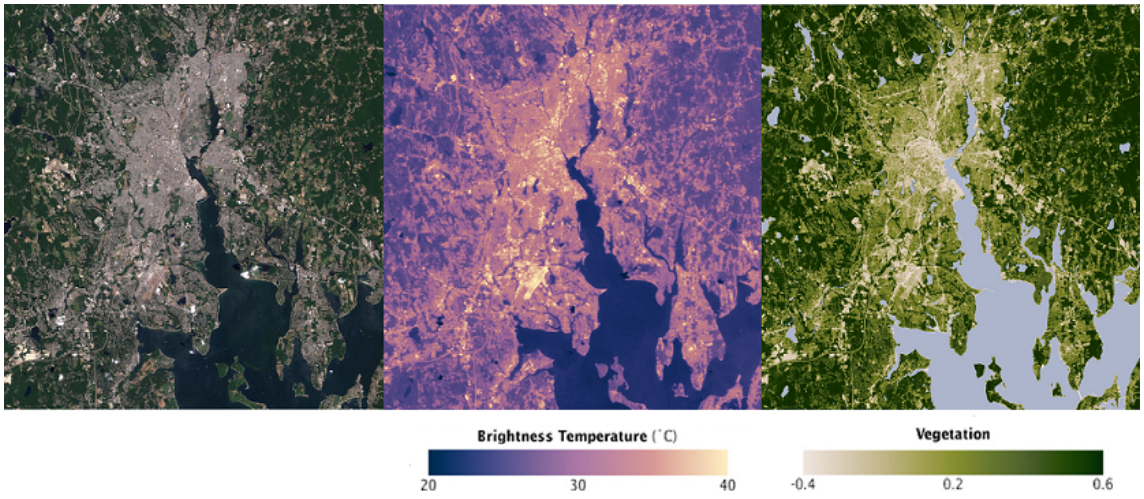


Figure 1.2: NASA satellite recordings of different surface characteristics of Providence, Rhode Island. From left to right: visible light, surface heat characteristics and the vegetation coverage in the city area [70].

Figure 1.2 presents an example of an urban heat island depicting different characteristics of the city of Providence, Rhode Island. The central part of Providence is located at the center of every view. The satellite view (left) and the vegetation cover view (right) show that the city is surrounded by green areas which are missing in its center. The image in the middle gives information about the temperature intensity that is indicated by the brightness of the color (e.g., buildings that absorb heat). It shows higher temperatures in the city's center with lower temperatures in the surrounding areas. Water surfaces are much cooler and can easily be distinguished from land surface.

The occurrence of the UHI effect has several socioeconomic and meteorological impacts. Most importantly, it has a direct negative impact on human health [77]. Depending on the geographic position of the city, different problems may arise: increased risk of malaria in South Africa, dengue fever in Thailand, hantavirus pulmonary syndrome in the Southwestern USA, cholera in Bangladesh, and more [81]. In European cities, seasonal heat-waves [28, 112] can affect human health resulting in dehydration, circulatory disorders, or exhaustion. During hot days, persistent exposure to heat causes an increased risk of mortality from heat stress, as well as other heat-related illnesses, like heat exhaustion and heat-stroke. This problem can even cause the death of urban citizens. [59]. Unfortunately, heat-related mortality is likely to increase in the future due to the ongoing urbanization process. Ambient heat exposure may result in thermal stress [66] as well as the degradation of air quality [90] that have a severe impact, especially on elderly people [77, 31, 92]. In combination with heat waves and global warming, the heat load of urban heat islands becomes even more detrimental [31]: The mortality risk increases due to an exacerbating intensity of the UHI effect. The probability of prolonged exposure to higher temperatures rises in the times of heat waves which would be less fierce if the UHI effect would be reduced [53, 28].

Apart from impaired health, another consequence of the UHI effect is the increase of energy consumption [39] that is caused by low-albedo surfaces. In street areas, industrial and commercial territories, the temperature is often higher than in residential areas and green spaces because of differences in solar radiation heating [89]. The roofing of buildings usually consists of dark materials that absorb more solar and infrared radiation than high-albedo surfaces [101, 17]. Consequently, this elevates the heat load, particularly of roofing. As a result, the cooling of these buildings requires a significantly larger amount of energy [52] and deteriorates the energy management of the city [3]. Such increased usage of air conditioning systems also increases air pollution [112] and greenhouse gas emissions [54, 44, 88]. In parallel, these causes have led to reduced air quality [65] and impaired water quality [4, 72]. All these factors result in a lower quality of life in urban areas.

The appearance of urban heat islands is subject to both the domain of urban climatology and city planning [64]. Due to technological progress, the amount of collected temperature data is increasing and available from different sources. For instance, governmental institutions are gathering high-quality data for meteorologic analysis. Unfortunately, these data sources have a very sparse geographic distribution. Therefore, they cannot be used to create acceptable models to analyze the temperature characteristics independently of the city size. The *Personal Weather Station* (PWS) network provides a multitude of meteorologically relevant data including temperature [108] from numerous connected and synchronized stations. This network is called *weather underground* (WU) and the dataset is publicly accessible via the WU API [109]. Compared to the weather stations of the German Meteorological Service (GMS) [24], this network contains more sensors. Consequently, more data is gathered which serves as a better foundation when creating geo-spatial temperature models as more data allows more precise analysis models.

In combination with land register information about the surface characteristics, provided by *open street map* (OSM) [78], and meteorologic data (wind speed, relative humidity, air pressure, soil temperature, precipitation, and cloudiness level), provided by the GMS, we can process contextual information and draw conclusions about the causes of the emergence of UHI in specific regions.

1.2 Research Objective

This thesis is intended to support domain experts with visual analytics techniques of weather and UHI-related data. A secondary objective is to develop a better understanding of the UHI problem. Therefore, we realize concepts to generate mitigation strategies regarding the consequences of the UHI effect. Eliciting influencing factors of UHIs is an essential aspect of this thesis. It is necessary to consider both the spatial and temporal dimension to reasonably address the problem of the UHI effect. We supply models to facilitate data exploration by applying methods of *knowledge discovery in databases* (KDD) [71] in combination with human interaction to support informed decision-making. The causes for UHIs depend on a complex composition of several parameters. We present a geographical information system that allows expert users to step into the analysis process to steer parameters of interest and improve the results of their analysis. Thus, heterogeneous data sources containing temperature from the *weather underground station network* (WUSN) and meteorological measurements as well as detailed land cover information have been integrated to be analyzed by classification models to predict the UHI level for different circumstances. We developed a glyph-visualization to encode information in its geospatial context. It can be adapted to the needs of the user with zoom and filter techniques. Users can dig into the details for insights about the reasons for the occurrence and extent of the UHI effect at locations of interest. We examine the benefits and shortcomings of our approach by presenting relevant use cases and conclude with potential improvements and missing features that are subject to future work.

1.3 Structure of the Thesis

This thesis introduces the problem of the UHI effect and focuses on the need for meaningful visualizations that help analysts in this domain by the usage of state-of-the-art analytics techniques to predict future trends. Therefore, in Chapter 2 we provide a detailed overview of the literature about existing UHI prediction approaches and the complex composition of many parameters that crucially influence the result of such analysis tasks. Afterwards, we state existing spatio-temporal visualization techniques to clarify our contribution to this domain. Chapter 3 contains a discussion about data sources which have been utilized for our implementation and have to pass multiple processing steps. The main components of our prototype and its functionality are introduced in Chapter 4. A thorough evaluation of several classification models is presented followed by various use cases in Chapter 5 to reveal strengths and weaknesses of our prediction visualization system. These are discussed in Chapter 6 accompanied by recommendations for future work that may improve the user experience when investigating event predictions of the UHI effect with our visualization system. With Chapter 7, we conclude the thesis by providing a summary of our visual analytics system as our primary contribution.

Chapter 2

Related Work

In this chapter, we give an overview of what has been subject to research in the past as part of the challenge when coping with the causes, effects and the prediction of Urban Heat Islands. We provide insights about previous scientific approaches addressing the problems resulting from the UHI effect. The topic of this thesis can be segmented into three fundamental components: UHI prediction, visualization of spatio-temporal prediction events, and predictive visual analytics for spatio-temporal data. First, we want to focus on approaches and models that model the prediction of the Urban Heat Island effect. By this, an overview is provided about what has to be considered to support the visual analysis task that is subject to specific requirements. Subsequently, different visualization techniques are compared concerning their strengths and weaknesses. Event prediction data visualizations can be divided into spatial, temporal and spatio-temporal event prediction. We will discuss related work for each subcategory and conclude with the positioning of our work and how it contributes to the analysis of the UHI effect.

2.1 Urban Heat Island Prediction

When it comes to the prediction of the UHI effect, we find a complex composition of multiple factors that are influencing the degree of the temperature level. The urbanization process results in a "massive land use change" of replacing vegetation by urban use areas (e.g., residential, industrial) [107]. Susca et al. discuss the "positive effects of vegetation" [97] for the mitigation of UHI. According to Lo and Quattrochi, land-use and land-cover characteristics are strongly related to the UHI level [57, 95, 96], since these changes affect "the degree of absorption of solar radiation, albedo, surface temperature, evaporation rates, transmission of heat to the soil [...]" [80]. Thus, surface characteristic data should be included in UHI prediction tasks. Shao et al. created a forecast model for UHI analysis applying artificial neural networks on eight factors: solar radiation, precipitation, sunshine duration, wind speed, evaporation, thermal conductivity, reflectivity, and heat capacity [94]. Xi and He provide a prediction model that is based on neural networks as well [117]. The

use of meteorological (relative humidity, air pressure, cloud level [...]) and human factors (urban green coverage ratio, industrial area, land use type [...]) are adequate to predict the urban heat island intensity [94]. Consequently, these parameters are essential for the assessment of urban heat island intensity and will be applied in the prediction model and the visual analytics process of this work. These systems lack an effective combination of prediction models and multiple visualizations that address different feature domains using state-of-the-art interaction techniques. We aim to close this gap by integrating visualization techniques that encode heterogeneous data. By this, we build an application that serves as a basis for the development of a visual analytics system to support domain experts during their planning and decision-making tasks. Due to the complex composition of multiple factors that lead to the UHI effect, it is essential to consider the geographic component of the UHI to be included in any visual analysis of this phenomenon. Shao et al. created a visualization model to depict the result of a prediction simulation on a 2D/3D map based on preset attribute values to support the spatial understanding. They use their map model to describe the "distribution of temperature through three-dimensional visualization techniques" [94]. Rajasekar and Weng follow a different approach to visualize UHI data for the city of Indianapolis geographically. They deploy remote sensed *Land Surface Temperature* (LST) data from NASA's Moderate Resolution Imaging Spectroradiometer (MODIS) satellite. Though this data source does not include a good temporal resolution, the satellite imagery provides images with an acceptable spatial resolution. The authors used this to create a visualization to get a better overview of the UHI effect [83].

2.2 Visualizing Spatio-Temporal Prediction Events

Since the primary objective of this thesis is to create an application that supports the visual analysis of UHI events, we can divide this prediction visualization into spatial and temporal events (see Figure 2.1).

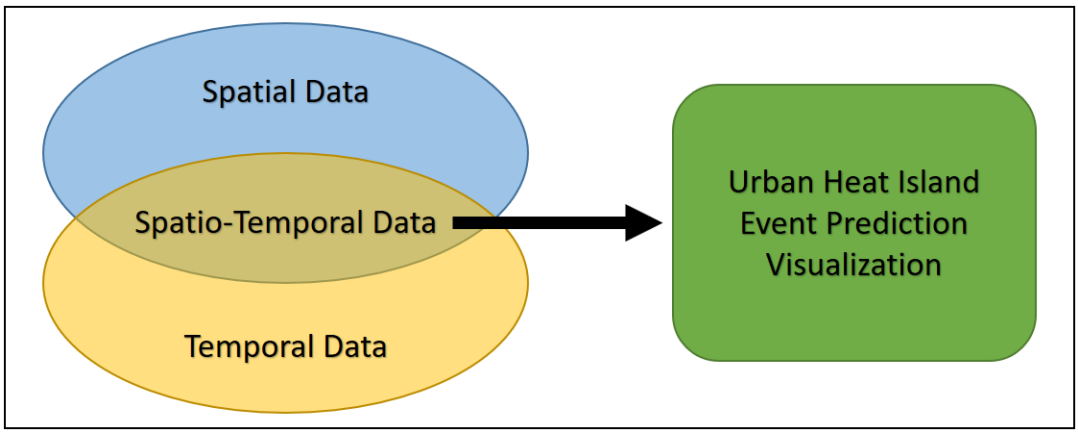


Figure 2.1: Spatial and temporal data serve as input for interactive visualizations of UHI event predictions.

When visualizing the UHI effect, we have to deal with event data containing geographic and time-sensitive characteristics. Information-enriched maps are the most apparent visualization type for visualizing spatially distributed event prediction data. Heat maps are often used to depict density and distribution of geospatial data [55, 60, 29]. For example, Oksanen et al. used a heat map to visualize trajectories of bicycle data to display the track density of a specific route [76]. Another example of the usage of heat maps is a central part of BirdVis, an application for the geospatial and temporal analysis of bird populations and their movements to identify patterns and correlations of interest [27]. Density estimate heat maps are a central part of the visual analytics application of Malik et al. for the predictive analysis of spatial crime data [62]. Besides heat maps, choropleth maps [103] are applied for encoding particular geographic elements in combination with the data that shall be explored. With the help of choropleth maps, Gimpel et al. portray the campaign contributions of American parties by applying aggregation based data on the postal (zip code) structure of the states [35]. An approach by Maciejewski et al. integrates a predictive visual analytics loop for the analysis of hotspots and includes multiple visualizations to simultaneously convey temporal and spatial information [61].

When it comes to the visualization of temporal data, radial layouts have gained popularity due to the data's periodic (e.g., yearly, monthly, daily and hourly measurements) characteristics. Fuchs et al. discovered that spiral visualization techniques [113, 7, 91] and dense-pixel displays [49] are more effective when trying to understand values at specific temporal locations [30]. Compared to the circular data presentations, earlier visualizations such as line graphs should be preferred for detecting peaks or trends within the data. Aigner et al. compare multiple time-dependent visualizations as TimeWheel [104] and ThemeRiver [42] that encode numerous attributes and rather the linear behavior of temporal data than the periodical patterns [2]. Aigner et al. see event-based visualizations as "promising means" for the design of applications to address user tasks. Calendar View visualizations use a temporal metaphor to encode time-dependent data and are "highly effective" for the analysis of clusters in combination with appropriate interaction techniques to detect correlations within the data [110].

There is data that contains both temporal and spatial characteristics. Thus, the simultaneous visualization of both dimensions is subject to many approaches regarding prediction data. In the domain of weather forecasting, Diehl et al. propose a visualization system "for the analysis of spatio-temporal patterns" [23]. They use multiple views, including a "timeline" that they linked with "geo-referenced maps". Hence, this framework can be used to recognize interesting patterns and weather trends by deploying effective visualizations. Similarly, to allow a visual analysis for temporal-geospatial data, Walsh et al. developed two new visualizations: The "Parallel Schedule View" shows the colocation of objects whereas the "Braille Plot" links information of a traditional 2D map with the "Parallel Schedule View". Consequently, each visualization focuses on either the time or space dimension and must be viewed simultaneously to understand all facets of the underlying data [111]. Unfortunately, the parallel operation of multiple visualization windows leads to an

increased cognitive load for users and should be avoided whenever possible [36, 16]. Since geographic data consists of two dimensions, and time can be seen as a linear, one-dimensional feature, the concurrent use of both components can be depicted in three dimensions. Based on these properties, Gatalsky et al. implemented the interactive visualization named "Space-Time Cube". Often it is referred to by the term 'time geography' that has been suggested by Torsten Hägerstrand [38]. This depiction allows the analysis of time-sensitive movement but potentially lack from overlapping visual elements due to the three-dimensional characteristics. Many other approaches have to deal with the same overlapping problem when visualizing data in 3D [105, 5, 14]. Alternatively, quantifiable data can be visualized by using the size of geometry shapes to indicate the value of a specific feature, i.e., glyphs [34, 56]. Using a similar approach, Tominski et al. provide an event-based approach to visualize 3D information of periodic temporal data proposing three-dimensional "pencil and helix icons" accompanied with additional "tunnel views" to overcome the problem of hidden information that results from overlapping. The spiral of the helix icon is an intuitive way of presenting the cyclic characteristic of temporal data [106]. Wickham et al. followed a glyph-maps approach to integrate the visualization of temporal patterns from climatic data immediately on a map. Thus, the display of glyph-plots on a map improves the understanding of the geo-temporal dependence. Another advantage of this approach is that the user only needs to interpret one visualization that contains all information that shall be conveyed [115]. These approaches serve as a starting position for our application by taking advantages into account while reducing disadvantages that exist in current concepts.

2.3 Spatio-Temporal Predictive Visual Analytics

The increasing progress in collecting vast amounts of data enables researchers to create models based on existing information to identify characteristic patterns within data to predict future states, trends, and outcomes. Compared to automatic analysis, predictive visual analytics applications do not only profit from large data foundations with different sources but can also be supervised by the user to focus on the relevant parts of a particular use-case. The integration of user-guided parameter steering into the predictive analytics pipeline significantly improves the informativeness of values presented in the visualizations. The use of appropriate interaction techniques allows to step into the analysis process and helps operators to explore the data and influence the analysis through effective and descriptive visualizations. Various research areas profit from prediction models and analytics methods to explain certain phenomena, e.g., sports [18], air traffic [13], firefighting [99]. The advance and refinement of interactive machine and classifier learning algorithms led to high accuracy levels of prediction results [58]. In the past, forecast models integrating classification, regression, clustering, or decision making have been found to achieve good results. To enhance traditional predictive analytics, including visual analytics components into this process reduces the complexity by replacing numeric-based models with adequate

representations of the underlying data. Every application domain has to deal with its challenges to help end users with decision-making and reasoning by enhanced perception and analysis performance. The area of analytics can be divided into three subcategories: descriptive, prescriptive and predictive analytics. Descriptive analytics includes the presentation of past and current states. Prescriptive analytics uses data to provide decision recommendations and to discover new opportunities. Eventually, predictive analytics combines descriptive and prescriptive models by providing knowledge about future trends that depend on decisions. Statistical techniques, clustering, and classification models are used to extract informative patterns and grouping within data.

Seebacher et al. present "Drosophigator", an interactive application integrating a visual analytics approach to investigate vast amounts of inhomogeneous data to provide better insights into the spreading behavior of invasive species. This approach processes heterogeneous spatio-temporal data to convey respective event predictions accompanied by presentations of temporal incidents and environmental attributes to support experts within their decision-making and examination processes. A glyph visualization is provided and evaluated to be comprehensible and helped experts to "interpret the results" and to "investigate hypothesis" due to the immediate linkage of geographic and temporal aspects [93, 115, 48].

2.4 Positioning of the Thesis

In this thesis, we introduce an application following a predictive visual analytics approach focusing on spatio-temporal data. Though spatial data indicates the use of heat maps, we follow different concepts to model useful visualizations. Amongst others, we present a glyph-visualization that takes time and space into consideration by providing environmental, meteorological data. This combination of multiple components is intended to be the primary contribution of this thesis. Due to the vast complexity of causes that crucially influence the occurrence of urban heat islands, we focus on the visual part of the analysis process. To retrieve more significant results, it is necessary to obtain complete real data such as light absorption levels of surfaces and three-dimensional models of buildings or streets parks. Because prediction visualizations of the UHI are independent of the data used, we declare that our approach can be extended by further data, once the number of potential data sources is increasing. We aim to fill this gap by linking temporal and geospatial characteristics through easy-to-understand depictions. Moreover, we implement interaction techniques to let the user steer relevant parameters and let him place the focus when predicting future outcomes.

Chapter 3

Data Foundation

Creating models that support the visual analysis of the urban heat island effect is crucial to select suitable data sets that can be analyzed and transformed into useful visualizations. In this chapter, we give an introduction of all data sets that serve as input to our visualization components. Moreover, data often is subject to preprocessing steps as cleaning and transformation to enhance the quality. In doing so, we aim to reduce potential misinterpretations of the consecutive visualizations to enable users to draw valid conclusions. Therefore, data processing is a central part of any analytics system dealing with large amounts of data. During this process, we follow the chronological order of the *Knowledge Discovery in Databases* (KDD) pipeline (see Figure 3.1) and describe every preparation step that is applied to the data to prepare the subsequent processing. To facilitate this operation, we used the platform KNIME of Berthold et al. [6] to create appropriate data processing workflows. Furthermore, KNIME provides the functionality needed to prepare and classify high-dimensional data, and to evaluate the significance of different classification models. We only focus on data from the year 2016 due to the vast amount of data provided by the WUSN for each year. We note that our approach can be applied to any dataset with desired time ranges when transformed as described in the following sections. Due to performance reasons of the classification step (detailed description

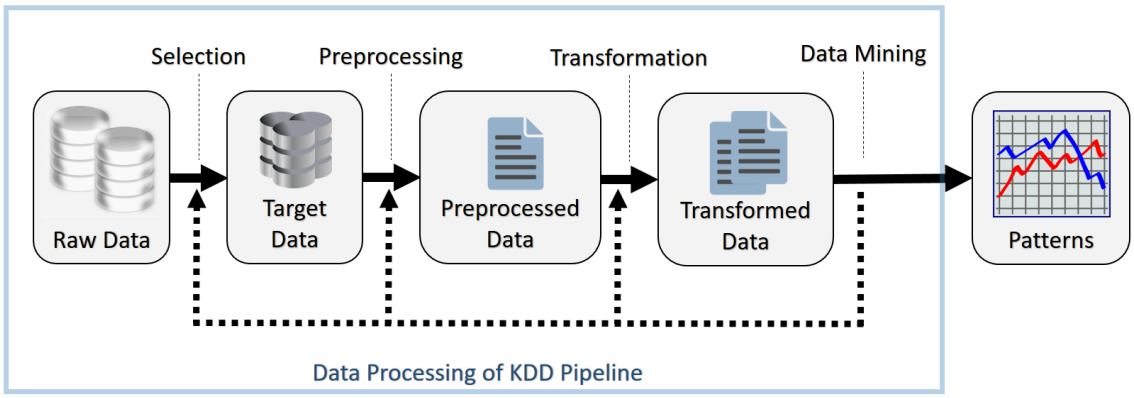


Figure 3.1: Knowledge discovery in databases (KDD) pipeline [26]. Data Processing as a central part of the KDD pipeline to prepare the data for visual analysis.

in Chapter 4), we also decided to focus on the city of Karlsruhe instead of analyzing all large cities in Germany. It is necessary to integrate geographic scalability into the provided system to broaden the scope of this geographic limitation. Despite this restriction, it is still possible to introduce our contribution and to present the interaction concept with our visualizations.

3.1 Data Description

For the implementation, our data foundation consists of three different sources (*Selection* in KDD Pipeline). As primary data, we use temperature data from the private Weather Underground Station Network (WUSN) [108] that has a better spatial distribution of stations than station networks from public authorities like the German Meteorological Service (GMS/DWD). The WUSN is a crowd-sourcing platform that provides meteorological data from its participating users to gather high-resolution data. Unfortunately, the WUSN dataset often contains invalid values that must be handled by cleaning, regression or deletion with KNIME. We found the meteorological variables provided by the GMS (DWD) to be useful in creating a feature vector containing heterogeneous data (see Table 3.5). We integrated a prediction model for future UHI development at given locations as a central component of our system. Therefore, it is essential to create a feature vector containing various variables capturing the complexity of the UHI phenomenon. This feature vector will be used as input to different machine learning algorithms to train a classifier that enables the identification of causes for the occurrence of UHI events. The DWD data underlies higher quality standards and is more reliable than the data from WUSN, which also provides meteorological records. Compared to weather, the surface characteristics can be directly influenced by humans. Therefore, we use land use data from OSM that provide high-resolution land surface information data [78]. The combination of these data sources offers an acceptable composition for a feature vector to deploy an extensive classification analysis.

