FAKULTÄT FÜR !NFORMATIK
Faculty of Informatics

# Analysis and Visualization of Vienna's Parking Enforcement Data

DIPLOMARBEIT

zur Erlangung des akademischen Grades

**Diplom-Ingenieur**

im Rahmen des Studiums

**Business Informatics**

eingereicht von

**Julian Lehner, BSc**
Matrikelnummer 01225997

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Univ.Prof. Dipl.-Ing. Dr.techn. Peter Filzmoser

Wien, 17. November 2019

| Julian Lehner | Peter Filzmoser |
|---|---|

FAKULTÄT FÜR !NFORMATIK
Faculty of Informatics

# Analysis and Visualization of Vienna's Parking Enforcement Data

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

**Diplom-Ingenieur**

in

**Business Informatics**

by

**Julian Lehner, BSc**
Registration Number 01225997

to the Faculty of Informatics

at the TU Wien

Advisor: Univ.Prof. Dipl.-Ing. Dr.techn. Peter Filzmoser

Vienna, 17th November, 2019

| Julian Lehner | Peter Filzmoser |
|---|---|

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.ac.at

# Erklärung zur Verfassung der Arbeit

Julian Lehner, BSc

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 17. November 2019

Julian Lehner

# Acknowledgements

I am very grateful to have the possibility to write this thesis at the Technical University of Vienna. I want to thank my academic advisor Prof. Peter Filzmoser for making it possible to realize this topic as a thesis.

I want to thank everyone who is involved in proof-reading this thesis and thanks for making this document something to be proud of.

Thanks to the responsible people of the City of Vienna to give me the possibility of using the data for this thesis.

A special thank you to Ronald who got me on the idea for choosing this topic for my thesis. And thanks to all other colleagues at my job, without your help I could never have been able to gain the necessary knowledge and know-how that is needed for this topic.

Thank you, to all the colleagues and awesome people I met at the university during my study. And a special thanks to those of you, whom I become friends.

I also want to thank my family, especially my parents Gabriele and Daniel, for all your patience, assistance and help. Especially thank you for your monetary support during my whole life, this made it possible to have at least one worry less.

Finally, thank you Claudia!
Thank you for the support you are giving me.
Thank you for the fun we are heaving together.
Thank you for every smile you give me.
Thank you for being with me.
I love you.

# Kurzfassung

Parkraumüberwachung ist ein, von Parkwächtern mit modernen elektronischen Geräten, unregelmäßig durchgeführter Prozess. Die gesammelten Daten sind in täglichen Protokolldateien für die Stadt Wien gespeichert und beinhalten persönliche Daten der Exekutive gemeinsam mit im Klartext vorhandenen Kennzeichen der Fahrzeuge. Im Jahr 2018 beträgt die Gesamtanzahl an Einträgen in den Protokolldateien mehr als 19 Millionen. Eine Übereinkunft mit der Stadt Wien berechtigt die Nutzung der Daten zur weiteren Analyse, solange bestimmte vordefinierte Bestimmungen erfüllt werden.

Zurzeit werden die Daten nur vereinfacht analysiert, aber die große Menge an Einträgen ermöglicht es zusätzliche Informationen über das Parkverhalten in der Stadt zu erlangen. Für zukünftige Entscheidungen und Planungen können solche Informationen von hohem Interesse sein. Das Ziel dieser Arbeit ist es, die verfügbaren Daten besser zu verstehen und geeignete Visualisierungsmethoden durch eine statistische Analyse bereitzustellen.

Die Daten werden während des Imports anonymisiert. Anschließend werden die Daten während der Verarbeitung mit offenen Daten um zusätzliche Standortinformationen erweitert. Für jede einzelne Zeile wird der nächstmögliche GPS-Koordinatenpunkt näherungsweise berechnet.

Für die Analyse und Visualisierung werden verschiedene Detailgrade in den Dimensionen Zeit und Ort definiert. Das Strafenverhältnis und die Herkunft der Autos sind die Hauptindikatoren und werden eingehender untersucht. Verschiedene auf der Literatur basierende Visualisierungen werden ausgewählt, um jede Eigenschaft in unterschiedlichen Kombinationen von Zeit und Ort darzustellen. Jedes einzelne Diagramm zeigt eine Herangehensweise um qualifizierte Aussagen zu treffen und zusätzliches Wissen über das Parkverhalten in einem bestimmten Bereich und Zeitraum zu erhalten.

Grundsätzlich ist die gegebene Art der Daten durchaus geeignet, um zusätzliche Informationen über das Parkverhalten in einer Stadt zu erhalten. Das Ergebnis hängt von der zur Verfügung stehenden Datenmenge, den individuellen Gegebenheiten und der damit verbundenen Herangehensweise ab.

# Abstract

Parking enforcement is an intermittent process performed by parking wardens with modern electronic devices. The collected data is stored in daily protocol files for the City of Vienna and contains personal data of the executive forces together with plain text license plates of cars. In the year 2018, the total amount of entries in the protocol files is more than 19 million. An agreement with the City of Vienna allows the usage of this data for further analysis if specific predefined conditions are fulfilled.

Currently the data is only analyzed in a simple way, but the huge amount of entries provides the possibility to gain additional knowledge about the parking behavior in the city. For future decision making and planing, such information can be of high interest. The goal of this thesis is to better understand the available data and to provide suitable visualization methods by performing a statistical analysis.

The data is anonymized during the import. Afterwards, the preprocessing phase is used to enhance the data with additional location information by using open data. For every available entry line, the closest possible GPS coordinate point is estimated.

Different levels of detail are defined in the dimensions of time and place for the analysis and visualization. The penalty ratio and the origin of cars are identified as the main indicators and are investigated in more detail. Various visualizations based on the literature are chosen to depict each feature in different combinations of time and place. Each diagram provides the approach for achieving qualified statements and gaining additional knowledge about the parking behavior in a specific area and time period.

Generally, the available type of data is definitely suitable to gain additional information about the parking behavior in a city. The outcome depends on the available amount of data, the individual circumstances and the related approaches.

# Contents

**Kurzfassung** **ix**

**Abstract** **xi**

**Contents** **xiii**

**1 Introduction** **1**
1.1 Motivation . . . 1
1.2 Problem Statement . . . 2
1.3 Aim of the Work . . . 2
1.4 Methodological Approach . . . 4
1.5 Structure of the Work . . . 4

**2 State of the Art** **7**
2.1 Working with Coordinates . . . 7
2.2 Parking . . . 9
2.3 Analysis and Visualization . . . 9
2.4 Time Series Analysis . . . 16
2.5 Data Preprocessing . . . 19

**3 Preprocessing the Data** **23**
3.1 Anonymization . . . 23
3.2 Data Source and Licensing . . . 25
3.3 Structure of the Data . . . 27
3.4 Data Transformation . . . 32

**4 Evaluation** **43**

**5 Analysis and Visualization** **49**
5.1 Relevant Indicators . . . 49
5.2 Levels of Detail . . . 52
5.3 Used Data Subsets . . . 56
5.4 Penalty Ratio . . . 58
5.5 Origin of the Cars . . . 78

5.6 Parking Density . . . . . 91

**6 Critical Reflection** **97**

**7 Summary** **101**
7.1 Future Work . . . . . 102

**List of Figures** **105**

**List of Tables** **109**

**Bibliography** **111**

CHAPTER 1

# Introduction

On the whole globe, the totally generated amount of data sums up to an enormous value of multiple zettabytes and this value continuous to increase. [1]

While professional companies illuminate every aspect of their data, it is often only a means for an end to achieve daily process goals for other organizations. In the field of public authorities, data is often available on open data platforms and hence it is provided to the public for further analysis. If the data is derived from the area of executive forces or contains personal data, the option of sharing the data does not exist.

## 1.1 Motivation

A field which is performed by an executive force and does also collect and rely on personal data is the process of parking enforcement. The author of this thesis is working at a company which is responsible for the backend server and the used software in the area of parking enforcement of the City of Vienna. In this city, the mobile parking solution is heavily used, which is the reason why nearly every car has to be checked electronically if it is parked correctly or not. These checks are performed by parking wardens on a specific control device. Every permission request and also all penalties are sent to the server. A daily chronological protocol file which contains all entries of all wardens is generated. More than 400 parking wardens are working on the streets of Vienna on some days and so the amount of entry lines in a single protocol file can be more than 100000 a day.

At the end of each day, the protocol file is forwarded to the City of Vienna. The transfered files are used for some basic analysis and simple conclusions by the officials. Thinking

[1] https://de.statista.com/statistik/daten/studie/267974/umfrage/prognose-zum-weltweit-generierten-datenvolumen/, Accessed 2019-10-15

about future decision making in the area of parking enforcement or planing parking lots in the city, additional information about the parking behavior of cars is of high interest.

The areas with specific parking rules in the city have been extended over time and also the amount of collected data is growing. This trend was the trigger for the main question that arises for the author, what kind of information could be gained by an extended analysis of this data. This question is the main motivation for this thesis and results in the formulation of the problem statement.

## 1.2 Problem Statement

At the end of each day, a protocol file, including all the requests for parking permissions and all penalty entries, is delivered to the City of Vienna by their responsible contractor company. This huge amount of data is currently only used for simple analysis, but there is a high potential of gaining additional knowledge about the parking situation in the city. For a further analysis and a better understanding of the data, at least the know-how, resources or necessary infrastructure is not available.

The data itself does only contain a limited set of fields, but it includes strictly personal information. On the one hand there is an indirect identifier of the warden's username and on the other hand it contains the complete license plate of a parked car. There is no open access to this data and therefore an agreement has been signed between the author and the city about conditions which have to be fulfilled when working with the data. This agreement includes the steps to anonymize the above data fields.

Another big issue is, that even if there are a lot of requests done by the wardens, there is no position information given for the permission requests. The reason is an internal prohibition to use the GPS functionality of the warden's electronic device. The entered location of a fined car of a penalty is the only available position information. Additional position data has high relevance because otherwise the entered district is the only real position information that is available, but an extra level of detail is of high interest. Furthermore, a good visualization of the results will make it easier for the City of Vienna to understand their own data better.

## 1.3 Aim of the Work

The main goal of this thesis is to provide mechanisms and approaches to gain additional knowledge about the parking behavior based on a parking enforcement process. Therefore, the data of the City of Vienna should be used to perform a statistical analysis and provide visualizations for a better understanding of the results.

This thesis describes in detail the process of preprocessing the source data to gain an enhanced and well formated dataset for further analysis. Afterwards it identifies suitable levels of detail in the data and by external circumstances. Then it performs a statistical

analysis and shows various visualizations by applying the data on different combinations of the dimensions of time and place.

One main research question arises for the described goals and will be answered by a set of sub-questions:

**What kind of information can be gained by analyzing the parking enforcement data about the parking behavior in the City of Vienna?**

**(1)** Are there possibilities to gain position information for all the data, especially for the permission requests in contrast to the penalty entries? To get an better outcome on this task, open data from the internet is used to enrich the information content and increase the entropy of the resulting data. Besides the algorithm for the position estimation of the permission requests, it is also important for how many entries no position information can be gained.

**(2)** What are the geographical and temporal levels of detail that can be described by the data? After the results of question (1) the levels of detail can be identified and described. It is also possible that the outcome will be different for various areas of the City or specific time periods.

**(3)** Which indicators can be identified in the data, what are their distributions and how do they change related to question (2)? To get an answer to this questions an analysis of the available attributes of each entry in the data is needed to combine them with the position and time information gained in the previous research questions. The outcome should be qualified statements about the parking behavior in the City of Vienna, supported by the application of statistical modeling techniques. Example conclusions could affect the origin of the car license plates or the relationship between the number of requests and the number of penalties.

**(4)** Is the type of data suitable to get an estimation of the parking density and to what extent can it be classified? A special indicator and a possible outcome of the analysis is the parking density in individual regions of the City for various periods of time. Depending on the available data, a feasible classification for the parking density could be Low, Medium and High, especially if the spreading of the calculated values differ too much.

**(5)** What are suitable visualizations of the results? To fulfill this task a widespread literature review has to be done to identify suitable Descriptive statistics and visualization methods to achieve the most meaningful representation of the data.

Referring to the data and the type of job that is done by the parking wardens of the City of Vienna, various assumptions will have to be made to harmonize the data for analysis to be able to gain proper results.

## 1.4 Methodological Approach

The methodological approach for satisfying the mentioned research questions consists of the following steps:

1. Performing a widespread literature review with detailed investigations on the following topics:
   - Studies and existing research about cars and traffic in cities, especially about parking
   - Working with position information like GPS coordinates and state of the art technologies in this area
   - Statistical modeling approaches, especially for the analysis of spatio-temporal data and time series
   - Descriptive statistics and information visualization methods for data applicable on a city or other geographical regions as well as for different time periods
   - Data transformation techniques for preprocessing the data
2. The implementation and evaluation of methods to achieve the following targets:
   - Anonymization of the provided source data to fit the customer needs defined in a separate contract
   - Preprocessing of the gained data to acquire as much position data as possible for later analysis. While preprocessing the data it can also be enriched with open data that is available on the internet to increase the entropy of the analyzable data
3. Identification of suitable indicators and analysis of the given data. Here the gained information from the previous steps is used to apply different statistical methods and modeling approaches on the data. The outcome is then transferred into formal statements about the characteristics of Vienna's parking behavior.
4. Selection of specific visualization methods based on the literature research to achieve the most appropriate representation of the data. This representations should be as easy to understand as possible. Additionally the visual parts of the statistical modeling approaches are placed in this section.

## 1.5 Structure of the Work

The thesis is structured as hereafter: Chapter 2 gives an overview of the widespread literature research with detailed investigations on various topics. It contains definitions in the context of working with coordinate points and shows previous research in the area of parking. Additionally investigated topics are about statistical analysis and visualization

techniques as well as time series analysis. At the end some existing research is presented about data preprocessing.

In the following Chapter 3 describes the whole process of preprocessing the data. It contains detailed information and descriptions of the structure of the available data, but also describes the applied processes of anonymization, gaining additional open data and the whole data transformation.

Subsequently, in Chapter 4 an evaluation of the data preprocessing is presented, which concentrates on the outcome of the coordinate point estimation for all the entry lines.

The next Chapter 5 contains the complete statistical analysis of the data including the related and applied visualization approaches. It describes the relevant indicators and the levels of detail. Afterwards, two separate sections are presenting the results for the penalty ratio and the origin of the cars before the last section concentrates on the parking density.

Chapter 6 critically reflects the different topics and the implementation of this thesis. It especially describes problem situations and mentions necessary assumptions in some parts of the document.

Finally, Chapter 7 concludes the thesis with the summary and also contains a subsection with a discussion on future work.

CHAPTER 2

# State of the Art

This chapter is used to introduce various topics that are relevant for this thesis and to present the respective and used literature. One of the targets of this thesis is obtaining additional position information for requested and fined car plates during the process of parking enforcement. Therefore the first section starts with an introduction on position coordinates and the corresponding standards in the applied areas. It is followed by current research in the context of parking and also other main areas are described. The next section focuses on various visualization methods and example applications which exist in the literature and how they could be used for analyzing the given data. Then a more detailed overview is given of relevant approaches in the area of time series analysis. During the parking and enforcement process, the daytime and exact timestamps can make the difference on whether someone is allowed to park or not. This is the reason why the topic of time series analysis has an own chapter, while it is basically an extension of the previous section. Finally, the chapter ends with a section about steps in the data preprocessing phase, which is necessary before starting the data analysis.

## 2.1 Working with Coordinates

The term "coordinates" is basically known from maths, where two values define a point on a two-dimensional plane surface. When working with position data on a map, this concept basically remains the same, even if sometimes an additional third dimension exists. This concept can also be applied to places in the real world, which are often shown on various maps. An example would be the start and end point of a street, or the midpoint of a house or property. The definition of such a real world position can be set in different ways, depending on the purpose and the person who is working with it. The main base for receiving coordinates of a position on an earth is a Global Navigation Satellite System (GNSS) like described in Hofmann-Wellenhof et al. [2007]. Such a system is the umbrella term for all the necessary parts that are involved in the process of determining a certain

position on earth. This includes among others the measuring stations on earth, the satellites in the orbit or the used maths and signals for communication. Another condition for a correct outcome in this process, as shown in Merkel [1996 (accessed April 20, 2019], is the need of a same used reference system for the measuring stations and the satellites. For the well known Global Positioning System (GPS), the definition of the reference system WGS84 of the World Geodetic System (WGS) is used, as stated in Görres [2010 (accessed April 20, 2019]. Besides GPS, also other GNSS exist, like GLONASS or Galileo [Hofmann-Wellenhof et al., 2007]. GPS is used for various kinds of navigation purposes as stated in Weber et al. [1995], but it is also known from smartphones or car navigation devices. The last step before working with coordinates is to think about the target coordinate system the values should be in. The standard for the GPS WGS84 system is EPSG:4326[1], which holds the coordinate values in the well known longitude and latitude format. But for displaying the coordinates in a map, in the same reference system the EPSG:3857[2] standard, also called Pseudo-Mercator or Spherical Mercator, is used. The main difference is that the former holds the geographic coordinate while the latter is a projected coordinate. There are many more other possible target coordinate systems, two of them are stated in the address service of the City of Vienna, which is used in this thesis:

- EPSG:25834[3] - A pan-European target coordinate system
- EPSG:31256[4] - The national coordinate system of east Austria

When using Open Government Data (OGD) provided by authorities, used data formats are most likely standardized. In case of the used address service[5], the standardized format that has been used is called GeoJSON. It is defined by the Internet Engineering Task Force (IETF) in Gillies et al. [2016] and it is based on JavaScript Object Notation (JSON). Basically, when querying a specific address of the address service, the successful response is a GeoJSON formatted object. This object holds various features which include details about the queried address. Among others, these features are the street name and number, district, category information and the postal code. It also includes a bounding box (bbox), a 2*n array which holds the coordinates of the address, where n is the dimension. The array values start with all axes of the most southwesterly point, followed by all axes of the more northeasterly point. This means that for each address there is a quadrangular area given instead of a single position.

[1] https://epsg.io/4326, Accessed 2019-03-31
[2] https://epsg.io/3857, Accessed 2019-03-31
[3] https://epsg.io/25834, Accessed 2019-03-31
[4] https://epsg.io/31256, Accessed 2019-03-31
[5] https://digitales.wien.gv.at/site/files/2019/01/adressservice-doku.pdf, Accessed 2019-03-31

## 2.2 Parking

Regarding the subject of parking, a lot of research has already be done in the past. This is mainly concentrating on two different domains where parking is of high interest:

- Understanding traffic flow in a city
- Gaining knowledge for town planning, especially transportation planning

As an example for the first domain there is a study in Shoup [2006] about the cruising of cars in a city. Cruising describes the driving of cars along the streets until a free parking space can be found on the curbs. These cars seem nearly invisible because all those cars are mixed up with others which are really going somewhere. Cruising produces more traffic and therefore more pollution problems. The study defines parameters and formulas to determine costs and time savings while cruising on the streets. These parameters and formulas are used to investigate, if parking a car for higher costs off-street or cruising longer until a cheaper parking space is found, is the correct choice.

In the other domain, which is about town planning, some other aspects are relevant in the research. The article in Wigan [1975] gives an overview about different models in the parking area which are related to stages in transportation planning. To highlight some of the addressed aspects, there is the flow capacity of traffic, especially at overload points, which needs to be handled. Another model considers the distribution and modal split, where the ratios of trips from one zone to another zone are taken into account. Parking and traffic restraints are also part of their analysis. In the end it also considers wider transport planning aspects like social influences.

## 2.3 Analysis and Visualization

Visualization is "The representation of an object, situation, or set of information as a chart or other image."[6]. At least according to an online dictionary, this is the main definition of the word. In a wider horizon, it is a process of analyzing a bunch of input data and transform it into a visual representation to give another or detailed view on the evaluated input.

There are multiple examples in history of the powerfulness and importance of visualizations in the context of maps and geographical data. Some of them make use of very simple methods to achieve their goal and others are working on a more complex level.

A so called spot map was already used very early in 1854/55 when the big cholera epidemic happened in London in 1854. As described in Brody et al. [2000], Dr. John Snow used a simple map of London to visualize his theory about the source of the cholera epidemic. He designed different versions of the map, but each included the city plan and

[6] `https://en.oxforddictionaries.com/definition/visualization`, Accessed 2019-04-03

different bars and markers for death cases in the studied area. With this approach, he tried to prove his theory, that the main source of the cholera epidemic is a water pump which provides polluted water. The main argument was, that the designated pump is the closest one to all the houses where death cases occurred, or at least all the dead people made some use of the water provided by this pump. In fact, after disabling the pump, the epidemic abated. At the beginning, a problem with the water was only a theory of Snow and he did not use the map directly to solve it, or at least just in his mind. He used the map to visualize the whole case to a committee of which he was a part of.

The cholera epidemic is an easy understandable use case of a visualization which can be very powerful and make a difficult situation much better understandable.

As described in Friendly [2002], in 1812 Napoleon's Russian Campaign takes place and many people died. 57 years later in 1869, Charles Joseph Minard made the depiction of Napoleon's March on Moscow. It shows the reduction of troop strength of Napoleon's army while conquering the Russian empire on a timeline, including the measured temperature in those days. This illustration shows in an emotional and depressing way the power, usefulness and possibilities of visualizing the given data. As it is described in Chih and Parker [2008], Tufte even called the depiction of Minard the "the best graphic ever produced" in his book [Tufte, 2001]. He also thought that the depiction would make an anti-war statement.

One of the more modern examples for a powerful visualization is shown in Vertesi [2005 (accessed April 20, 2019]. The author describes the normal map of London's Underground as an interface between the city and the public. A user study has been done with the people to gain some insights about their knowledge of the city of London. Different interviews have been made and all started with the question "Draw me London". They found out that basic above-ground city concepts of the people are influenced by the Tube Map. For example how they navigate through the city or how distances and time ranges are measured.

By comparing all of the previous examples, they show in their own way, that a normal visual representation can change peoples view on a specific situation. It can even influence their behavior, by using it as an argument in discussions or just be part of their daily life, like the London Tube Map.

In their book, Card et. al proposed multiple ways in which visualization can help people to get better cognition [Card et al., 1999]. Some of those ways are:

- By using visualization the memory and processing resources that are available to the users are increased.
- The time and effort that is needed for searching information are reduced.
- Visual representations make the detection of patterns easier.

For presenting data that can be projected on a map, different use cases already exist in the research. One of the features that is used in many of them is a two-dimensional local street map where the analyzed data is displayed by a defined color scheme. The purpose of these kind of visualization can be really different, but the idea is always the same. A variation in colors describes the differences of an observed value at one specific point of time. If somewhere no defined color is visible, then there is no value available or no observations have been made. Placing all the different colors on the normal street map has the logical purpose of defining the place where the measurement has been done, which is their geographical position. This visualization method is often called a heatmap, because of the color of an area, in which the value is higher, which is most of the time more red or more hot.

In Lécué et al. [2014] the traffic flow in a city has been investigated. A heat map has been chosen to highlight the quality of traffic flow at a specific point of time on every street in a special area of the city. If the color of a street part is green or even light green, the traffic flow is basically free. On the other hand, an orange or even red color means that the traffic flow is moderate or heavy. Another application of a heat map is shown in Hoang et al. [2013], where the different colors describe the amount of pollution measured in a hexagonal part of the city. With this approach it is fast and easily visible which area of the map has a big amount of pollution at the measured point of time and where the values are low. The coloring scheme is again the same, and the highest value is represented by the red/orange colors and the lowest values are in green/light green. A hexagonal grid has been chosen for displaying the different values instead of the more widely used rectangular grid. Therefore, the outlines of groups in the grid can form more different shapes than in a rectangular grid. As another example for using a heat map for visualizing data on a map, there is the study of Tecchia et al. [2002]. It has investigated the crowd and their behavior in real time. The addressed heatmaps in this example are used to visualize the amount of thoughts or behavioral changes performed of a person at a specific place. With this approach, by making use of different colors like in the previous examples, it is possible to show that there are areas where the investigated crowd behaves noticeably different. Namely, the people are more attracted by points of interest, because in this case their thoughts and behavioral changes are different than if they just walk around.