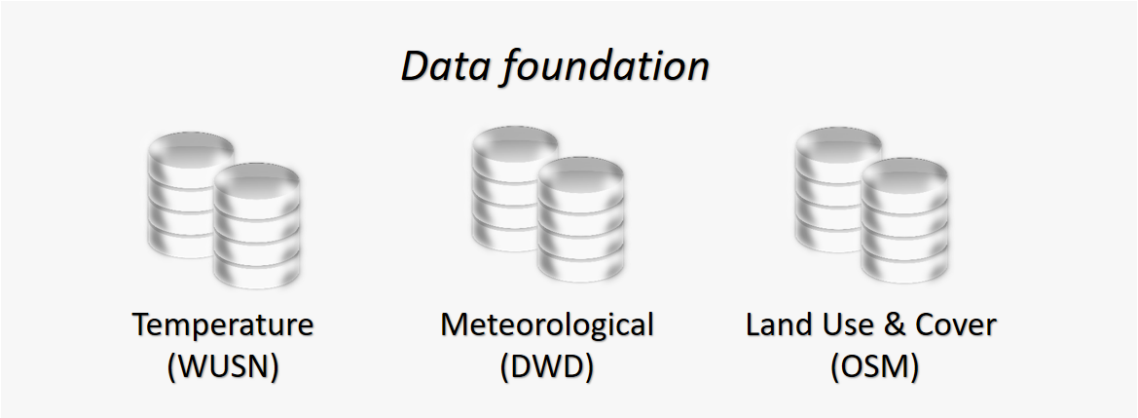


Figure 3.2: The data foundation used in our prototype consists of three different datasets: temperature data (WUSN), meteorological data (DWD) and land use/cover data (OSM).

3.1.1 Weather Underground Station Network

In Germany, the *Weather Underground Station Network* (WUSN) consists of 25298 stations (1832 stations within city borders) that provide multivariate data with a high temporal resolution of up to 2-3 measurements per hour (see Table 3.1). In the preprocessing step, we replaced missing and invalid (out of possible range) values for every station by using the average value of the previous and next recording to keep valuable information. We removed all meteorological attributes that are provided since temperature is sufficient to decide whether the location of a station is warmer than its surrounding. We decided to average the available temperature data hourly for each station which results in a dataset containing about 16,5 million data rows for the year of 2016.

Variable	**Description**
Station ID	Unique WU Station Identifier
Location	Geographic position of the station consisting of longitude and latitude
Time	Timestamp of the recording with an accuracy of one minute
Temperature	Provided in °C with an accuracy of one decimal place
Other	Dataset contains information as software type and other meteorological variables such as wind, precipitation, air pressure

Table 3.1: Description of the data provided by WU stations.

For the identification of local urban heat islands (hotspots), we considered the temporal and spatial attributes of the WU stations. First, for every station that lies inside the city borders, the closest (up to 10) neighbor stations have been identified using a maximum distance of 5 km. Then, a temperature value for the respective surrounding of each city WU station was calculated using a bilinear strategy considering the distance d as weighting factor for the temperature value of the respective by the formula

$$AST(t,n) := \frac{\sum_{i=0}^{n} \frac{d_{\max}}{d_i} \cdot T_i(t)}{\sum_{i=0}^{n} \frac{d_{\max}}{d_i}}.$$

Figure 3.3: WU station distribution in and around Karlsruhe.

The quantity AST is called bilinear Averaged Surrounding Temperature, where $T_i(t)$ is the temperature recorded by station i at time t, and n denotes the number of stations in the neighborhood. The maximum distance $d_{\max}$ of all considered neighbor stations is used to weight the respective temperature value. By this, we ensure that the stations which are closer have a stronger influence on the surrounding temperature than stations that are located farther away. Eventually, we can set the class (occurrence of the UHI effect) for each data record. Due to preliminary testing, we decided to consider the location of a station as a hotspot (local UHI) when the center temperature was at least 1 kelvin warmer than the bilinear averaged temperature of the surrounding. Hence, this transformation and aggregation lead to an extended description table of the WU dataset. Table 3.2 includes all relevant features of the preprocessed temperature data and is the starting point for a selection of several variables that will be extended by the other datasets.

Variable	**Description**
UHI class	**Serves as indicator whether this station is a hotspot (local UHI) at this location and point in time**
Station ID	Unique WU Station Identifier
Location	Geographic position of the station consisting of longitude and latitude
Time	Timestamp of the recording with an accuracy of one minute
Temperature	Provided in °C with an accuracy of one decimal place
Surrounding Temperature	**Bilinear interpolated temperature of WU neighbor stations using the distance to the central station as weighting factor**
Neighbor Stations	List containing Stations IDs up to 10 nearby stations

Table 3.2: Extended WUSN data description. The UHI Index indicates the occurrence of hotspots. Further information is given about the nearby WU stations that were considered to calculate the UHI index.

3.1.2 German Meteorological Service

The German Meteorological Service (GMS/DWD) [24] provides diverse and high-quality meteorological data that is crucial when analyzing the UHI phenomena. Therefore, we processed the DWD dataset as part of the feature vector. We are going to evaluate the finally composed dataset with KNIME's integrated classification functionality to analyze its capability to predict UHI development appropriately. The DWD measurements required less data cleaning steps compared to the WUSN records. Instead, the reduced spatial distribution (two DWD stations in the city of Karlsruhe but 28 WU stations in the same area) required a strategic mapping of the features to receive a reasonable combination of both sets. Only 1600 DWD stations exist in Germany. This not an issue regarding the meteorological aspects of a region, since parameters as wind, pressure, and temperature solely change gradually with spatial distribution. Table 3.3 explains all features that are provided within the historical data of the DWD station network. Some variables include different levels: For example, soil temperature (see Table 3.3) is measured at different depth values. To make it applicable to our input feature selection, we decided to use its mean to consider it in our application. In total, the DWD dataset contains about 1,6 million rows for the year of 2016. Since we only focus on the city of Karlsruhe, merely a fraction of this number is used to create our visualization application.

Variable	**Description**
Station ID	Unambiguous identification number
Measure Time	Timestamp of the recording with an accuracy of one minute
Location	Geographic position consisting of longitude and latitude
Air Temperature	Provided in °C with an accuracy of one decimal place
Relative Humidity	Percentage rate of the relative air humidity
Cloudiness	Coverage level (from 1 to 8); -1 if cloudiness could not be identified
Precipitation	Hourly precipitation height
Air Pressure	Mean sea level pressure & air pressure at station height
Soil Temperature	temperature measurements taken in °C at a depth of 2cm, 5cm, 10cm, 20cm, 50cm, and 100cm
Sunshine duration	hourly sunshine duration in minutes
Wind Direction	Given in degree from 0° to 360°; 0°→ north
Wind Speed	Strength of mean wind speed in $\frac{\mathrm{m}}{\mathrm{s}}$

Table 3.3: Parameter list of meteorological features provided by DWD network [24].

3.1.3 Open Street Map Data

The open street map (OSM) [78] data source provide high detailed polygon data including land use and cover information (geographic basis data). Since the surface characteristics are essential for the estimation of a UHI, we integrate information about the land usage of the neighborhood of each WU station that can be investigated in our application. First, we used the geographical information system QGIS [102] to analyze the polygon distribution in the surrounding of each station.

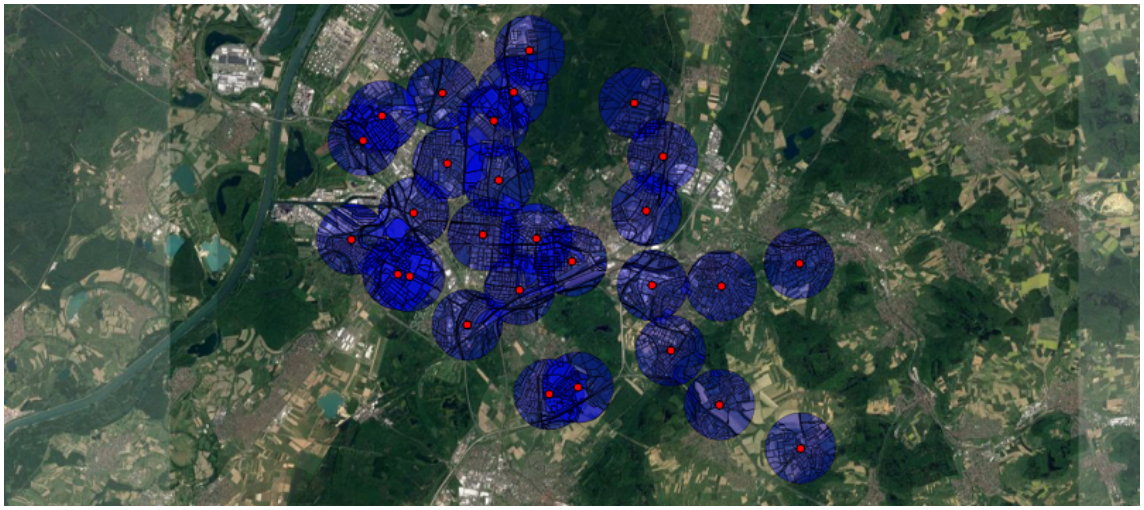


Figure 3.4: Red dots indicate WU stations within the city borders of Karlsruhe. The blue circles around the WU stations show the detailed distribution of land use and coverage data from the open street map (OSM) data set within a radius of 1km around each station.

Figure 3.4 shows the geometry information contained in the OSM data source. We consider the area with a radius of 1km around each WU station to be relevant for the UHI index. To create a meaningful input feature selection, we selected five surface categories to be of interest for the UHI scenario: water surfaces, green spaces, sand/stone area, residential/institutional buildings, and industrial territory (see Table 3.4). This dataset provides land use data that we aggregated into these five subcategories. It was necessary to map 38 different sub-features of these five categories to find a general classification and to proceed with the visualization models (detailed description of the OSM tag categorization is provided in Table B.1 in the appendix). Thus, we ensure improved interaction opportunities when working with this dataset. Unfortunately, due to a too high number, it is not possible to let the user decide which area features are relevant. Consequently, it was essential to simplify this dataset while preserving its general class information. The OSM dataset is the third pillar of the data foundation that has been used to model the UHI effect problem.

Variable	Description
Water Area	Water area surface coverage in %
Green Space	Greening area surface coverage in %
Sand/Stone Area	Sand/Stone area surface coverage in %
Residential or Institutional buildings	Residential and Institutional area surface coverage in %
Industrial Territory	Amount of industrial area in %

Table 3.4: OSM: variable selection and description. Area ratio (in %) result from WU station's surrounding (within a radius of 1km).

3.2 Machine Learning Classification Models

To identify locations, in our case within the city of Karlsruhe in 2016 where the UHI effect potentially appears, we train a model based on existing machine learning algorithms. For this step, we use a feature vector (described in Table 3.5) composed of the datasets collected from WUSN, GMS(DWD), and OSM. The classification training analyzes which composition of features influences the occurrence of the UHI effect at specific dates (date ranges). Before training our classification model, we need to create a definition of a UHI in the context of our work. As described in Section 3.1.1, we consider the temperature level of the immediate surrounding (1km radius) of a location to be a specific indication whether a UHI event is present or not. We follow a bilinear calculation approach that combines the temperature values of different stations using the inverse distance of the neighbor stations as weights. Based on this weighted averaging strategy, we introduce the following definition of a UHI event: A UHI event is present if a location is warmer by 1 kelvin than its surrounding.

3.2.1 Random Forest

For our application, we chose Breiman and Cutler's *Random Forest* (RF) algorithm which is an ensemble classifier and "an effective tool in prediction" [11]. This machine learning algorithm is based on an arbitrary number of tree-structured classifiers (ensemble of fully grown decision trees) that takes the mode of the result of all classes the individual trees return. The RF algorithm randomly (controlled variation) selects a subset (called "the random subspace" [11]) of the available variables from the feature vector to create a collection of decision trees. This strategy produces a highly accurate classifier for our dataset. Besides this fact, RF has the advantage that missing values do not affect the result significantly, and the accuracy is maintained, even if many data records are not available. Another advantage is that random forests are robust against overfitting if a large dataset is used during the classifier model training process. To minimize the bias of the classification result, RF uses unpruned trees to retrieve a minimal bias initially. The trees and are independently trained (not correlated with the other decision trees) to reduce the variance.

3.2.2 Stacked Generalization

As an alternative, we compared the RF algorithm with the Stacked Generalization (SG) algorithm, proposed by Wolpert [116]. One advantage of this ensemble learning technique, also called *Stacking*, lies in the combination of the predictions of several different learning algorithms. The resulting prediction is more accurate than any single outcome of the underlying models while variance and bias are decreasing. Unfortunately, the execution time increases drastically, compared with the RF algorithm. Even though we observed that the SG outperforms RF, the resulting predictions of RF are still very useful. Consequently, we decided to make the compromise by applying the RF algorithm for our prediction model instead of SG. In Chapter 5, we dive into the details of the evaluation of the machine learning algorithms that we compared against each other and provide exact details of the precision of each model.

Variable	**Dataset**
UHI Index	WUSN
WU Station ID	WUSN
Year	WUSN
Month	WUSN
Day	WUSN
Hour	WUSN
Temperature (Min, Max, Avg)	WUSN
Absolute Temperatures of 1-12 hours before	WUSN
Delta Temperatures of 1-12 hours before	WUSN
Relative Humidity	DWD
Cloud Coverage	DWD
Precipitation Height	DWD
Sunshine Duration	DWD
Wind Speed	DWD
Wind Direction	DWD
Mean Sea Pressure	DWD
Station Pressure	DWD
Soil Temperature	DWD
Water Surface Area	OSM
Green Spaces Area	OSM
Sand/Stone Area	OSM
Residential/Institutional buildings area	OSM
Industrial Territory	OSM

Table 3.5: Final feature vector used to create the RF prediction model with WEKA.

Chapter 4

Visual Analytics of UHI

Our application system is intended to support expert analysts with sense-making visualizations to interactively explore data that is related to the occurrence of the UHI effect. Primarily, our system provides an intuitive way to efficiently extract information and to set up hypotheses when working with data that can individually be transformed and filtered in multiple ways. We integrate Card et al.'s Information Visualization Reference Model [15] (see Figure 4.1) as a standard to realize our application by applying its concepts. Visualizations are mappings of data to a visual form that supports human interaction in a workspace to generate new knowledge about the data that is visualized. We use the preprocessed data (see Chapter 3) as raw data input to realize and evaluate our application providing multiple visualizations that are linked. All data is integrated to adequately convey helpful information and complete each other by covering different aspects of every scenario. Consequently, extracting semantic information from a data set containing heterogeneous attributes can be a difficult task. Therefore, our contribution consists of a thought-out combination and composition of several visualizations resulting in a visual analytics workspace. Users

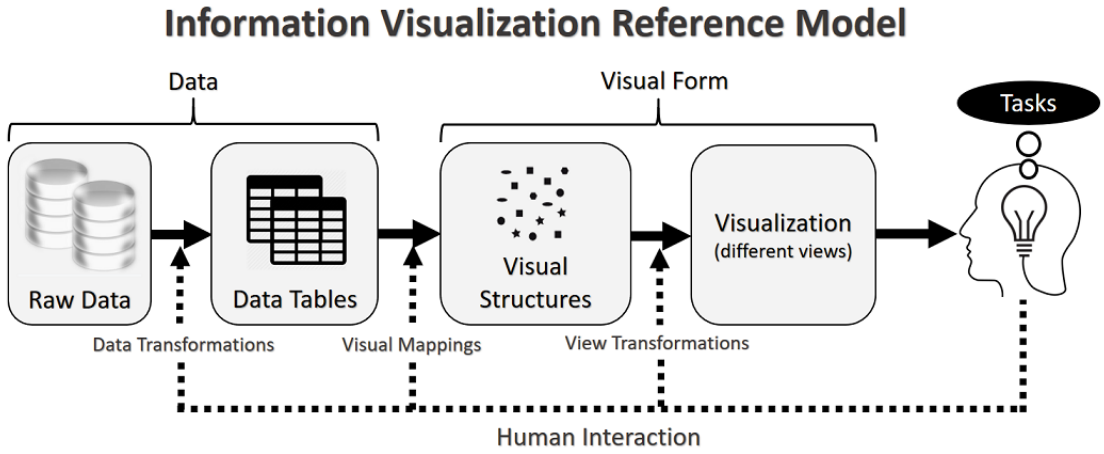


Figure 4.1: Reference model for information visualization [15]. Human interaction as an integral part of a visual system that maps data structures to visualizations while providing different views.

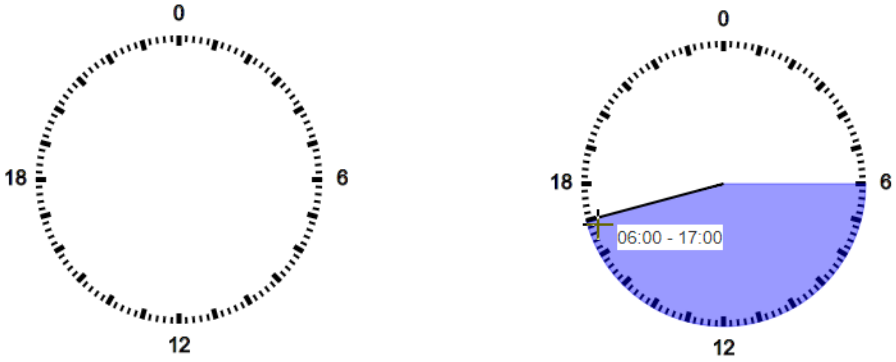
can steer different parameters and transform the views as necessary for a specific use case. Using different views enhances the experience when working with our UHI analysis application. This is crucial due to the impossibility of presenting all potentially relevant information in a single global view. Appropriate techniques are used to retrieve both overview and detail information about the dataset to improve the users understanding of the feature composition. We follow Keim et al.'s *Visual Analytics Mantra*: "Analyze First, Show the important, Zoom, Filter and Analyze Further, Details on Demand" that was proposed to create effective visual analytics information systems. It is a profound way to define how information seeking tasks can be supported to provide meaningful data analytics applications by integrating "automatic analysis methods before and after the interactive visual representation is used". In this chapter, we introduce all components that were implemented to enable exploration and to allow users to execute prediction tasks. Our prototype was designed to address three major aspects: geospatial, temporal and high-dimensional data. Each part has its challenges. Moreover, we implemented a concept to combine time and space into a single representation directly. Therefore, we provide different elements that will be described in detail in the subsequent sections. We distinguish between components that can only be used to adjust different views, and other components simultaneously conveying important information and serving as input selectors.
The consistent employment of visual variables such as color and shape plays a fundamental role. While color is well-suited for the visual encoding of nominal groups, size is more accurate for visual variables to provide efficient comparison opportunities [32]. The design of each visualization should be easy to understand without steep learning curves. The same color is consistently used throughout all visualizations if the same data dimension is presented. Additionally, we take semantic color-mapping into account to link familiar concepts with corresponding color schemes (water is represented in blue, and green area is displayed by green color, etc.). By this, the user can link similar concepts which are supported by appropriate filters and highlight interaction techniques such as linking and brushing [50]. This improves the combined use of multiple components supporting visual data analysis tasks.

4.1 Temporal Components

The visual analysis of UHI requires the reasonable representation of time. In the realm of UHI analysis, the temporal dimension plays a substantial role that must be addressed accordingly. Consequently, we integrated two components to enable filtering for both hourly intervals and the selection of arbitrary date ranges. We implemented the *Time Range Selector* as a filter for hourly intervals and complement it with a *Calendar Overview* visualization that serves as a date selection panel. In combination, all relevant temporal dimensions can be individually adjusted. Both components are referred to as *Input Selectors* since the selection changes the other visualizations (linking).

4.1.1 Time Range Selection

The **time range selector** (TRS) is a temporal component that functions as a filter for hourly intervals. This input filter can be applied at any time during the visual analysis. This component is based on Häussler's jQueryClockIntervalPicker plugin [41] that utilizes the radial clock metaphor as a basis concept (see Figure 4.2). In our application, we only use the 24h layout, because for UHI analysis tasks it is not crucial to specifically focus on forenoon and afternoon separately. We integrated this component, since using the clock metaphor is an intuitive way to apply temporal adjustments on other visualizations.



(a) Default presentation of the *Time Range Selector*: no interval selected.

(b) Hourly interval filter selection example: 6 a.m. to 5 p.m.

Figure 4.2: jQueryClockIntervalPicker: time range selector [41].

4.1.2 Calendar Visualization and Date Selection Panel

The second temporal component is the **Calendar View Visualization** that is based on Bostock's *Calendar View* [9] implemented using the D^3 library [8]. This visualization provides an overview of the current year, 2016 in our case. Of course, this can be extended to any year range desired. Figure 4.4 depicts an overview of the UHI dataset of 2016. Each square represents a day with a month-label that is located above. The visualization is designed to give an overview of the number of hotspots of each day (blue outer part) as well as the average temperature difference to the surrounding of each WU stations (yellow-red inner part). Mouseover interactions are available to get immediate feedback of the underlying data: In Figure 4.4, April 21 in

Figure 4.3: Calendar view: the inner part of the square measures the **average UHI hotspot intensity** ($> 50\%$) of all stations that have been identified as hotspots on a respective day using a continuous yellow-red color scale. The **number of identified hotspots** is given by a continuous white-blue color scale. In combination, both parameters convey UHI information of the hotspot intensity.

2016 is selected. On this day, seventeen hotspots were identified with an average UHI intensity of 67 % and an absolute temperature of 15.21°C of all hotspot stations. A station is considered as a hotspot, if the UHI intensity for the station is at least 50 % on the respective day. We used two continuous color schemes to indicate the level of each parameter. They can be used to identify particular dates by the opacity of each color (see Figure 4.3). This visualization is intended to give overview information of the temporal UHI situation of the city that is subject to analysis instead of conveying exact numeral information of each hotspot. Hence, dates (ranges) of interest can be visually determined. This component does not only provide an overview but can be used as **Date Selector** to retrieve deeper insight of the visualized dataset.

In Figure 4.5, an example date selection (from 05/01/2016 to 08/31/2016) is presented. The other visualization components of the application are temporally filtered and only display data of this date range. Therefore, two different concepts are encoded within a single component: overview information can be extracted, and a temporal exploration is facilitated by intuitive date filter functionality.

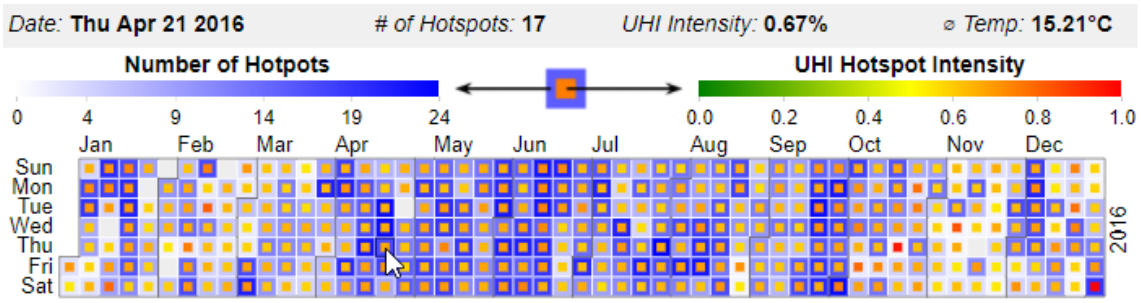


Figure 4.4: The calendar overview visualization uses a weekly structure to convey a temporal overview that is easy to understand. Each day is represented by square that simultaneously provides both information about the UHI intensity and the number of hotpots.

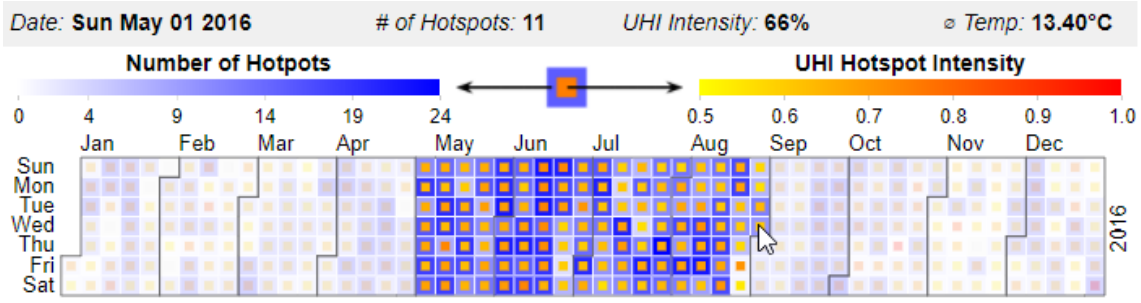
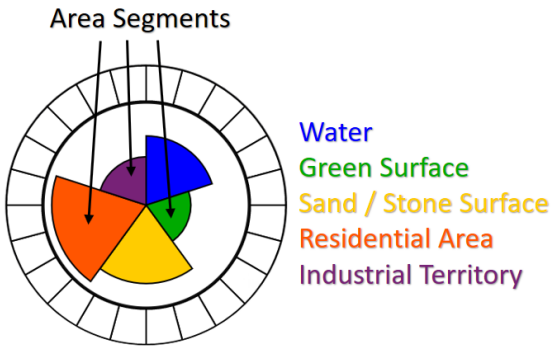


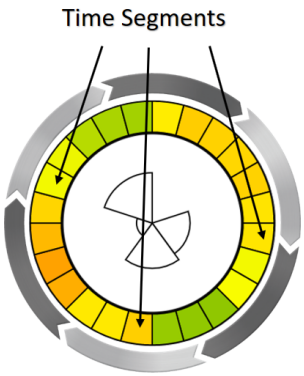
Figure 4.5: The calendar view component can be used to select particular date ranges. In this example, the date range *from 05/01/2016 to 08/31/2016* is selected. Other visualizations are temporally filtered by this component.

4.2 Spatio-Temporal: Pie Chart Glyph

Besides the temporal dimension, spatial characteristics are necessary to estimate the impact of the geographic location of a station on the outcome. For this purpose, we conceptualized a **two-level pie chart glyph** (PCG) that includes both spatial and temporal information. To directly link these dimensions, we put them into a joint context. This pie chart glyph is radially arranged and divided into an inner and outer group. Each glyph represents specific attributes of one WU station. Figure 4.6 provides a graphical explanation of both groups. We consider the area located around a WU station within a distance (within the radius) of 1 km. Therefore, the maximum size of any area is 3.14km^2. The inner part of the pie diagram represents the size of each area type within this 1km radius. In our application, we summarized all area types into five different area categories: water, greening surfaces, industrial territory, residential areas and dry (sand and stone) surfaces (see Figure 4.6 (a)). The angles of the inner group are fixed, whereas the percentage value of an area type of the station's surrounding is only encoded in the radius of the pie element. This facilitates the comparison of the area proportion between different stations. Hence, geographical surface information of the stations' surroundings can immediately be analyzed. If desired, the segment number can be extended to any arbitrary number of different area types and solely serves as a specific example of how this glyph can be used. The outer group visualizes the number of UHI events based on the selected temporal aggregation method and represents the temporal progress of the UHI intensity of a particular station of interest. The number of segments depends on the current temporal aggregation selection. If *hourly* is selected, the outer group consists of 24



(a) The inner group visualizes the geographic section of the glyph representation. Every segment indicates the amount of a certain area, that is located within a radius of 1km around a WU station. The colored text is the legend for five different area types.



(b) The outer group visualizes the temporal progress of the UHI index using a clockwise ordering. Every segment indicates the UHI intensity of the respective temporal unit (hour, day, month, date range). Here, 24 hours are displayed: one segment per hour.

Figure 4.6: A visual explanation of the pie chart glyph visualization.

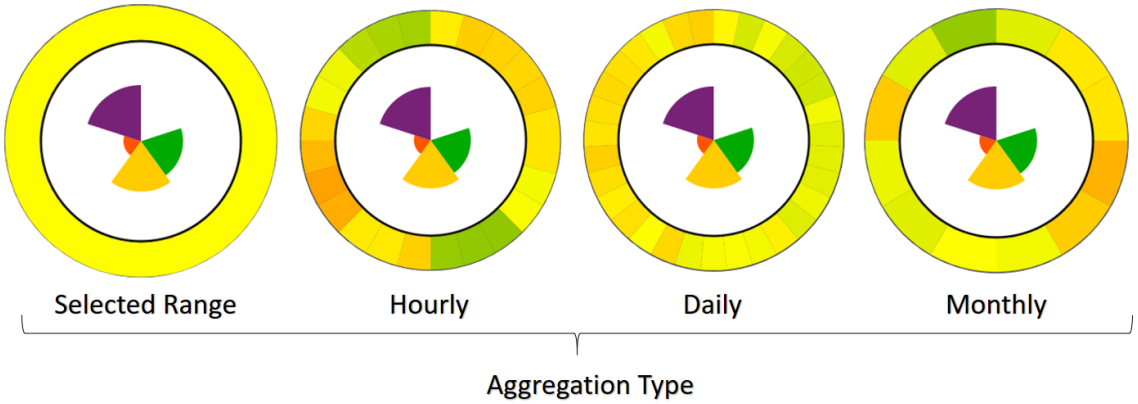


Figure 4.7: Four different temporal aggregation types of the pie diagram glyph. The outer group can represent different time units using the respective aggregation method (from left to right): total selected date range, hourly, daily, or monthly.

segments. Analogously, 31 segments are depicted if the *daily* aggregation is selected. The aggregation type can be individually adjusted: the user can select between hourly, daily or monthly aggregation. Moreover, an average aggregation of the total selected date and time range can be chosen, if the user is not interested in the temporal progress of the UHI intensity level but only wants an overall value. In this case, the outer group becomes a single-colored ring. A continuous green-yellow-red color scale is used to indicate the UHI intensity ratio. Thereby green color represents the absence of UHI incidences, whereas red color encodes the occurrence frequency of UHI hotspots. We emphasize that the color does not convey any information about the temperature difference between viewed and neighbor stations.

In Figure 4.7, four different aggregation types are presented. The temporal glyph part (outer group) contains 24 segments for hourly aggregation, 31 segments using daily aggregation and 12 segments if the data is monthly aggregated. The average UHI hotspot occurrence value is linearly mapped to the continuous green-yellow-red scale. Depending on the current analysis task, the user can identify the progress of the UHI index over time for each station separately. This can be helpful to recognize hidden UHI intensity patterns that may exist for a station. The UHI intensity represents the average of all UHI hotspot occurrences. Red color indicates high UHI intensity, and green colors the opposite.

4.3 Multidimensional: Parallel Coordinates Plot

Besides temporal and spatial data, the dataset contains other features (see Table 3.5) that are covered by neither the temporal components nor the PCG. These features consist of meteorological variables (e.g., wind, cloud level, air pressure) and are a central part of UHI analysis tasks. We implemented a **parallel coordinates plot** (PCP) [10] that has been proposed by Inselberg and Dimsdale [47] to visualize multidimensional datasets. To put these features into a geo-temporal context, we added the spatial and temporal variables into the PCP (see Figure 4.8). Although this

leads to a redundant representation of the dataset, it facilitates the understanding of the relationship between the depicted features. Each data record is represented by a line that crosses each dimension scale at the value that the record has for the respective dimension. The PCP has been used to visually analyze correlations of multidimensional features [114]. The color of each line indicates whether a UHI exists (red) or if the station is not of particular interest for the UHI analysis (green) since it has not been identified as a hotspot location for this data record. Information about the correlation between the different dimensions can be retrieved without the need of horizontal reordering of the vertical axes (feature dimensions). If the user is interested in correlations of different features, it is possible to reorder all dimensions concerning the user's preferences individually.

The interaction technique *Brushing* was implemented on each dimension of the PCP similar to Hauser et al.s brushing technique for Extended Parallel Coordinates [40]. Any arbitrary range of every dimension can be selected to set an explicit focus on a subset of the presented data records during information visualization. Figure 4.9 pictures two examples of this technique that has been applied to the *Class* dimension that represent whether a record has been identified as a hotspot (image above with red lines) or not (image below with green polylines). Moreover, depending on the number of displayed lines, an opacity strategy was implemented: The opacity value increases with a higher number of displayed lines. In this way, PCP sectors can be visually identified where many lines converge, since the opacity of overlaying lines is summed up (e.g., the *precipitation* dimension).

The default ordering of the PCP starts with the *WU station ID* and its respective hotspot indicator (*class*) (light gray) followed by the temporal characteristics (hour, day and month in dark gray). The meteorological variable dimensions are colored in light blue and contain the highest number dimensions (10) presented in the PCP. The spatial characteristics of the stations are depicted as separate dimensions on the right-hand side each with the respective color that is used in the PCG to improve visual grouping. To equip the PCP with meaningful temporal analysis functional, we added seven temporal aggregation methods which are similar to the temporal grouping that has been introduced with the PCG. These methods are divided into *unary*, *binary*, and *total selected range* aggregation. Detailed examples of these aggregation types are provided in the appendix (see Figures A.1 - A.3 in the appendix)

The aggregation methods provide the opportunity to hide a specific temporal dimension while preserving other characteristic temporal aspects. As a consequence, users can detect particular patterns in each time dimension (hour, day, month). In the same window of the PCP, a simple info panel (see Figure 4.10) provides information about the currently selected date range, the total number of data records shown in the PCP, and the number of records that apply to the individual dimension filters. Using the aggregation type dropdown, the user can execute desired aggregation methods to the PCP. The data records that are visualized within the Parallel Coordinate Plot are a small sample (approximately 1000 data records selected using linear sampling) of the total number of data records that are aggregated and visualized by the other visualizations (206589 data records in total for Karlsruhe in 2016). Consequently,

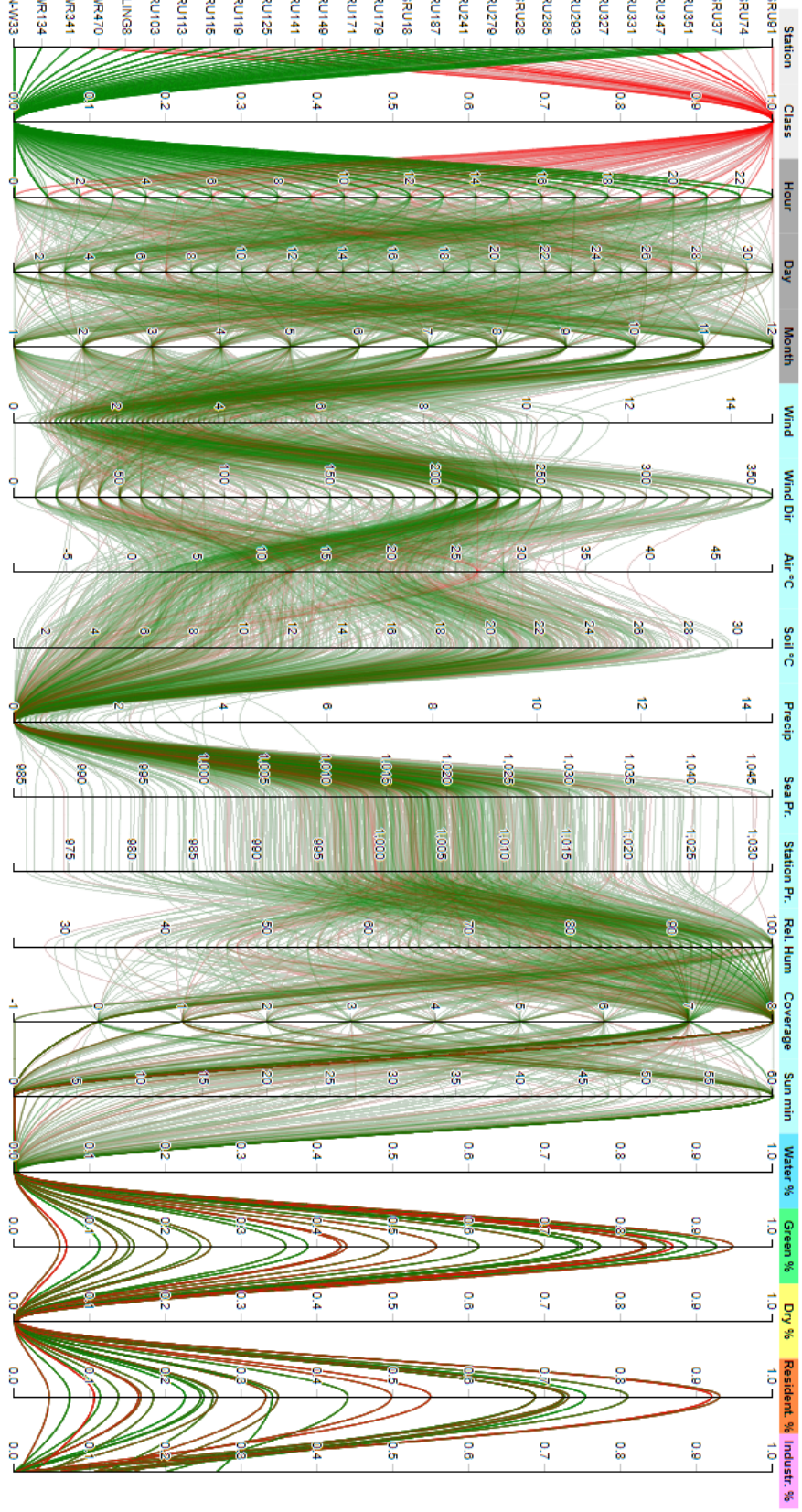


Figure 4.8: Parallel coordinates plot: WU station information, time and date section (dark gray), meteorological (light blue) and area (color consistent with PCG) section. All sections combined consist of 20 different dimensions.

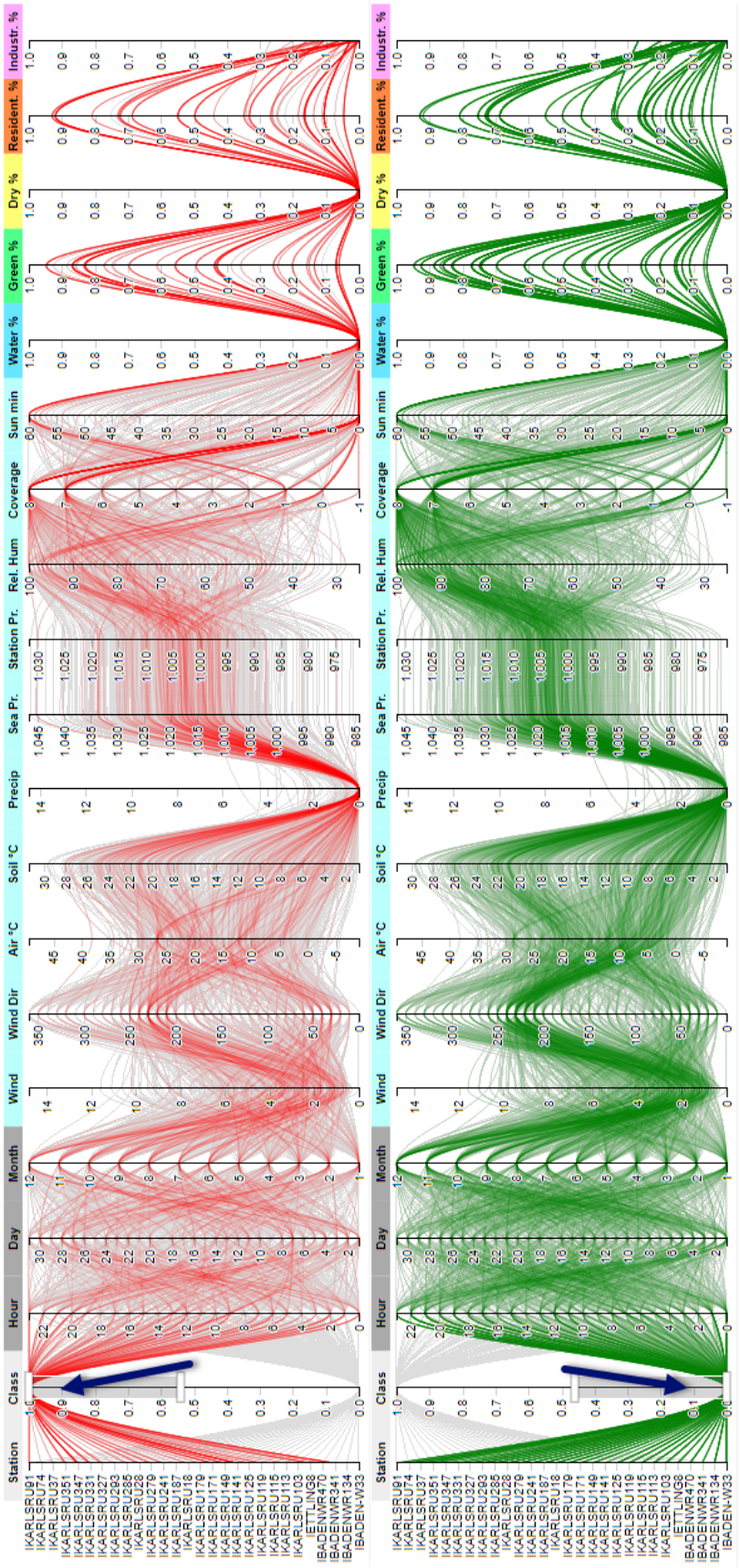


Figure 4.9: Brushing is a visual filter technique to focus only on a desired subset of the data [40]. These are two examples that show how the filter is applied to the *class* dimension (see blue arrows). Data records which do not apply to the filter are grayed out.

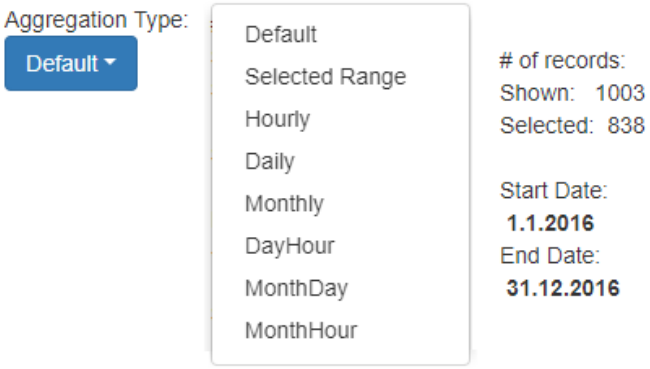


Figure 4.10: Information panel of the status of the visualizations. Within the same window of the PCP, information is given about the current number of selected (filtered) data records of the PCP. Additionally, the currently selected date range is stated. An aggregation dropdown list can be used to change the aggregation method that leads to different PCP representations (see Figures A.1 - A.3 in the appendix).

the PCP does not provide a complete overview of the dataset but rather highlights interesting patterns that may exist within the data. Nevertheless, this restriction does not prevent users from retrieving deeper insights when executing UHI analysis tasks. Figure 4.11 displays a PCP using the unary aggregation type *monthly*. Using the brushing filter on the sunshine duration dimension shows that longer sunshine duration times correlates with higher UHI intensity. Moreover, this example displays that increased sunshine durations happen more often during the summer time which indicates that direct sun radiation is responsible for the appearance of UHI hotspots. Consequently, the extensive functionality makes the PCP a central part of our application for the analysis of characteristic feature patterns.

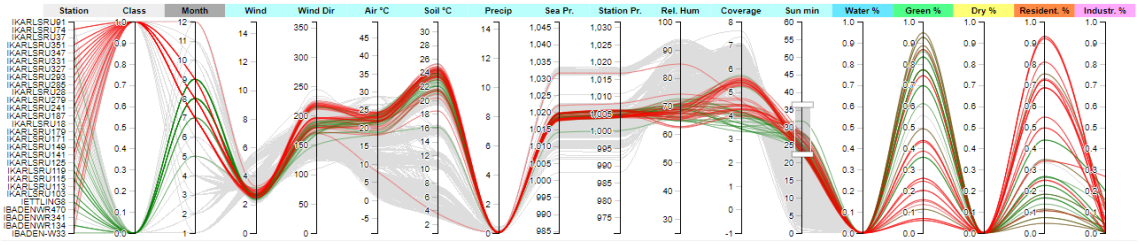


Figure 4.11: Unary PCP aggregation example: monthly. The distribution of the UHI intensity for each variable can be identified without the need of further interaction.

4.4 Prediction Input Feature Selection Panel

The last visual component of our application is essential for the execution of every prediction that can be processed by the application. The user can define several parameters by adjusting parameter sliders of meteorological and spatial filters. Temporal changes are not necessary since all visualization provide effective temporal exploration techniques. We included these temporal characteristics in the forecast process. Figure 4.12 displays the default layout of the **input feature prediction**

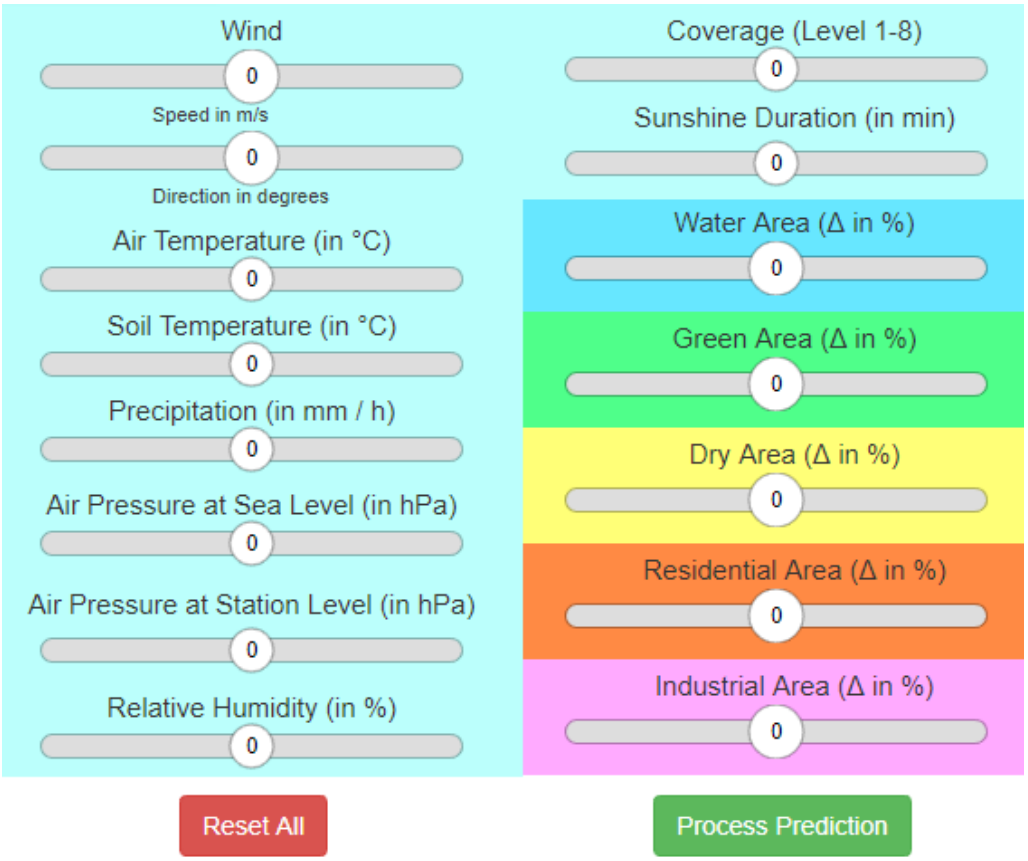


Figure 4.12: IFPP: Steer multiple variables for prediction processing. The IFPP consists of two main sections: Meteorological (light blue: from *wind* to *sun minutes*) variables and geographic (multicolored; *water area* to *industrial area*) features.

panel (IFPP). The IFPP is divided into two parameter sections. The feature sliders with the light blue background adjust the input values of meteorological variables whereas the remaining five sliders adjust the area value of the area types that have been introduced with the PCG and the PCP visualizations. Since the surrounding area can not surpass 100%, the area sliders are adjusted to its relative amount to preserve that the sum of all areas complies with the exact amount of area that has to be in the surrounding of each location. Therefore, the percentage assignment is changed of all sliders (including the slider that has been manually changed). By this, the user is not responsible for making sure that the area values correspond to values that can exist in reality. Hence, the result is automatically better than if the user would need to ensure that the area selection is valid. As before, the color selection is also consistent for the area types in all visualizations throughout the entire application. By adjusting the handle of each parameter slider a *delta vector* is created that will be applied to *all data records* before prediction execution. Hence, the *adjustment delta vector* provides *relative value* changes for every dimension instead of setting *absolute* values for all records. This strategy enables the analysis of changes of environmental variables and provides reasonable forecasting methods. Individual changes can be applied, and an investigation can be executed to examine the impact

of the single features as well as any desired combination of all specified prediction parameters. After specifying a parameter selection of interest (an example is provided in Figure 4.13), by operating the process button, the prediction is executed, and after a short period, all visualizations are updated to present the forecast data. In advance, we calculated all possible minimum and maximum values for all dimensions to prevent impossible feature configurations. If a minimum or maximum threshold is exceeded, the threshold value of the respective dimension is assigned instead of the value that results from the feature delta vector. Therefore, the user can trust the result to be in the realm of real data except that certain combinations may be possible, that would not occur in reality: e.g., 60 minutes sunshine and a high amount of precipitation. We emphasize that the feature vector is applied to every presented station in the same way and cannot be individually adjusted. Nonetheless, by using this technique, users can put forward hypotheses about the occurrence of UHI hotspots.