When talking about data gained from parking enforcement there are two main factors:

- The place where a car was checked or fined.
- The point of time at which it is done.

These two factors lead directly to the research topic of working with spatio-temporal data. This neither means that the existing data is just a bunch of different values that are documented for a single point of time, nor does it mean that only a single place is observed alternately after a certain time period. It means that the data contains various information from different places for multiple points of time.

Some basic examples for abundant and easily obtained spatio-temporal data are given in Pebesma et al. [2012], like satellite images of the earth or election results of consecutive elections from multiple districts. In this paper also different graphs are described which can be used to visualize spatio-temporal data. In the following some of these suggestions and others are described in more detail and how they are used in research.

One of the most well known methods is the time series plot. It shows the change of an observed value over time, where the value is depicted with the same single symbol for each point of time, for example a cross or a bigger dot. Often the symbols are then connected with lines or curves to highlight their slopes. In the context of spatio-temporal data, the time series plot can even be used to display multiple places in one single diagram by using different colors. The changes of values can thereby be directly seen by comparing the different colored lines.

An advanced application of time series plots can be seen in Lee et al. [2015], where also a normal time series with two different lines is depicted in the paper. It shows the values of two measured observations, the series $T_a$ and $T_b$ and how their value on the y-axis changes for each point of time that is depicted on the x-axis. Other different time series plots are used in Fernandes et al. [2017] to show the measured values of wearable EEG sensors over time. In this example, the time is also plotted on the x-axis and the measured value on the y-axis. In one of their plots, they even used multiple time series above each other to visualize the values of measured brainwave patterns.

A multi-panel plot, or also called multi-panel graph, or multi-panel display is a very simple approach for presenting data from various observations like multiple places. The idea is that several, relatively simple graphs are plotted next to each other. This can be realized by multiple columns and/or multiple rows. With this approach it is for example possible to show the change of a measured value over time in a single graph and display the same value for various places next to each other.

According to Bacon [1996] the arrangement of the panels itself is one of the interesting design issues. The example of a multi-panel display in the article, shows the data of six different segments in their own dot chart, organized into two rows and three columns. Each dot chart shows the importance of performance attributes of multiple service attributes. The article also shows a second example, with four even three-dimensional plots in a 2x2 arrangement which shows the demand in kilotons for a pork flavored yogurt. Another application of this visualization method can be seen in Khalsa [2012], where the birds diversity has been investigated. At first the multi-panel approach has been used to depict the species richness every year for 14 different routes, where each route has its own graph. Additionally, the same visualization is then used for showing the observations of specific bird species on these 14 routes every year.

Another very useful visualization approach for spatio-temporal data are animated plots or also called animated graphics. Additionally, these graphics can also be interactive. An suitable application can be seen in Hocking et al. [2015, (accessed April 20, 2019], where they introduce two new terms on animated plots. It is about which value is selected and

consequently what data is shown based on the selection. One of the examples shows an interactive partial map of America divided into multiple small squares. If the user makes a click on one of these areas, the color of all the squares changes. The animated map uses the selected square as a reference value and changes the color of all other areas by displaying the temperature difference between every square and the selected one. An example without interactivity for animated plots can be seen in Moll et al. [2011], where vortices on the solar surface have been investigated. The article shows a snapshot of their animated plot, which shows the swirling strength of the vortices at a specific point of time. They even referenced to their electronic version of the journal, where the whole animated plot can be found.

This previous sentence leads directly to the main problem of animated plots. The technique is highly applicable for spatio-temporal data because it can depict multiple observations one after another in an animation process. As interesting and powerful it is for digital journals or visualizations in movies or presentations, it is normally not usefully applicable in documents because the animation process cannot be printed. The only way to achieve a depiction of animated plots in documents is also shown in the two former citations, by adding individual snapshots or a set of snapshots. Depending on the complexity of the animation, these snapshots would consequently lead to a loss of benefits. The viewer will just get some temporary pictures instead of the complete impression of an animated plot.

Besides all the highly specialized representations, regarding Pebesma et al. [2012] also the so called space-time plots exist. Those diagrams just consist of a grid, where the point of time is depicted on one axis, and the space is plotted on the other axis. A downside of the grid style is, that only one value can be displayed for every combination of time and space and it is written directly next to the crossing point of the two dimensions.

The trough and ridges diagram, introduced by Hovmöller in 1949 in Hovmöller [1949] is another advanced representation for specific space and time data. He is also the name giver for this kind of diagrams. The time is plotted on the y-axis in the diagram, particularly the whole month of November 1945 in daily steps. The x-axis is used to depict the different locations, namely by the longitude in degrees. By the usage of hatching, the troughs and ridges are highlighted. This visualization gives a very good overview of how the troughs and ridges are enveloping and moving through the whole month.

Besides the classical usage of Hovmöller diagrams in the area of meteorology, there are also other application areas where they are useful. Together with better technological possibilities, the application areas for this type of diagram has been grown. Additional modern features include for example the usage of colors or even animated versions like on weather maps.

An example for a modern application of these diagrams can be seen in Shamir et al. [2018], where wave simulation solutions have been tested. The diagrams are shown in a panel plot with 4 columns representing different algorithms and their parameters. Each Hovmöller

diagram has the longitude of a wave plotted on the y-axis and the corresponding wave period as the time on the x-axis. To depict the measured value at a certain longitude and time combination, instead of the previously described hatching, now a color gradient ranging from blue to yellow has been used. This color gradient depicts the value range from -1 to +1 in steps of 0.2 Without this step size, the developing areas would otherwise not be that clearly visible. Therefore, the propagation of a wave can easily be shown by the diagrams.

In the paper Lécué et al. [2014] about the traffic flow investigation, also another visualization example is shown, which they use for their given spatio-temporal data. It is called parallel chart or parallel coordinate plot for visualizing multiple observed groups and their different properties in one single diagram. In the mentioned paper, the diagram is used to depict the different traffic flow qualities together with their related properties like minimum travel time or distance. This visualization mechanism for multidimensional data was introduced by Inselberg and Dimsdale [1990] in 1990. As mentioned in Graham and Kennedy [2003], this method has been modified in various ways for gaining additional features. Their own paper suggests to use curves and spreading to get advanced applications of the parallel coordinate visualization method.

Comparing the previously shown time series plot and the just explained parallel coordinate visualization method, they seem very similar on the first sight. Both share their depiction of several differently colored lines seemingly developing from left to right, but their meaning is rather different. In fact, the time series plot could eventually be seen as a special type of a parallel coordinate plot, where every point of time is a different position at the y-axis. This would mean that every depicted property of the parallel coordinate plot shows the same type of value, but at different points of time, sorted in ascending order. Normal applications of both visualization methods show, that their structure is absolutely not the same. A time series plot depicts the change of a value over time and can include multiple observation points in one diagram by using different colored lines, as described before. On the other hand, the parallel coordinate plot is also used for including multiple observation points by different colors, but it does not show their change over time. Instead it depicts different properties that occur at a specific point of time of a single observation next to each other. By using different colors for multiple observation points, it is thereby possible to compare multiple attributes of various groups at the same time.

Another presentation technique for statistical data is shown in Potter et al. [2006], namely the box plot. It is used to quickly summarize the distribution of a dataset by visualizing the minimum and maximum range values, the upper and lower quartiles and the median. Additionally to the normal representation, some modifications of box plots exist. These could be for showing the density of the distribution over the whole box area, like a violin plot described in Hintze and Nelson [1998], or changing the shape by using notches or variable box widths. Outliers are sometimes also represented differently, for example by using symbols, whereby the meaning of the minimum and maximum range values changes.

Box plots can also be used for visualizing multiple properties or distributions of different time periods by printing them side-by-side. In this way the various distributions can be compared visually and differences of key values can be seen directly. An example for side-by-side box plots is depicted in Benjamini [1988], where the telephone bill costs distribution is shown. Multiple box plots besides each other depict the same observation but with a different set of data, based on how long people lived in Chicago in years. Each box plot additionally has notches and also the width of the boxes is based on the group size of the data set.

Reese [2005] shows another example for multiple box plots in one diagram. It visualizes the distributions of fish length by species above each other in ascending order sorted by the median. The diagram makes it possible to get a very fast impression of species length differences and if the variance is high or low. This simple version of a box plot does not show the density and also the sample size is only indicated as a label in brackets instead of using different box widths.

After comparing multiple applications of box plots it is obvious that there are many different modifications which can be used for various visualizations. It is also clear that, for each type of data and distribution, the used modifications of the box plots have to be considered individually. An example modification is the change of the box width proportional to the sample size of the group. This feature is really helpful when comparing a small amount of plots side-by-side. It is possible that the difference in width is pretty small and a lot of plots are presented next to each other. The differences between various width values of the boxes are then not identifiable for the beholder and therefore the whole modification is unnecessary.

Another visualization method for the values of single variables, mentioned in Crawley [2007], is the histogram. A histogram can be used to visualize the statistical distribution of a feature, as stated in Runkler [2012]. It is a bar chart depicting all values of the data, which are therefore classified and assigned to a bin number. Each bin represents an interval of the complete distribution of the data. Various formulas exist in the literature for calculating the number of bins. These include a calculation based on the number of data by Sturges [1926], based on the standard deviation by SCOTT [1979] or based on quantiles by Freedman and Diaconis [1981]. Each of those approaches has its own optimal application area. The main characteristic of an histogram is that the surface of the bars is proportionally depicted based on the number of elements in the bin. Therefore, the y-axis is not as important as it is in other visualization methods because its meaning depends on the prior performed calculations to model the histogram bars. In some representations the height of the bars directly depicts the amount of elements in a bin, while it shows the relative frequency on other applications. The proportional surfaces also allow another visual modification, namely individual interval widths for the bars. This makes it possible to have finer granularity of intervals in areas with higher data density.

A modern application of histograms is shown in Sun et al. [2005], where it is used for analyzing pixel values of images. The printed histograms show the intervals of various pixel values on the x-axis and their corresponding number of pixels on the y-axis. The

used data for the diagrams is modified by different algorithms and the various resulting distributions can be recognized by the differently looking histograms. By applying this visualization method, the researchers recognize their insufficient solution by analyzing the printed diagrams and therefore know that a new algorithm has to be developed for their needs.

## 2.4 Time Series Analysis

The term "time series analysis" covers several topics that are relevant in the process of analyzing and understanding the data and the graph of a time series. During the analysis of a time series it can be the objective to develop a suitable mathematical model that provides a plausible description of the data as stated in Shumway and Stoffer [2017]. Each time series is considered as a sequence of values, $x_1$, $x_2$, $x_3$,..., where the series value that is taken on the first point of time is $x_1$ and at the second point of time $x_2$ and so on. Typically the index $t$ of a value is considered to be a discrete integer value like ...,−1,0,+1,+2,... or some other subset. Conventionally the time is plotted on the x-axis and the values of adjacent points of time are connected to visualize the continuous progress of the series.

When starting the development of a suitable mathematical model for a time series, many approaches and issues have to be considered. Some easier models like a linear or squared model could be enough for specific time series to describe the data correctly, especially in combination with other mathematical concepts like viewing at the logarithmic values. Usually more advanced approaches are needed. Based on Cowpertwait and Metcalfe [2009], some relevant terms are described in the following:

- **Stationary time series**

  The term "strictly stationary", or often just called stationarity, is relevant before defining the following time series models. Stationarity means that the joint statistical distribution of the series remains the same for each random set of adjacent $n$ values of the series. A time shift of the series therefore don't changes the distribution and it implies that the mean and variance are constant in time. If a trend or a seasonal pattern exist, those components are signs for non-stationarity. Such a series is also called an integrated series. Removing trends can be achieved by differentiation of the series for one or more times. The series is *integrated* of order $d$, where $d$ is the number of differentiation steps.

- **Autoregressive models**

  An autoregressive model is the name for a model that is based on an autoregressive process. This is the umbrella term for a time series, where a certain number of $p$ preceding values define the current term. The most recent points thereby have the highest impact, what is realized by a set of $p$ additional parameters. It is then called an autoregressive process of order $p$, also written as AR($p$).

- **Moving average models**

  The basic approach for the moving average process is similar to the previous one by taking the recent points into account. The baseline value is the mean of the series and instead of weighting the set of recent values itself, here the error of each recent value is weighted by a set of parameters. The number of considered points is defined as $q$, whereby the name of a moving average process of order $q$ is written as MA($q$).

- **Mixed Models**

  Very useful model classes are obtained when combining the autoregressive process and the moving average process. The autoregressive moving average process of the orders ($p$, $q$) is denoted as ARMA($p$, $q$). Stated in Shumway and Stoffer [2017], if the model is stationary, such a time series $\{x_t; t = 0, \pm1, \pm2, ...\}$ is defined as:

$$x_t = \alpha + \phi_1 x_{t-1} + ... + \phi_p x_{t-p} + w_t + \theta_1 w_{t-1} + ... + \theta_q w_{t-q} \tag{2.1}$$

  Here $\alpha$ is a calculated constant based on the autoregressive process, $\phi_1$, ..., $\phi_p$ are the autoregressive process parameters, $\theta_1$, ..., $\theta_q$ are the moving average process parameters and $\{w_t\}$ is white noise. If the time series is non-stationary, the ARMA process is extended to include the multiple differentiation steps. This is known as autoregressive integrated moving average process of the orders ($p$, $d$, $q$) and is written as ARIMA($p$, $d$, $q$).

Besides the mathematical model development of a time series, also other topics can be of interest for an analysis. One of these topics have already been relevant in the previous section in the context of stationarity, namely the trend or seasonal patterns of a time series. Determining these components is called the decomposition of a time series. For this approach it is assumed that a simple additive decomposition model, shown in Cowpertwait and Metcalfe [2009], is given as:

$$x_t = m_t + s_t + z_t \tag{2.2}$$

The equation defines the observed series $x_t$ by the time $t$, where $m_t$ is the trend, $s_t$ the seasonal effect and $z_t$ the error term.

After decomposition the outcome is an estimation of the trend $m_t$ and the seasonal effect $s_t$ at the time $t$. Calculating these estimations can be achieved in different ways. An example for estimating the trend would be to calculate a moving average value centered on $x_t$. This means that if a monthly time series is given, the trend value at time $t$ is an average of $n$ values before and after the designated month and the value of the month $x_t$ itself. If the calculation is continued for all time periods, the result is a series of average values that can then be linearly interpolated. The moving average calculation is an example of a smoothing process, which is used to identify an underlying signal or trend.

Besides the moving average approach, more advanced algorithms exist. One of these is a locally weighted regression technique that is called loess, which is another smoothing mechanism that can be applied on time series. Loess uses an iterative process, where each repetition is used to enhance the weight that is specified for each of the closest points to the designated value. This algorithm is used in the Seasonal-Trend Decomposition Procedure Based on Loess function, shortly named STL, described by Cleveland et al. [1990]. After applying this function on a time series, its visualization shows the decomposition results in a plot, starting with the original data, followed by the trend, seasonal effect and a remainder in four separate diagrams above each other. The time is depicted on the x-axis, while the y-axis shows the corresponding value of each of the four diagrams.

Another topic in the area of time series analysis is the correlation, especially the autocorrelation. Stated in Cowpertwait and Metcalfe [2009], correlation is a measure for the association of variable pairs. Therefore, autocorrelation in the context of time series is used to determine how similar the series is to itself after a certain time period. This period is called the lag $k$ and the corresponding function is the autocorrelation function "acf". The correlogram is the visualization mechanism that is used to show the autocorrelation of a time series. It depicts the dimensionless correlation value on the y-axis and the time period lag on the x-axis.

A simple example for such a correlogram visualization could be a sine wave. If the time series would be a repetitive sine wave, where the zero point is reached every 10 time steps and the whole cycle needs 20 time steps, the correlogram would also have the form of a sine wave. Except of a decreasing amplitude over time, every lag of 20 steps the autocorrelation value is very high, nearly up to plus 1. On the other hand there are the lags at 10, 30, 50,.. and so on, where the value is very low, nearly up to minus 1. The reason for a high value is, that the similarity of the time series is very high at this point, but if it is negative, the resulting values are inverted.

Additionally to the similarity of a series to itself, the cross-correlation function "ccf" exists, like described in Cowpertwait and Metcalfe [2009]. It focuses on the comparison of two separate time series. The associated visualization is the cross-correlogram, which looks exactly like the previously described correlogram, but the meaning is different. A high value in a cross-correlogram means that one time series leads, or lags the other by lag $k$ time steps.

An example for a high value in a cross-correlogram could be a process which needs one week, like the check of a proposal. Comparing the daily time series of the amount of handed in proposals and the daily time series of the amount of confirmations of proposals, the cross-correlogram would show a high value at lag 7. If it is assumed that most of the proposals are always good, one week after the proposals were handed in, the number of confirmations change in the same way.

There are existing many well known approaches regarding time series analysis. The programming language R provides an environment for data analysis and graphics, including different packages with implementations for the prior stated approaches, like described

in the book Crawley [2007]. It contains various application examples in written code and their related illustrations for analyzing time series and many other topics that could be solved by the R language. Some of the provided functions are stl, acf and ccf, which have been described before.

## 2.5 Data Preprocessing

The term "data preprocessing" is well known in the field of data mining as many algorithms assume that the underlying data is free of disturbances, like stated in García et al. [2016]. Real world data is far from being complete or clean and therefore many additional steps are needed. Besides the data mining field, every data analysis process has a special need for investigations on their starting data. As shown in Runkler [2012], this includes preparation and preprocessing of the data before the proper analysis and visualization start. The necessary approaches that have to be applied during the process can differ, depending on the structure of the given data and the defined goal that should be achieved. Various applications of different preprocessing steps exist in the literature, of which some are described based on García et al. [2016], Runkler [2012] and Singhal and Jena [2013]:

- **Data integration**

  If the given data does not contain all information that is necessary to achieve a defined goal, it can be helpful to integrate data from other resources to the main data set. The source of these information is thereby not limited to internal data sets, but can also be obtained from external pages or services. The way of how this is realized also depends on different factors, such as data source, structure and complexity of the integrated data, as well as the existing data. It could be enough to just iterate once over a list of data entries and add an additional boolean flag in one situation. On the other hand it is necessary to query a web service, manipulate the response and explicitly adjust the result for each data entry. After this preprocessing step, the main data is enriched with additional information and therefore the possible outcome of an analysis can be more significant.

- **Errors and missing values**

  In optimal conditions all the values in the whole data set are complete and error-free. The definition of an error depends on the usage that is planned for the values of a certain field in the data set, which should be illustrated in the following example. If a field has a numerical type, an empty value will not be a problem if counting all the occurrences or calculating the sum of this field is the target. Is the value set to "Not a Number" (NaN), the same process of counting could lead to an error during the analysis. A short workaround for correctly modifying the data to fit the described needs could be to set all the unknown values to zero or the mean. If the values are zero, there will be no more errors when calculating the sum. A zero value seems the right choice for unknown values, but if the goal looks different it

can cause additional trouble. When analyzing the ratio of two values, the division by zero means another undefined result.

Depending on the type of analysis, outliers can also be seen as errors, especially if they are generated by incorrect or missing measurements. In these situations, the removal or the replacement of specific parts of the data set could be the correct approach. On the other hand, the detection of outliers can be one of the certain goals of the analysis process and every modification during the preprocessing steps would be a big mistake. After this step, the quality of the main data set is increased by taking the given target into account.

- **Classification and binning**

  The classification preprocessing step is a very wide spread area with different applications. The basic target is that the given values or even whole data entries should be assigned to one of many special types, namely the classes. For example, if a given entry is a set of human health features, like blood pressure and body temperature, the classification is used to assign it to either the group of healthy people or sick people. A more easier classification approach is the categorization of all the values from a single field, also known as binning. This is for instance applied on the household income which should be categorized in classes. Therefore various parameters have to be considered. One parameter is the number of classes, like poor and rich or low, medium, high, school grades or any other possibility. This is followed by the boundaries of each class. It could be evenly spaced by the amount or percentage of instances, by density or by natural boundaries. Also the length of the number could be a boundary like 10.000, 100.000 and one million.

  Each categorization approach has its own advantages and disadvantages, which could influence the later analysis. If the granularity of the classes is finer, the number of instances per class is smaller. This leads to a common problem of this part of the preprocessing step. It states, that classification often leads to a small set of classes, where the majority of instances is assigned to a part of it and leaves the rest of the classes with only a tiny percentage of instances.

- **Data transformation**

  The term "data transformation" contains multiple topics of the preprocessing phase including some parts of the already described approaches. Besides the previous cases, there are more general actions that are considered in the step of transformation. Not only the goal of the data analysis is important in this context, it is also necessary to know which tool is used for the later realization. The data transformation focuses on mapping the data elements from the source system to the destination data system by considering all the needed steps in between. It is not just a change of the format, it also includes a data reduction to get a smaller size, but still includes all the relevant information for the result. A classical approach that fits to this topic are the different representations of decimal numbers in various languages. If the input data source are files where numbers are written with a comma as decimal

separator, but the destination system needs a point as the separator, the numbers have to be transformed correctly. The decimal separator is just a simple example of various points that have to be considered during the data transformation.

# CHAPTER 3

# Preprocessing the Data

In the following section all the necessary steps are described before the analysis and visualization part can start. While the main outcome of the thesis focuses on the analysis of parking enforcement data in common, this chapter includes multiple parts which are relevant for the given use case, the data of the City of Vienna. Some specially adapted parts also use references to open data provided by the City of Vienna and therefore would also be needed for any other parking enforcement analysis. The chapter includes a detailed description of the source data, the applied changes, the data integration steps and the provided destination format. Many of the performed steps are implemented by algorithms in code which is not presented in this thesis, but even if all the code parts are not written down, the applied concepts and transformations are described. Because of the fact that some data integration is based on open data, also a short section about licensing information is part of this chapter. The following often used terms have a special definition in the context of this thesis:

- **Entry** has the meaning of any line in the data set, regardless of being a request or a penalty.
- **Request** means a simple call to the server performed by the parking warden to receive the parking permission information for a specific license plate.
- **Penalty** means the line of a fined car and beside the license plate it contains additional details about the location where the car was parked.

## 3.1 Anonymization

The source data that is used as the basis for working on this thesis contains highly personal data and therefore a special permission is needed to work with it. This permission includes some conditions which have to be fulfilled in advance. One of the conditions includes

that the thesis should not show a detailed analysis of the complete city but rather point out special situations of different districts. The main condition handles the use of the highly personal parts of the data and how it has to be anonymized. These parts and their possible risks are described in the following sections. Additionally to the data handling itself, it is also stated that the anonymization step has to be performed in the area of influence of the City of Vienna, to make sure that the non anonymized data remains secure. As a result of this last condition, some special precaution is also needed for the algorithms and implementations, which is shown in Section 3.4.

### 3.1.1 Warden Username

One of the fields in the source data contains information about the warden who was performing this request or penalty. While the value of this field is not directly the real username of the warden, it is a unique hash value that is based on the real username. This means that there is a direct link between the person itself and all performed actions that are listed in the source data. By considering this information it would be a possible source of data misuse and by all means has to be eliminated. Some examples of how this information could be misused are listed in the following:

- If only one penalty and the associated time stamp are known from a warden, it will directly reveal every other action that is performed by this warden.
- There is the possibility for a certain warden to get located and this makes the person vulnerable for criminal acts.
- It could reveal relationships between the warden and specific car plates, like the own car, if the warden checks the own car repeatedly.

It should also be mentioned that a parking warden is not popular by drivers, because the warden is the person who fines them. Additionally the driver often feels disadvantaged, especially if there are problems concerning the parking tickets. In the past there have also been attacks upon the parking wardens.

All these points lead to the condition that the warden username has to be somehow anonymized and must not be part of the data that is used for analysis. The information of who has performed a certain entry is of high interest. Therefore it is needed to connect all entries that are performed by one warden in a defined time period to a completely anonymized sequence of entries. The information of linked entries is not lost, but the username hash value can be dropped completely, which is described in more detail in Section 3.4.

### 3.1.2 License Plate

The second data field which contains personal information is the license plate. Besides the hashed username value, in this case the value is clearly written down in text and

contains no type of obfuscation at all. This leads directly to the next list of possible misuse of the information:

- Patterns can be detected of a specific car plate which gives details about when a car will be where.
- If the car position can be estimated it is vulnerable against criminal acts, especially anonymized because the car is not guarded.
- Affiliated people can control the location of the cars to check if they are parked where persons pretend to be, for example in the context of employee tracking.

Of course the data is not available live and the stated problems are less relevant afterwards, but although each of the described cases would be a misuse of the data. The license plate information has not such a big impact on the analysis like the warden username, but it also provides suitable features that should be part of the analysis. The main feature which can be derived from the license plate is the origin of the cars. This can give the opportunity to compare certain observations and their distributions based on the origin of the cars.

As certain information given by the license plate should not be lost, a classification is performed on all the values and the plate is split into separated fields based on their format. Therefore the original value is dropped, but the feature information is still available for the analysis, which is also described in more detail in Section 3.4.

## 3.2 Data Source and Licensing

The main source data, which contains all the parking enforcement data, is directly provided by the City of Vienna, after fulfilling all the requirements that are described in the anonymization part Section 3.1. The implemented algorithm can access directly the data via a path to the directory where all the files described in Section 3.3 are placed. During the preprocessing phase the initial data is enriched with additional information coming from different external sources, which are described in the following.

### 3.2.1 Address Service

The initial data contains a big amount of penalty lines. Each penalty line has an additional location information, which is described by an combination of street name and an optional house number. For more detailed analysis steps it can be necessary to know the position of the fined car as accurate as possible and therefore it is one of the research questions to determine the coordinate points for as many locations as possible. The initial data does not contain coordinate points and therefore an implementation is needed to enrich it with this kind of information.

For the City of Vienna, there exists an own web service which provides the possibility to query for additional address information based on the street name and the house number. It is described on the open data website of Austria and is named the OGDAddressService[1]. The detailed parameters which are available on this service are described in the documentation file[2]. Usage and implementation of the web service for integrating the data is stated in Subsection 3.4.4.

The obtained data which is queried from the web service is free but under a certain license, stated in the terms of use[3], which makes it necessary to name it at this place: "Datenquelle: Stadt Wien - data.wien.gv.at".

### 3.2.2 Street Graph File

Finding the coordinate points of the penalty locations is not the only goal of this thesis. Additionally the location of the request entries between the penalties should also be estimated as good as possible. For this task the street graph of the City of Vienna is used to gain knowledge about all streets of the city where a car could possibly be parked. The street graph is a single file containing all coordinate points of the complete street system and it can be downloaded in different versions based on a few parameters. It is also available via the open data website of Austria[4,5]. The usage of this data source is described in Section 3.4.

The street graph file is from the same website as the address service and therefore the same terms of use are applicable, which are already stated in Subection 3.2.1.

### 3.2.3 Additional Knowledge Sources

For every step that is performed when working with data, having knowledge about the given context or the data itself is of great value. Knowledge, but also know-how of the processes is a key factor for the data modification steps and analysis of the data. These important information can be gained in different ways, internally and externally.

In case of parking enforcement, the author of this thesis is working in the contracting company of the City of Vienna, which is responsible for the backend system and the used software. During this work, a lot of knowledge is gained about the process of parking enforcement, the involved stakeholder and the whole workflow.

[1] https://www.data.gv.at/katalog/dataset/c223b93a-2634-4f06-ac73-8709b9e16888, Accessed 2019-03-07

[2] https://digitales.wien.gv.at/site/files/2019/01/adressservice-doku.pdf, Accessed 2019-03-31

[3] https://digitales.wien.gv.at/site/open-data/ogd-nutzungsbedingungen, Accessed 2019-03-07

[4] https://www.data.gv.at/katalog/dataset/1039ed7e-97fb-435f-b6cc-f6a105ba5e09, Accessed 2019-03-07

[5] https://data.wien.gv.at/daten/geo?service=WFS&request=GetFeature&version=1.1.0&typeName=ogdwien:STRASSENGRAPHOGD&srsName=EPSG:4326&outputFormat=json, Accessed 2019-03-07

Parking in a city is often restricted and follows predefined rules which are defined by the city council over time. Such rules can be different for various areas in a city and even be limited and individually defined for single streets. The interpretation of the amount of data and distributions during the analysis is heavily influenced by the knowledge of the specific rules which are applicable in the investigated area. For the City of Vienna there exists a website[6] which contains detailed information about the parking rules in each district and describes the differences of various areas. Beside standard rules for short-term parking in Vienna, also special rules and possibilities for permanent parking of residents exist.

When working with a big amount of coordinate points of a city, the unusual structure of the numbers, which often differ only after many decimal places, are difficult to understand in the beginning. During development and usage of those points it is convenient to verify the coordinates by visualizing them on an online platform like OpenStreetMap[7], Google Maps[8] or the city map of Vienna[9]. Beside a representation of the coordinate positions, also the street names and especially the street numbers are depicted on those platforms.

## 3.3 Structure of the Data

The provided source data for this thesis contains the daily protocols of parking enforcement of the complete year 2018. At the end of each day, the contracting company delivers the daily protocol file to the City of Vienna who is responsible for the data and stores it for legal reasons. For the context of this thesis the data is provided for analysis under special conditions described in Section 3.1.

The daily protocol file is generated on the backend server in a kind of a log file where all the entries are repetitively written over time with comma separated values. Each penalty line contains the same data fields like the request, extended with additional location information of the fined car. The data fields which are part of each entry and the penalty extensions are described in the next sections. Then the enforcement process is described to explain the generation process of the protocol file and to clarify the changes made after the anonymization. Additionally the requirement for different output formats for the analysis is stated.

### 3.3.1 Entry Lines

Whenever a parking warden is performing a request or a penalty, the same basic data fields are generated for each entry. These entries contain information of five different categories separated to eight data fields, which are described in the following:

- **Date and Time**

[6] https://www.wien.gv.at/verkehr/parken/, Accessed 2019-04-30
[7] https://www.openstreetmap.org/, Accessed 2019-04-30
[8] https://www.google.at/maps, Accessed 2019-04-30
[9] https://www.wien.gv.at/stadtplan/, Accessed 2019-04-30

The first two data fields of each entry contain the date and the timestamp when the action was performed by the warden. As a protocol file is generated on a daily basis, the date value is always the same in each file. The timestamp on the other hand is spread over the whole work time of the parking wardens and basically ordered sequentially. The accuracy of the time is given in seconds, which makes it possible that multiple entries exist for the same timestamp. As the protocol is written sequentially, the assumption is close that the list of entries is completely in an ascending order, which is wrong. In some situations the timestamps are not ordered, because the backend server uses multiple threads to process the warden requests and therefore due to differences in processing time it is possible that errors occur.

- **Warden Identifier**

  The next data field contains a hash value of the real username that is unique for each parking warden. This approach leads to a direct link between the provided information and the real warden who has performed the requests. Due to various reasons described in Section 3.1 it is necessary to drop this field during the anonymization step.

- **License Plate**

  The initial license plate information is split into two data fields, namely the main value itself and the country of the plates origin. The country is given in a short form containing one to three characters. The main value of the license plate is separable because it contains a hyphen. For Austrian license plates the hyphen can be used to split the value into the former identification letters, also called authority or district, followed by the number and letter sequence. Is the origin of the license plate not Austria, the part before the hyphen can, but does not have to mean the same. This is because of different rules for license plates all over the world. For the value of the license plate some anonymization conditions have to be fulfilled, which are also mentioned in Section 3.1. Apart from correct license plates of cars, there also exists one test value, which can be used every time by the wardens for checking the availability of the system.

- **Basic Location**

  The basic location information basically is made of two fields, whereas their content is the same. The fields are named district and rayon, but the rayon is not used explicitly for parking enforcement. Therefore the meaning of the two fields is the same in the context of this thesis. This leaves the only location information that is given initially for each request as the district of the city.

- **Request Type**

  The last data field in the source data is the request type, which gives details about how the warden has scanned the car in front of him. The values only contain the characters M or R. The value M stands for manual and means that the warden has

entered the license plate of the car manually with the keyboard on his control device. If instead the value is R, it means that the warden has scanned the radio-frequency identification (RFID) tag of the car. This data field implicitly gives much more information about the car than it seems, because every car which has a RFID tag indicates that it is or was used for permanent parking at some time in the past.

### 3.3.2 Penalty Extension

If a warden fines a car, a penalty is generated on the server and besides the normal entry, additional information is added to the protocol file. This extended data contains location details of the parked car. It is separated into four data fields:

- **Street and Number**

  The first data field is a combination of two details, the name of the street and a directly added house number. The warden has to choose the street name from a predefined list, which prevents errors in this data field. The house number is entered manually and is not mandatory for a warden. If there is no house number given in this field, some additional text has to be entered in the next data field.

- **Additional Text**

  This data field is a free text area, where the warden can enter additional information about the parked car. If no house number is entered previously, this data field is mandatory. The usage of this data field is very differently handled by the wardens and the diversity of all the texts is very high. While this texts can also be used to determine a house number for a given penalty, it is difficult and often not explicitly clear. Sometimes this data field also contains manually entered street names of adjacent streets, but everything is typed in manually and therefore the content cannot be automatically parsed by an algorithm due to high diversity and errors.

- **Opposite Flag**

  Besides the street name and number, two additional flags exist. The first one is the opposite flag, which is set by the warden if the car is parked on the opposite side of the street, but the entered house number is the one reachable.

- **Frontage Road Flag**

  The second flag means that the car is parked on the frontage road instead of the main lane. This data field is needed because in some situations the parking rules for the frontage road can be different than for the main lane.

An additional fact for each penalty entry is that the request type is always set to M. This is because the penalty can never be triggered automatically by any mechanism and always has to be confirmed additionally by the warden.

### 3.3.3 Enforcement Process

The parking enforcement process follows a specific scheme involving wardens, streets and cars. A parking warden is a person who has the job to walk around on a specific route or in an area and check the parking permissions of the parked cars. This area consists of various streets in a city and the warden checks every car he passes by. In Figure 3.1 an example of two parking wardens is shown. The left side depicts the timeline of each warden and marks the different entry types, requests and penalties, with two different symbols. A warden is going on the streets and every time he passes a car it is checked for parking permissions, shown by the requests. If a car does not have a valid permission, it is fined and therefore a penalty is send to the server. On the timeline it can also be seen that issuing a penalty needs more time than checking cars.

On the right side of the image a part of the protocol file is shown which lists all the entries, placed on the two timelines, in ascending order by their timestamp. The protocol part explicitly shows the differences between requests and penalties by adding the street information. Additionally the image shows the case where two entries have the same timestamp, which is also possible because many wardens are working contemporary.

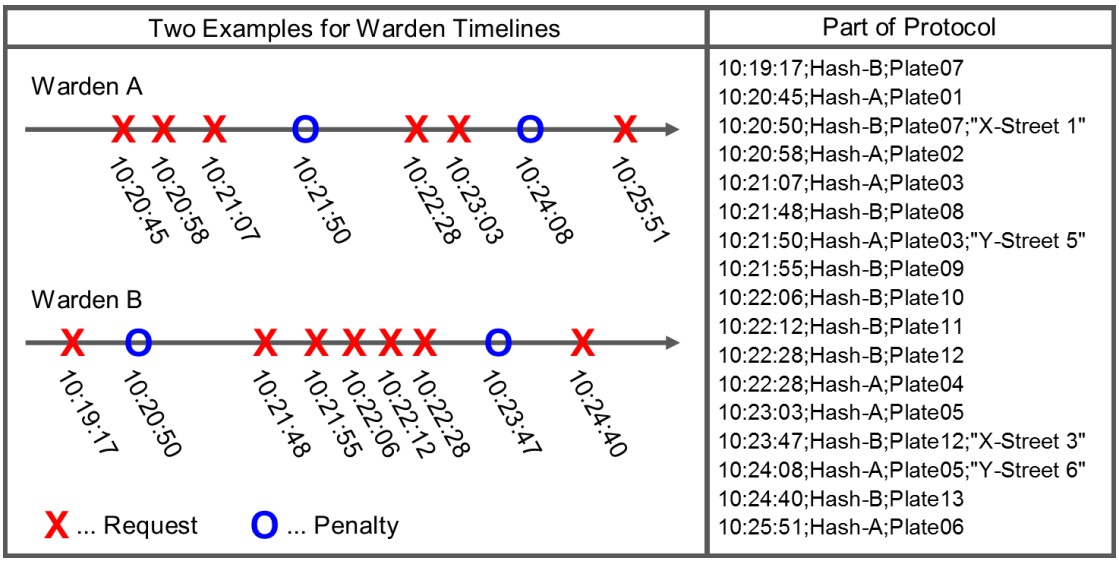


Figure 3.1: An example timeline of two parking wardens and a part of the corresponding protocol. It depicts the mixture of requests and penalties of multiple wardens in the protocol file.

While the wardens are working outside and they use modern software for performing their checks, it would be easier using the GPS of their devices and forward the current position to the server on every request or penalty. This approach, however, is prohibited because the works council of the parking wardens is strictly against this type of employee tracking.

### 3.3.4 After the Anonymization

As already mentioned previously, the first step that is performed on the source data is the anonymization. It includes the discarding of the warden hash values and the classification of the license plate values. The second point of modifying the license plate information is a simple categorization based on the characters of the value. Besides the origin country and the identification letters, the number and letter sequence is used to determine the category of the value. It is mainly allocated to one of four categories:

- Standard license plates of Austria, where the value starts with numbers and ends with letters.
- Personalized license plates of Austria, where the numbers and letters are inverted.
- Special license plates of Austria, which can have special values, like only numbers or only letters.
- Foreign license plates, where no further graduation is done, because the diversity is too high and many different rules exist in various countries.

After this categorization the initial value of the license plate is not needed anymore because all relevant information have already been filtered.

The discarding of the warden hash value requires another approach to make sure that the main information doesn't get lost. This main information is the connection between all the requests and penalties that are performed by the same warden. These connections should not be lost, but the hash value of the warden username must not be kept. Therefore it is necessary to group all entries of an individual warden of a certain time period together. For determining the borders of a group, a time period is defined which has to pass without any warden request before a new group starts. Each group of entries is then independent of the warden hash value, but still contains all the relevant connections between requests and penalties.

### 3.3.5 Output Format

The source data consists of millions of entries and during the analysis various features in different time periods are investigated. The view on the data changes for different analysis approaches and therefore also individual output formats are needed. In one example it can be necessary to provide data for the whole year which includes aggregated values for the amount of requests and penalties. For another analysis a weekly list of all entries where coordinate points have been found could be the source. The individual approaches for the output formats additionally give the advantage of smaller data sets with less size and therefore are easier to handle. The data which is used for a certain analysis is described in detail in Chapter 5.

## 3.4 Data Transformation

The data transformation process consists of multiple steps, which start with the source data and external resources and end up with individual output formats for different analysis. This section describes the transformation steps and states special cases, like assumptions which have been made in certain situations or problems that have been occurred. A depiction of the performed steps is provided by Figure 3.2. It also includes the integrated data from external resources, which are the street graph file and the address web service.

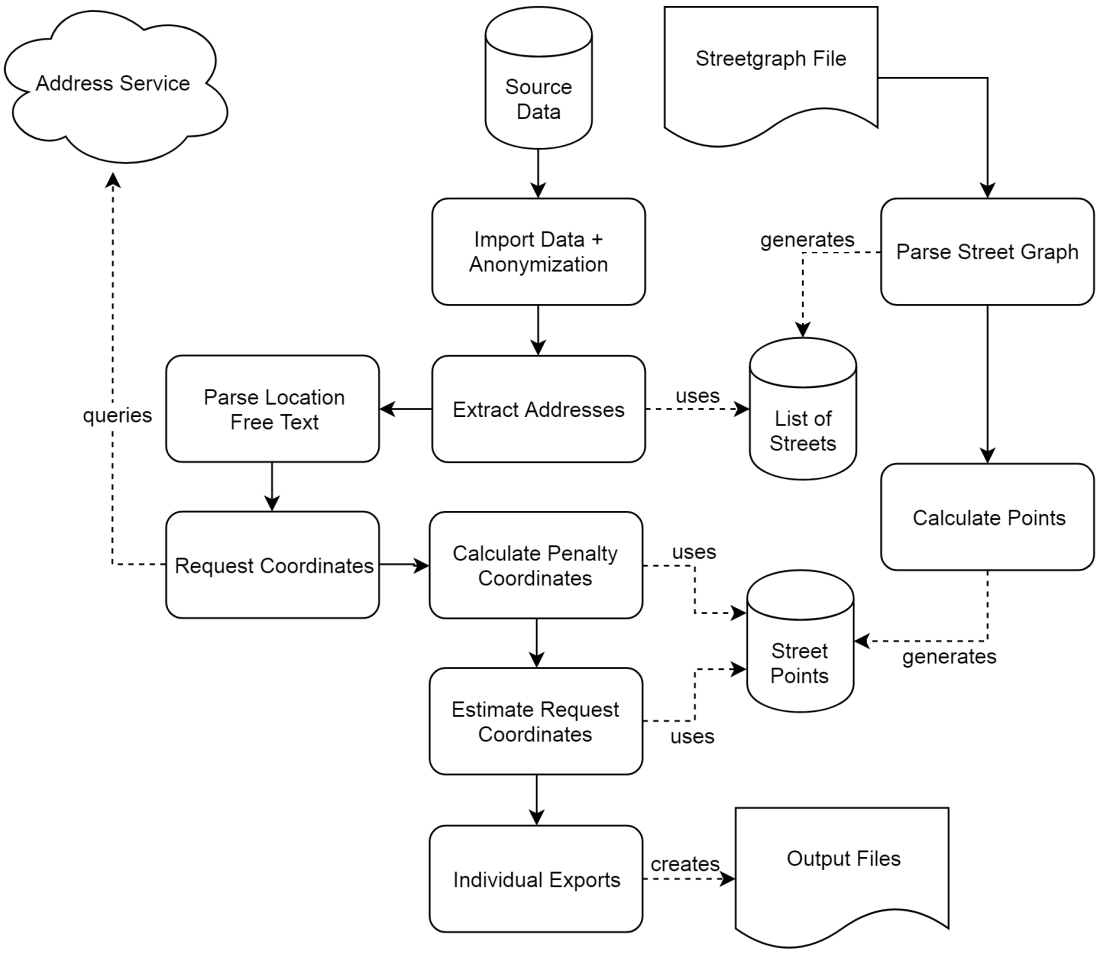


Figure 3.2: Overview of the data transformation steps and the included external resources, the street graph file and the address web service.

### 3.4.1 Data Import

The first step in the data transformation process is the import of the source data. As already mentioned in Section 3.1 it is mandatory that the anonymization is performed during the import and the described conditions are fulfilled. This includes two data fields, the value of the license plate and the hash value of the warden username. As described in Subsection 3.3.4, the license plate is split into different relevant parts, based on the origin country. The hash value of the warden username has to be dropped completely and therefore it is necessary to group together all the entries which are from the same warden in a certain time period. This period is measured as the distance between two entries of the same warden and is set to 15 minutes. This value is big enough to cover all cases where a warden needs longer for a single action, but is also small enough to consider breaks or long traveled distances. The common perfect value does not exist for this period, because it is completely different between each warden and the area he is working in. The decision for choosing 15 minutes is an estimation and based on knowledge and logical thinking.

Besides the anonymized data fields, all the other ones are read normally by considering the specific formats like date, time, strings and numbers. The structure of the source data is described in detail in Section 3.3.

The import from the source protocol files also brings up the first errors. About a hand full out of millions of entries were incomplete and started or ended with unreadable characters. It seems that there was an error when writing these lines to the protocol file. The number is so small that it should not have any impact on the outcome.

### 3.4.2 Importing the Street Graph File

The street graph file and its source has already been described in Subsection 3.2.2 and includes additional information for all streets of the city. The file is available in different formats and includes a long list of objects describing the various parts of streets and contains coordinate points for each segment. It has two main purposes in the context of this thesis. The first one is to provide a list of all streets of the city, which can then be used to match with given locations. Therefore the whole street graph file is read and for each street a unique object is generated.

The second purpose of the street graph file is to provide a complete dataset of all possible locations of the city. For each unique street object, all given street parts are added and every part consists of multiple coordinate points provided by the street graph file. A street part describes the course of a road between two junctions and if it is a straight line, only the start and the end coordinate points are given. If there is a curve, the whole part is split into segments and thus provides as much coordinate points as needed for describing the path. The distance between two points is sometimes really long. Theoretically, all possible locations for a checked car are anywhere on or between such given coordinate points.

To limit the amount of locations to a smaller number, which is discrete and faster processable, the street part segments are split into additional points if the distance is too long. This partition is continued until the distance between two coordinate points is smaller than about ten meters. The value of ten meters is chosen, because on a real street a divergence of this size does not have any negative impact in relation to the length of a street and the size of buildings. Additionally, the amount of calculated estimation points is kept relatively small with this parameters. The value is not exactly ten meters because the coordinates are given in the longitude and latitude format and the calculation of the distance is based on the earth radius and the chosen latitude value, which are both variable. The used distance between two latitudes is 111,325 kilometers, the used latitude is the 48th, which is the closest to Vienna. The equation for the threshold calculation is:

$$th = \frac{10}{cos(\frac{\pi * 48}{180})111325} \tag{3.1}$$

This threshold is the reference value which decides if the distance between two points is too big and is split again or if the distance is small enough. The distance itself is calculated as the euclidean distance of the coordinate points. The results of these calculations conclude in about 7000 streets with 28000 street parts and 400000 street points based on the street graph file from the City of Vienna.

### 3.4.3 Extract Addresses of Location Data

The address extraction process consists of two consecutive parts. It has the aim to gain as many pairs of a street name and a house number from the location information of all penalties as possible. The pairs are needed for requesting coordinate points from the address service where combination of street name and house number is essential.

In the first step of the address extraction only the main street data field of the location information is used. It is necessary to determine if the string value contains only a street name or also the optional house number. Therefore the list of all street names is used from the street graph file. The implementation checks the begin of the string value and tries to match it with any entry of the street list.

After recognizing inconsistencies in the spelling of some street names with diacritical marks, a normalization of the values on both sides resulted in a match for nearly every given street value. The number of impossible matches is including 870 duplicates and contains 27 unique street names. Some of these names are just walk ways or green areas where addresses exist but usually no car should be parked and some others also contain encoding errors from the source file. For those entries where the street name matches, the leftover of the string value is considered as the house number of the location. If no leftover is available, the entry is forwarded to the second step.

All location values, where no pair of street name and house number could be found in the first step, are the input for the second step. It has the aim to filter the house number from the location free text, which is entered manually by the warden and has no predefined

format. As this data field is filled manually by the warden, it can include all the well known problems which can arise in such cases. These are for example misspelling, typing errors or different meanings for various abbreviations.

Beside different formats of the free text value, it is also not mandatory for the wardens to type in any house number. This can be the case if there is simply no house close by or if another description of the location is more accurate. Some examples are places of interest, prominent artworks, statues or even numbers of trees.

In multiple steps the implementation uses regular expressions to simplify the location free text to be able to determine a house number:

- At first some specific verbalizations, which are a combination of a word and a following number, like "tree number 12345", are removed from the value.
- Afterwards a check is performed if there is still any digit left in the remaining string. Because if there is no digit left, no house number will be extractable.
- As the meaning of the free text values can be really different, the only chance to determine the house number is that the warden has entered it at the beginning or nearly at the beginning of the text. Therefore a static list of key words in different notations is iteratively removed from the begin of the value. If a number is the first part of the string after this removal, it is chosen as the house number. The list contains for example additional descriptions like "left", "before", "between", simple nouns like "church", "memorial", "entrance" or redundant text like "house number".

In some situations multiple house numbers are entered and thereby up to two possible occurrences are chosen. Later this approach provides a higher probability of requesting any coordinate point for the entered location.

The Table 3.1 shows some examples of location free texts where the house number could be determined or not. In negative cases also a short description is provided why it is not possible. It also shows the special case of manually entered street names which are adjacent to the known one. Basically it would be possible to extract this street name and the house number and use it as the new location pair, but these cases are not considered because the spelling of the street names are too different and there would be too much errors.