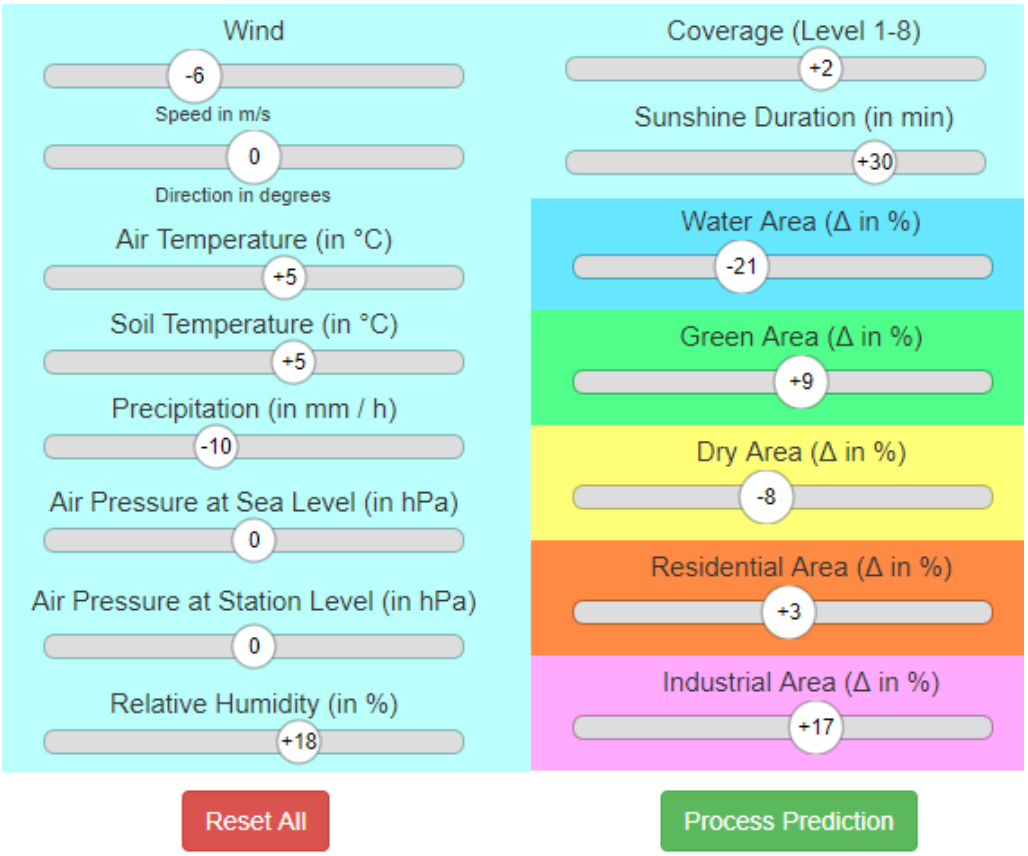


Figure 4.13: The following feature changes are applied by this selection: wind speed ($-6\frac{\text{m}}{\text{s}}$), air and soil temperature (+5 °C), precipitation (-10$\frac{\text{mm}}{\text{h}}$), relative humidity (+18 %), cloud coverage (+2), sun level (+30 $\frac{\text{min}}{\text{h}}$); water area (-21 %), green area (+9 %), dry area (-8 %), residential area (+3 %), and industrial area (+17 %).

4.5 Adaptive Visual Workspace

To combine the visualizations that we introduced in this chapter, we created a user interface that integrates and links different concepts. Figure 4.14 portrays the default arrangement of the components described in the preceding sections in a graphical user interface. Based on restful web services, we developed a system through a server-client structure that lets users directly access our system via standard web browsers without the need for a particular setup. Due to the employment of standard web development technologies as HTML, CSS, and JS, we were able to create our visual analytics prototype with the help of multiple powerful JS libraries. Besides Leaflet [1] that we use for the map view, and d3.js [8], which provides powerful data-driven based functionality that we employed for both the calendar view and the parallel coordinate plot, we determined GoldenLayout [20] to be a suitable library for the layout of our visual analytics framework. GoldenLayout is a reasonable solution that allows a flexible presentation of multiple windows. This multi-screen organization library enables users to individually resize and rearrange the single views concerning the particular requirements of a use case or the users' preferences.

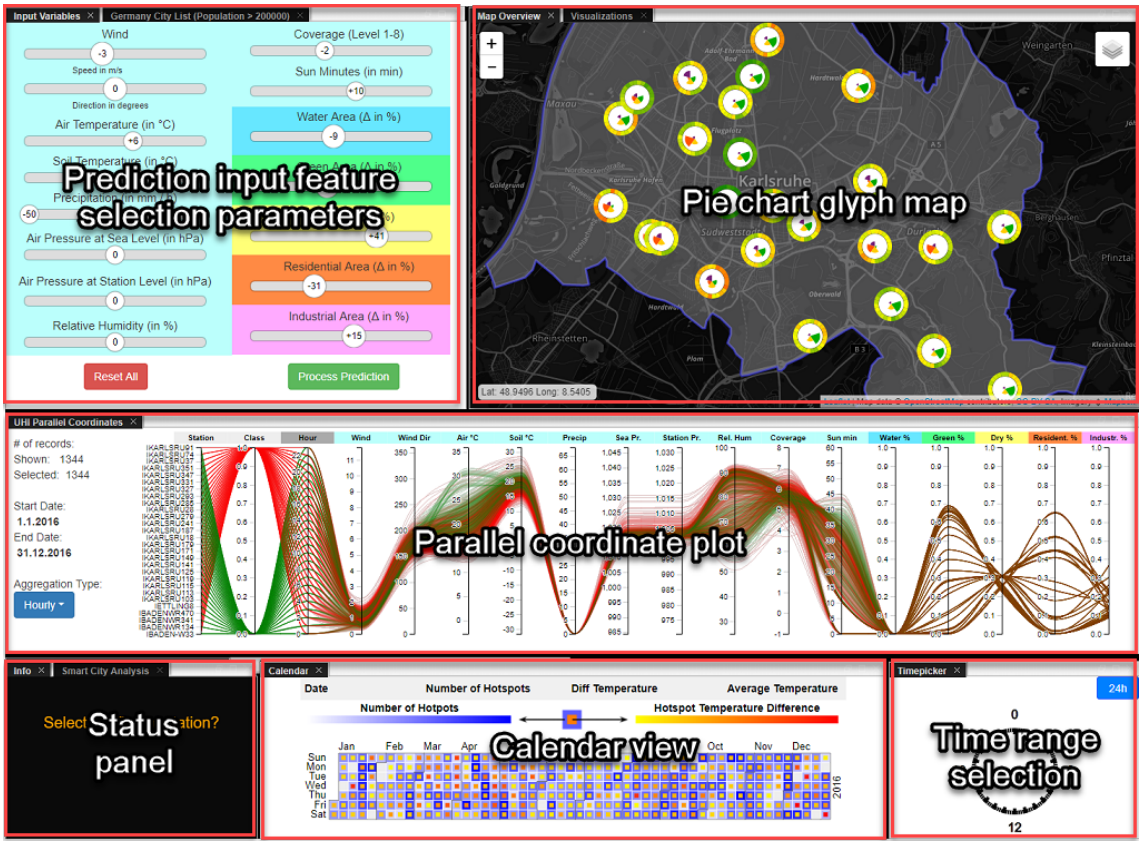


Figure 4.14: The proposed visual analytics framework consists of six different components (red highlighting) in total. The composition of the visualizations enable an effective UHI analysis by providing heterogeneous information about multiple features through the deployment of different views. In the bottom left corner, the *status panel* conveys helpful information about current system states, in contrast to the other components.

For instance, if a user-task is data-centered, the PCP can be enlarged to facilitate users to identify differences in the high-dimensional visualization more easily (see Figure 4.15). If geographic characteristics are dominating, the user's attention can be drawn to the map visualization by resizing affected layout components accordingly. Consequently, the adaptive user workspace supports task-oriented visual analysis processes, which serves as a viable foundation for future adjustments, since new visualization components can be added, if needed.

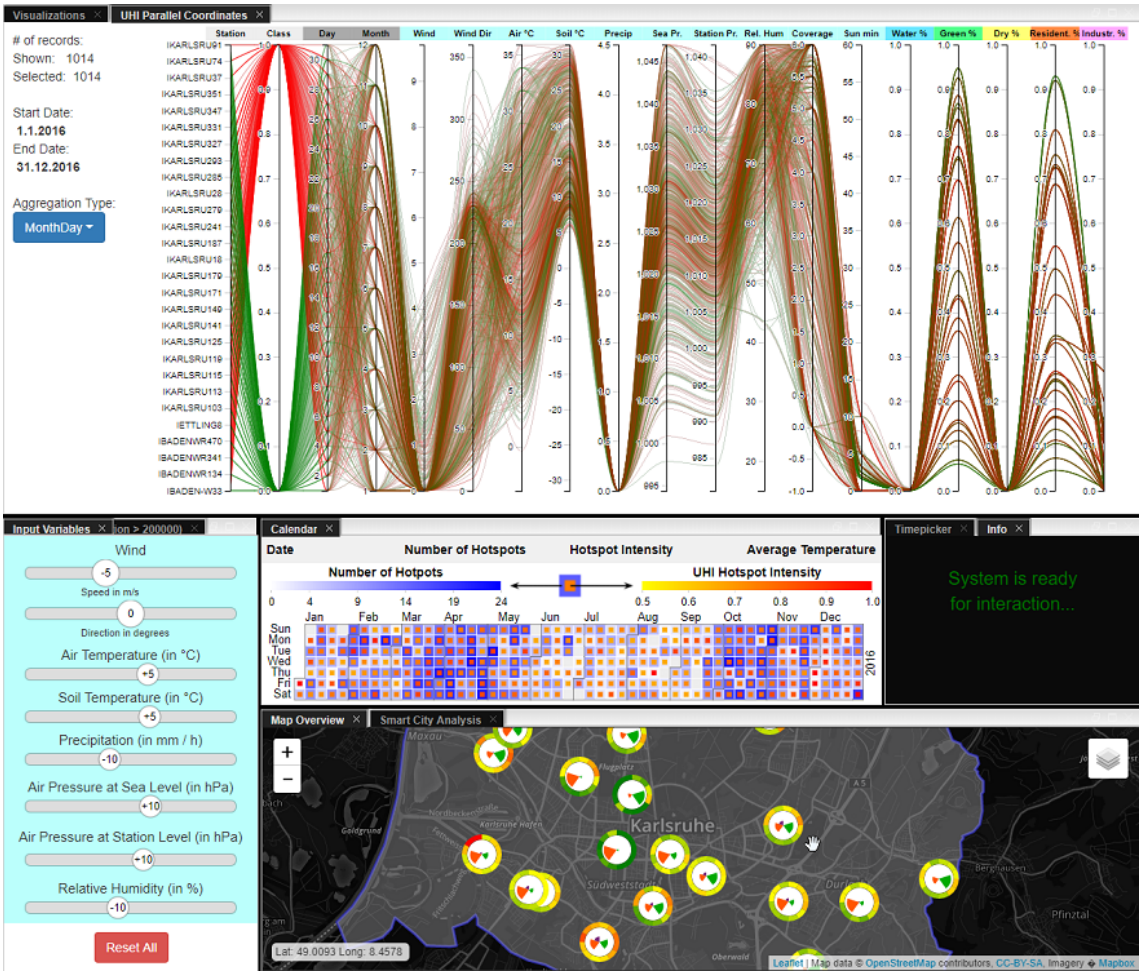


Figure 4.15: All components can arbitrarily be rearranged and resized by the user providing task-oriented flexibility. Thus, the visual framework facilitates users to focus on the most relevant components when examining particular use cases.

Chapter 5

Evaluation & Use Cases

To create a useful model for the prediction of the UHI effect, we need an expressive data foundation to perform classification methods that enable this prediction. Therefore, in Section 3.1, we provided a detailed description of all data components that we used to evaluate the informativeness of existing correlations within the data. In this chapter, we give a detailed argumentation why we choose *random forest classification* for our application instead of other classification methods. Subsequently, we present three use case scenarios to illustrate the usefulness of our application and how information can be extracted.

5.1 Classification Learning Evaluation

To find a classifier method that fits our visualization system, we investigated and compared different learning models. We considered two specific performance attributes to be crucial for the selection of the classification model: accuracy and speed of the classification training and prediction execution, respectively. Before evaluation, we removed the variable *WU Station ID* from the training set to avoid undesired overfitting that otherwise may have been introduced into our prediction results. For the evaluation process, we selected KNIME [6] to execute and compare several classification methods. The complete dataset that is the starting point of the assessment contains 16 507 022 rows. To improve the performance, we reduced this number to only 200 000 data entries by using random downsampling including 100 000 records of each UHI prediction class. We randomly divided this sample into a training set containing 80 % of all entries and used the remaining 20 % entries (display in Figure 5.1) to examine the accuracy of single classifiers. Moreover, we used WEKA nodes [45] of KNIME for the classifier comparison, to exploit WEKA's effective data mining methods for data analysis.

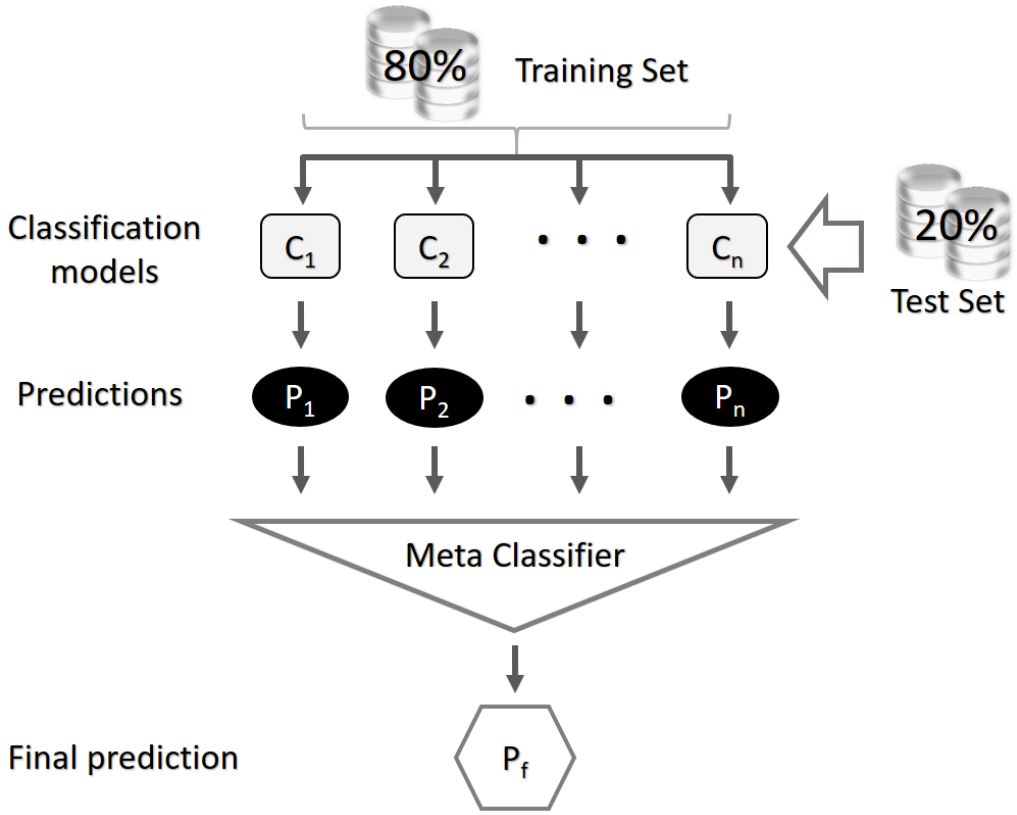


Figure 5.1: Concept of a stacking classifier [84].

Stacking Classification is an ensemble-learning meta-classifier that combines the predictions of an arbitrary number of different classification models [116]. This strategy aims to improve the prediction result compared to the performance of each model. Figure 5.1 shows the concept of a stacking classifier and how it is configured. This strategy joins n classification models by individual training and verification. The single predictions are forwarded to the second level (meta-classifier) that is based on the desired regression model. Eventually, the meta-classifier node calculates a final prediction with higher accuracy. For this, the logistic regression node of WEKA 3.7 *(multinomial logistic regression)* is used for the combination of the first level classifiers. We selected seven different first-level classification models to test them against each other: *Random Forest* (RF), *Decision Tree* (J48), *Multilayer Perceptron* (MLP), *k-Nearest-Neighbor* (kNN with $k = \{1, 3, 5\}$), and *Naive Bayes*. First, these classifiers are trained individually (as indicated in Figure 5.1): the Decision Tree model (C4.5 standard [82]) is configured to be pruned with pruning confidence of 0.25 and at least two instances per leaf. The k-Nearest-Neighbor classifications were conducted with KNIME's *IBk (3.7)* node by a linear nearest neighbor search algorithm of WEKA. The only parameter that we changed was the number of neighbors k. MLP is an artificial neural network classifier that utilizes backpropagation during the training process [43].

Classification Model	Accuracy	Cohen's κ
Stacking Classifier (SC)	96.3 %	.926
Random Forest (RF)	82.3 %	.647
Decision Tree (J48)	79.2 %	.584
Multilayer Perceptron (MLP)	75.9 %	.519
3-Nearest-Neighbor (3NN)	71.7 %	.433
1-Nearest-Neighbor (1NN)	71.5 %	.43
5-Nearest-Neighbor (5NN)	71.5 %	.43
Naive Bayes (NB)	52.7 %	.055

Table 5.1: Classification model evaluation results sorted by accuracy.

In Table 5.1 the prediction results of the individual models are displayed. As expected, SC shows the best results with an accuracy of 96.3 percent. It contains the composed prediction results of all other models. Therefore, the performance of SC depends on the combined execution time of all models. Consequently, SC is the slowest classification method which may be a reasonable cost for some applications. The second-best classifier is RF, which is a first-level model compared to SC (second-level). RF needs significantly less execution time that is necessary to retrieve appropriate results compared to the SC strategy. One reason being, that RF is one of those algorithms that are so-called "embarrassingly parallel" since the subtrees can be calculated entirely independent of each other. The remaining first-level classification models perform worse than RF and SC. Therefore, we did not consider those classifiers. Even though the accuracy of RF is about 14 percent worse than that of SC, we state that building moderate visual analytics systems requires models that are fast enough to compute to ensure interactivity. Since RF's execution speed is much faster, we make the compromise to work with predictions of slightly worse quality than dealing with long waiting times. Short feedback cycles are an essential measure for interactive visualization systems to improve user experience.

In Table 5.2 we provide a detailed overview of the performance of all classification models for each UHI class (0 = no UHI hotspot identified, 1 = UHI hotspot detected) since accuracy is not sufficient to retrieve significant information about the performance of each classifier. The right column of Table 5.2 contains the precision, recall, specificity, and F-score of those rows that have been identified as UHI hotspot. Even though RF's recall (sensitivity) score is second-best (0.035 below J48's recall score) which means that J48 is better in correctly identifying a higher rate of UHI hotspots, RF is has a higher precision. Thus, RF's proportion of the detected UHI hotspots that are correctly identified is higher than for J48 (0.8 above J48's precision score). Except for SC, which is the best classifiers regarding all measures, RF's precision and F-score scores highest of the remaining first-level classification models. Analogously, RF scores highest in specificity: the rate of mistakenly identified hotspots is the lowest compared to the other first-level classifiers. Hence, using RF as the classification model for the visualization system, considering these aspects is the best decision.

	Class = 0				Class = 1			
Classifier	Precision	Recall	Specificity	F-score	Precision	Recall	Specificity	F-score
SC	.965	.960	.966	.963	.960	.966	.960	.963
RF	.785	.890	.757	.834	.873	.757	.890	.811
J48	.792	.792	.792	.792	.792	.792	.792	.792
MLP	.717	.856	.663	.780	.822	.663	.856	.734
1NN	.723	.697	.733	.710	.708	.733	.697	.720
3NN	.725	.698	.736	.711	.709	.736	.698	.722
5NN	.721	.700	.730	.711	.709	.730	.700	.719
NB	.520	.699	.355	.597	.542	.355	.699	.429

Table 5.2: Detailed evaluation measures of the classification models.

5.2 Confirmatory and Exploratory Analysis

In this section, we will examine different UHI scenarios to discover typical correlations between variables. The following example use cases show how our visual system can be used to extract previously unknown information as well as confirming common hypotheses. In parallel, the strengths and drawbacks of each visualization are explained while analyzing the effect of various features. We stress that our system is not limited to these use cases but only show a narrow portion of possible interaction opportunities when working with the visualization models. Consequently, further scenarios can be flexibly investigated that are not presented here, too.