In total nearly 1.28 million penalty entries with location data have been analyzed, which also includes 67000 free text value checks. After eliminating duplicates, the list of values contains about 22300 unique strings, which are the input for the described step two. Slightly more than 50 percent, namely 11700 additional house numbers could be extracted from the free texts.

The diversity of location free text values is very high, and there is still a risk of false positive extractions of house numbers during the removal process described before. Due to the

| Values with determined house number(s) highlighted | |
|---|---|
| **146** a **144**<br>between **9**a.**11**<br>next to **7** before school<br>between opposite **15** and street name 69<br>nearest to house number **14-16** before park | |
| **Values without a determined house number, inclusive description** | |
| right of tree no. 1234 | No house number exist |
| next to church row 2 | "row" is not removed and therefore "2" is correctly not determined as a house number |
| nearest to street name 13a | Manually entered street names are not checked. "Street name 13a" would be the correct location here |
| station forecourt next to number 43 | House number "43" would be correct here, but "station forecourt" only occurs one time and therefore the word is not added to the static list |

Table 3.1: Examples of location free text values with positive determined house numbers and negative cases where no house number could be determined.

nature of the process of entering values into this data field, the described implementation is correct in nearly every case. After the execution of the final implementation, a few hundreds of extractions have been checked together with the original location free text and no false positive example has been found.

### 3.4.4 Request Coordinate Points

In the previous section the location data is analyzed to gain pairs of street name and house number. These value pairs are now used to gain additional location information for the penalty entries in the form of coordinate points. The source of the coordinate points is a web service described in Subsection 3.2.1, which needs a combination of street name and house number as a parameter.

Before the web service is called, all pairs of street name and house number are checked and duplicates are eliminated to minimize the number of requests which have to be sent to the web service. As already stated previously, the spelling of the street name can differ in the data sources and also the extracted house number of a location free text field can be more than one number. These facts make it necessary to provide different combinations of a street name and a house number for each data pair. If the response of the first combination sent is insufficient, the next available combination is tried. The

number of extracted locations is about 150000 still including the duplicates, i.e. this is the number of street name and house number pairs. It does not include different combinations, especially if there are two house numbers available.

After all the requests are finished, the total sum of addresses with coordinate points is 63700 and basically does not include duplicates. The number of data pairs which could not get any coordinates is about 6000, i.e., four percent of all extracted location pairs of street name and house number could not get any coordinate. The reasons for the lack of coordinate points are very diverse. Sometimes the extracted address does not exist, which can be an error of the warden or of the previous implementation steps. In other cases the address service has no information about the given address and the response is insufficient, even if it exists in the real world and is a correct address.

The response of the web service is in the GeoJSON format and contains a collection of predefined features. If the parameters are not clearly defined, the web service responds with a list of similar addresses. As an indicator for the suitability of the response, every entry contains a "Ranking" value. It ranges from 0.0 for unique results, to very high values, like 400, if only the street name could be matched and the house number is unknown. In this cases, the closest possible address, based on the house number, is the provided response with a high "Ranking" value. At first, the ranking is used to determine if the response is suitable, afterwards the values of the returned "StreetName" and "StreetNumber" are checked individually.

The response of the web service is very clear in most of the cases. During the implementation, for some pairs of street name and house number, some undocumented inconsistencies could be found regarding the content of the "StreetNumber" feature. It can contain for example additional information like stall, school or stair values, but there is no consistent formatting for these extensions and additional parsing would be necessary.

If a suitable feature collection is found for a pair of street name and house number, the features are further processed. The main information is the bounding box which includes the coordinate points of the area of the received address. The GeoJSON object provides additional information of the requested location, for example the district, postal code, municipality, subdivision and street type.

### 3.4.5 Calculate Penalty Coordinate Points

The results of the previous step are the bounding box areas of all the locations where coordinate points could be gained. Each bounding box represents the area of a specific address next to a street and it depends on the location and the size of the property or the building. For each bounding box a point on the street, which is close by, has to be defined. A bounding box consists of the most southwesterly point and the most northeasterly point, and therefore the center point is calculated by these two points. Based on the center point of the bounding box, the closest street point of all calculated points is used as the "point on street" of an address. As a limitation for the calculated points, only those points are considered, which are situated on the street that is entered

by the parking warden, i.e., if a point of another street is closer, it is not considered. This "point on street" is thereby the estimated location of a fined car, based on the entered location information of a parking warden.

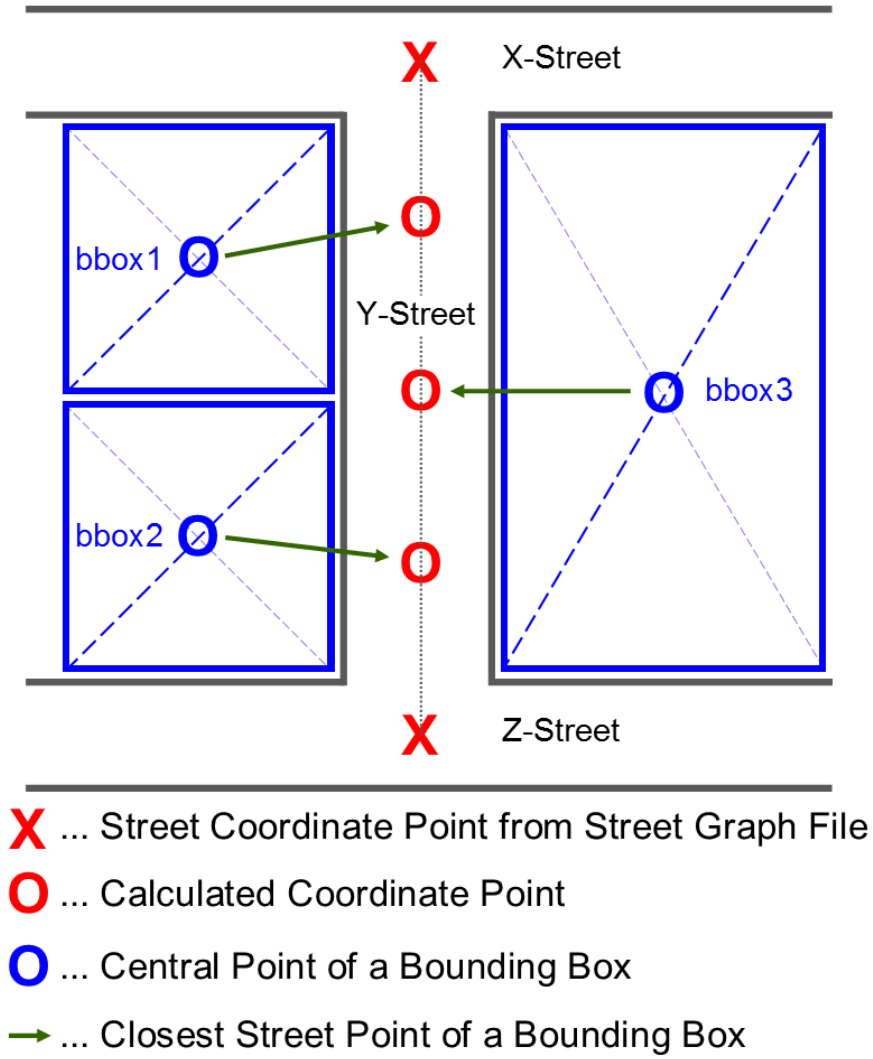


Figure 3.3: Simplified representation of three bounding boxes and the calculated points of Y-Street. The center point of a bounding box is calculated by the bottom-left and top-right point of a bounding box. The addresses lie on Y-Street and the green arrows mark the closest "point on street" of an address.

A simplified example assignment of a "point on street" is shown in Figure 3.3. It contains three bounding boxes of addresses, which are lying on the Y-Street. The Y-Street itself is defined in the street graph by two coordinate points, marked with the red crosses. The distance of the red crosses is too far and by splitting, three more coordinate points are

calculated, shown by the red circles. Next to the street, three addresses exist, represented by their bounding boxes with blue squares, which can differ in size. The center point of each bounding box is calculated by the bottom-left and top-right point, shown by the diagonals and the blue circle. To gain a "point on street" for an address, the distance from the center point of a bounding box to all points of Y-Street is compared and the closest point is chosen, which is depicted by the green arrows.

While this method can lead to good results if the street and all the addresses are mainly standardized, it also has its limitations. If the bounding box of an address is very large and consumes a lot of space next to a street, it is always assigned to one "point on street", which is approximately in the middle. The entered address of a parking warden states that the car is parked next to a house number, but sometimes also contains more additional information. This additional information is not interpreted by the implementation and therefore a more precise location description gets lost. As a consequence, all coordinate points that are assigned for cars are only approximations with the highest possible accuracy based on the described approach. Considering the previously available location information, which is only limited to the district and maybe the street name and house number, the described coordinate point approximation provides additional benefit to the data.

### 3.4.6 Estimate Request Coordinates

In the section before, an approximation is calculated for the addresses that are given of a fined car. The next step is to use this additional information for gaining more knowledge about the location of all the requests which are in between the penalties. As already mentioned, it is not possible to gain coordinate points for all penalties due to different reasons. In the following estimation process a penalty without coordinate points is considered as a normal request, to get an estimation as well.

Every sequence of entries generated by a warden is unique and does not follow any prerequisites, i.e., the amount of requests and penalties is completely random. As a consequence, the type of coordinate points estimation depends on the position of an entry in the sequence of a warden. In total, four different types exist and are described in the following:

- **Calculated Penalty Coordinate Points**

  The penalty coordinate points mean the points on the street, which are calculated based on the bounding box of the given address. It is the most accurate type of coordinate point estimation and is described in detail in Subsection 3.4.4.

- **Estimated Coordinates by Street Points**

  For being able to use this estimation based on street points, some preconditions have to be fulfilled. It is only applicable if the request is between two penalty entries, both lying on the same street and if they already have penalty coordinate points.

The number of requests in the same situation do not influence the estimation. At first, the distances of all the street points which lie between the two penalty entries are calculated. This also explains the precondition for the same street, because otherwise a traveling salesman problem occurs and all the street points of all other streets have to be considered. Then the timestamp of the request is used for calculating the distance to the penalties in seconds. After calculating the ratio of the complete distance for both types of distance value, the street point with the closest value is chosen as the estimated coordinate point for a request. This is the most accurate, possible coordinate point estimation for a warden request, since there is no location information provided by the source data.

- **Closest Penalty Points**

  In many situations, a request is not lying between two penalties on the same street. This can be either because it is at the border of a sequence, or if the penalties before and after a request are on different streets. No matter which case it is, the closest known coordinate point is used as an estimation. The distance is again calculated by comparing the timestamp of the request and the penalties before and after it. If the request is at the beginning of a sequence, the coordinate point of the first penalty is used and at the end of a sequence, the last penalty is used. Depending on the structure of a sequence and the time between a request and the used penalty, the result can be more or less good. For increasing the significance of this estimation type, the distance in seconds is also stored together with the coordinate points. This approach gives a further analysis the possibility to consider the weight of coordinate points, which are generated by this type of estimation. When storing the closest penalty coordinate points, additionally the street name and the house number of the closest penalty are also added.

- **Absolutely no Coordinate Information**

  There exists special cases where absolutely no coordinate information can be gained for the entries of a sequence. This can be either the case if the whole sequence does not contain any penalty, or if none of the existing penalties could get any coordinate point.

In Figure 3.4, the process of gaining the estimated coordinate points for requests is shown. Step 1 depicts the randomly iterating requests and penalty entries generated by a parking warden. At this moment all the penalties are considered as equal. In the next step, the penalty without coordinate points is excluded and from that moment on it is interpreted as a request, illustrated by the red coloring. In step 3, for each penalty the address information, consisting of street name and house number, is added. The red penalty also has some address information, but no house number could be gained in the previous implementation steps, which is also the reason why no coordinate point is available. Before the next step, all requests which are not between two penalties of the same street are skipped and the closest penalty coordinate point will be calculated

afterwards. In step 4, for each list of requests between two penalties of the same street, the distances are calculated and the closest street points are assigned as the estimated coordinate points.

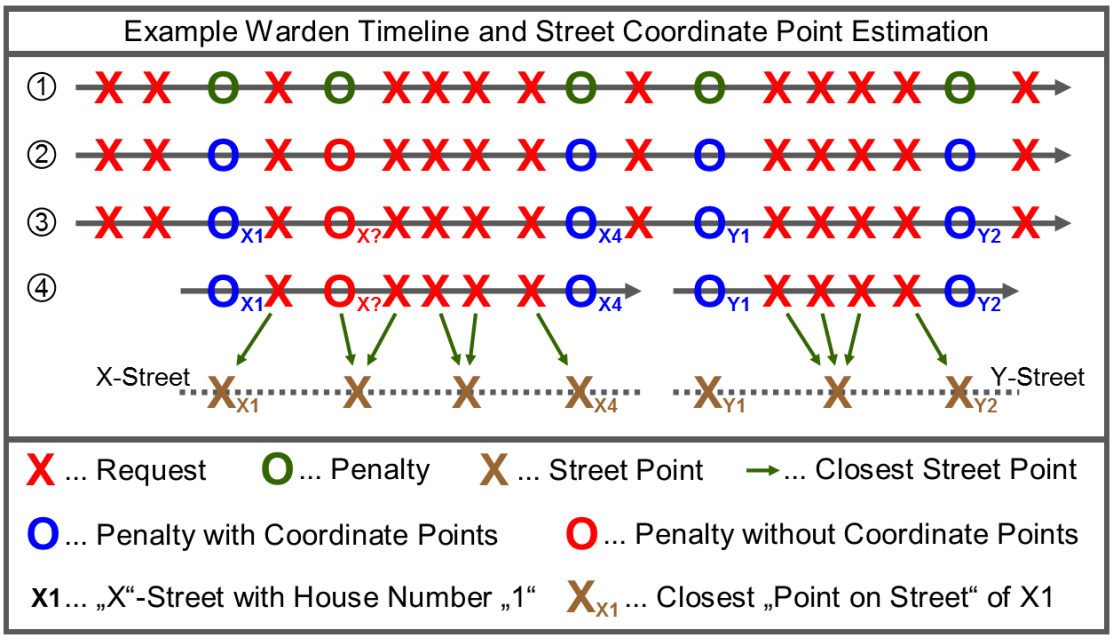


Figure 3.4: Example timeline of a warden, including illustration of gaining the estimated coordinate points for requests between two penalties of the same street.

The four named types of coordinate estimations, starting with the calculated penalty coordinate points and ending with absolutely no coordinate information are the main answer to research question one, stated in Section 1.3. When the data is used for further analysis in Chapter 5, additional details like the amount of possible estimations of each type are provided.

### 3.4.7 Individual Exports

As already stated in Subsection 3.3.5, it is necessary that the implementation provides different output files depending on the later usage. The output files are then considered as the input files for the analysis tool and therefore need to be in a special format. For the given data provided by the City of Vienna, which includes millions of entries, in some cases an iterative approach is used for gaining the data of the whole year. Therefore the size of the processed data is smaller and the output can be handled easier. If only the data of a single district is used, another file is generated which only includes a subset of the entries and the needed properties. Which data is used for further analysis is explicitly described when used in Chapter 5. Additionally to the used data subset, also the format of the provided files can be different.

CHAPTER 4

# Evaluation

The evaluation concentrates on the outcome of the whole data transformation process, which is starting with the data import and ends with the data in the correct format for further analysis. The main key of this process is to gain coordinate point information for penalty lines, because they are then used to estimate the position of all the other entry lines. In this chapter, a map of an example district of the City of Vienna is used to demonstrate the possibilities and the outcome of the presented implementation, but also to show the weak spots of the whole process.

The chosen district for this evaluation is district six, which lies close to the center of the city and generally has the same parking rules in the whole area. Additional numbers and facts about this district are given in Chapter 4. At first, Figure 4.1 shows a standard map of the sixth district of Vienna including border lines to the neighborhood districts. The source of the map is the website of the City of Vienna[1] and it is provided by Stadt Wien - ViennaGIS[2]. It does not contain any additional markers or symbols to get a better understanding of the appearance of the district.

For the next diagram, the data of the whole year 2018 is used to show the results of the data transformation process. The exported entry lines are filtered, to show all penalty entries which contain successfully calculated coordinate point information. In Figure 4.2, all the described points are printed on a two-dimensional plot. It depicts the longitude coordinate point values on the x-axis and the latitude on the y-axis, exactly like on a real map. All points together form a grid-like arrangement of lines, which stands for the streets of the district. The outer borders of the district area are also nicely shaped and the positions of all the points seem to follow the correct ways.

The presented Figure 4.2 itself only allows a rough guess that the implementation and the coordinate point calculation is correct, but the rightness can be illustrated much

[1] https://www.wien.gv.at/stadtplan/, Accessed 2019-08-10
[2] http://www.wien.gv.at/viennagis/, Accessed 2019-08-10

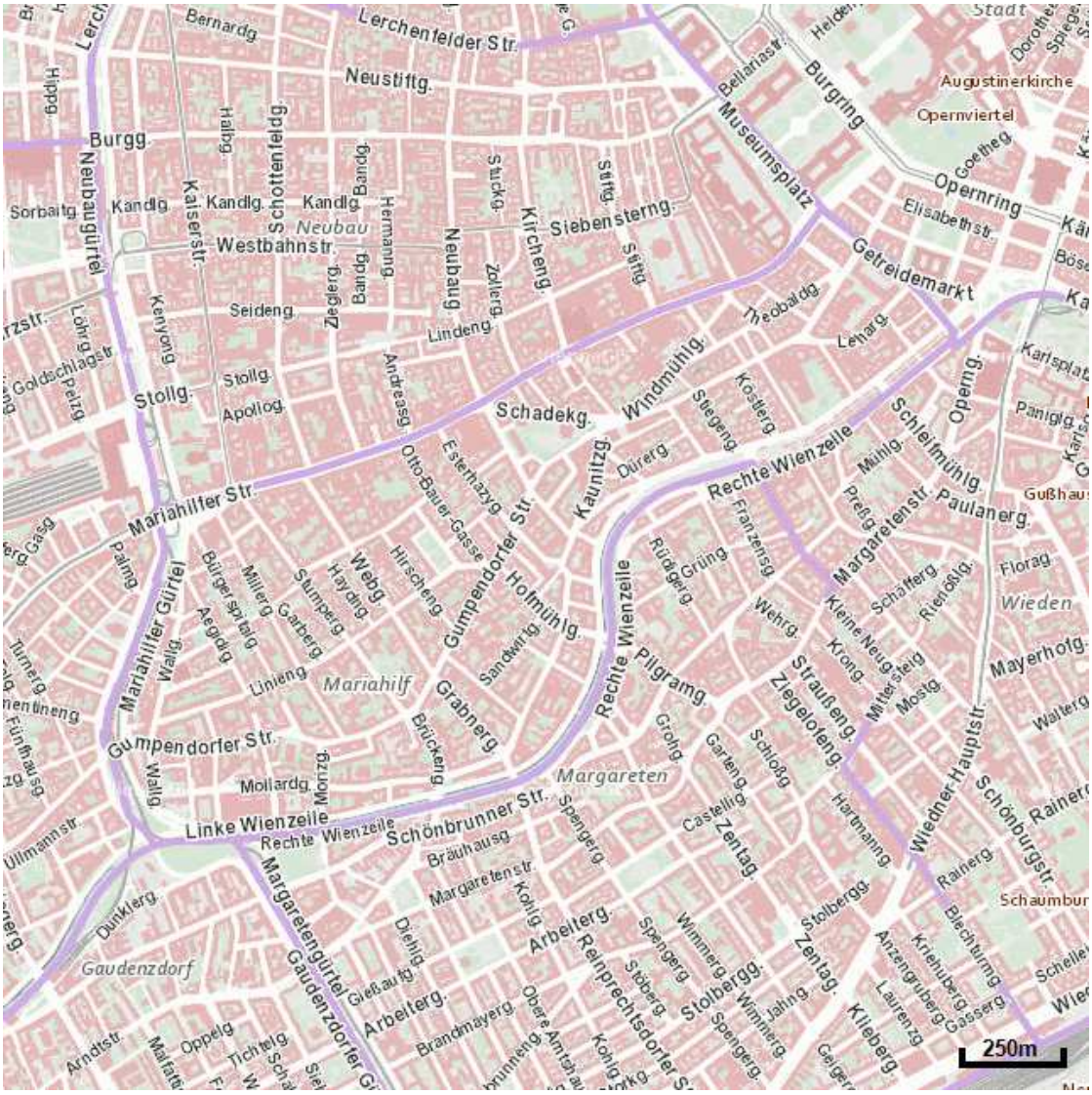


Figure 4.1: A map of district six in Vienna with border lines included.

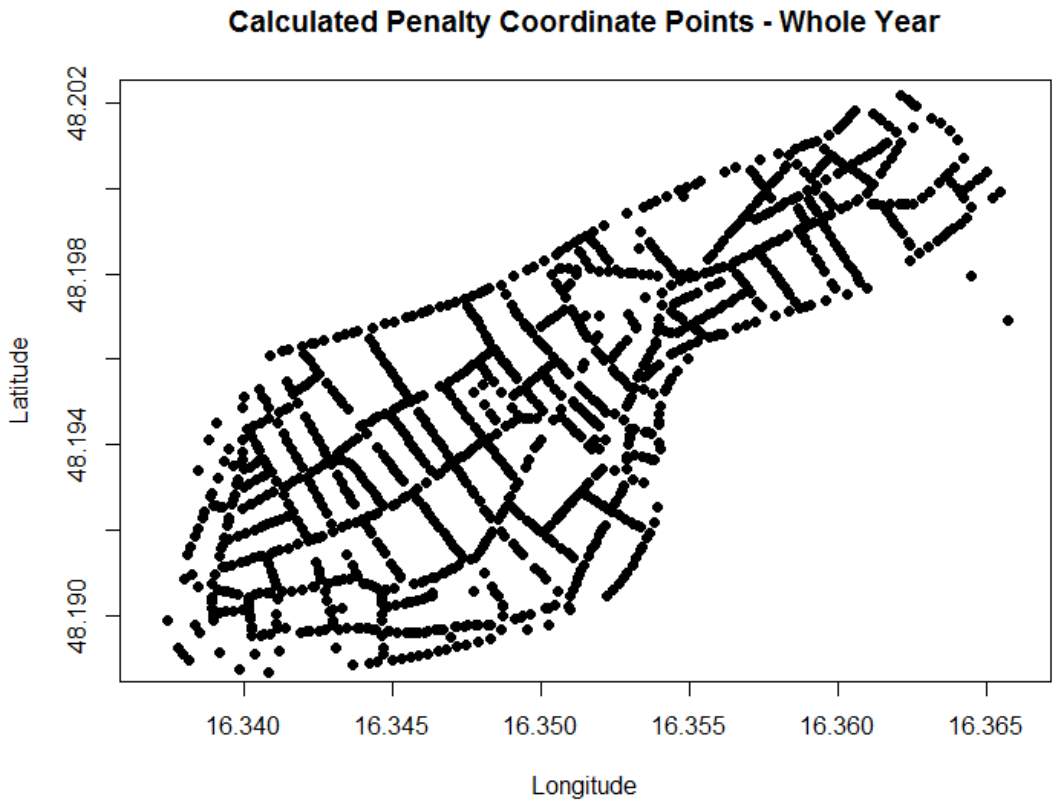


Figure 4.2: Plot of all calculated penalty coordinate points of district six.

better by an additional visualization. Therefore, Figure 4.3 combines both previous images together to present the correct position of all the penalty entries which are directly situated on the streets of the standard map. The points of the diagram have not been manipulated except of some scaling to directly fit the size of the underlying street map, but no single points have been removed or edited. The largest part of the streets grid is nearly completely covered by black dots, although all the streets have been divided into predefined segments and only these are used as calculated coordinate point positions. In general, the map shows that the outcome is very good, fulfills all the needs and the data can be used for further analysis.

Besides the successfully pictured penalty coordinate points, Figure 4.3 is also used to describe some of the problem situations and weaknesses of the used approaches during the data transformation. The image has eight marked areas in it, which are named with letters from A to H. Those markers are used to describe some example situations where the applied processes work very well, but also has its limitations when trying to gain coordinate points for penalty entries. The positive, as well as negative examples are described in the following list:

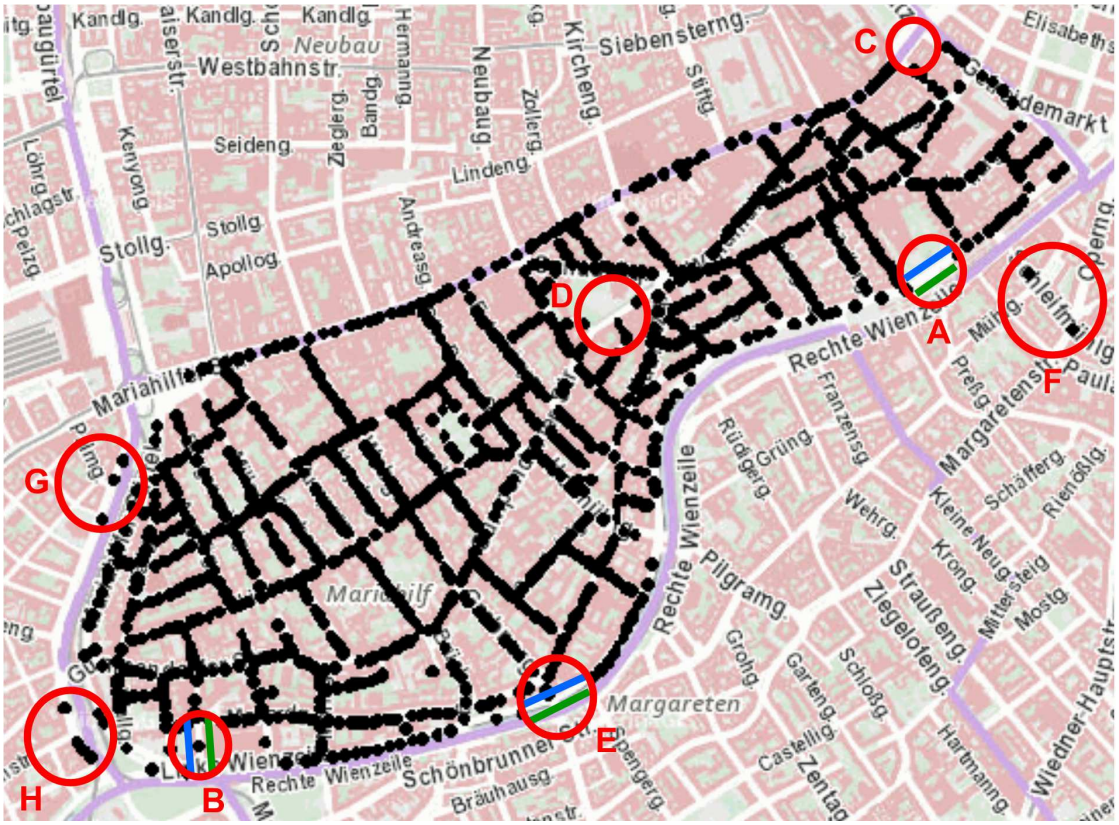


Figure 4.3: Map of district six including coordinate points of all penalty entries. Additionally some special areas are marked and named with letters.

- **A - No House Number** There are many parking spaces in the region of marker A, but if a car is fined there, no coordinate point can be gained in this case. The reason is that all calculated coordinate points for penalty entries depend on an available and successfully detected combination of street name and house number, but in the presented case there is no normal house close by. On one side, which is indicated by the blue line, there is some kind of public green space. On the other side, which is indicated by the green line, there is a market with only small houses which do not have their own official house number. As a consequence, a parking warden cannot enter a detectable street name and house number combination and penalties which would have been in this area are stored without coordinate points.

- **B - Big Buildings with One Address** In the small street at marker B, there is only one single black dot, while theoretically the area would be much bigger and all penalties in this area could be distributed along the road. Furthermore, there are multiple parking spaces in the small street. The implemented algorithm for gaining coordinate points depends on combinations of street name and house number and therefore a black dot can only be placed in front of a house. Besides

the existence of a house, it is also necessary that it has its own house number at the position of the street. The situation at the presented marker is, that on the side of the blue line, there is one big block of a house, but it just has an accessible door on the side of this road and hence there is no house number which can be used by a parking warden. On the other side, indicated with the green line, there is another big house, which also only has one single door on the side of this road. As a consequence, all fined cars in this street are at the same position, which is the only available combination of street name and house number, or for the other fined cars a warden has entered something else which is not recognized by the algorithm.

- **C - No Normal Parking Spaces** At the position of marker C is no known penalty entry coordinate point because the only parking space in this area is limited to taxi drivers. If a normal car would have been parked there and gets fined, there is a house number available which could have been used by a parking warden. Such a case is not present in the data.
- **D & E - No Parking Spaces** In the areas of marker D and E no black dot for a penalty coordinate point is shown. The explanation is rather simple, because on both examples there is just no parking space in the area of the streets. For marker D the segment of the street is very narrow but although has two to three lines, while for marker E the situation is similar. On the side of the blue line there is a sidewalk with fences instead of parking spaces and on the other side with the green line, there is a flowing river without enough space in between to park cars.
- **F - Error in Source Data** Each parking warden is selecting the district he is working in by its own and as a consequence this is a possible situation for errors made by humans. Such a case is shown at marker F, where two penalty entries do have coordinate points which are totally outside of the districts area, but both entries do have district six written in their respective field. As these two entries are very close to the border of the sixth district, it is possible that the warden just forgot to change the district on his device to district four.
- **G - Error or Ignored Opposite Flag** The area of marker G also shows three black dots outside of the border of district six. One reason can be the same like marker F, which would mean that the parking warden has forgotten to change the district field. It is also possible that the warden uses the house number of the other street side and sets the opposite flag, which is not considered in the data transformation in this thesis. For most of the streets it does not matter if the opposite flag is dropped, because the street normally consists of one line in the street graph. In case of marker G the street is very wide and consists of two lines in the street graph. As a consequence, an entry line with opposite flag set will place the coordinate point directly in front of the house on the other side street line and like in this situation can be outside of the districts border.
- **H - Parking Space on District Border** At marker H the described situations of marker F and G can also be the reason for the appearance, but additionally there

is a small parking lot directly on the border of the districts, but there is no extra house number at this location. This means that a parking warden has to choose either the house number of the nearest house in district six or a house number of the other side. In some cases it is possible that the warden uses another description and does not enter a street name and house number combination, which is then not recognized by the used algorithm.

Besides all the previous points, for sure it is also possible that on some locations there just was no fined car in the whole year of 2018 and this can also be the reason why no black dot is shown on some specific spots on the map.

CHAPTER 5

# Analysis and Visualization

This chapter combines the findings of Chapter 2 and the outcome of the implementation described in Chapter 3 and applies it for answering the research questions, stated in Section 1.3. The first section describes the relevant indicators and data fields which are investigated in the analysis. Afterwards, all the logical levels of detail are stated, both for the dimension of time and for the place. The combination of different indicators and all the various levels of detail result in a high amount of combinations and therefore not all of them will be handled in this thesis. As a consequence, the usage of visualizations is also customized, based on the current combination, which should maximize the visibility of the characteristic outcomes. The next section describes the chosen subsets of data and the reason for their selection. Afterwards the main analysis and visualization of the chosen indicators follows.

## 5.1 Relevant Indicators

The data fields which are available for each entry have already been described in Section 3.3 including the penalty extension containing the location information. Some of the features are very interesting for further analysis, while some others are indirectly in use. There are also some data fields which can not be used or are not useful due to different reasons. The following list describes all the mentioned data fields of the previous chapters and if they can be used.

- **Date and Time**

  The date and time information itself as a feature is not directly useful for further analysis. It is although mandatory because otherwise no further separation about the level of detail in the dimension of time would be possible in the following section.

- **Warden Identifier**

  The warden identifier is only available for the very first anonymization step and therefore is no option for analysis.

- **License Plate**

  The license plate information is also part of the anonymization, but is still available as categories. The country is part of the license plate and the origin of the cars can be of high interest and should be investigated more closely. Besides the country, for Austrian license plates the identification letters can additionally be used for examining the origin of the cars.

- **Basic Location**

  The basic location mainly refers to the district where the car is parked when the warden performs the request. It has the same characteristic as the date and time, because it can be used as a level of detail of the place dimension.

- **Request Type**

  Theoretically the request type contains the information if the car is already known or not, because of the RFID tag. Indeed this difference can be seen for all the request lines. All penalty lines however, have the same request type and therefore this information is not available for all entry lines. An analysis of all request lines and the ratio of cars with a RFID tag can although be performed.

- **Street and Number**

  The pairs of street name and house number depend on the outcome of the data transformation and as a result are not available for every penalty entry. Only the street name is given for every penalty and it can be used as an additional level of detail of the place dimension.

- **Additional Text**

  The additional location free text is already used in the data transformation to gain a higher amount of pairs of street name and house number. While the diversity of all the texts is very high, it can not be used for further analysis.

- **Opposite and Frontage Road Flag**

  As the whole process of finding coordinate points is focusing on the center of a street, both flags can not be used.

- **Additional GeoJSON Data Fields**

  After requesting the coordinate points from the address web service, some additional information, like municipality, subdivision and street type is gained. Some of the new data fields have the same meaning as the district and others are not useful for further analysis. Additionally only the penalty entries, where coordinate points have been found include this information and all other entry lines do not.

- **Different Coordinate Points**

  During the data transformation different types of coordinate points have been gained for penalties and requests. These coordinate points are the most possible accurate location estimation and can be used as a level of detail in the place dimension.

Besides the main data fields itself, there is additional information given for each entry, which has not been stated yet. This extra details are included in the other data, but for analysis reasons they are extracted and stated explicitly to be more consistent.

- **IsPenalty Flag**

  This flag directly contains the information if the current entry line is a penalty or not. It is generated if a penalty extension exists.

- **IsPenaltyWithCoordinates Flag**

  This is an additional flag, which states if for the current penalty a coordinate point could be gained or not. It just means the direct coordinate points and does not include the closest coordinate point type.

The whole design of the parking enforcement process and the origin of the source data generates an additional problem before starting the analysis. The simple amount of occurrences of a feature in a defined level of detail does not have high significance without further modifications. This is because there is no predefined interval or structure when the parking enforcement is performed.

As an example, the statement "There were 100 penalties in district X and 200 penalties in district Y.", after analyzing the given input files, can clearly show the main problem. Even if the time period of the measurements are the same and if both districts X and Y have the same size, it still does not have a high significance. The main reason is, that it completely depends on the amount of performed parking enforcement actions in the evaluated time period and district. If the time period is one week and in district X only one parking warden is working and he has done 1000 requests in the complete week, the 100 penalties are a lot, namely 10 percent. In district Y, on the other hand, are working ten wardens and they perform 10000 requests in the same time period, then the 200 penalties are few, namely 2 percent.

As a consequence of this findings, the type of features which can be analyzed, based on the input files from the parking enforcement process always have to be a ratio of the total amount of observations in the same place and time. With this knowledge gained, after looking at the provided list of data fields, a new list of two main relevant indicators for analysis is stated as:

- **Penalty Ratio**

The term "penalty ratio" means the proportion of penalty entries to the total amount of requests which have been performed. This indicator therefore gives a feeling of how likely it is that a car gets fined depending on the different levels of detail. By analyzing this feature the outcome shows the differences for various time periods and location areas of the penalty ratio.

- **Origin of Cars**

  In the context of the cars origin, various possibilities are given for further analysis. One aspect can be the ratio between foreign license plates and Austrian cars, as well as the ratio of license plates which are from Vienna in comparison to all the other categories. Additionally to the investigations of all entry lines, the analysis of fined cars in this context can be much more interesting.

It has already been stated at the begin of this chapter, that there are many combinations of levels of detail in both dimensions and the provided list of analyzable data fields. Therefore some specific visualizations and analyses are only performed in some special cases to highlight a particular behavior or characteristic. In the majority of cases and if nothing else is stated, all the approaches can also be used for other investigated features. Before a feature is analyzed in detail, sometimes the distribution is also provided to illustrate the reason for the chosen combination of the levels of detail.

A special case of analysis is provided by the research question about the parking density. Therefore new approaches have to be used, but all the necessary changes are described in detail in the respective section.

## 5.2 Levels of Detail

The input files contain spatio-temporal data and therefore the levels of detail are parted in two dimensions, namely in time and place. In both dimensions, many different classifications are possible, which are described particularly in the following subsections. There are two main reasons for these dimensions, which are coming from different point of views:

- **Considering Parking Rules**

  As already mentioned in Subsection 3.2.3 the decisions during the enforcement process are highly influenced by the applied parking rules in an area. A parking rule is usually defined as a combination of a time period and a certain area in which it is applicable. The rules state directly if a car driver needs a parking ticket or not, i.e., if he will be fined in case of an enforcement check or not. Therefore these prevailing circumstances are already predetermining the two dimensions which are sensible for analysis.

- **Investigating Driver Behavior**

Another aspect, which can also be considered independently of the parking rules, is the behavior of car drivers. As it is already well known from other research areas, the span of attention, reaction rate, potential for aggressions and other personal characteristics have high variations during the day and in certain situations. Some examples are being tired in the morning, having a higher potential for aggressions after a workday or feeling more relaxed on the weekend. These characteristics can influence the parking behavior of a driver and therefore are a reason why the time dimension is of high interest.

Personal needs are also affecting the place where a car driver is parking his car, based on different concerns in the city area. This can include orders during a workday, going shopping at different points of time or visiting relatives and friends on the weekends. All of these personal issues additionally influence the places and areas where people go with their cars.

### 5.2.1 Dimension of Time

Before considering the possible separations in the dimension of time, it is important to state again that the provided source data consists of all entry lines of the whole year 2018. Therefore the data is limited to a period of one year, and beyond that no further classification is useful because no data is available. A selection of possible separations is provided in the following list:

- **The whole Year**

  For the whole year the absolute amount of occurrences of all the features can be provided.

- **Half a Year**

  In the economy a half-year report is common, here it would result in two value sets because the source data is only available for one year. The separation could be performed directly at the half of the year, or by splitting the data into a cold and a warm season of the year. If the second option is chosen, there are different ways when the split could be performed, for example by the name of the months or at the middle of a solstice. Similar considerations are needed for the four part splitting. Also the two sets of six months in a row with the highest difference can be an option.

- **Four Part Splitting**

  When considering four parts of a year, there are many possibilities which can be considered. In the economy the year is split into four quarters, each containing three months, starting at the first of January. Another option are the four seasons of the year, spring, summer, autumn and winter but also therefor two alternatives exist. There is the meteorological classification, consisting of three months, starting with the first of March. On the other hand is the astronomical classification which

considers the real position of the sun and for example has the start of spring approximately on the 20th of March.

In the context of this thesis, the astronomical approach is not considered. The data is easier to handle if the months start and end with the first or the last day of a month, which is applicable for the meteorological and the economic approach. The same statement can also be applied for a half of a year.

- **Monthly View**

  The monthly separation is straight forward and consists of the 12 months of a year. Summarized values of all entries in a month can also be used as the data for smaller levels of details in contrast to the total view of the whole year. For example by comparing the months with the highest amount of data.

- **Weekly View**

  Similar to the monthly view, the weekly separation just splits the year into all the weeks. As the year 2018 starts with a Monday and the last day of the year does not have any enforcement data, this approach results in 52 weeks. The summarized values can also be applied on smaller levels of detail, for example by looking at all weeks of the month with the highest amount of data.

- **The daily Level**

  A day is the standard unit, which is used for an aggregation of the entry lines. This means that a year shows 365 days, a month up to 31 days and a week shows seven days. Besides this classical approach, a day can be used in much more ways.

  It is possible to consider all weekdays based on a smaller level of detail, for example all Mondays of the whole year. Each weekday can be compared to all the other days of a week, like differences between Mondays and Wednesdays. The average business day can be considered in comparison to the average weekend day, for different time periods. Basically also the number of a day in a months can be investigated, but this should not have any significance away from the regular findings.

- **Considering Time of the Day**

  Besides the standard approach of using the data of the complete day, also the parts of a day can be analyzed furthermore. It can be classically split by an amount of hours, or in dependence of the usual working times, when before and afterwards naturally most of the traffic takes place.

The list contains many possible levels of detail and as a consequence the analysis sections do not always depict each combination of a feature and the levels of detail. Most of the approaches which are only present in specific sections, can also be applied for other features by doing smaller changes, if not stated otherwise.

Besides the theoretical levels of detail, the analysis should also show the limitations of the source data, in the meaning of the given data fields as well as by the amount of entry

lines. Even if the source data contains millions of lines, the parking wardens are not continuously present on each street to directly request every car which is parking and therefore some time periods will not contain many data lines.

### 5.2.2 Dimension of Place

In contrast to the large list of possible separations in the time dimension, the options in the dimension of place are less. The reason is the structure of the data, which only provides limited information for the location of a requested car. The source data is only available for the City of Vienna, which makes the whole city the smallest level of detail. Nevertheless, in the following, possible levels of detail are described:

- **The whole City**

  The whole city is the smallest possible level of detail and consists of all the districts. Due to conditions specified by the City of Vienna, a common overview of all the entries from all districts of the whole city is not allowed. Instead, a detailed comparison of different districts can be made. The highly personal data and the public interest are the reason for this limitation.

- **Districts**

  The district is the first level of detail in this dimension, which can be analyzed without limitations. It is provided for every entry line and the only location description for the requests. In the City of Vienna, besides some special streets, parking rules are defined per district and can vary greatly.

- **Rayon**

  A rayon is a more detailed part of a district, but it is not provided correctly in the source data and only contains the district information, as already mentioned in Section 3.3. Therefore it is not relevant for further analysis.

- **Street Names**

  The street names are entered manually by the warden, but are only available for penalty lines. By the results of the data transformation process it is also estimated for request lines. A street can be very short, but also be very long, it can even run through several districts. The combination of street name and district limits the area, but it is still very large for some streets. There are thousands of streets in Vienna, which is the reason why it is no useful classification.

- **House Number**

  The house number, in combination with the street name, is in the source data only available for some penalty lines. After the data transformation the amount is increased and also an approximation exists for a certain amount of entry lines. It also limits the area to a smaller region, but the number of unique location

combinations of street name and house number is very large and therefore it cannot be used as it is for further analysis.

- **Coordinate Points**

  At the end of the data transformation process a high amount of entry lines includes some classified coordinate point information, as described in Subsection 3.4.6. Therefore a two dimensional place definition is provided including a ranking of the relevance of the value. This information can be used for detailed analysis and visualizations for various levels of detail. In different sizes, grids can depict the chosen features for certain areas.

## 5.3 Used Data Subsets

As already mentioned before, one of the conditions for using the data of the City of Vienna is that only different districts should be described in detail in the context of this thesis. Therefore a specific district is chosen by some criteria to perform the main part of the analysis.

For being able to get the best results regarding the research questions, the chosen district is the one with the most available data from the source files. As the number of entry lines from a district does not say much about the real density of parking enforcement, the total amount is divided by the size of the district. The data for the size of the districts is taken from the website of the City of Vienna[1]. Table 5.1 depicts the first three districts with the highest total amount of entry lines and the three biggest districts based on their area size together with the three districts with the highest ratio of entry lines to district size. It has to be considered that some districts have no special parking rules in their area and sometimes they just apply for a part of the district. The four large districts 13, 21, 22 and 23 are skipped because there are no special parking rules regarding to the website of the City of Vienna[2] and thus the amount of parking enforcement is very low. In total, the City of Vienna consists of 23 districts. The table shows that the biggest selected districts are ranked on the last three places of the ratio ranking of entries to district size, even district ten, which is the one with the highest amount of entries.

Besides the ratio between the number of entry lines and the size of a district, the ratio between the number of requests and penalties is also of high interest based on the identified indicators in Section 5.1. Therefore Table 5.2 provides a list of the previously seen top three districts of entries to district size ratio including the added number of penalty lines. Additionally the ratio of entry lines to penalties is shown, which should not be mixed up with the requests to penalties ratio that is considered in the later analysis.

Based on the numbers shown in the table, district one and district six are chosen as the main base districts for further analysis. Both have a rather small area with a high

[1] https://www.wien.gv.at/statistik/lebensraum/tabellen/nutzungsklassen-bez.html, Accessed 2019-05-30

[2] https://www.wien.gv.at/verkehr/parken/kurzparkzonen/, Accessed 2019-06-06

| Comparison of Special Districts - Entries per District Size | | | | | | |
|---|---|---|---|---|---|---|
| District | Entries | Entries # | Size (ha) | Size # | Ratio (≈ %) | Ratio # |
| 5 | 1335278 | 5 | 201 | 19 | 6643.00 | 1 |
| 6 | 873551 | 10 | 145 | 22 | 6024.00 | 2 |
| 1 | 1697057 | 2 | 287 | 18 | 5913.00 | 3 |
| 3 | 1610261 | 3 | 740 | 13 | 2176.00 | 9 |
| 10 | 2642882 | 1 | 3182 | 6 | 830.60 | 14 |
| 14 | 720737 | 13 | 3376 | 4 | 213.00 | 17 |
| 19 | 84813 | 21 | 2494 | 7 | 34.01 | 19 |

Table 5.1: Amount of entry lines in comparison to the size of a district, sorted by the amount of entries per ha of the district size.

| Comparison of Special Districts - Entries to Penalties Ratio | | | | | | |
|---|---|---|---|---|---|---|
| District | Entries | Entries # | Penalties | Penalties # | Ratio (≈ %) | Ratio # |
| 1 | 1697057 | 1 | 158930 | 1 | 9.37 | 1 |
| 6 | 873551 | 3 | 50298 | 3 | 5.76 | 2 |
| 5 | 1335278 | 2 | 55909 | 2 | 4.19 | 3 |

Table 5.2: Amount of entry lines in comparison to the number of penalty lines, sorted by the ratio of entries to penalties. Listing only the districts with the highest entries to district size ratio.

amount of entry lines available. District six consists of a clear street structure with a few exceptions which can also be seen on the map[3]. The streets follow a grid like scheme with only a few exceptions consisting of diagonal, curvy or deadlock streets. It also has many small properties like residential buildings with own house numbers and just a few big, campus like areas, which only have a single house number. Additionally, there are some shopping streets on the borders, where the parking behavior can be different.

In contrast, district one, as the central district of Vienna, consists of various special situations regarding the structure of the streets, which is also visible on the map[4]. Parts of the district are featured by many points of interest and are highly touristic.

Before the main analysis of the data goes on it has to be mentioned that there are special time periods in the data. These periods have hardly any additional information and are the reason for non existing source data on some days of the year:

[3] https://www.wien.gv.at/verkehr/parken/kurzparkzonen/bezirk06.html, Accessed 2019-06-06

[4] https://www.wien.gv.at/verkehr/parken/kurzparkzonen/bezirk01.html, Accessed 2019-06-06

- In the middle of June, on the weekend of the 16th, a part of the enforcement system changed and therefore no parking enforcement took place.
- Due to an technical error in September starting with the 14th, no protocol file was generated for three days.
- On some special holidays, like Christmas and the first and last day of the year, no parking enforcement is performed.

Some of these facts are clearly visible in various parts of the analysis, but some others do not influence the results. If those special situations are visible, it is also described as a part of the analysis.

## 5.4 Penalty Ratio

As already described in Section 5.1, the penalty ratio means the ratio between the penalty lines and the amount of requests. This feature is investigated for different levels of detail.

### 5.4.1 Data Amounts

At first, Table 5.3 depicts the results of the data transformation process of all entry lines from the source data of the whole city. The data fields are in respect to the chosen feature of the penalty ratio and therefore consists of multiple numbers of penalty and request lines for different classifications. Some of the provided values need to be highlighted particularly:

- The total amount of available entry lines is more than 19 million.
- The ratio of fined penalties in contrast to the performed requests is 6.6 percent.
- For 96.7 percent of all penalties a coordinate point could be gained, which is the base for further estimations.
- The average amount of requests with estimated coordinate points is about a fifths of all requests.
- Absolutely no location information could be gained for only 1.3 percent of all entries, in respect to the applied implementation.