5.2.1 Typical UHI Characteristics

When it comes to the UHI phenomenon, there are events, which are quite characteristic. For example, high temperatures have the potential to intensify the UHI effect and evoke hotspots. As well, low *precipitation* and increased *sunshine duration* levels raise the probability of hotspots. Using the hourly aggregation type on the PCP visualization, we retrieve a quite informative overview (consisting of samples using a linear sampling strategy) of the correlations between the features and the UHI class of the whole year. We present some examples by selecting all records of 2016 illustrating the stated correlations in Figure 5.2 denoted by respective numbering: The *wind speed* dimension (1) indicates that lower *wind* levels positively correlate with the occurrence of UHIs, whereas higher wind speed reduces the UHI intensity. Analogously, the *cloud coverage* level dimension (2) conveys that increased clouding decreases the number of existing hotspots, while decreased *cloud coverage* leads to a higher probability of UHI occurrences. Understandably, the *sunshine duration*, which is negatively correlated with *cloud coverage* level, influences the appearance of UHI intensity: Long periods of *sunshine* (3) lead to increased solar radiation levels

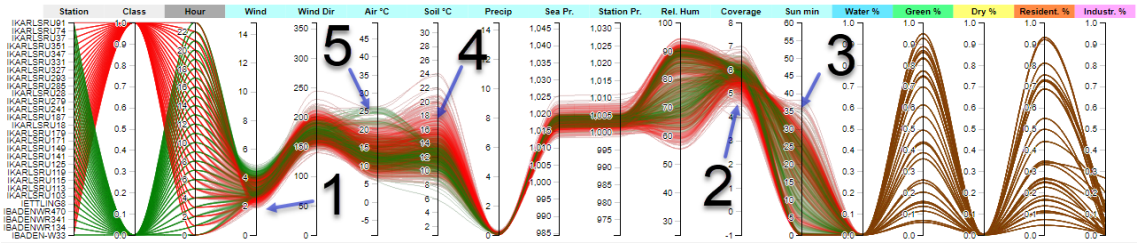


Figure 5.2: PCP: Hourly Aggregation of the UHI Class Data.

and consequently cause UHI hotspots to exist more frequent than short sunshine periods. A further characteristic of the *sunshine duration* dimension can be identified by using the brush filter selecting only low values (few sunshine minutes per hour): From evening to morning the *sunshine* level must be low independent from the *cloud coverage* level (see Figure 5.3). Consequently, during night times UHI hotspots are positively correlated with a shortened *sunshine* period (up to 15 minutes in the morning/evening hours). As displayed in Figure 5.2, the dimensions *soil tempera-*

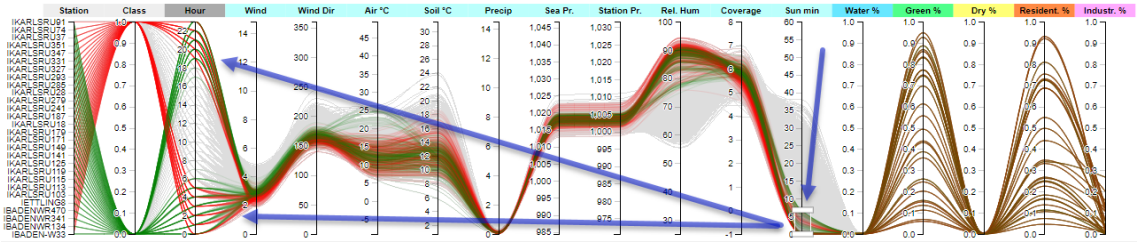


Figure 5.3: PCP: applied brush filter on sunshine duration dimension when hourly aggregation is selected. UHI intensity can be large during night times, though sunshine duration is low.

ture (4) and *air temperature* (5) have a positive correlation. Higher temperatures lead to increased UHI intensity. In this example, there are some outliers (high air temperature, but no UHI appearance; UHI class is 0). Using the same brushing filter as before on this dimension (see Figure 5.4), we detect that the outliers only belong to two stations that is only a small non-dominant fraction. Therefore, it does not contradict the previous hypothesis that high *soil* and *air* temperatures lead to emerging UHI hotspots. Informal interviews with experts from the UHI analysis domain yield to information about another typical property of UHIs: Increased temperatures lead to elevated UHI intensity values that may cause hotspots to last through the night until morning (dawn). The UHI intensity increases in the evening

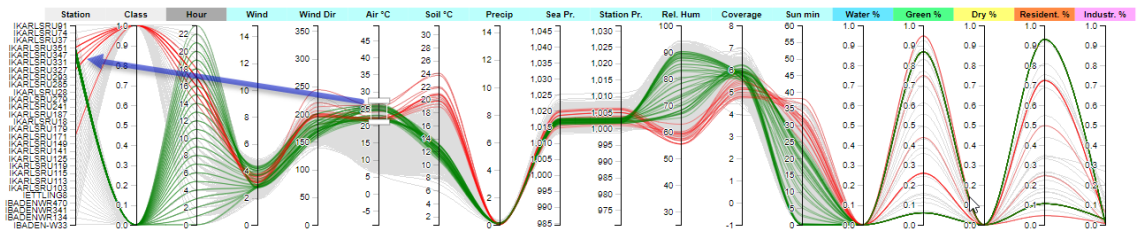


Figure 5.4: PCP: brush filter applied on air temperature dimension using hourly aggregation. UHI intensity outliers are related to only two WU stations.

until midnight or up to two hours later and slowly diminishes until dawn. The red arrows in Figure 5.5 represent this UHI behavior for multiple stations. This degree of this attribute individually depends on each weather station and its properties such as surrounding area and location. Also, characteristic surface radiation properties could not be taken into account due to missing data sources.

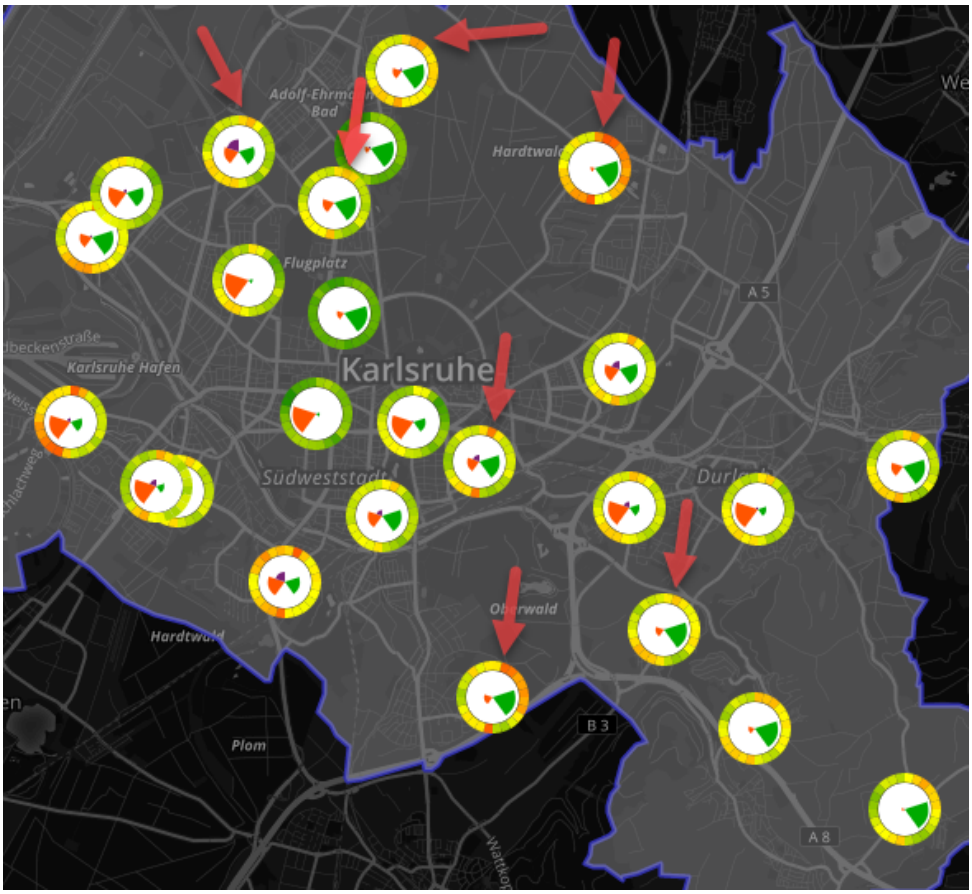


Figure 5.5: Red arrows show a characteristic property of WU stations that have an increasing UHI intensity that weakens during night until dawn.

5.2.2 Increase of Industry Area Level

To analyze the influence of the *industry area* on the UHI intensity, we process a prediction increasing the industry level of all stations by 60 percent. The area level of the remaining area types is automatically balanced equally (15 percent each) preserving the maximum allowed area level in the surrounding of each station. Some stations cannot reduce the area level of some features. In this case, the remaining area reduction is repetitively balanced to the other area categories. Figure 5.6 presents the change of the UHI intensity of the WU stations. Red circles are highlighting a WU station (Station ID: `IKARLSRU351`) that is strongly influenced by the prediction.

Figure 5.6: These PCG maps are showing the change of the UHI intensity for the WU station `IKARLSRU351` (indicated by red circles) when increasing the industry area level by 60 percent while the remaining four area types are equally reduced (balanced) by 15 percent. The left image shows the UHI intensity situation before processing the prediction. The right image presents the prediction result.

Looking at the UHI intensity levels by applying the different temporal aggregation types (hourly, daily, and monthly), we can see the increase of the UHI intensity at 12 a.m. to 8 p.m. by 30 percent to 35 percent (see Figure 5.7). Moreover, the prediction specifies that the average UHI intensity level rises in all months. Additionally, the UHI intensity is almost inverting for the monthly aggregation: Before the prognostication process, the UHI intensity is high from August to September, whereas afterward the UHI intensity is low in these months, but high UHI intensities are stated in the time of December to July. In Figure 5.7, the inner groups of the PCG display the reduction of the green area level from 6 percent to 0.7 percent and of the residential area level from 92 percent to 70 percent.

This prediction scenario suggests a positive correlation between the size of the paved surface or number of buildings or industry and the emergence of a UHI. Moreover, this prediction model enables users to detect area level thresholds, indicating the influence of additional residential or industrial areas. Alternatively, other area properties can be examined to identify changes that could cause an improvement of the UHI situation in a particular UHI-affected area: e.g., build parks in the vicinity of industrial buildings.

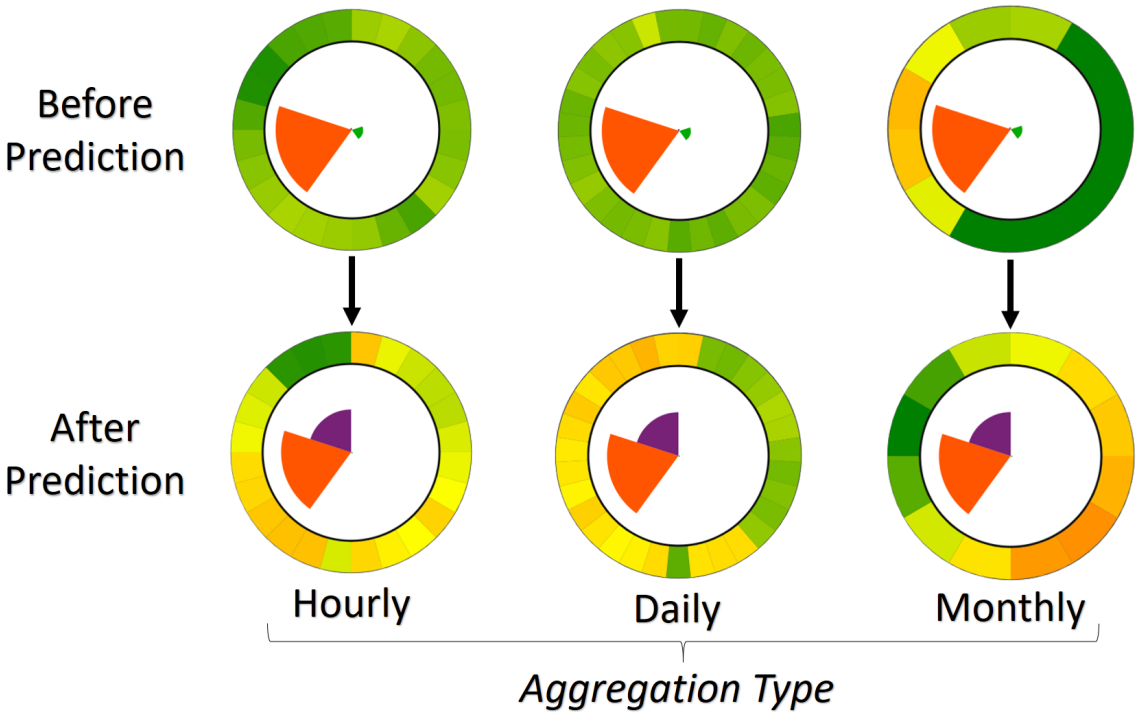


Figure 5.7: PCGs of station `IKARLSRU351`: prediction changes of the UHI intensity for hourly, daily, and monthly aggregation (left to right).

5.2.3 Influence of Weather and Time

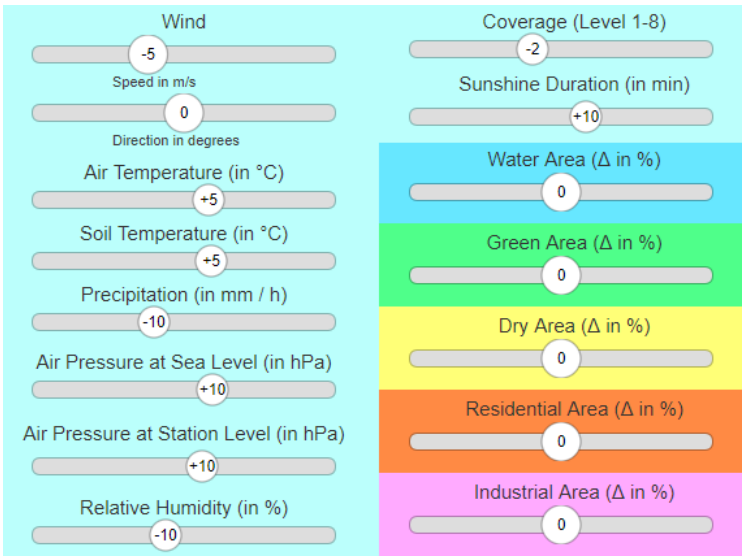


Figure 5.8: Adjusted feature parameters to simulate a warmer and dryer weather situation.

To analyze the effect of changing weather conditions, we select a scenario using the IFPP to simultaneously adjust different weather-related parameters that are not independent in reality but change concurrently. For example, rainy weather situations cause dropping temperature which in turn leads to decreasing air pressure. Consequently, the air pressure imbalance results in rising wind speed levels and movements of air layers with different temperature levels. To address the opposite scenario of warm temperatures and low wind and precipitation levels, we applied a corresponding feature vector (see Figure 5.8). The calendar visualization and the interactive time range selector are

helpful to examine this weather prediction simulation temporally. We will compare the simulated results with the standard values (no feature value changes) to detect differences of the UHI intensity that can be attributed to the changes that were applied to the variables with the IFPP.

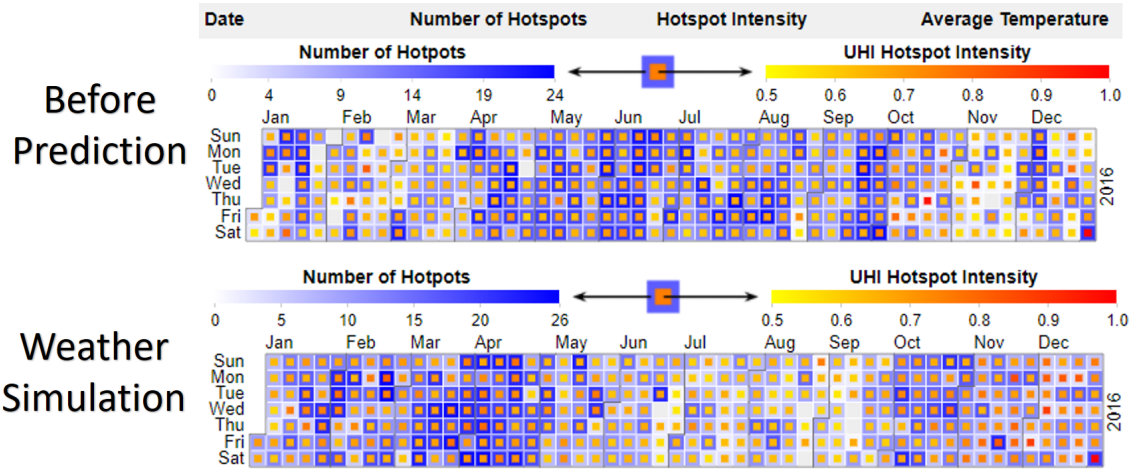


Figure 5.9: Yearly overview of simulation changes for warmer weather scenario.

Figure 5.9 displays both the UHI intensity before and after the prediction process and provides a yearly overview for each day combining the UHI intensity of all stations into a single factor. At first sight, the warming has a strong influence on the months from January to April. Both the number of hotspots and UHI intensity are increasing significantly, whereas the situation of the months from May to September changes in the opposite direction. The values for October underly no drastic changes, except that the UHI intensity slightly decreases for some days (e.g., October 20, from 96 % to 74 %) that seem to be outliers, since before simulation on that day only one hotspot was detected. Focusing on specific times that are characteristic of the emergence of UHIs is interesting to analyze the influence of daytime on the result. For instance, UHI areas cool down more slowly than the suburban surrounding that often includes more vegetation. Therefore, we apply the time range selector to compare both day and night times for the simulation choosing day hours from 7 a.m. to 8 p.m. and night hours from 8 p.m. to 7 a.m.. Especially in winter, an inherent property of the UHI effect is that city areas are warmer compared to the surrounding areas. Similarly, in summer the air does not cool down in the night as much as areas that contain more vegetation. To analyze the simulation regarding this property, we applied the respective day and night filter on our prediction result. The calendar charts in Figure 5.10 depict principal differences in UHI intensity between night and day times. Predominantly, during day times (7 a.m. - 8 p.m.), the UHI intensity is evenly distributed. However, in April a higher number of hotspots appears. In November/December, less hotspots having a higher average UHI intensity exist compared to other months. In contrast to daytimes, both the hotspot amount and the UHI intensity is significantly higher from October to April compared to the rest of the year. Consequently, the simulation of warmer weather indicates a strong influence on the UHI hotspots in the colder seasons, whereas the summer season

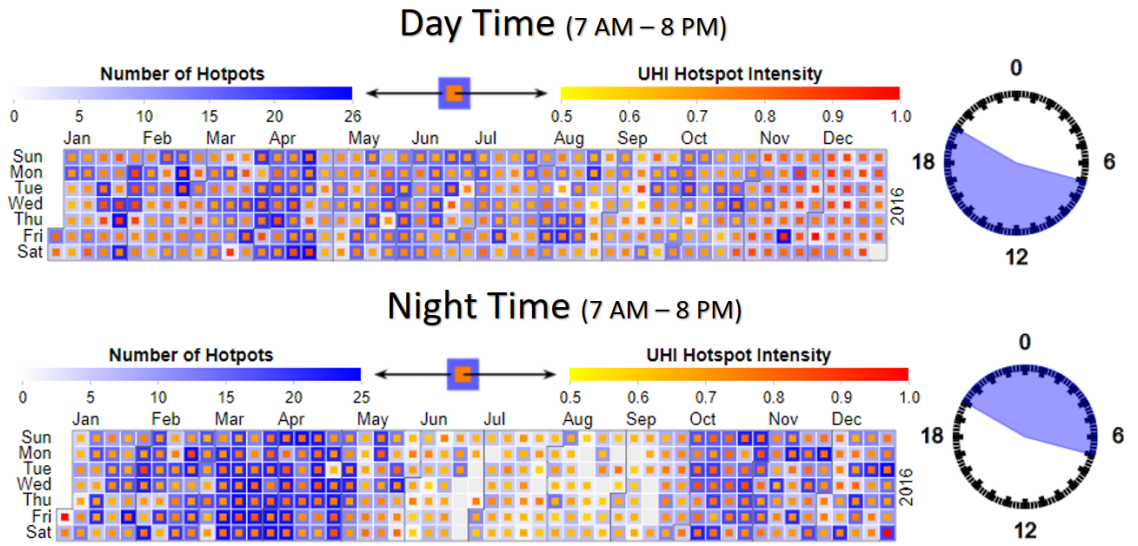


Figure 5.10: Calendar view: UHI intensity comparison of day and night times.

"profits" from this meteorological change. To examine this situation for the night hours further, we subdivide this time range into two halves: we select the time frame from 8 p.m. to 1 a.m. as the first half of the night, and from 2 a.m. to 7 a.m. as second half (see Figure 5.11). There are only small differences for the warmer months (May - September), whereas the UHI effect intensifies from November to April. The number of hotspots increases in the second half of the night in these months, which is a typical characteristic for UHIs.

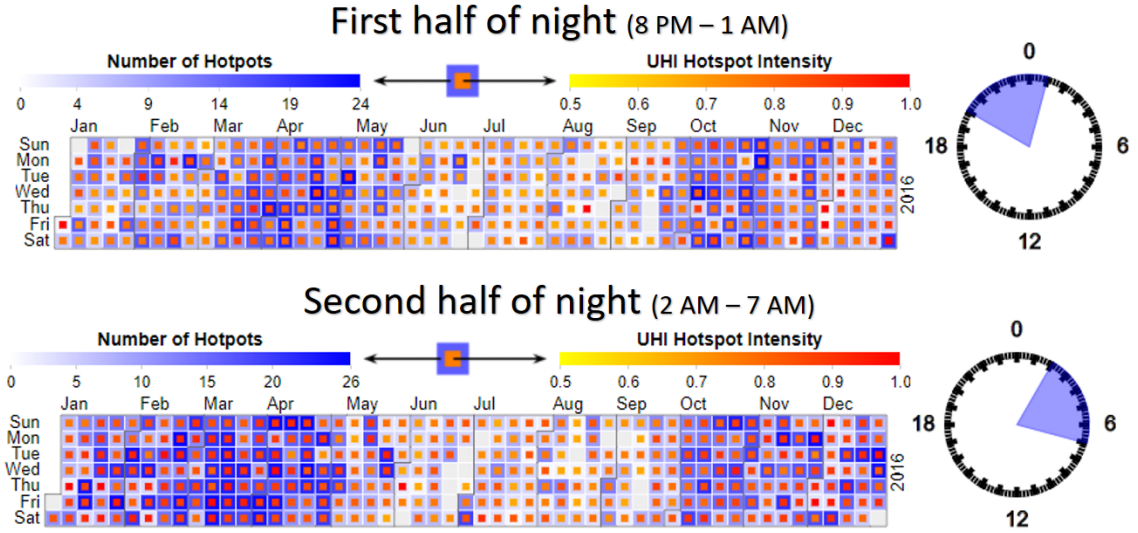


Figure 5.11: Comparison of UHI hotspots characteristics for the first and second half of the night.

This use case in combination with the prediction model on area parameter changes in Section 5.2.2, meteorologists can examine long-term weather trend changes by entering their information. In this way, expert users can identify relevant UHI situations at an early stage and analyze possible mitigation strategies by changing characteristic surface features such as green areas.

Chapter 6

Discussion & Future Work

In this chapter, we discuss several key topics presented in this thesis: Visual Analytics for UHI, limitations of our approach, and areas of future work.

6.1 Visual Analytics of UHI

By the example of the city of Karlsruhe, we illustrate some use cases of how event predictions can be investigated depending on several parameters that address domains such as meteorology, environmental development of urban areas, land-usage planning, and surface texture analysis. The combination of visualizations that cover a variety of aspects facilitates to focus on specific attributes. The user can steer use-case-relevant features to retrieve information about possible UHI scenarios. As presented in the previous chapter, our application supports different kinds of use cases to analyze the effect of multiple parameters in various scenarios. The provided examples outline the potential of our approach by presenting the usefulness of the available components in practice.

6.1.1 Use Cases

Frequently, the first step of an investigation consists of examining typical characteristics to confirm existing hypotheses that are more comprehensible. In Section 5.2.1, several examples are depicted that validate the effectiveness of the designed visualizations. Feature correlations between multiple variables can be identified by the PCP. The brushing filter, which can be flexibly applied to an arbitrary number of feature dimensions, is a potent interaction method to let users quickly focus on desired parts of the dataset. Multiple hypotheses confirmation examples are depicted: The emergence of UHI hotspots correlates with lower cloud coverage level, reduced wind speed levels, and increased temperature. When examining the relationship of the sunshine duration level with the occurrences of UHI hotspots, we realized that many records indicate that during periods of no sunshine, often UHI hotspots are present. At first glance, this finding seems to be counterintuitive, since more

sunshine implicates higher UHI intensity levels. This is true except for night hours: At night, the sunshine duration is at its minimum at all stations. Some areas underly slower cooling, which implies that UHI hotspots are also occurring at night times. This explanation is supported by the PCP, since nearly all records are correlated with the hours from 8 p.m. to 6 a.m., confirming the previous explanatory approach. The dimensions displayed by the PCP can be flexibly rearranged using drag and drop to reveal potentially related features that are not positioned next to each other in the default PCP dimension order. Hence, the PCP provides informative insights when the analysis is focused on multi-dimensional data.