The chosen districts for further analysis are district one and six. In the following, two additional tables of the same structure are provided showing the values restricted to the results of the chosen districts. In Table 5.4 the values of district one are shown. There are some interesting differences to the values of the whole city. These are a higher ratio of penalties with nearly a tenths, a higher percentage value of penalties with coordinate points of 98.5 and also a higher amount of requests with estimated coordinate points of

| City of Vienna Results - Penalty Ratio and Coordinate Points | | | | |
|---|---|---|---|---|
| | | Amount | ≈ % | Percent of |
| 01 | Entries | 19283877 | | |
| 02 | Requests | 18005719 | 93.400 | of Entries |
| 03 | Penalties | 1278158 | 6.630 | of Entries |
| 04 | Penalties with Coordinate Points | 1235969 | 6.410 | of Entries |
| 05 | | | 96.700 | of Penalties |
| 06 | Penalties with Estimated Coordinate P. | 4815 | 0.025 | of Entries |
| 07 | | | 0.377 | of Penalties |
| 08 | Penalties with Closest Coordinate P. | 31592 | 0.164 | of Entries |
| 09 | | | 2.470 | of Penalties |
| 10 | Requests with Estimated Coordinate P. | 3908745 | 20.300 | of Entries |
| 11 | | | 21.700 | of Requests |
| 12 | Requests with Closest Coordinate P. | 13844810 | 71.800 | of Entries |
| 13 | | | 76.900 | of Requests |
| 14 | Entries without any Coordinate Point | 257946 | 1.340 | of Entries |
| 15 | Penalties without any Coordinate P. | 5782 | 0.030 | of Entries |
| 16 | | | 0.452 | of Penalties |
| 17 | | | 2.240 | of Entries without |
| 18 | Requests without any Coordinate P. | 252164 | 1.310 | of Entries |
| 19 | | | 1.401 | of Requests |
| 20 | | | 97.800 | of Entries without |

Table 5.3: Results of the data transformation process of the whole City of Vienna. Chosen data fields are in respect to the penalty ratio.

nearly a quarter instead of a fifths. As the gaining of coordinate points is depending on the penalty location data, a higher ratio of penalties consequently gives a better chance of more requests with estimated coordinate points. The number of entries without any location information is below one percent.

In Table 5.5 the same values are shown for district six, which also contain some deviations. As already seen, the penalty ratio is much smaller than in district one and even below the whole city percentage value. The number of penalties with coordinate points are still above 98 percent, but the amount of requests with estimated coordinate points is nearly still average. While the amount of entries without location information is not conspicuous, the part of the penalties without any coordinate point is very low.

The presented amounts of data are the source for the following analysis, which is applied on the defined levels of detail from Section 5.2. As stated, the dimension of time provides multiple different levels, with various possible interpretations, but the dimension of place

| District One Results - Penalty Ratio and Coordinate Points | | | | |
|---|---|---|---|---|
| | | Amount | ≈ % | Percent of |
| 01 | Entries | 1697057 | | |
| 02 | Requests | 1538127 | 90.600 | of Entries |
| 03 | Penalties | 158930 | 9.370 | of Entries |
| 04 | Penalties with Coordinate Points | 156468 | 9.220 | of Entries |
| 05 | | | 98.500 | of Penalties |
| 06 | Penalties with Estimated Coordinate P. | 410 | 0.024 | of Entries |
| 07 | | | 0.258 | of Penalties |
| 08 | Penalties with Closest Coordinate P. | 1814 | 0.107 | of Entries |
| 09 | | | 1.140 | of Penalties |
| 10 | Requests with Estimated Coordinate P. | 390998 | 23.040 | of Entries |
| 11 | | | 25.400 | of Requests |
| 12 | Requests with Closest Coordinate P. | 1133778 | 66.800 | of Entries |
| 13 | | | 73.700 | of Requests |
| 14 | Entries without any Coordinate Point | 13589 | 0.801 | of Entries |
| 15 | Penalties without any Coordinate P. | 238 | 0.014 | of Entries |
| 16 | | | 0.150 | of Penalties |
| 17 | | | 1.750 | of Entries without |
| 18 | Requests without any Coordinate P. | 13351 | 0.787 | of Entries |
| 19 | | | 0.868 | of Requests |
| 20 | | | 98.200 | of Entries without |

Table 5.4: Results of the data transformation process of district one. Chosen data fields are in respect to the penalty ratio.

is restricted to two main levels, except of the whole city. These two levels are the districts and the coordinate points.

### 5.4.2 Per Level of Detail

The chosen approach for the selection of the levels is to start with the most general one and to continue with more detailed views. Therefore the first visualizations consider the data of single districts in different levels of time, while the coordinate point based approach follows afterwards.

#### Half of a Year

Figure 5.1 depicts the six possible ways to split the year by month into two parts. The data is from district one. Each colored line consists of two data points, representing a

| District Six Results - Penalty Ratio and Coordinate Points | | | | |
|---|---|---|---|---|
| | | Amount | ≈ % | Percent of |
| 01 | Entries | 873551 | | |
| 02 | Requests | 823253 | 94.200 | of Entries |
| 03 | Penalties | 50298 | 5.760 | of Entries |
| 04 | Penalties with Coordinate Points | 49412 | 5.660 | of Entries |
| 05 | | | 98.200 | of Penalties |
| 06 | Penalties with Estimated Coordinate P. | 180 | 0.021 | of Entries |
| 07 | | | 0.358 | of Penalties |
| 08 | Penalties with Closest Coordinate P. | 677 | 0.078 | of Entries |
| 09 | | | 1.346 | of Penalties |
| 10 | Requests with Estimated Coordinate P. | 181797 | 20.810 | of Entries |
| 11 | | | 22.080 | of Requests |
| 12 | Requests with Closest Coordinate P. | 632725 | 72.400 | of Entries |
| 13 | | | 76.900 | of Requests |
| 14 | Entries without any Coordinate Point | 8760 | 1.003 | of Entries |
| 15 | Penalties without any Coordinate P. | 29 | 0.003 | of Entries |
| 16 | | | 0.058 | of Penalties |
| 17 | | | 0.331 | of Entries without |
| 18 | Requests without any Coordinate P. | 8731 | 0.999 | of Entries |
| 19 | | | 1.061 | of Requests |
| 20 | | | 99.700 | of Entries without |

Table 5.5: Results of the data transformation process of district six. Chosen data fields are in respect to the penalty ratio.

value in the first and the second half of the year. The various colors denote the different ways of splitting the year into two halves, and the corresponding label names the first month of the first half of the year by its number. For example the label 'M2' means that the year starts with February and the first half ends with July. As the source data only includes the year of 2018, a half year which reaches over the New Year's Eve uses the months of the same year for completing the series. The figure shows two different applications of this approach. The left one contains the total numbers of entry lines in each half of a year. It is clearly visible that the highest difference is given between the first six months and the last six months of the year 2018, but if the split is made before June, the difference is the smallest. The right plot depicts the ratio of requests to penalty lines and again the splitting approach with the biggest gap between the two halves is clearly visible. In the half year from December to May, the rate of penalties is highest in comparison to all other combinations of half years. The difference is just about 0.5 percent.

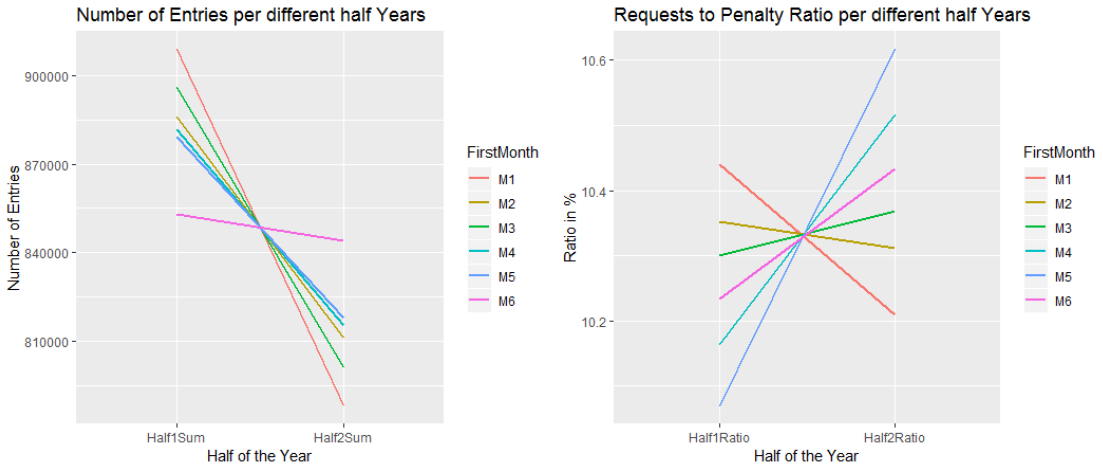


Figure 5.1: Comparison of the six different half year splitting possibilities of district one of the total number of entry lines and the ratio of requests to penalties. MX labels the begin of the first half of the year, where the 'X' is the number of the month, i.e., M2 means that the first half of the year starts with February.

By comparing the 'M5' split classification with the standard economy approach of using the first six months as the first half year, the difference is minimally. The half with the highest ratio has about 10.62 percent in comparison to about 10.44 percent of requests to penalties. On the half with the smallest ratio it is about 10.07 compared to about 10.21, while the economy approach is the bigger one. The difference of the likeliness to be fined when parking a car is negligible with respect to the stated values.

### Quarterly View

The next more granular level of detail is the quarterly view of the data. As already mentioned in Section 5.2, there is a main difference between the economy approach, of starting with the first month of a year as the first month of the first quarter and the meteorological point of view.

Figure 5.2 shows in a parallel chart two possibilities for dividing the year into quarters and four related values in separate diagrams. The 'Eco' or economic approach uses the first day of the year, as the start of the first quarter, while the meteorological approach 'Meteo' starts with the first of March. The used data is from district one and the presented values are the total number of entries, the amount of requests and penalties and the ratio of requests to penalties. As there are two very high values in the 'Ratio per Quarter' diagram, these are the observations which stick out the most. They contain Q1, the first three month of the year, in the economic approach and Q4, December, January and February, in the meteorological approach. This seems like January and February do have a high impact on an increase of the requests to penalty ratio, which should be investigated in more detail in the monthly analysis. Comparing the three diagrams of the


Figure 5.2: Comparison of the two different quarter splitting possibilities of district one, showing the four values of total number of entries, number of requests, number of penalties and the ratio of requests to penalties. 'Eco' is the economic approach where the first quarter starts with the first month of the year, while 'Meteo' is the meteorological approach where the first quarter starts with the first of March.

standard values of entries, requests and penalties it seems like they are having similar graphs. By a more detailed consideration it is visible that the requests graph depicts the highest differences between both approaches on each quarter. On the other hand, the graph of the penalties do have the smallest divergences, which consequently lefts the first graph of all entries as the tradeoff. Additionally, the 'Entries per Quarter' diagram gives a good first overview of when the most enforcement entries are performed, namely in 'Eco' Q2 and 'Meteo' Q1. This means that April and May are for both approaches in the quarter with the most entries. While for the requests to penalties ratio it is not obvious, the extreme values of the other diagrams are more distinctive when applying the meteorological approach on the data and therefore seems to be the better choice.

### Monthly View

After the quarterly view, the next level of detail is considering all the single months as they are. Besides the previous diagrams, which only depict district one, Figure 5.3 shows the same data fields of district one, as well as of district six. The data fields are the numbers of entries, requests, penalties and the ratio of requests to penalties. Both districts are visualized in the figure, to illustrate the clear difference between the graphs and the values of each data field for various districts. As already stated in the tables before, the amount of performed entries in district one is much higher than in district six, which is also visible in the diagrams, as the amount of requests and the amount of

penalties is less. Additionally the difference of 4 to 5 percent of requests to penalties in the ratio diagram is a presented feature. For district one, the amount of performed requests is highest in the months March, April and October, which is the same for the penalties. In the quarterly view, both quarters which included January and February are the one with the highest ratio of requests to penalties and this is also approved by the diagrams with the monthly data. For district six, February is also noticeably the month with the highest penalty ratio. The lowest percentage of the penalty ratio is for both districts given in autumn, namely in September and October. The most striking value is the very low amount of entry lines in December of district one, in relation to the other months. One of the reasons for this behavior can be the holidays during the months, where no or less parking enforcement is performed, but on the other hand this is also applicable for other districts like district six. The latter, however, does not have such a low peek in this month. It is not even the lowest value, which is February.

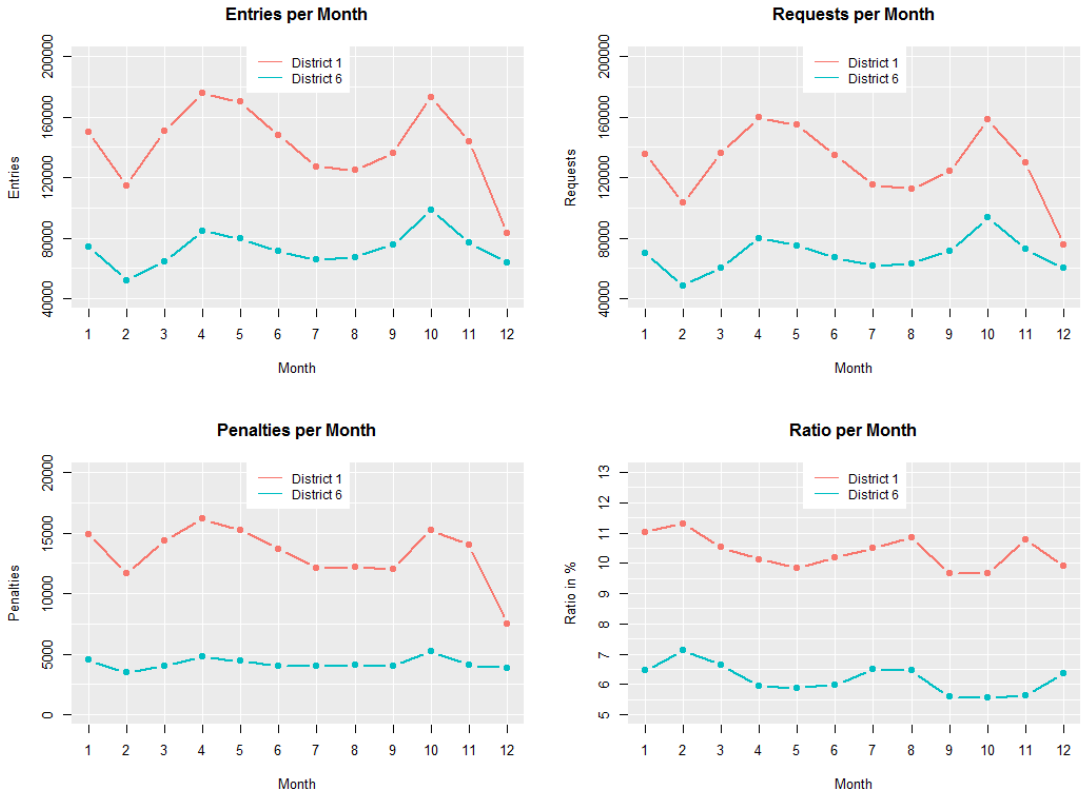


Figure 5.3: Comparison of the yearly data, based on monthly values for district one and six. The four diagrams are showing the number of entries, requests, penalties and the requests to penalties ratio.

When looking at the three diagrams of the absolute values in Figure 5.3 and comparing it with the graph of the ratio, it seems for some months that there is a relationship

between the ratio series and the amount of entries. Specifically, when the number of entries per month is increasing, the requests to penalties ratio is decreasing. For example in the summer period of July and August of district one, the number of requests has a local minimum, while the penalty ratio diagram shows an increase of percentages. This behavior can be verified by using the cross-correlation function, which is used to determine the relationship between two series.

The cross-correlation of the number of requests to the penalty ratio is shown in Figure 5.4. It depicts district one and six and illustrates the relationship between the two considered series. It can be seen in the right diagram that the previous described behavior of the inversely increasing penalty ratio is indeed applicable for district six, as the cross-correlation value is about minus 0.8 at lag 0. This means that a higher amount of requests leads to a lower ratio of requests to penalties in the considered months for district six. The diagram of district one does not show such a behavior, which is a reason for further investigations on this feature in next levels of detail. Additionally, the cross-correlation diagram of district one shows two significant values at lag 2 and 3, but as this series consider whole months, no series should lead the other and therefore is also investigated in the next levels of detail.

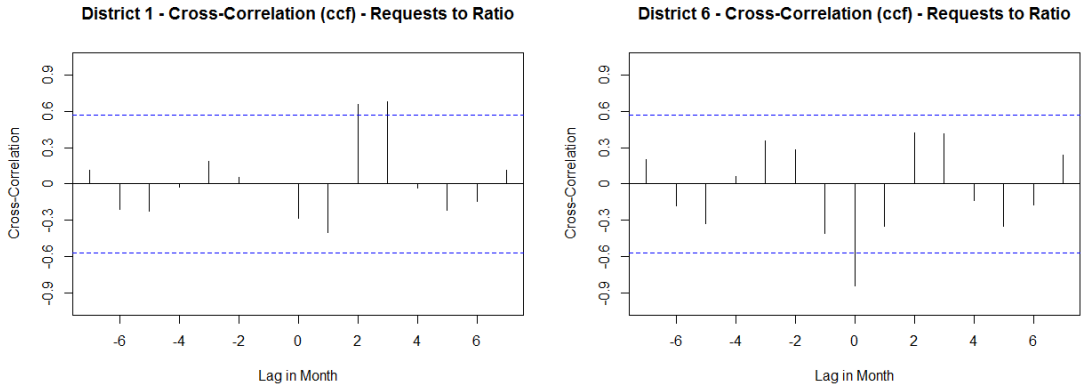


Figure 5.4: Cross-Correlation functions of district one and six for showing the relation of the number of requests and the penalty ratio.

### Weekly View

The next level of detail after the monthly view considers the summed up values of whole weeks. As there is no enforcement data available for the last day of the year and 2018 starts with a Monday, the number of weeks concludes to exactly 52. These weeks are shown in Figure 5.5, which depicts the four values of the number of entries, requests and penalties, as well as the requests to penalties ratio. The three diagrams with the total numbers do not have many special features. There are two weeks with very low values in both districts, which is week nine and the last week of the year. For district one, a decrease of performed requests and penalties can be seen in the time from about week 23

to 33, which is a part of the summer time. Considering the penalty ratio, there are a few weeks with higher values for both districts, but only week nine of district one is at 14 percent, which is a gap of one percent to the second highest value. The lowest requests to penalties ratio cannot be identified directly, since there are many points on the same level for both districts.


Figure 5.5: Comparison of the yearly data, based on weekly values for district one and six. The four diagrams are showing the number of entries, requests, penalties and the requests to penalties ratio.

As already mentioned for the monthly data, the result of the cross-correlation function and the relationship between the two series is different for each district. For clarifying the findings, the same diagrams are generated for the weekly data. They are shown in Figure 5.6, which provides a more clear confirmation of the described behavior. For both districts, the value of the penalty ratio is inversely increasing as the number of requests does, even if the correlation value is higher for district six. As a consequence, a higher amount of performed parking enforcement requests lead to a lower ratio of requests to penalties and therefore a smaller amount of penalties in relation to the number of requests. At this point, it should be added, that a parking warden normally does not fine a car, if it has already been fined at the same day and place, which can have a small

impact on the behavior, that is described by the cross-correlation diagrams.

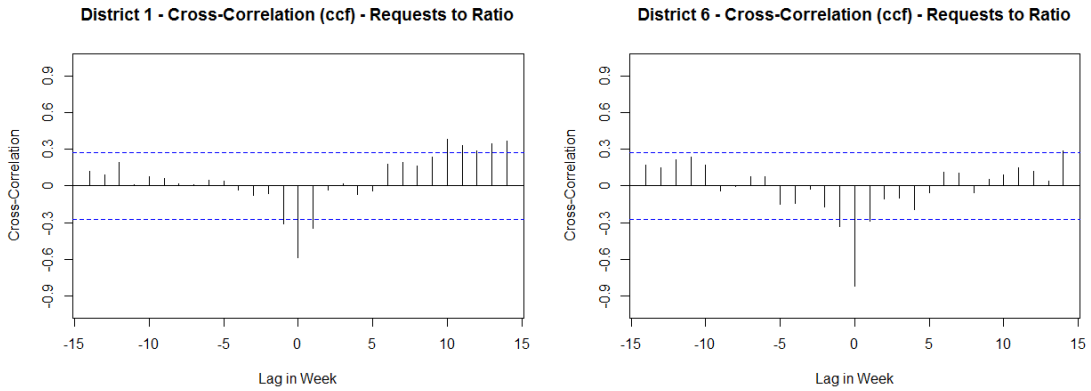


Figure 5.6: Cross-Correlation functions of district one and six for showing the relation of the number of requests and the penalty ratio.

### Days of a Year

By considering all days of the year, it is the finest level of detail which only uses the date of the entries in the dimension of time. The next level with more fineness also has to consider the timestamp of the entries. As already mentioned, the amount of parking enforcement that is performed on every day can be completely different, and there are also some days when no work is done. This is mainly influenced by public holidays, weekends or the prevailing weather conditions, like snowfall.

The penalty ratio completely depends on the number of performed requests and the amount of penalties, i.e., on a day without parking enforcement or when the data source only consists of a few entries, the outcome of the penalty ratio value is questioning. Therefore the amount of requests is checked to determine a border for which all days with a value below this border are standardized. It depends on the later purpose how this standardization looks like.

In some situations, such missing data values are filled with the mean of all the values, which will prevent the generation of outliers and does not destroy the real meaning of a graph. The parking enforcement data already has days without any value present in the source data, which will lead to zero values in a graph of requests and entries or no valid value for the penalty ratio. Another possible way of handling invalid values is to set it to zero and especially care about the penalty ratio by also setting it to zero if no valid number of requests or penalties is available. Before the standardization, a border has to be defined under which value a daily number of requests is manipulated. As a visual tool for making this decision, a histogram is used to display the different amounts of requests of all the days.

The histogram in Figure 5.7 depicts the amount of requests per day of district one and groups them together to show the number of days with a similar count. All days without any request present in the source data are excluded from this diagram. This condition fits to 15 days, which would be in the very first bin. The bin size of 50 requests is chosen to visualize the relatively high amount of days with only a few performed requests. After looking at the histogram and by checking the values of all days below 400 requests, a decision is made to standardize the first two bins, i.e., all request amounts below 100. This should avoid high influence on the outcome by special days, when for example the parking enforcement is only performed on single places, like a parking space, without any overall context. The distribution of the other pins shows, that the amount of performed requests can be completely different and is much spread.

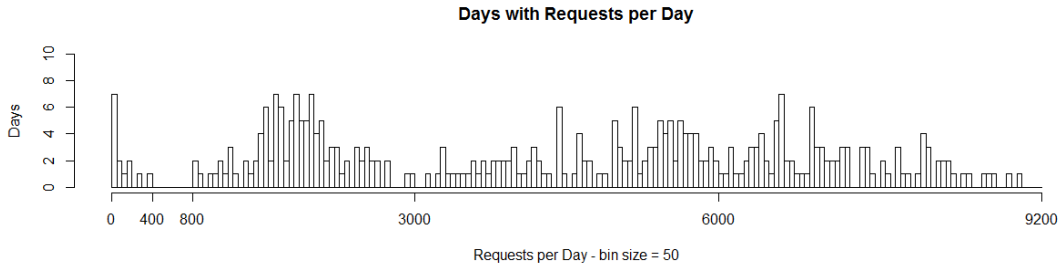


Figure 5.7: Histogram showing the number of requests per day of district one. All 15 days which have no performed request in the source data are excluded.