Since PCP does not supply any geographic information, we designed the pie chart glyph as an efficient way to simultaneously present both spatial and temporal data. Owing to this combination of space and time into a single glyph, the cognitive user load can be decreased. Thus, user-experience can be enhanced while taking advantage of the available space. For instance, the use case presented in Section 5.2.2 "Increase of Industry Area Level" includes the prediction of increasing industry area for all stations. The WU station `IKARLSRU351` signals an increasing occurrence of UHI events from 12 a.m. to 10 p.m., whereas concurrently the UHI hotspot level of remaining stations are subject to only minor variations. Hence, this station seems to be an outlier: further analysis of the immediate surrounding shows that this station is located directly next to the Nottingham Facility, a green area situated in the center of Karlsruhe. This could explain the result presented in the PCG map. Nonetheless, the PCG map is a suitable representation to support preattentive information to the user.

The use case "Influence of Weather and Time" in Section 5.2.3 indicates how a change in multiple meteorological features influences the prediction result. In combination with the calendar view, the time range selection filter is a useful tool to analyze different daytimes sequentially. Thus, differences between night and day can be detected. Based on this example, the UHI intensity levels are stronger in colder seasons of the year. Since the UHI intensity directly depends on temperature values, one could expect worse UHI values for the summer times as well, which could not be confirmed by the analysis. This result states that the occurrence of UHI hotspots decreases as precipitation, relative humidity and cloud coverage decrease and temperature, sunshine minutes and air pressure increase.

The simultaneous change of multiple variables (feature vector that has been applied for the prediction) originates from the fact that changing single features does not provide apparent variations in the prediction outcomes. Consequently, we selected a reasonable combination of multiple variable changes that depend on each other and often happen concurrently. The forecast outcome explains, that in winter and spring season the UHI level is more intense, especially at night. Compared to the summer season, warmer temperatures in colder seasons have a positive effect on energy consumption since expenditure for heating decreases. Based on this forecast, in summer a temperature increase may affect not only the inner parts of a city but also the corresponding surrounding. Hence, the level of the UHI effect is relativized. Our visualization only indicates that the UHI behavior is reduced. This does not

mean that the general UHI situation improves, but conversely deteriorates the overall condition. Our simulation strategy underlies certain limitations: we expected an equal change for all days when executing this scenario as well as a degradation (not improvements) of the UHI intensity situation for the months from May to September. Therefore, it could be worthwhile to improve the model for our definition of a UHI by integrating further sources to cover more relevant features to retrieve more precise predictions.

6.1.2 Feature Categories

Filtering techniques as the selection of a WU station or a particular value range within a variable dimension draw the user's attention to relevant attributes. The features can be divided into two main types. The first type contains aspects that are directly suggestible by city planners. Thus, decision-making tasks primarily focus on this feature type, since mitigation strategies can more efficiently be elaborated. The second type contains parameters that can not be immediately modified by decision-makers such as the meteorologic situation of a city. Local weather depends on the geographic location which makes it impossible to influence features such as precipitation, clouding, air pressure, relative humidity, and sunshine hours. Nevertheless, the observation of these features is crucial to identify undesired scenarios that might happen if particular features are subject to change. Our prototype takes both feature types into account and provides an indication which parameters should be under surveillance.

6.1.3 Data Analytics

As illustrated in Figure 6.1, there are three different analytics types within the realm of data analytics: descriptive, predictive and prescriptive. Up until now, the first two types have been addressed in our visual framework. Analyzing gathered data from the past is an integral part of descriptive analytics to provide insights on past events. Apart from that, the employment of statistical classification strategies to create prediction models that are forecasting possible scenarios is the primary objective of predictive analytics. It gets more complicated if a system is intended to

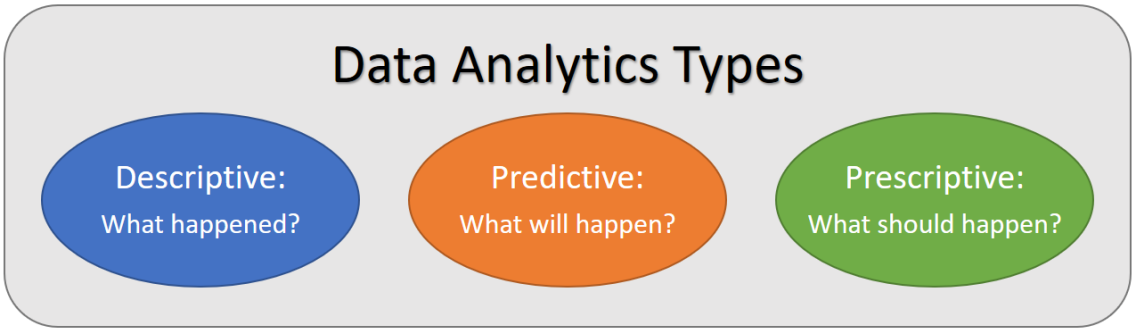


Figure 6.1: There are three different types of data analytics: descriptive, predictive, and prescriptive analytics [21].

automatically propose mitigation strategies for particular scenarios. This method is called *prescriptive analytics* and does not only describe how specific adjustments of features would affect future outcomes but additionally provides models to present strategies to reduce undesired effects. Our prototype only indicates how steering of particular parameters affects the predicted result. In this way, users have to identify possible avoidance measures without support. Our approach does not include such recommendations since it requires more sophisticated models to generalize the influence of different features on the outcome. Developing UHI mitigation strategies is another crucial branch within the realm of UHI research which requires intricate models that allow incorporating expert knowledge to retrieve useful suggestions.

6.2 Limitations

The complex composition of various factors is hard to be wholly included in such a visualization system. The meaningfulness of the analysis is limited due to essential information that could not be integrated yet and due to the spatial distribution of temperature stations. For the detection of existing UHI hotspots, we only applied temperature data from the WUSN and verified this data using DWD data that is subject to higher quality standards.

So far, our application lacks techniques to explicitly guide the user's attention to relevant sections of the visualization to emphasize outliers or extreme values by highlighting or animations that must be currently identified by the user. This may consist of recommendations for scenarios that are vulnerable for the UHI effect in certain situations such as reduced wind levels or days with a high amount of sunshine minutes. Currently, the results of the analysis only rely on the user's interactions which is not an acceptable assumption for users that are not familiar with the subject of the UHI effect.

The presentations are subject to varying levels of uncertainty if classification models for the prediction of particular scenarios are used. The classifier evaluation in Chapter 5 provides further information about the conciseness of the single models. Even though the results have reasonable performance, the uncertainty of the forecasting process should be reflected in the visualizations to increase the trustworthiness of the depicted information. So far, this aspect has not been integrated limiting the certainty of conclusions that users could draw.

The current prototype version is designed for presenting our visualizations at the example of a single city and only for the year 2016. The integration of spatial and temporal scalability would expand the potential application areas of our approach. Since UHI hotspots are only relevant for larger cities, considering cities with a minimum population of 50 000 would cover a large number of areas that may be subject to heat stress. The separate analysis of different city districts may reveal further information about the correlation between geographic characteristics and emerging UHIs. Additionally, domain experts may benefit from extending the temporal range for many years depending on available datasets to retrieve a better

chronological overview. In this way, characteristic patterns for the single seasons and months of the year could be detected. Unmistakably, the presentation of data of a single year is a limitation of temporal scalability concealing potentially relevant long-term trends. Therefore, extended geographic and temporal scalability would be beneficial for an informative visual analytics approach. Better scalability enhances the investigation of UHI hotspots by considering greater time ranges and geographic representations.

6.3 Expert Feedback

To integrate research field knowledge into an analytics framework, inquiring research field experts can be profitable to retrieve domain-specific information for designing meaningful models that are applicable in real scenarios. For this reason, we conducted informal feedback sessions with Dr. Saskia Buchholz[1]. Her research is situated in the domain of environmental impact assessment, climate change, and urban planning within the *Climate and Environment Consultancy* department of the German Meteorological Service also called DWD [25] and Prof. Dr. Stefan Emeis[2], who is the group leader of "Regional Coupling of Ecosystem-Atmosphere Processes" at the Institute of Meteorology, Climate, and Atmospheric Environmental Research of the Karlsruhe Institute of Technology [46]. We gained valuable knowledge from their feedback about multiple parameters that are influencing the emergence of UHIs. We describe several aspects that are relevant for UHI analysis in the following sections.

6.3.1 UHI vs. UHI Hotspots

Both experts pointed out that there is a significant difference between the analysis of a UHI, which deals with comparisons of temperature levels between a city area and its colder surrounding, and the investigation of UHI hotspots that we also call "intra urban heat islands". Buchholz also denoted, that usually UHI hotspot and UHI are exchangeably used in this research field. Therefore, we emphasize that in this thesis, the definitions of the terms "UHI hotspot" and "UHI" vary. The first case mainly results from a colder surrounding including more vegetation, whereas UHI hotspots often additionally depend on other factors such as anthropogenic heat release, radiation properties of the surfaces, and traffic. According to the experts, the temperature difference mentioned before can reach up to ten kelvin if corresponding conditions such as prolonged sunshine periods and low wind speeds are met. To retrieve a UHI overview of a city, it is necessary to compare temperature measurements from stations that are situated within a city with those that are in the rural surrounding. Therefore, it is essential to consider surface information of a station's surrounding to choose stations that are located in green areas to

[1]Dr. Saskia Buchholz – saskia.buchholz@dwd.de
[2]Prof. Dr. Stefan Emeis – stefan.emeis@kit.edu

retrieve a representative value for the colder surrounding area. Explicitly, Buchholz recommended using the temperature data of the DWD stations that are subject to high-quality measurement standards in opposition to the stations of the WUSN. We emphasize that we used the WU stations to calculate UHI intensity values due to a higher geographic resolution. This is an acceptable approach since we preprocessed this dataset to increase the data quality by dealing with missing values (linear interpolation) and removing outliers. The current version of our prototype does not differentiate between stations that are within or outside of the city. This is critical when comparing the temperature level of the overall city with the suburban areas. Currently, only local hotspots are subject to our visual analysis system, although it is not sufficient for many use cases that analyze the temperature situation of a whole city.

6.3.2 Influencing Factors

As mentioned in previous chapters, Emeis confirmed that the temperature level depends on a variety of parameters. Air temperature levels depend on the geometry of buildings, the thermal properties of building fabric, and the radiation characteristics of the surfaces. Moreover, the temperature is affected by anthropogenic heat-release such as air conditioning systems, industry, traffic, and domestic heating. We were advised that the UHI shape primarily depends on the building density and the degree of surface sealing [12].

A characteristic feature of the UHI phenomenon is that a UHI usually has its highest intensity during night times. Typically, at daytimes, the UHI temperature differences reach 1–4 kelvin. In contrast to this, the UHI intensity reaches its maximum during the night with a temperature difference not exceeding twelve kelvin. We could confirm these characteristic features with the help of the PCG map in one of the example use cases as described in Figure 5.5 provided in the previous chapter.

Due to further feedback provided during the informal interview, it became apparent that the UHI effect's inherent complexity does not merely depend on the temperature difference between two weather stations. Therefore, we implemented a method that calculates a distance-weighted measure by multiple surrounding stations. Nevertheless, this is incomplete, since the UHI effect even depends on further parameters such as the morphology and the physical properties of urban surfaces (e.g., roughness) [51]. According to Buchholz, air temperature is the decisive factor for the identification of a UHI effect, especially at night. Furthermore, surface temperatures are an essential part as well as the respective surface characteristics: surfaces of darker colors absorb more solar radiation resulting in increasing warmth. Additionally, the thermal properties of building materials influence the temperature in cities as well as the radiation properties of surfaces, and anthropogenic heat release from domestic heating systems, transportation, and industry. Emeis pointed out that evaporation processes reduce the UHI effect since a lot of energy is necessary for this process. Due to him, in the domain of urban meteorology, experts currently try to "solve this complex topic by numerical simulation equations" based on physical

equations of balance for impulse and energy. Only if these fundamental aspects are considered within a model, they can be used to predict scenarios of future conditions.

6.3.3 Potential Fields of Application

We were informed by Buchholz that the application's objective is heading in the right direction since many communities would be interested in visualizations incorporated in our visual framework. Nonetheless, before publishing, it is essential to revise methodical and climatological concepts. Interactive visualization tools are prominent, especially if every relevant municipality (e.g., all cities with a minimum population of 50 000) would be included. In many cases, expert users in the domain of city planning are in need of a climatological evaluation to understand the existence of a particular hotspot within a city. This is especially true for disaster management, and hotspots in the vicinity of hospitals, retirement homes, and schools. Consequently, analyzing concurrent conditions is a task of particular interest to estimate whether countermeasures should be initiated to avoid a problematic situation. The expert feedback included that multiple visualizations are useful because every user has individual preferences regarding what representation is more comprehensible. The domain expert stated that working with symbols or glyphs will also enhance user experience.

Due to missing data sources, the following features have not been considered yet. For instance, the transportation system may also play a fundamental role regarding the analysis of urban temperatures. However, it could be easily integrated into our system once available. Therefore, we claim that our application is a viable visual analytics solution for this research field. Another factor is the deployment ratio of sustainable technologies to find potential improvement suggestions. For simplicity, we focused on the visualization of temporal and geographic attributes that are a fundamental part of other datasets and a crucial point when adding UHI relevant information. Nonetheless, the evaluation of various classification models indicates that we use a strong data foundation as a basis for the application. The implementation of user interaction techniques as filtering, selection, linking and brushing concerning the knowledge discovery pipeline provides a groundwork that can be extended by further features to improve the validity of the analysis results.

6.4 Future Work

The discussion of our visual analytics framework reveals several limitations that should be addressed in the future to remove weaknesses and add further functionality. As described, we already use meteorological and land-usage data. Further data sources should be considered to integrate in the future, to improve the informativeness of the presented visualizations. For instance, MODIS satellite imagery by NASA [69] could be utilized to gain better geographical resolution, though lacking an appropriate temporal resolution since a satellite traverses over the same area only twice a day.

Consequently, to solely use the MODIS' temperature data is not sufficient to analyze temperature changes on an hourly basis. Nevertheless, MODIS data could enrich the current UHI model by 'filling the gaps' between the WU stations by building an interpolation model to reconstruct possible temperature levels for times when data records are missing for a particular geographic position. Thus, the usefulness of a data source depends on the specific use case. Hence, a new visualization model would be necessary to interpret the user's needs and provide the correct data or a combination of both. This results in a complex composition of heterogeneous data that requires innovative visualization models. The MODIS data source has not been considered in this thesis since it surpasses the scope of this work. Such an approach would integrate domain-relevant information with a higher spatial resolution and coverage since satellites can remotely sense spatially distributed surface temperature data more easily than weather stations with a fixed location. In the future, the existing visualizations and models could be extended regarding new challenges by integrating new datasets to cover further features such as surface radiation and traffic levels.

As part of this thesis, data records were manually gathered to serve as data foundation for our visualization system. The analytics system could profit from an automatic data gathering, processing, and integration pipeline that regularly collects new data to keep the displayed models up to date. By this, expert users do not have to wait for manual data updates but could benefit from a future-oriented application that can easily be accessed via a web browser without the need of a complicated setup process.

Our PCP is currently limited to present singular associations that afford dimension reordering (active user interaction) to depict particular feature relationships that are not visible in the default PCP view. Moreover, only a sample of $\sim$1000 data records is presented without the application of temporal aggregations. As a consequence, many data records are not visualized, if this number is exceeded. If presenting significantly more than 1000 data records, the visualization interactions cannot be smoothly executed any further. The unary and binary aggregation types provide only average information, that may contain various inaccuracies. Integrating a binning technique that can be individually applied to each PCP dimension could overcome the drawback of just presenting a sample since the user can flexibly choose the aggregation. Thus, characteristic patterns that correlate with particular dimensions and are hidden so far can be identified. This would emphasize correlations of the single sections of a dimension that get lost by the current averaging techniques due to under-representation. Alternatively, a correlation matrix could be added and linked to the PCP to facilitate the identification of correlations between features without losing the exact values of each record.

Until now, our application contains no functionality to track the user interaction to enable workflow traceability. Users still need to memorize executed feature selection as well as the respective results. Prospectively, users may greatly benefit from an automatized tracing of their actions providing a history of recent changes to the visualization. In this way, users can restore past application states which

facilitates repeatability of particular filter and view selections. Apart from that, an interaction flow overview may enhance the user's interpretation ability of the visual representations. To enable users to explicitly compare the results of different predictions with the base data set, it is necessary to implement techniques that support comparability. In this manner, various prediction outcomes can be compared and reduce the cognitive load of finding the differences of single features. Further models could be implemented that allow the comparison of the appearance of the UHI effect for different years. Moreover, diverse semantic geographic grouping concepts could improve the understanding of spatial attributes of emerging hotspots. Extended comparison visualizations and models are needed to enable efficient analysis to extract common factors of occurring UHI events. User experience is a substantial measure when it comes to interactive analysis systems. Thus, it can be improved in the future by introducing guided highlighting to draw the users' attention to relevant elements of the representations.

To support users better during their decision-making tasks, it is crucial to convey information about the trustworthiness of provided representations. Due to the limited scope of this thesis, we did not include strategies to display the uncertainty of classification results. Future developments should contain concepts of how uncertainty visualizations can be added to the workspace we developed, for instance by employing techniques as random noise [13] or texture opacity [86]. Ensuring that users understand the uncertainty of visualized information is a key to providing a reliable system[87]. Visualizations become more effective if users comprehend the significance of the different prediction probabilities they are working with.

Eventually, an extensive domain expert study is part of additional future work to assess the strengths and weaknesses of the proposed system. In doing so, we can determine existing challenges and improve the understanding of essential elements that are in the center of UHI analysis but are yet missing. Further feedback from domain experts is a valuable input for the development process and may reveal possible defects of our approach.

Future extensions, adaptations, and additions of novel forecasting and visualization techniques will expand the opportunities for usability. Despite this potential for improvement, our approach can be generalized and transferred to other application fields that have to deal with spatial, temporal and high-dimensional data. This could enable experts to gain insight into various complex application areas, such as, but not limited to, crime data, purchasing behavior, and public transport optimization.

Chapter 7

Conclusion

The main contribution of this thesis was to provide an application that supports interactive visual analysis tasks including event predictions in the domain of Urban Heat Island analysis. The combination of heterogeneous visualizations that take geographical, temporal and multidimensional characteristics into account enables analysts to investigate large datasets regarding correlations of diverse variables of different domains such as meteorology and city planning. The integration of an evaluated prediction model assists end-users with hypothesis creation regarding various features that influence the UHI effect. The deployment of real data enhances the understanding of feature correlations that affect the emergence of UHI hotspots within city areas. An extended classification evaluation was conducted to show the reliability of the results. Additionally, the evaluation demonstrates the validity of the datasets that were used to build the application. We developed a visualization model called pie chart glyph that displays both spatial characteristics of geographic features and temporal event information. By placing them at the respective position on a map connects temporal and geographic attributes. In consideration of profound information-seeking concepts and data exploration methods, the implementation of our application is an elaborated basis within the domain of urban heat island analysis. Nevertheless, the effectiveness of our system should be examined thoroughly to discover strengths and weaknesses of this approach. Particularly, constructive feedback from domain experts is indispensable to develop a system that is useful for end-users to facilitate their daily work. The presentation of concrete use cases shows how our models can be applied and in which areas of application our approach can be used. The event prediction section allows the detection of specific patterns and how parameters change the course of incidences. Implicitly, decision tasks are supported by providing indications of how a situation may change when executing actual mitigation strategies. Nevertheless, the field of prescriptive analytics is part of future work as it's inherent complexity that could not be covered by this work.

Bibliography

[1] Vladimir Agafonkin. *Leaflet – An open-source JavaScript library for mobile-friendly interactive maps*. Accessed February 20, 2018. Retrieved from `http://leafletjs.com/`. 2011.

[2] Wolfgang Aigner et al. "Visual Methods for Analyzing Time-Oriented Data". In: *IEEE Transactions on Visualization and Computer Graphics* 14.1 (2008), pp. 47–60.

[3] Hashem Akbari, Steven J. Konopacki, and Melvin Pomerantz. "Cooling energy savings potential of reflective roofs for residential and commercial buildings in the United States". In: *Energy* 24.5 (1999), pp. 391–407.

[4] Kent B. Barnes, J. Morgan, and Martin Roberge. "Impervious Surfaces and the Quality of Natural and Built Environments". In: *Baltimore: Department of Geography and Environmental Planning, Towson University* (2001).

[5] David A. Bennett, Keith D. Hutchison, and Steven C. Albers. "Preliminary results from Polar-Orbiting Satellite Data Assimilation into LAPS with Applications to Mesoscale Modeling of the San Francisco Bay Area". In: (2000).

[6] Michael R. Berthold et al. "KNIME: The Konstanz Information Miner". In: *Data Analysis , Machine Learning and Applications : Proceedings of the 31st Annual Conference of the Gesellschaft für Klassifikation e. V., Albert-Ludwigs-Universität Freiburg, March 7 9 , 2007*. New York: Springer, 2007.

[7] Enrico Bertini, Patrick Hertzog, and Denis Lalanne. "SpiralView: Towards Security Policies Assessment through Visual Correlation of Network Resources with Evolution of Alarms". In: *2007 IEEE Symposium on Visual Analytics Science and Technology*. Oct. 2007, pp. 139–146.

[8] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. "D^3 Data-Driven Documents". In: *IEEE Transactions on Visualization and Computer Graphics* 17.12 (2011), pp. 2301–2309.

[9] Mike Bostock. *D^3 – Calendar View*. Accessed January 29, 2018. Retrieved from `https://bl.ocks.org/mbostock/4063318`. 2017.

[10] Mike Bostock. *D^3 – Parallel Coordinates*. Accessed January 29, 2018. Retrieved from `https://bl.ocks.org/jasondavies/1341281`. 2017.

[11] Leo Breiman and Adele Cutler. "Random Forests". In: *Machine Learning* 45.1 (Oct. 2001), pp. 5–32.

[12] Saskia Buchholz and Meinolf Kossmann. "Research note. Visualisation of summer heat intensity for different settlement types and varying surface fraction partitioning". In: *Landscape and Urban Planning* 144 (2015), pp. 59–64.

[13] Juri Buchmüller et al. "Visual Analytics for Exploring Local Impact of Air Traffic". In: *Computer Graphics Forum*. Vol. 34. 3. Wiley Online Library. 2015, pp. 181–190.

[14] Stefan Buschmann et al. "Hardware-Accelerated Attribute Mapping for Interactive Visualization of Complex 3D Trajectoriess". In: *2014 International Conference on Information Visualization Theory and Applications (IVAPP)*. 2014, pp. 356–363.

[15] Stuart K. Card, Jock D. Mackinlay, and Ben Shneiderman. *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann Publishers Inc., 1999.

[16] M. Sheelagh T. Carpendale, David J. Cowperthwaite, and F. David Fracchia. "3-Dimensional Pliable Surfaces: For the Effective Presentation of Visual Information". In: *Proceedings of the 8th Annual ACM Symposium on User Interface and Software Technology*. UIST '95. ACM, 1995, pp. 217–226.

[17] Mei Zhu Chen, Wei Wei, and Shao Peng Wu. "On Cold Materials of Pavement and High-Temperature Performance of Asphalt Concrete". In: *Eco-Materials Processing and Design X*. Vol. 620. Materials Science Forum. Trans Tech Publications, Sept. 2009, pp. 379–382.

[18] Thomas H. Davenport. "Analytics in Sports: The New Science of Winning". In: *International Institute for Analytics* 2 (2014), pp. 1–28.

[19] Neil Debbage and J. Marshall Shepherd. "The urban heat island effect and city contiguity". In: *Computers, Environment and Urban Systems* 54 (2015), pp. 181–194.

[20] DeepstreamHub. *GoldenLayout – A multi-screen layout manager for webapps*. Accessed February 20, 2017. Retrieved from `http://golden-layout.com/`. 2016.

[21] Dursun Delen and Haluk Demirkan. *Data, information and analytics as services*. 2013.

[22] UN Desa. "World urbanization prospects, the 2013 revision". In: *Final Report with Annex Tables. New York, NY: United Nations Department of Economic and Social Affairs* (2014).

[23] Alexandra Diehl et al. "Visual Analysis of Spatio-Temporal Data: Applications in Weather Forecasting". In: *Computer Graphics Forum*. Vol. 34. 3. Wiley Online Library. 2015, pp. 381–390.

[24] DWD. *Deutscher Wetterdienst - Wetterstationen*. Accessed June 12, 2017. Retrieved from `https://www.dwd.de/DE/leistungen/klimadatendeutschland/stationsliste.html`.

[25] DWD. *German Meteorological Service – Climate and Environment Consultancy*. Accessed February 23, 2018. Retrieved from `https://www.dwd.de/EN/climate_environment/consultancy/consultancy_node.html`.

[26] Usama M. Fayyad et al. *Advances in Knowledge Discovery and Data Mining*. Menlo Park, CA, USA: American Association for Artificial Intelligence, 1996.

[27] Nivan Ferreira et al. "BirdVis: Visualizing and Understanding Bird Populations". In: *IEEE Transactions on Visualization and Computer Graphics* 17.12 (2011), pp. 2374–2383.

[28] Erich M. Fischer et al. "Soil Moisture – Atmosphere Interactions during the 2003 European Summer Heat Wave". In: *Journal of Climate* 20.20 (2007), pp. 5081–5099.

[29] Danyel Fisher. "Hotmap: Looking at Geographic Attention". In: *IEEE Transactions on Visualization and Computer Graphics* 13.6 (Nov. 2007), pp. 1184–1191.

[30] Johannes Fuchs et al. "Evaluation of Alternative Glyph Designs for Time Series Data in a Small Multiple Setting". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM. 2013, pp. 3237–3246.

[31] Katharina M.A. Gabriel and Wilfried R. Endlicher. "Urban and rural mortality rates during heat waves in Berlin and Brandenburg, Germany". In: *Environmental Pollution* 159.8 (2011), pp. 2044–2050.

[32] Simone Garlandini and Sara Irina Fabrikant. "Evaluating the Effectiveness and Efficiency of Visual Variables for Geographic Information Visualization". In: *International Conference on Spatial Information Theory*. Springer. 2009, pp. 195–211.

[33] Lisa Gartland. *Heat Islands: Understanding and Mitigating Heat in Urban Areas*. Routledge, 2012.

[34] Peter Gatalsky, Natalia Andrienko, and Gennady Andrienko. “Interactive Analysis of Event Data Using Space-Time Cube”. In: *Information Visualisation, 2004. IV 2004. Proceedings. Eighth International Conference on*. IEEE. 2004, pp. 145–152.

[35] James G. Gimpel, Frances E. Lee, and Joshua Kaminski. “The Political Geography of Campaign Contributions in American Politics”. In: *Journal of Politics* 68.3 (2006), pp. 626–639.

[36] Jonathan Grudin. “Partitioning Digital Worlds: Focal and Peripheral Awareness in Multiple Monitor Use”. In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM. 2001, pp. 458–465.

[37] L. Haddad and Z. Aouachria. “Impact of the Transport on the Urban Heat Island”. In: *World Academy of Science, Engineering and Technology, International Journal of Environmental, Chemical, Ecological, Geological and Geophysical Engineering* 9.8 (2015), pp. 968–973.

[38] Torsten Hägerstraand. “What about people in regional science?” In: *Papers in Regional Science* 24.1 (1970), pp. 7–24.

[39] S. Hassid et al. “The effect of the Athens heat island on air conditioning load”. In: *Energy and Buildings* 32.2 (2000), pp. 131–141.

[40] Helwig Hauser, Florian Ledermann, and Helmut Doleisch. “Angular Brushing of Extended Parallel Coordinates”. In: *Information Visualization, 2002. INFOVIS 2002. IEEE Symposium on*. IEEE. 2002, pp. 127–130.

[41] Johannes Häussler. *jQueryClockIntervalPicker: jQuery plugin time selection picker clock*. Retrieved from `http://johae.github.io/jQueryClockIntervalPicker/`. Accessed January 29, 2018. 2016.

[42] Susan Havre, Beth Hetzler, and Lucy Nowell. “ThemeRiver: Visualizing theme changes over time”. In: *Information Visualization, 2000. InfoVis 2000. IEEE Symposium on*. IEEE. 2000, pp. 115–123.

[43] Simon Haykin. “Neural Networks: A Comprehensive Foundation”. In: *Neural Networks* 2.2004 (2004), p. 41.

[44] Y. Hirano and T. Fujita. “Evaluation of the impact of the urban heat island on residential and commercial energy consumption in Tokyo”. In: *Energy* 37.1 (2012), pp. 371–383.

[45] Geoffrey Holmes, Andrew Donkin, and Ian H. Witten. “WEKA: A Machine Learning Workbench”. In: *Intelligent Information Systems, 1994. Proceedings of the 1994 Second Australian and New Zealand Conference on*. IEEE. 1994, pp. 357–361.

[46] IMK – IFU. *Institute of Meteorology and Climate Research Atmospheric Environmental Research (IMK-IFU), Campus Alpin, Garmisch-Partenkirchen*. Accessed February 24, 2018. Retrieved from `https://www.imk-ifu.kit.edu/index.php`. 2018.

[47] Alfred Inselberg and Bernard Dimsdale. “Parallel Coordinates for Visualizing Multi-Dimensional Geometry”. In: *Computer Graphics 1987*. Springer, 1987, pp. 25–44.

[48] Alexander Kachkaev, Jo Wood, and Jason Dykes. “Glyphs for Exploring Crowd-sourced Subjective Survey Classification”. In: *Computer Graphics Forum*. Vol. 33. 3. Wiley Online Library. 2014, pp. 311–320.

[49] Daniel A. Keim. “Designing Pixel-Oriented Visualization Techniques: Theory and Applications”. In: *IEEE Transactions on Visualization and Computer Graphics* 6.1 (2000), pp. 59–78.

[50] Daniel A. Keim. “Information Visualization and Visual Data Mining”. In: *IEEE Transactions on Visualization and Computer Graphics* 8.1 (2002), pp. 1–8.

[51] Christine Ketterer et al. “Comparison of selected approaches for urban roughness determination based on voronoi cells”. In: *International Journal of Biometeorology* 61.1 (Jan. 2017), pp. 189–198.

[52] Yukihiro Kikegawa et al. “Impacts of city-block-scale countermeasures against urban heat-island phenomena upon a building’s energy-consumption for air-conditioning”. In: *Applied Energy* 83.6 (2006), pp. 649–668.

[53] Edwin M. Kilbourne. “Heat waves and hot environments”. In: *The Public Health Consequences of Disasters* (1997), pp. 245–269.

[54] Steven J. Konopacki and Hashem Akbari. “Energy Savings of Heat-Island Reduction Strategies in Chicago and Houston (Including Updates for Baton Rouge, Sacramento, and Salt Lake City)”. In: *Lawrence Berkeley National Laboratory* (2002).

[55] Silviya Korpilo, Tarmo Virtanen, and Susanna Lehvävirta. “Smartphone GPS tracking – Inexpensive and efficient data collection on recreational movement”. In: *Landscape and Urban Planning* 157 (2017), pp. 608–617.

[56] Menno-Jan Kraak. “The Space-Time Cube Revisited from a Geovisualization Perspective”. In: *Proc. 21st International Cartographic Conference*. 2003, pp. 1988–1996.

[57] C.P. Lo and Dale A. Quattrochi. “Land-Use and Land-Cover Change, Urban Heat Island Phenomenon, and Health Implications”. In: *Photogrammetric Engineering & Remote Sensing* 69.9 (2003), pp. 1053–1063.

[58] Yafeng Lu et al. “The State-of-the-Art in Predictive Visual Analytics”. In: *Computer Graphics Forum* 36.3 (2017), pp. 539–562.

[59] George Luber and Michael McGeehin. “Climate Change and Extreme Heat Events”. In: *American Journal of Preventive Medicine* 35.5 (2008). Theme Issue: Climate Change and the Health of the Public, pp. 429–435.

[60] Ross Maciejewski et al. “A Visual Analytics Approach to Understanding Spatiotemporal Hotspots”. In: *IEEE Transactions on Visualization and Computer Graphics* 16.2 (Mar. 2010), pp. 205–220.

[61] Ross Maciejewski et al. “Forecasting Hotspots – A Predictive Analytics Approach”. In: *IEEE Transactions on Visualization and Computer Graphics* 17.4 (Apr. 2011), pp. 440–453.

[62] Abish Malik et al. “Visual Analytics Law Enforcement Toolkit”. In: *Technologies for Homeland Security (HST), 2010 IEEE International Conference on*. IEEE. 2010, pp. 222–228.

[63] Rizwan Ahmed Memon, Dennis Y.C. Leung, and Liu Chunho. “A review on the generation, determination and mitigation of Urban Heat Island”. In: *Journal of Environmental Sciences* 20.1 (2008), pp. 120–128.

[64] Parham A. Mirzaei. “Recent Challenges in Modelling of Urban Heat Island”. In: *Sustainable Cities and Society* 19 (2015), pp. 200–206.

[65] Parham A. Mirzaei and Fariborz Haghighat. “Approaches to study Urban Heat Island – Abilities and limitations”. In: *Building and Environment* 45.10 (2010), pp. 2192–2201.

[66] Asit Kumar Mishra and Maddali Ramgopal. “Field studies on human thermal comfort – An overview”. In: *Building and Environment* 64 (2013), pp. 94–106.

[67] Madalena Vaz Monteiro et al. “The impact of greenspace size on the extent of local nocturnal air temperature cooling in London”. In: *Urban Forestry & Urban Greening* 16 (2016), pp. 160–169.

[68] C.J.G. Morris, Ian Simmonds, and Neil Plummer. “Quantification of the Influences of Wind and Cloud on the Nocturnal Urban Heat Island of a Large City”. In: *Journal of Applied Meteorology* 40.2 (2001), pp. 169–182.

[69] NASA. *MODIS - Moderate Resolution Imaging Spectroradiometer*. Accessed October 10, 2017. Retrieved from `https://modis.gsfc.nasa.gov/`. 2016.

[70] NASA. *Satellites Pinpoint Drivers of Urban Heat Islands in the Northeast*. Accessed October 10, 2017. Retrieved from `https://www.nasa.gov/topics/earth/features/heat-island-sprawl.html`. 2010.

[71] M. Jay Norton. "Knowledge Discovery in Databases". In: *Library Trends* 48.1 (1999), p. 9.

[72] MO Obiakor, CD Ezeonyejiaku, and TC Mogbo. "Effects of Vegetated and Synthetic (Impervious) Surfaces on the Microclimate of Urban Area". In: *Journal of Applied Sciences and Environmental Management* 16.1 (2012), pp. 85–94.

[73] Timothy R. Oke. *Boundary Layer Climates*. p. 291. Routledge, 2002.

[74] Timothy R. Oke. "City Size and the Urban Heat Island". In: *Atmospheric Environment (1967)* 7.8 (1973), pp. 769–779.

[75] Timothy R. Oke. "The energetic basis of the urban heat island". In: *Quarterly Journal of the Royal Meteorological Society* 108.455 (1982), pp. 1–24.

[76] Juha Oksanen et al. "Methods for deriving and calibrating privacy-preserving heat maps from mobile sports tracking application data". In: *Journal of Transport Geography* 48 (2015), pp. 135–144.

[77] Marie S. O'Neill and Kristie L. Ebi. "Temperature Extremes and Health: Impacts of Climate Variability and Change in the United States". In: *Journal of Occupational and Environmental Medicine* 51.1 (2009), pp. 13–25.

[78] OpenStreetMap contributors. *OpenStreetMap*. Accessed October 10, 2017. Retrieved from `https://www.openstreetmap.org`. 2004.

[79] World Health Organization et al. "Urban population growth". In: *Global Health Observatory. World Health Organization.* (2013).

[80] Swades Pal and Sk. Ziaul. "Detection of land use and land cover change and land surface temperature in English Bazar urban centre". In: *The Egyptian Journal of Remote Sensing and Space Science* 20.1 (2017), pp. 125–145.

[81] Jonathan A. Patz et al. "Impact of regional climate change on human health". In: *Nature* 438.7066 (2005), p. 310.

[82] J. Ross Quinlan. *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann Publishers, 1993.

[83] Umamaheshwaran Rajasekar and Qihao Weng. "Urban heat island monitoring and analysis using a non-parametric model: A case study of Indianapolis". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 64.1 (2009), pp. 86–96.

[84] Raschka, Sebastian. *StackingClassifier – Ensemble-learning meta-classifier for stacking*. Accessed February 7, 2018. Retrieved from `https://rasbt.github.io/mlxtend/user_guide/classifier/StackingClassifier/`. 2018.

[85] Jinyoung Rhee, Seonyoung Park, and Zhenyu Lu. "Relationship between land cover patterns and surface temperature in urban areas". In: *GIScience & remote sensing* 51.5 (2014), pp. 521–536.

[86] Philip J. Rhodes et al. "Uncertainty Visualization Methods in Isosurface Volume Rendering". In: *Eurographics 2003, Short Papers*. 2003, pp. 83–88.

[87] Maria Riveiro. "Evaluation of Uncertainty Visualization Techniques for Information Fusion". In: *2007 10th International Conference on Information Fusion*. July 2007, pp. 1–8.

[88] F. Salamanca et al. "Assessing summertime urban air conditioning consumption in a semiarid environment". In: *Environmental Research Letters* 8.3 (2013).

[89] M. Santamouris, A. Synnefa, and T. Karlessi. “Using advanced cool materials in the urban built environment to mitigate heat islands and improve thermal comfort conditions”. In: *Solar Energy* 85.12 (2011). Progress in Solar Energy 2, pp. 3085–3102.

[90] C. Sarrat et al. “Impact of urban heat island on regional atmospheric pollution”. In: *Atmospheric Environment* 40.10 (2006), pp. 1743–1758.

[91] Amit P. Sawant, Matti Vanninen, and Christopher G. Healey. “PerfViz: A Visualization Tool for Analyzing, Exploring, and Comparing Storage Controller Performance Data.” In: *Visualization and Data Analysis*. Vol. 6495. 2007, p. 07.

[92] Dieter Scherer et al. “Quantification of heat-stress related mortality hazard, vulnerability and risk in Berlin, Germany”. In: *Journal of the Geographical Society of Berlin* 144.3-4 (2014), pp. 238–259.

[93] Daniel Seebacher et al. “Visual Analysis of Spatio-Temporal Event Predictions: Investigating the Spread Dynamics of Invasive Species”. In: *Symposium on Visualization in Data Science (VDS) at IEEE VIS 2017*. 2017.

[94] Bin Shao et al. “Prediction and Visualization for Urban Heat Island Simulation”. In: *Transactions on Edutainment VI*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 1–11.

[95] Marina Stathopoulou and Constantinos Cartalis. “Daytime urban heat islands from Landsat ETM+ and Corine land cover data: An application to major cities in Greece”. In: *Solar Energy* 81.3 (2007), pp. 358–368.

[96] Weizhong Su, Chaolin Gu, and Guishan Yang. “Assessing the Impact of Land Use/Land Cover on Urban Heat Island Pattern in Nanjing City, China”. In: *Journal of Urban Planning and Development* 136.4 (2010), pp. 365–372.

[97] T. Susca, S.R. Gaffin, and G.R. Dell’Osso. “Positive effects of vegetation: Urban heat island and green roofs”. In: *Environmental Pollution* 159.8 (2011). Selected papers from the conference Urban Environmental Pollution: Overcoming Obstacles to Sustainability and Quality of Life (UEP2010), 20-23 June 2010, Boston, USA, pp. 2119–2126.

[98] A. Synnefa, M. Santamouris, and K. Apostolakis. “On the development, optical properties and thermal performance of cool colored coatings for the urban environment”. In: *Solar Energy* 81.4 (2007), pp. 488–497.

[99] Alexandra D. Syphard et al. “Predicting spatial patterns of fire on a southern California landscape”. In: *International Journal of Wildland Fire* 17.5 (2008), pp. 602–613.

[100] Helmut E. Landsberg. “The Urban Heat Island”. In: *The Urban Climate*. Vol. 28. International Geophysics. Academic Press, 1981, pp. 83–126.

[101] Haider Taha et al. “Residential cooling loads and the urban heat island – the effects of albedo”. In: *Building and environment* 23.4 (1988), pp. 271–283.

[102] Quantum GIS Development Team. *QGIS Geographic Information System. Open Source Geospatial Foundation*. 2013.

[103] Waldo R. Tobler. “Choropleth Maps Without Class Intervals?” In: *Geographical analysis* 5.3 (1973), pp. 262–265.

[104] Christian Tominski, James Abello, and Heidrun Schumann. “Axes-Based Visualizations with Radial Layouts”. In: *Proceedings of the 2004 ACM symposium on Applied computing*. ACM. 2004, pp. 1242–1247.

[105] Christian Tominski, Jonathan F. Donges, and Thomas Nocke. “Information Visualization in Climate Research”. In: *Information Visualisation (IV), 2011 15th International Conference on*. IEEE. 2011, pp. 298–305.

[106] Christian Tominski, Petra Schulze-Wollgast, and Heidrun Schumann. "3D Information Visualization for Time Dependent Data on Maps". In: *Ninth International Conference on Information Visualisation (IV'05)*. July 2005.

[107] Andhang Rakhmat Trihamdani et al. "Impacts of Land use Changes on Urban Heat Islands in Hanoi, Vietnam: Scenario Analysis". In: *Procedia Engineering* 198 (2017). Urban Transitions Conference, Shanghai, September 2016, pp. 525–529.

[108] Weather Underground. *Personal Weather Station Network*. Accessed November 10, 2017. Retrieved from `https://www.wunderground.com/`. 2016.

[109] Weather Underground. *Weather Underground API*. Accessed November 10, 2017. Retrieved from `https://www.wunderground.com/weather/api`. 2016.

[110] Jarke J. Van Wijk and Edward R. Van Selow. "Cluster and Calendar based Visualization of Time Series Data". In: *Information Visualization, 1999. (Info Vis '99) Proceedings. 1999 IEEE Symposium on*. IEEE. 1999, pp. 4–9, 140.

[111] James A. Walsh et al. "Temporal-Geospatial Cooperative Visual Analysis". In: *Big Data Visual Analytics (BDVA), 2016*. IEEE. 2016, pp. 1–8.

[112] Kathrin Ward et al. "Heat waves and urban heat islands in Europe: A review of relevant drivers". In: *Science of The Total Environment* 569-570 (2016), pp. 527–539.

[113] Marc Weber, Marc Alexa, and Wolfgang Müller. "Visualizing Time-Series on Spirals". In: *Proceedings of the IEEE Symposium on Information Visualization 2001 (INFOVIS'01)*. 2001.

[114] Edward J. Wegman. "Hyperdimensional Data Analysis Using Parallel Coordinates". In: *Journal of the American Statistical Association* 85.411 (1990), pp. 664–675. ISSN: 01621459. URL: `http://www.jstor.org/stable/2290001`.

[115] Hadley Wickham et al. "Glyph-maps for Visually Exploring Temporal Patterns in Climate Data and Models". In: *Environmetrics* 23.5 (2012), pp. 382–393.

[116] David H. Wolpert. "Stacked generalization". In: *Neural networks* 5.2 (1992), pp. 241–259.

[117] Wujun Xi and Ping He. "Prediction of Urban Heat Island Intensity in Chuxiong City with Backpropagation Neural Network". In: *Advances in Neural Network Research and Applications*. Springer, 2010, pp. 29–36.

[118] Li Yang et al. "Research on Urban Heat-Island Effect". In: *Procedia Engineering* 169 (2016), pp. 11–18.

Appendix

Appendix A

PCP: Aggregation Types

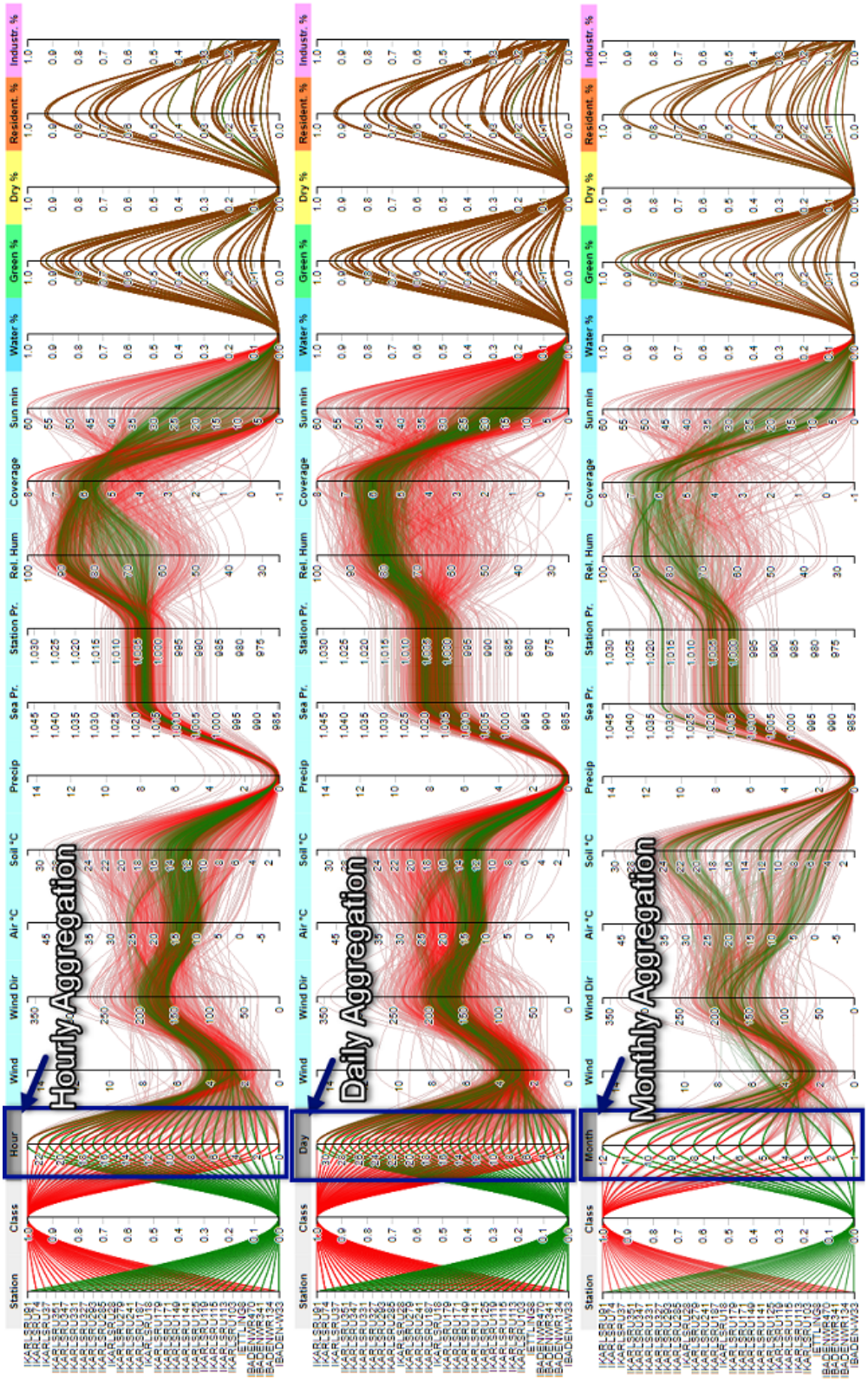


Figure A.1: Unary temporal aggregation methods for PCP: hourly, daily, and monthly aggregation.