For the standardization, all dates of the days with less than 100 requests, also including the days with zero values, are stored for getting filtered later when calculating the penalty ratio values. Therefore all undefined results coming from the penalty ratio division can be successfully replaced. For a comparison of both standardization processes, an additional visualization is used to determine the accuracy and usefulness of each approach.

In Figure 5.8, two similar plots are presented, with the daily calculated ratio of requests to penalties used as the source data. As already described, all days with less than 100 requests are standardized before calculating the ratio. The plot on the left side is modified to have more zero values, while on the right side the mean is used. Each side shows a Seasonal-Trend Decomposition Procedure Based on Loess (STL), which includes a seasonal, trend and remainder part in addition to the input data of all days of the year. The used frequency is seven days, i.e., a weekly period is the base for the seasonal part.

Both STL plots depict a clear structure of the seasonal part, also if the change is small in relation to the input data. On the right side, the initial bar on the right border of the input data is bigger than on the left side, i.e., the seasonal change is a little bit more clear on the plot with the mean values for standardization. Basically it means that on each weekend the penalty ratio becomes slightly higher than at the begin of the week in both plots. The scale of the trend part is closer to the input data in both plots, but it is more obvious for the left plot. By considering the relative trend on the right plot, it does

not have a clear structure, beside of some ups and downs. On the other side the trend of the left plot has much more impact and especially at the end of the year, the modified zero values cause an enormous decrease of penalty ratio. Besides the seasonal and the trend part, the remainder is an indicator for the fitness of the underlying model which is used for the plots, i.e., it shows how accurate it is for example to use the depicted trend part as the general conclusion of the analysis. While there are more big peeks in the remainder of the left plot, both sides are not having a very well fitted underlying model, which is used by the STL algorithm. As a conclusion for the two used standardization methods, after analyzing both diagrams, the usage of mean values is a little bit more suitable in case of the penalty ratio, because the divergences are smaller.

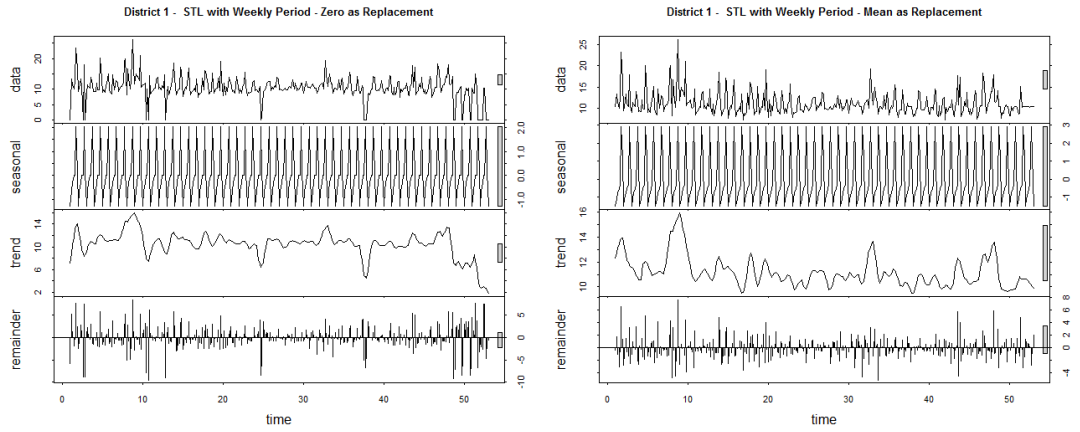


Figure 5.8: Two standard seasonal decomposition time series by Loess (STL) plots of district one, including data values, seasonal, trend and remainder part. The used standardizations are zero values on the left side and the mean on the right side.

The underlying models in the two plots of Figure 5.8 are not fitting very well, but it still seems that there are clear differences between the weekends and the other days of a week. Therefore, the differences between the days of a week are investigated in the following.

Figure 5.9 shows two different boxplots to visualize the distribution of the penalty ratio of district one, based on the day of the week. On the left side, the values are grouped by all seven weekdays, while on the right side the separation is done between the accumulated workdays from Monday to Friday and the weekend days. Holidays are not handled explicitly. The input data has not been standardized and therefore the penalty ratio calculation results of all days are included. To have a better scale on the plot, all penalty ratio values above 35 percent are modified and distributed closely around this border value. In total, the values of eight days have been changed.

The right diagram clearly illustrates that the ratio of requests to penalties is higher on the weekends. The difference is nearly five percentage points. A more detailed look to the left plot emphasizes this observation. Especially the sixth day of a week, which is Saturday, has the highest mean value of all days of the week. While on Saturday and

Sunday the interquartile range is high and the whiskers are stretched a lot, the five weekdays are very compact and similar. When comparing the single working days of a week, there is a small increase of the mean values from Monday to Friday, which means that the penalty ratio is increasing during the work week.

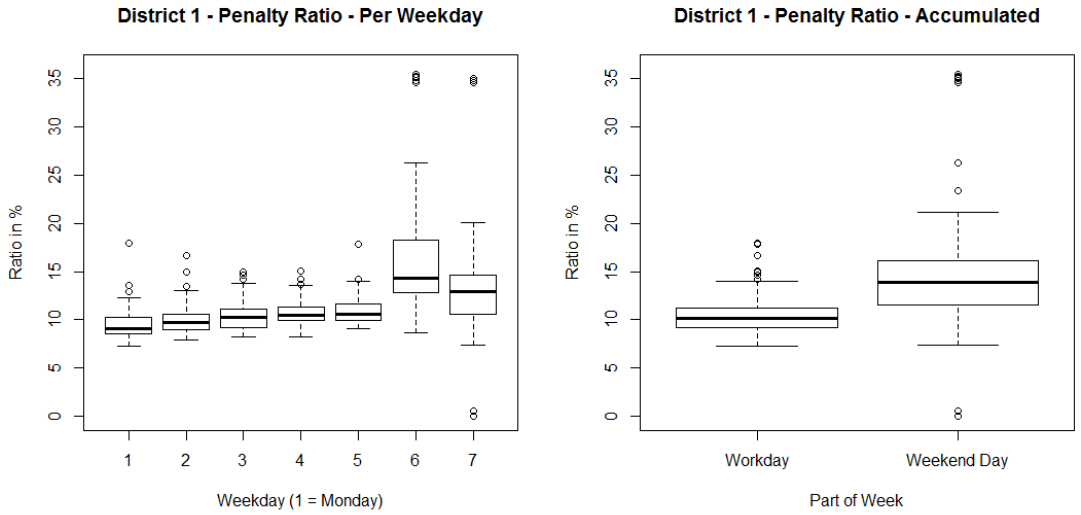


Figure 5.9: Two boxplots depicting the penalty ratio of district one of all days grouped by the day of the week in the left plot and a workday to weekend day in comparison on the right side. The input data is not standardized, but y-axis values above 35 percent are distributed around this border value.

### Time of a Day

This level of detail considers the timestamp as well as the date of an entry. It should give an overview about the performed parking enforcement at different time periods of a day. The separation of daytime can be done in many ways, like equally splitting by hours or explicitly highlight special periods to emphasize their difference.

The two boxplots in Figure 5.10 show the number of requests and the penalty ratio of district one, on every working day grouped by certain time periods of the day. The borders of the four groups should reflect different periods of a day, namely night, morning, midday to afternoon and afternoon to evening. On the left side, the amount of requests is presented for all four groups. If there does not exist a single entry for a group on a day, no zero value has been added and therefore the amount of values for each group is different. As there are many days with only a few requests available in the night, this group is not present in the right diagram, which shows the calculated penalty ratio in percent. To make the right boxplot better readable, all penalty ratio values which are higher than 25 percent are modified and distributed closely around this value, i.e., six values have been changed.

The four groups on the left diagram depict the various amounts of requests which are performed in these time periods. The night, which is represented as the group from 10:00 PM to 4 AM, only has a few days with a few hundreds of requests. The boxplots of the other three groups have wide spread whiskers, but the interquartile range is formed with 500 to 1000 requests around certain points. The median of the morning group is slightly above 1000 requests, while the other two groups have 1000 requests more. The group from 10 AM to 4 PM has the highest median. It is also visible in the diagram, that the boxplot with the highest median does have the biggest interquartile range and the longest whiskers. On the other hand, the morning group, as the group with the lowest normal median, does have the smallest interquartile range and the shortest whiskers.

The diagram on the right side, which shows the penalty ratio only for the three groups of morning to evening, depicts a slight increase of the median from the earliest to the latest time period. The interquartile range for the percentage values is in all groups only two to three percent and the distance between the minimum and maximum values of the whiskers is less than ten percent. In the group with the most requests, from 10 AM to 4 PM, the ratio between requests and penalties is spread the littlest.

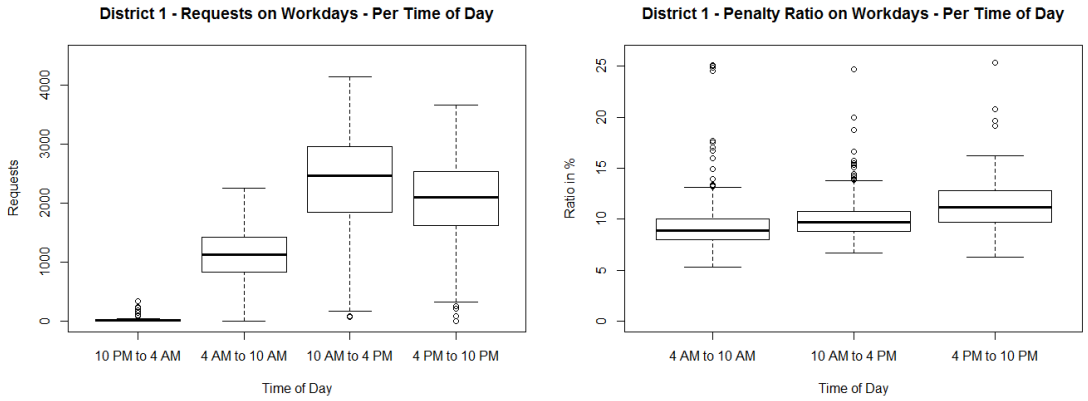


Figure 5.10: Two boxplots showing the number of requests and the penalty ratio of district one on every working day in certain time periods of the day. The input data is not standardized, but y-axis values above 25 percent in the right diagram are distributed around this border value.

The separation by the time periods can also be done in smaller steps, but this representation is one possible approach and each can be used to describe appearing characteristics. Both diagrams in Figure 5.10 only use the data of working days, but this visualization can also be applied for weekend days or all together. Therefore, the possible combinations can result in multiple diagrams, and each can demonstrate other special characteristics, but it is also possible that different levels of detail result in different findings.

All the previous visualization approaches do have one big thing in common, which is the change of the dimension of time. This is because in Section 5.2, it is shown that there are

many more different levels of detail in this dimension. In the following section, a change in the dimension of place is performed as well and possible visualizations are presented.

### Place by Coordinate Points

In Section 5.2, the useful and possible levels of detail in the dimension of place are mainly limited to the district and to the gained coordinate points. The district, as the first level, is heavily used for the different time periods in the previous sections. The following approaches are focusing on the application of gained coordinate points for the dimension of place.

District one is chosen as the input, because the percentages of gained coordinate points is relatively high. Besides some entries, for which absolutely no coordinate point information could be achieved, there are two categories of coordinate points which will be distinguished, based on the definitions in Subsection 3.4.6:

- **Calculated and Estimated Coordinate Points** This category combines two different types of coordinate points, namely the calculated penalty coordinate points and the estimated coordinate points of entries which are lying between penalties on the same street. This category includes about 32 percent of all entries, which can be seen in Subsection 5.4.1.
- **Closest Penalty Points** By considering the closest penalty points information in the second group, the distance in time to the next penalty with calculated coordinate points is given as an additional parameter in this category. Maybe a border or weighting of this additional parameter value is needed for improving the outcome. While the prior category only includes one third of all entries, this one considers theoretically about 99 percent of all entries.

As the different categories of coordinate points consist of unequal amounts of requests and penalties, the penalty ratio is also different for each category. The total numbers of district one are already stated in Table 5.4, but the penalty ratio is not explicitly mentioned for each type of coordinate points. The meaningfulness of these values highly depends on the amount of included data, but the following numbers are important in the context of the next visualizations. For the first category of calculated and estimated coordinate points, the amount of penalties is 156878 and the requests are 390998, which results in a penalty ratio of about 40 percent. When considering the closest penalty points, the category consists of about 99 percent of all entries, and therefore the ratio of requests to penalty entries is about 10 percent.

Before the coordinate points are used to visualize the data, Figure 5.11 shows a map of the chosen district one, which can be used as a reference to better understand the presented visualizations. The source of the map is the website of the City of Vienna[5]

[5] https://www.wien.gv.at/stadtplan/, Accessed 2019-07-26

and the map is provided by Stadt Wien - ViennaGIS[6]. The area that is chosen for the diagrams is a square, which fits to the borders of the district, but therefore also areas which are outside of the district are included. Especially the corners of the square include many areas which belong to other districts.

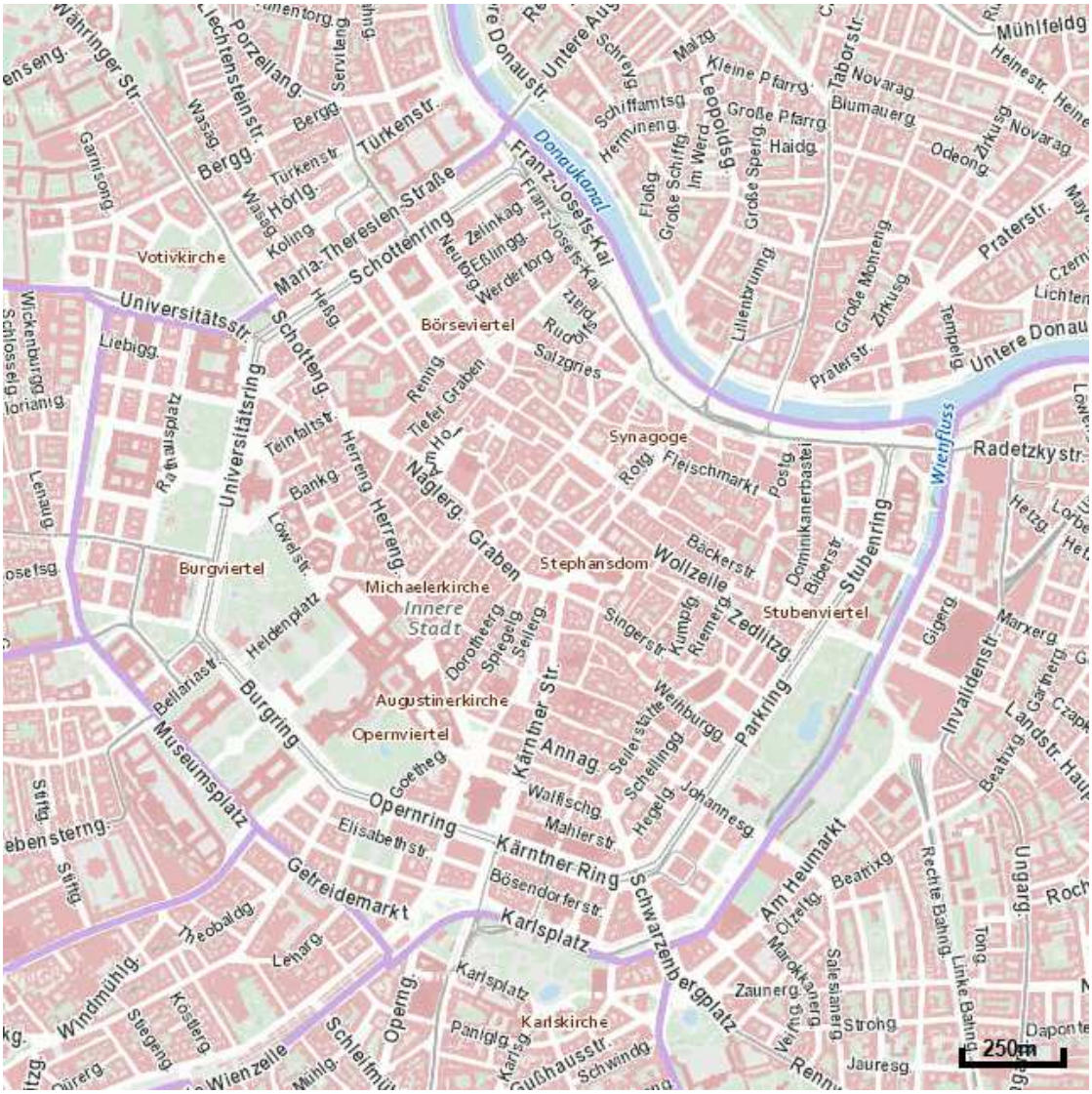


Figure 5.11: A map of district one in Vienna with border lines included.

There also exist some data points in the graphics which are outside of the district area, but included in the chosen data set. There are various reasons for this. It can be an error of the data transformation step and the gained coordinate point of an address is outside

[6] http://www.wien.gv.at/viennagis/, Accessed 2019-07-26

of the district because the combination of street name and house number is wrongly detected. Alternatively, it can also be an error in the input data and the district is set wrong by the parking warden, which is easily possible because they have to change the district manually.

The first visualization which uses the gained coordinate points is Figure 5.12 and consists of three pairs of tile maps. Tile maps are chosen as a simple version of heat maps, to show the amount of values for various regions. Each pair uses a defined number of bins to calculate the local value of each tile in the map and the number of bins define the level of detail of the presentation. The left side shows the amount of entries and the right side depicts the ratio of requests to penalties. For getting a better scale with the penalty ratio, all values above 100 percent are limited to this values. The amount of changed values is very small and does not influence the main appearance of the diagrams. The source data for the figure includes the calculated and estimated coordinate points as described before of the complete year 2018.

The first row gives a very brief overview of the area of district one, but the corners of the left diagram already show the lack of data in these regions, because they are positioned outside of the district. The percentage values of the penalty ratio is very high in the whole figure and in many bins, because the amount of included requests is low. In the second map on the left side, the characteristics of the district becomes visible, as there is a big free area in the bottom-left of the center. This area means that there is no enforcement data available and the reason is that there are no parking spaces, many green areas, big buildings and pedestrian-only areas. On the right side of the second row, there are many bins with high penalty ratio. In the last pair of tile maps, the number of bins is high enough for being able to recognize single streets in many areas of the district. The regions without any enforcement data entries are much more conspicuous with this high level of detail. The penalty ratio map of this level of detail now also depicts many clear differences of spots with very low and very high values. A disadvantage of the very detailed map is the small size of each single bin and as a consequence, some differences often can not be distinguished.

The following Figure 5.13 has the same characteristics than Figure 5.12 and is created in nearly the same way. It also consists of six tile maps, depicting the amount of entries and the penalty ratio in multiple bins with the same levels of detail. There are two main differences in the creation of Figure 5.13:

- The source data which is used for the diagrams consists of all entries with any coordinate points, including the closest penalty points.
- As the number of entries is heavily increased, the penalty ratio is much smaller and therefore all values above 35 percent are limited and set to this value. This leads to a much better scale for the tile maps depicting the penalty value and does not influence the main outcome.

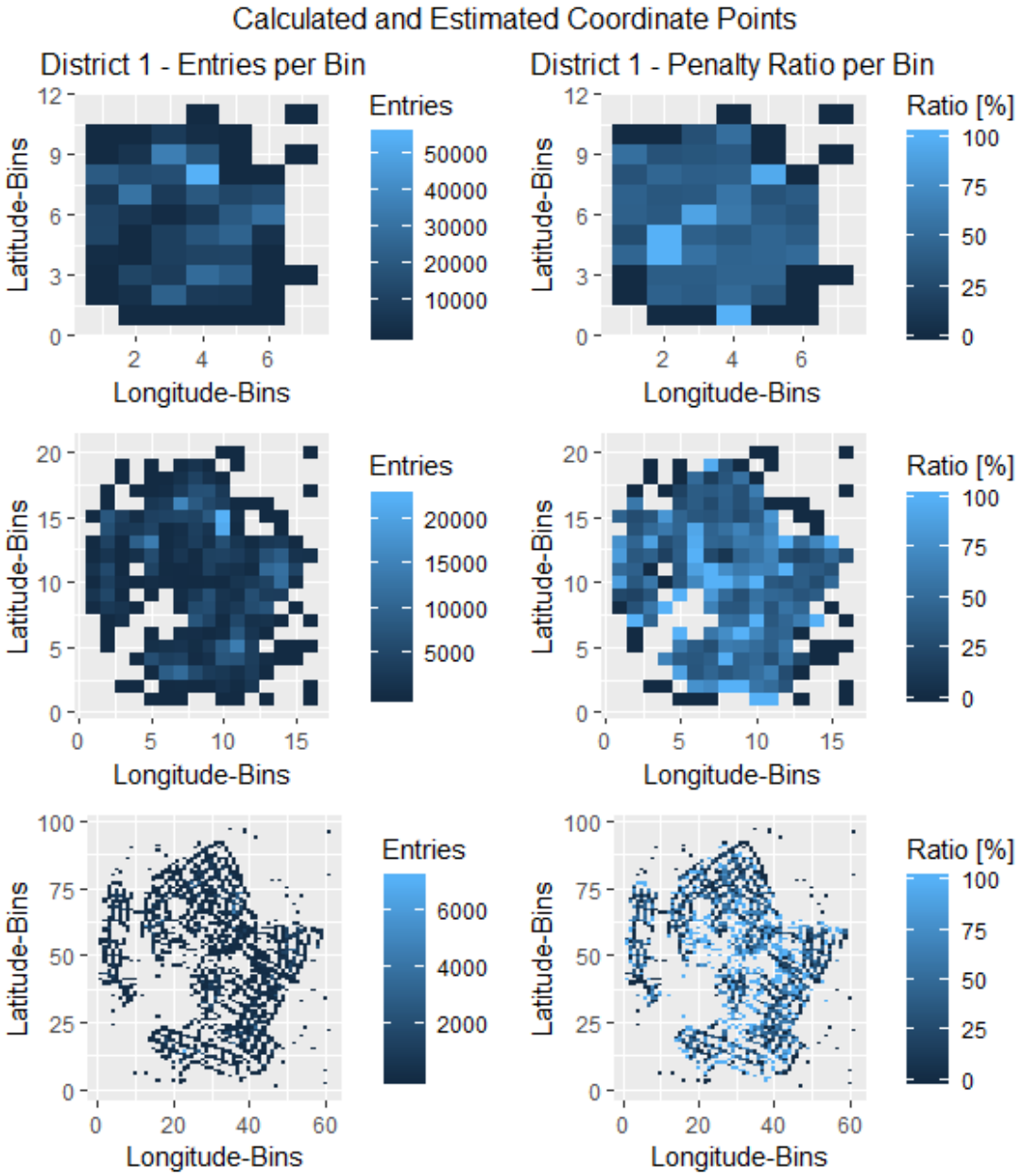


Figure 5.12: Tile maps of district one entries with calculated and estimated coordinate points with different amounts of bins in each row. The source is the complete year. The left side depicts the number of entries in a bin and the right side depicts the calculated penalty ratio. Ratio values above 100 are set to 100.

The data now consists of about 99 percent of all entries of district one and this has a big impact on the appearance of the resulting tile map. Logically, the amount of entries in the left maps is now higher, but the spreading of the colors has not really changed. Areas with higher values in the previous figure also appear similar in the new figure. The main difference can be recognized in the right visualizations, because the penalty ratio values appear more standardized now and is more often much closer to the calculated average 10 percent than before. Although if there is a limit of 35 for the penalty ratio, this setting only manipulates five bins out of 264 in the second row. The figure also shows much more sparse values, especially in the corners of the tile maps, which is also referable to the higher amount of considered entries.

Both presented figures have many things in common, but the second one, which includes the closest penalty points information is much closer to the totally calculated values. A disadvantage is the big inaccuracy in the data transformation process when gaining this type of coordinate points, because the distance value is not used in this type of visualization and therefore the closest penalty point information counts like a normal place definition.