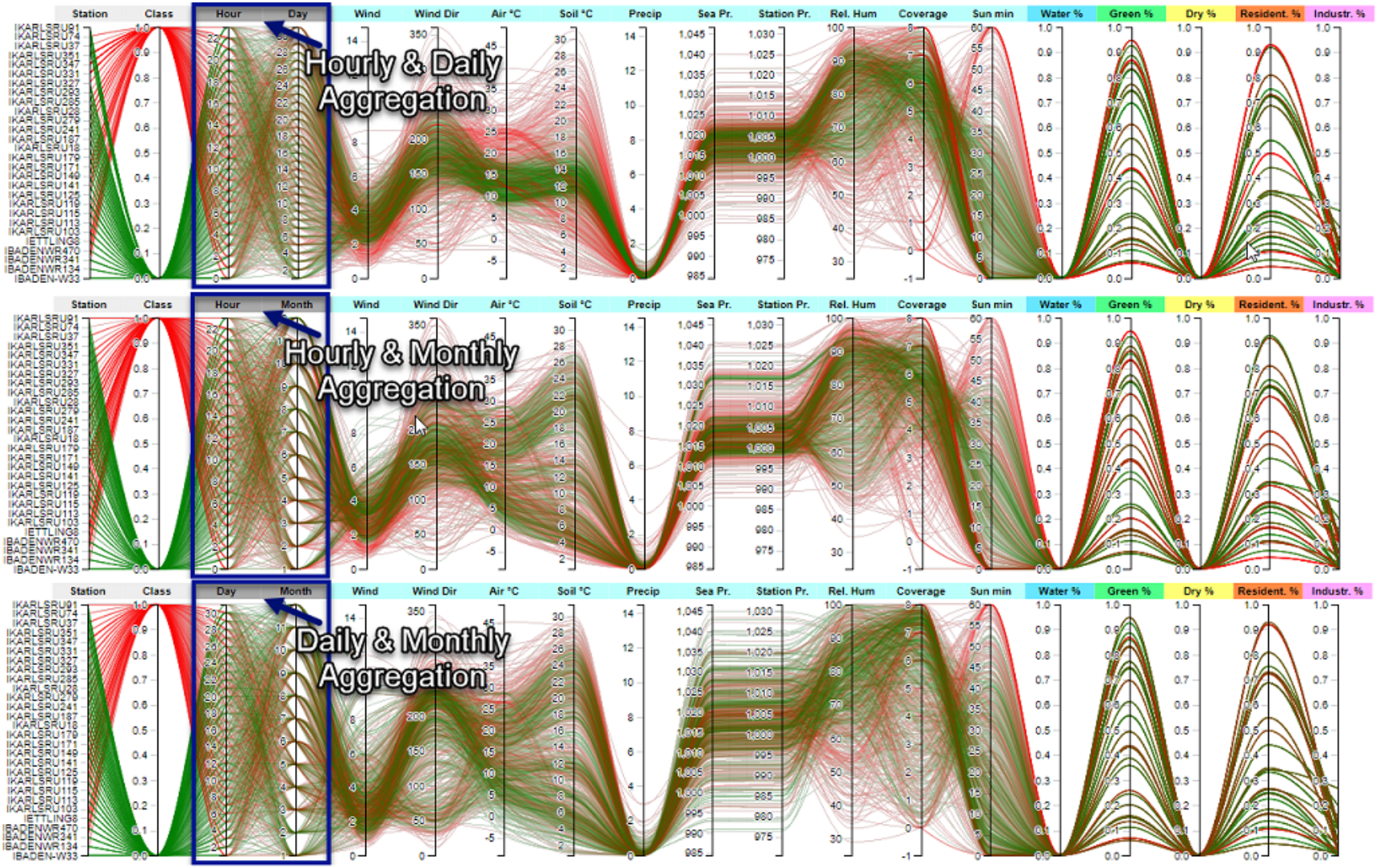


Figure A.2: Binary temporal aggregation methods for PCP: hourly-daily, hourly-monthly, and daily-monthly aggregation.

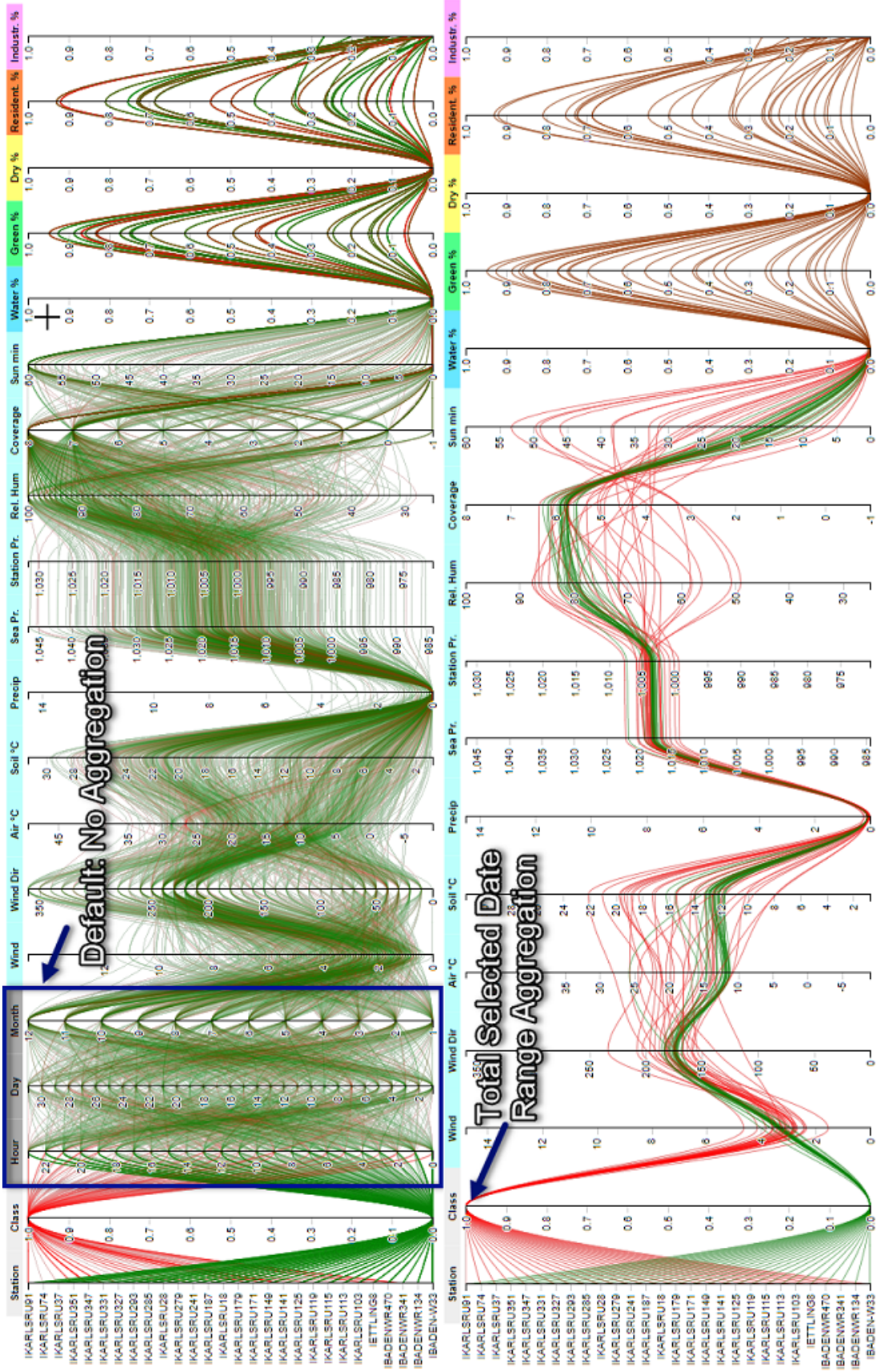


Figure A.3: Aggregation methods for PCP: by default no aggregation is applied. If a date range is selected, no temporal dimension is displayed and only a single value per class is visualized.

Appendix B

OSM: categorization of area types

Area Type	Name	OSM tag *(with hyperlink)*
Green Area	Allotments	landuse=allotments
	Cemetery	landuse=cemetery
	Farmland	landuse=farmland
	Forest	landuse=forest
	Grass	landuse=grass
	Garden	landuse=grass
	Greenfield	landuse=greenfield
	Greenhouse Horticulture	landuse=greenhouse_horticulture
	Meadow	landuse=meadow
	Orchard	landuse=orchard
	Park	leisure=park
	Plant Nursery	landuse=plant_nursery
	Recreation Ground	landuse=recreation_ground
	Scrub	natural=scrub
	Village Green	landuse=village_green
	Vineyard	landuse=vineyard
Water Area	Basin	landuse=basin
	Reservoir	landuse=reservoir
Residential Area	Churchyard	landuse=churchyard
	Commercial	landuse=commercial
	Farm	landuse=farm
	Farmyard	landuse=farmyard
	Office	building=office
	Public Administration	landuse=civic
	Religious	landuse=religious
	Residential	landuse=residential
	Retail	landuse=retail
	Traffic Island	landuse=traffic_island
Industrial Area	Construction	landuse=construction
	Garages	landuse=garages
	Industrial	landuse=industrial
	Military	military
	Railways	railways
Dry Area	Animal Keeping	landuse=animal_keeping
	Brownfield	landuse=brownfield
	Highway	highway
	Landfill	waste_processing
	Quarry	landuse=quarry

Table B.1: Division of OSM tags into five area categories.

Appendix C

Seasonal UHI scenario overviews

The figures that are displayed in this appendix section present the prediction result for different meteorological scenarios. To provide a seasonal overview, we separately selected day and night hours for each season by considering the sunset and sunrise hours. Every overview visualization is divided into a season section, and they are subdivided into day and night hours for the respective season of the year. For this, we applied the time filter using suitable hour ranges on the calendar visualization (see Table C.1). These were typical seasonal hour ranges in Germany in 2016.

Season	**Day hours**	**Night hours**
Spring	7 a.m. - 8 p.m.	8 p.m. - 7 a.m.
Summer	6 a.m. - 9 p.m.	9 p.m. - 6 a.m.
Autumn	7 a.m. - 6 p.m.	6 p.m. - 7 a.m.
Winter	8 a.m. - 5 p.m.	5 p.m. - 8 a.m.

Table C.1: Suitable hour range selection for every season of the year.

Figure C.1 provides the calendar presentation that shows the UHI data without processing a prediction. In contrast, Figure C.2 provides the calendar overview for a scenario that simulates a colder and more rainy weather situation (see Table C.2 for exact feature changes).

Wind speed	*Air temperature*	*Soil temperature*	*Precipitation*	*Air pressure (station level)*	*Air pressure (sea level)*	*Relative humidty*	*Cloud coverage*	*Sunshine duration*
-5 $\frac{m}{s}$	+5 °C	+5 °C	-10 $\frac{mm}{h}$	+10 hPa	+10 hPa	-10 %	-2	+10 min

Table C.2: Feature vector for predicting colder and more rainy weather scenario.

Analogously, we selected a feature set for a reverse situation: warmer and dryer weather scenario. Table C.3 contains the exact variable changes that were applied to receive the result that is presented in Figure C.3. In both cases, missing features indicate accordingly, that these values were not changed (including area features, wind direction).

Wind speed	*Air temperature*	*Soil temperature*	*Precipitation*	*Air pressure (station level)*	*Air pressure (sea level)*	*Relative humidty*	*Cloud coverage*	*Sunshine duration*
+5 $\frac{m}{s}$	-5 °C	-5 °C	+8 $\frac{mm}{h}$	-10 hPa	-10 hPa	+15 %	+2	-15 min

Table C.3: Feature vector for predicting warmer and dryer weather scenario.

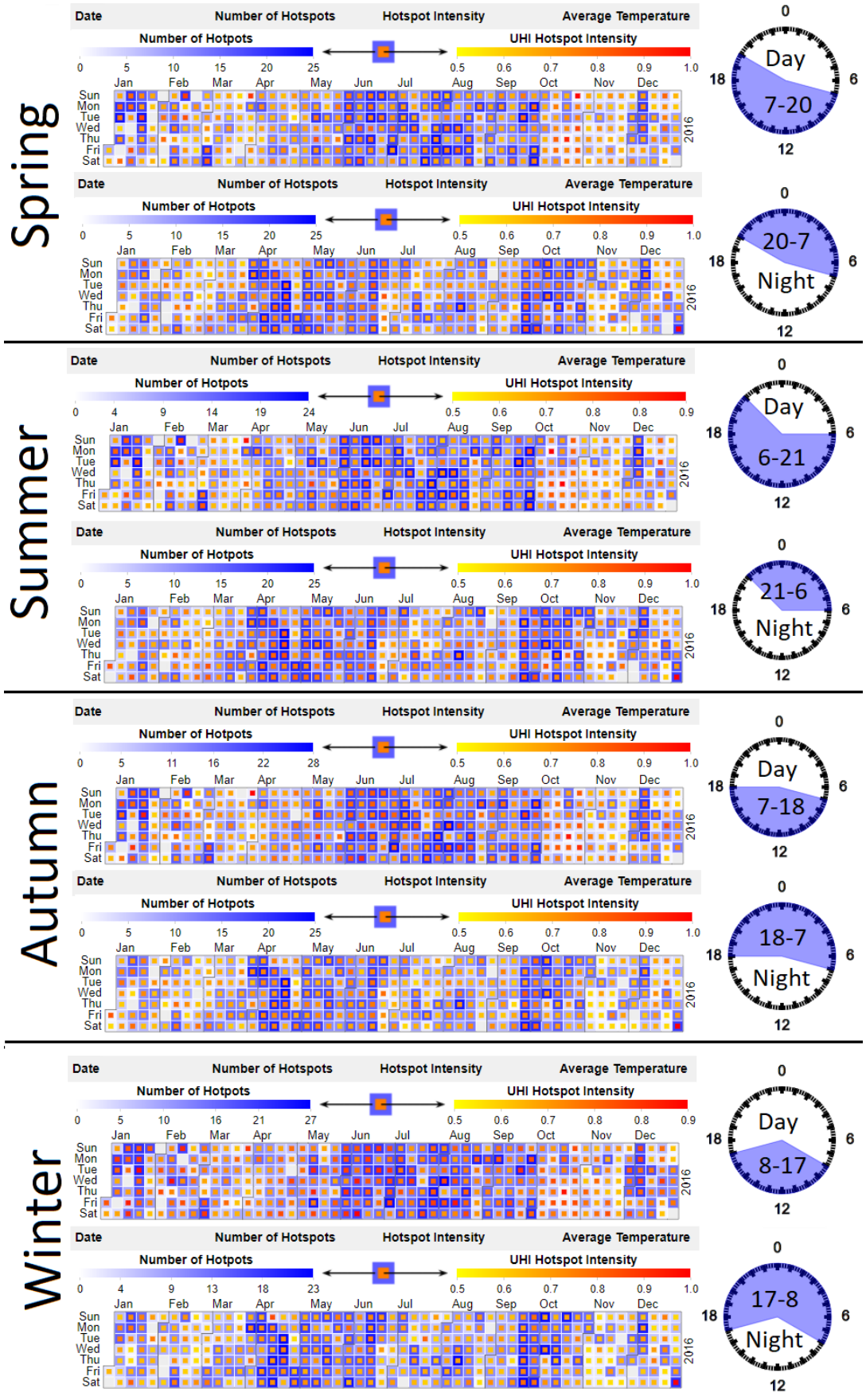


Figure C.1: Standard features (no prediction)

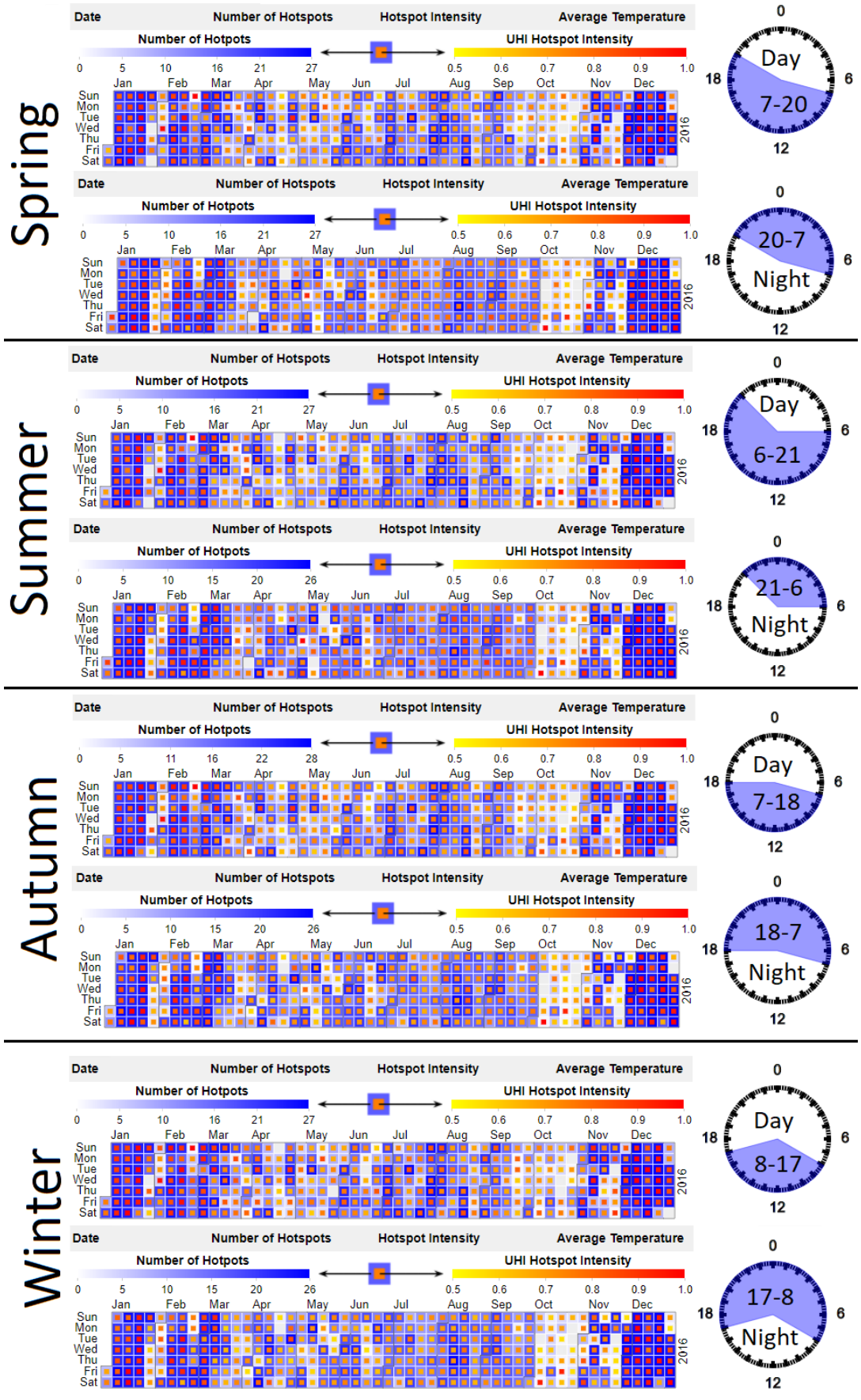


Figure C.2: Decreased temperature and more precipitation (prediction).

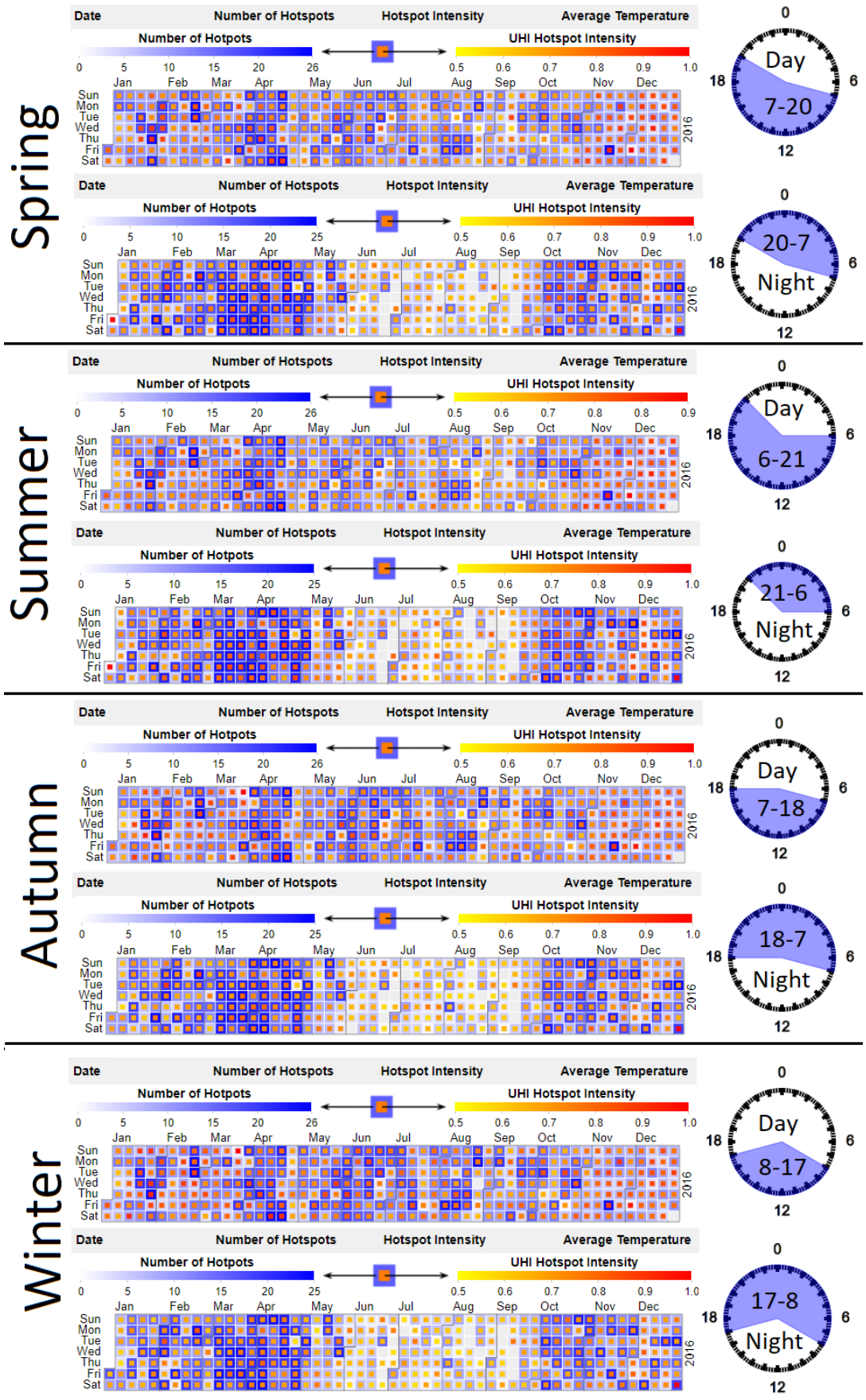


Figure C.3: Increased temperature, less precipitation (prediction).