Besides all the described features of Figure 5.12 and Figure 5.13 so far, both do not consider any additional time dimension, because the complete input data of the whole year is used. Therefore the next figures are used to combine the most detailed level of the dimension of place together with different time levels, which have already been handled in the previous sections.

The two visualizations of Figure 5.14 and Figure 5.15 are showing the number of entries and the penalty ratio for different levels in the dimension of time. They use the entries including the closest penalty point information. The used source data for each level is the one occurrence with the most entries in the level of district one, as stated in the following list:

- The weekday with the highest amount entries of the whole year is Wednesday.
- The month with the most entries is April.
- The week in the year 2018 with the most entries is the 16th week, which is from April 16th to April 22nd.
- The day with the highest amount of entries is the 16th of April.
- The morning with the most entries, which includes the daytime from midnight to 12 o'clock is June 14th.
- For the daytime from 12 o'clock to midnight, the day with the highest amount of entries is January 29th.

The amount of bins in each tile map is nearly the same and is similar to the bottom row of Figure 5.12 and Figure 5.13.

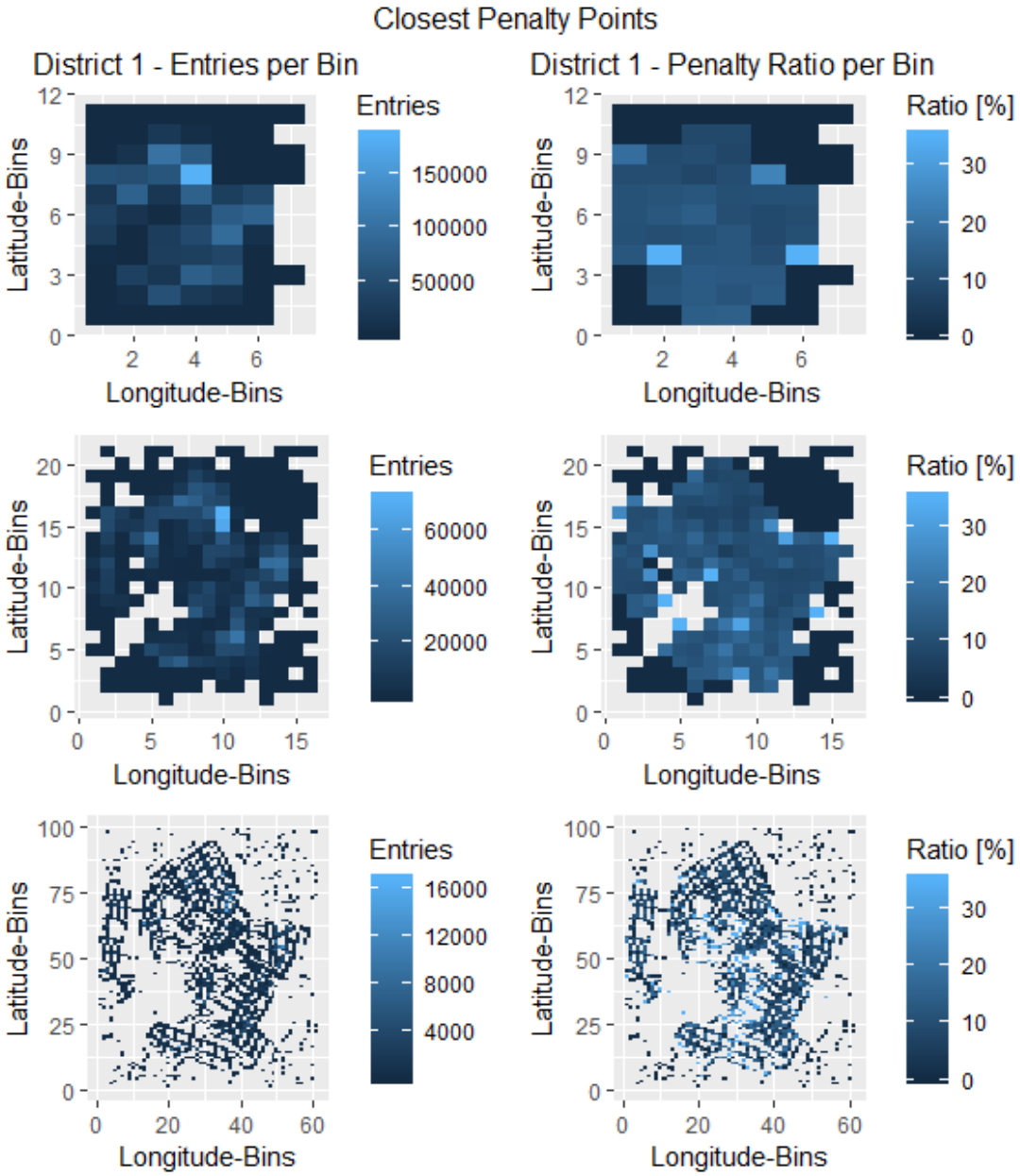


Figure 5.13: Tile maps of district one entries including closest penalty points information with different amounts of bins in each row. The source is the complete year. The left side depicts the number of entries in a bin and the right side depicts the calculated penalty ratio. Ratio values above 35 percent are set to 35.

By considering the amount of entries in Figure 5.14, the presented tile maps show big differences for the various levels of detail. As there are only seven possible weekdays, the top-left tile map depicts the most entries of all six maps. The appearance of the top-right tile map, which is including the data of a single month is also similar, but for some areas no data seems to be available. Both maps in the top row make it possible to have a presentiment of the structure of the district, like it has already been shown in the figure with the data of the whole year. The third map, which only includes the data of the week with the most entries already has multiple areas where the structure cannot be determined. On the daily level, the day with the highest amount of entries starts to have only sparsely defined tiles in the map. By performing an additional separation by the daytime, the last tile map with entries after 12 PM looks similar to the whole day map, but the amount of entries before 12 PM is much more sparse than all other maps.

Regarding the ratio of requests to penalties, the structures of the tile maps are the same, as they are depending on the available amount of entries. Like in the previous visualizations, the penalty ratio is again limited. For all six diagrams it is defined by 40 percent and all values above this limit are set to the value for having a better scale. In the first row, which includes the weekday and monthly maps, there are multiple areas where the higher values of penalty ratio can be recognized normally. Starting with the third diagram and the data of one week, it becomes more difficult to identify those areas. The three tile maps on the daily level also consist of tiles with higher and lower penalty ratio, but the distribution is unsystematic and the sparsely available data gives no chance to detect certain features in the maps.

An important detail at the end of this section is, that the amount of bins which is used for each tile map can always be chosen completely different. The outcome of each diagram can vary based on the amount of bins. For example it can result in very nice looking tile maps that obviously present different features of the data, or only a few bins more or less can influence the outcome in a way, that the differences are too small to recognize the main features. This is the reason why normally a heat map is used, which continuously shows the various regions where the values are higher or smaller. For visualizing the penalty ratio, this approach is not useful applicable, because the ratio has to be calculated separately for each predefined area. Although, the tile map is a fast and simple alternative for the given use case.

## 5.5 Origin of the Cars

The second section of the analysis considers the origin of cars in various applications due to different possibilities. As already mentioned in Section 5.1, the origin of cars can be considered in many ways, for example:

- Comparing the number of foreign and Austrian license plates and their penalty ratios.

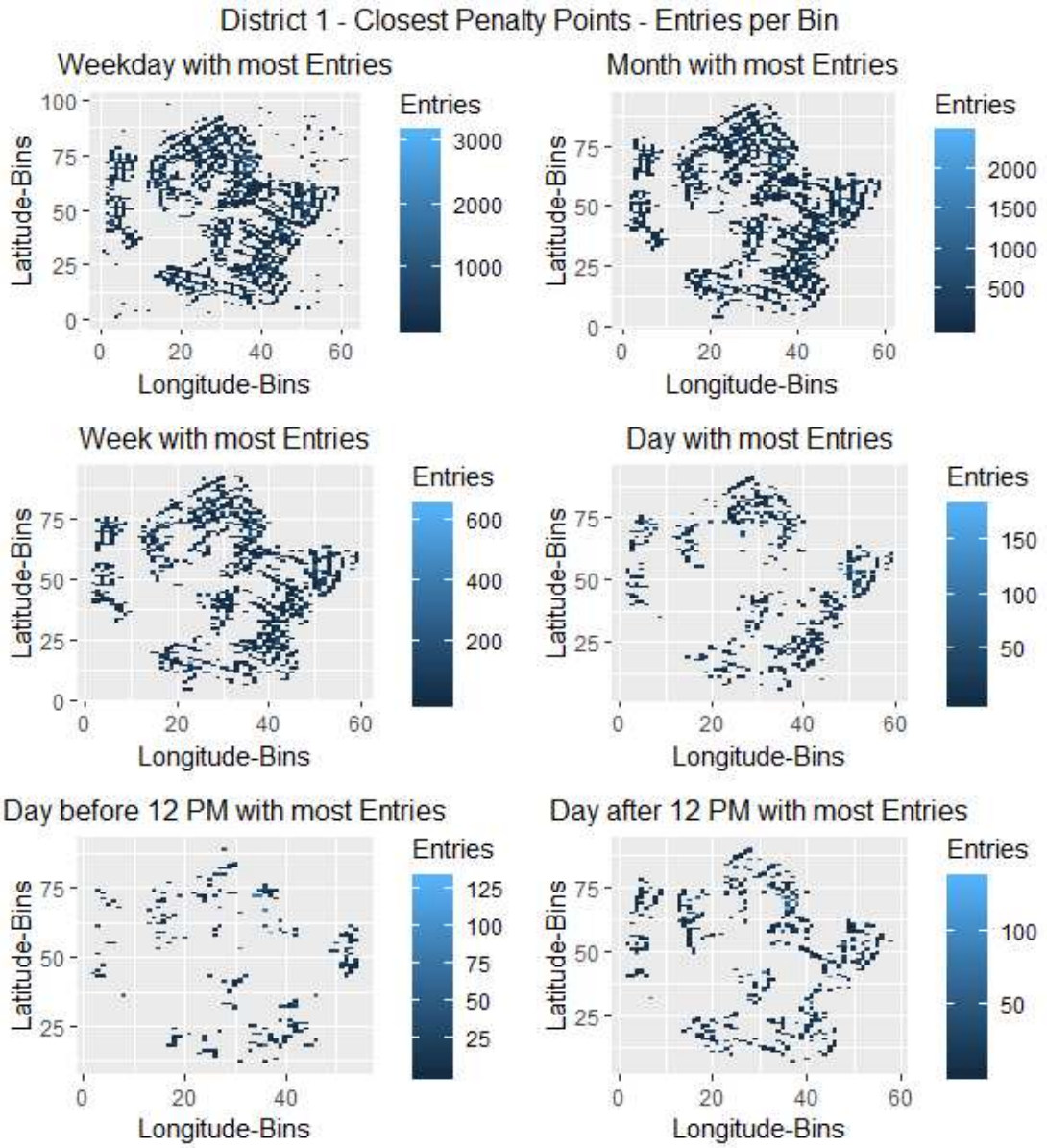


Figure 5.14: Tile maps of district one entries including closest penalty point information with similar bin sizes showing the number of entries, but for different time levels.

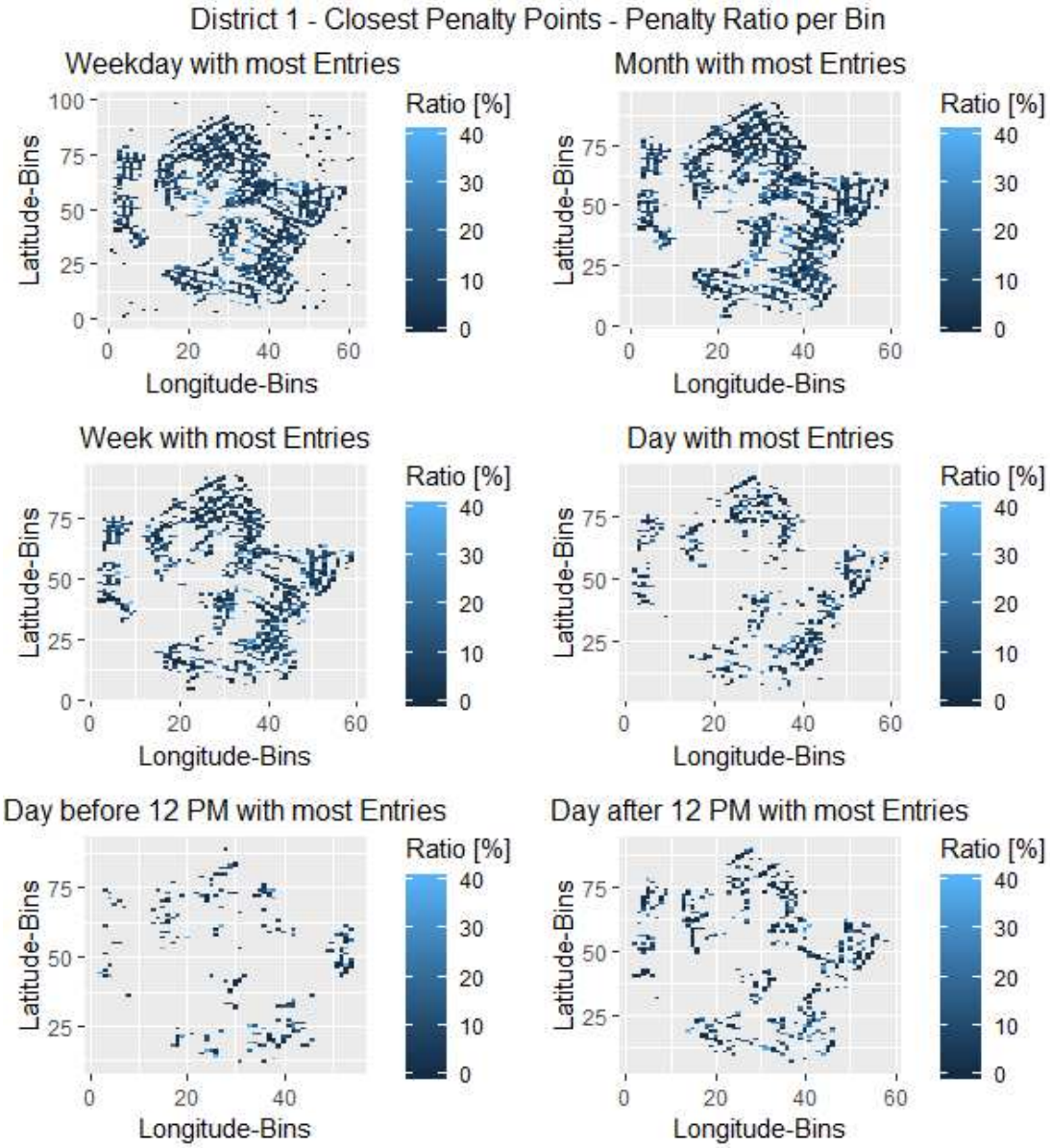


Figure 5.15: Tile maps of district one entries including closest penalty point information with similar bin sizes showing the penalty ratio, but for different time levels. Ratio values above 40 percent are set to this value.

- Comparing the number of Viennese and all the other Austrian license plates and their penalty ratios.
- Showing differences in request and penalty locations of cars from different origins.

After presenting the amount of data in the next section, some specific levels of detail are chosen to visualize the data. The Section 5.1 already contains visualizations for all levels of detail and therefore in this section only a subset is used to present special characteristics of this indicator. Furthermore, the multiple available categories given by the origin of cars provide new possibilities for visualization methods.

### 5.5.1 Data Amounts

As district one of the City of Vienna is the most famous and touristic one, it is chosen as the main data source for further analysis about the origin of cars. Additionally, district six is chosen for the coordinate point based analysis about the origin of cars, because the geographical location gives other possibilities for further analysis. In the following, three tables are presented, stating the total amount of entries for each category of the whole City of Vienna, the first and the sixth district.

Table 5.7 shows the results of the data transformation process with the data amounts of the whole City of Vienna. The chosen data fields are in respect to the actual indicator of the origin of cars and consist of multiple values and their respect ratio to the total amounts. Some values can be highlighted explicitly:

- The ratio of entries with an Austrian license plate to the total number of entries is 98.3 percent.
- Only about eight percent of all entries with Austrian plates are from outside of Vienna.
- A quarter of all penalties are from entries with Austrian license plate, but not from Vienna.
- Less than two percent of all entries do have foreign license plates, but 10 percent of all penalties are from this category.

One of the chosen districts for further analysis is district one and therefore an additional table with the same structure is provided by Table 5.7. The main difference between the values of district one and the complete city is the definite swivel to have more entries with license plate from outside of Vienna. The number of entries with Austrian plates increases from eight percent up to 15 and the amount of foreign license plates increases by one percent.

Table 5.8 is the final one and shows again the same structure as the previous two. One of the differences is the smaller amount of entries, which also influences the requests and

| City of Vienna Results - Origin of Cars | | | | |
|---|---|---|---|---|
| | | Amount | ≈ % | Percent of |
| 01 | Entries | 19283877 | | |
| 02 | Requests | 18005719 | 93.400 | of Entries |
| 03 | Penalties | 1278158 | 6.630 | of Entries |
| 04 | Entries with AT license plate ('A') | 18951765 | 98.300 | of Entries |
| 05 | Entries with AT but Non-Vienna plate | 1640180 | 8.510 | of Entries |
| 06 | | | 8.650 | of AT Entries |
| 07 | Penalties with AT license plate | 1149085 | 5.960 | of Entries |
| 08 | | | 89.900 | of Penalties |
| 09 | | | 6.063 | of AT Entries |
| 10 | Penalties with AT but non-Vienna plate | 313144 | 1.620 | of Entries |
| 11 | | | 24.500 | of Penalties |
| 12 | | | 1.650 | of AT Entries |
| 13 | | | 27.300 | of AT Penalties |
| 14 | Entries with foreign license plate | 332112 | 1.720 | of Entries |
| 15 | Penalties with foreign license plate | 129073 | 0.669 | of Entries |
| 16 | | | 10.098 | of Penalties |
| 17 | | | 38.900 | of foreign Entries |

Table 5.6: Results of the data transformation process of the whole City of Vienna. Chosen data fields are in respect to the origin of cars.

penalties. Beside that, the percental value of the total penalties in contrast to district one is lower, but also Austrian license plates do have a lower penalty percentage, while foreign car plate penalties have increased.

One important detail in this whole section is the naming, meaning and usage of the license plate categories. It is crucial that the number of 'Austrian license plates' or 'foreign license plates' do not mean the number of cars with the respective license plate. It means that there are entries in the source data where the data fields for the country and also the identification letters of the license plate are part of those categories. The process of how the parking enforcement is performed, the preconditions for working with the data, which are described in Section 3.1, and the performed data transformations only allow to work with entry information and the respective categorized data fields. This process allows situations in which multiple entries in one considered time period are meaning the same car with the same license plate, but are counted multiple times. As a consequence, the total amount of cars which are checked for parking tickets or are fined can be completely different from the number of requests or penalties.

| District One Results - Origin of Cars | | | | |
|---|---|---|---|---|
| | | Amount | ≈ % | Percent of |
| 01 | Entries | 1697057 | | |
| 02 | Requests | 1538127 | 90.600 | of Entries |
| 03 | Penalties | 158930 | 9.370 | of Entries |
| 04 | Entries with AT license plate ('A') | 1647470 | 97.080 | of Entries |
| 05 | Entries with AT but Non-Vienna plate | 258892 | 15.300 | of Entries |
| 06 | | | 15.700 | of AT Entries |
| 07 | Penalties with AT license plate | 138422 | 8.160 | of Entries |
| 08 | | | 87.100 | of Penalties |
| 09 | | | 8.402 | of AT Entries |
| 10 | Penalties with AT but non-Vienna plate | 45506 | 2.680 | of Entries |
| 11 | | | 28.600 | of Penalties |
| 12 | | | 2.760 | of AT Entries |
| 13 | | | 32.900 | of AT Penalties |
| 14 | Entries with foreign license plate | 49587 | 2.920 | of Entries |
| 15 | Penalties with foreign license plate | 20508 | 1.208 | of Entries |
| 16 | | | 12.900 | of Penalties |
| 17 | | | 41.400 | of foreign Entries |

Table 5.7: Results of the data transformation process of district one. Chosen data fields are in respect to the origin of cars.

### 5.5.2 Selection of specific Levels of Detail

This section visualizes particular combinations of special values in different levels of detail in the context of the cars origins. Like in the previous section, the first diagrams are considering wider levels of detail, while the later figures are also including the coordinate point information.

#### Weekly View of Categories Ratios

In Figure 5.16, the data of district one is used to visualize two ratio values of different carplate origin categories on a weekly basis of the whole year. The ratio values describe two features and are present in the figure with the same order:

- **Request Ratio per Carplates Origin** This feature is referring to the calculated ratio of the number of requests in this category and all available requests.
- **Penalty Ratio per Carplates Origin** In difference to the the already mentioned penalty ratio, this feature follows the same pattern as the request ratio. It is

| District Six Results - Origin of Cars | | | | |
|---|---|---|---|---|
| | | Amount | ≈ % | Percent of |
| 01 | Entries | 873551 | | |
| 02 | Requests | 823253 | 94.200 | of Entries |
| 03 | Penalties | 50298 | 5.760 | of Entries |
| 04 | Entries with AT license plate ('A') | 855485 | 97.900 | of Entries |
| 05 | Entries with AT but Non-Vienna plate | 79143 | 9.060 | of Entries |
| 06 | | | 9.25 | of AT Entries |
| 07 | Penalties with AT license plate | 43524 | 4.980 | of Entries |
| 08 | | | 86.53 | of Penalties |
| 09 | | | 5.088 | of AT Entries |
| 10 | Penalties with AT but non-Vienna plate | 13861 | 1.590 | of Entries |
| 11 | | | 27.600 | of Penalties |
| 12 | | | 1.620 | of AT Entries |
| 13 | | | 31.800 | of AT Penalties |
| 14 | Entries with foreign license plate | 18066 | 2.068 | of Entries |
| 15 | Penalties with foreign license plate | 6774 | 0.775 | of Entries |
| 16 | | | 13.500 | of Penalties |
| 17 | | | 37.500 | of foreign Entries |

Table 5.8: Results of the data transformation process of district six. Chosen data fields are in respect to the origin of cars.

calculated as the ratio of the number of penalties in this category and all penalties in the data.

Each of the two diagrams depicts five variables and some of them are non-independent, which means that the total number of percentages for a week is more than a hundred. Every variable pictures a certain category for the origin of license plates, which are:

- all variants of Austrian license plates
- foreign license plates, which is referring to every non-Austrian
- only Viennese carplates
- Austrian license plates but excluding those from Vienna
- every carplate but without those from Vienna

The idea of including and excluding the license plates from Vienna comes from the fact that drivers of cars with this type of license plate are the majority and basically should know their parking rules, while other cars maybe don't know them in the same way.

Both diagrams of Figure 5.16 show completely different behaviors for the five values. In the first diagram, all values are having a more or less straight line and are only moving up and down for few percentages. The category of requests with Austrian license plates, for example, is always very close to a hundred percent during the whole year and there are only two time periods with clearly visible negative differences. The first period is the time before and after new year, while the second time period is in the middle of the year during summer time with the lowest percentage value in week 33. An additional observation in the first diagram is a drop of the percentage of Viennese license plates in the weeks of December. This drop is nearly completely smoothed by carplates from other Austrian regions, because there is hardly no increase of cars with a foreign license plate.

The second diagram, which depicts the penalty ratio of each carplates origin category consists of much more spreading of the percentage values. Two time periods are showing the highest difference, which are again the time before and after new year and the summer time with the highest peak in week 33. While the penalty ratio for cars with Austrian license plates is usually around 90 percent, the values go down to 75 percent in week 33, while obviously at the same time the foreign license plate ratio increases to 25 percent.

### One Month with Subset of Authorities

The next diagram gets more in detail and uses the data of a single month of district one, which is April because it has the highest number of entries. Figure 5.17 shows a stacked bar chart with twenty bars and each contains the number of requests, the number of penalties and the calculated penalty ratio. Only authorities from Austrian license plates are considered and those with the most entries in this month are selected. As most of the cars do have Viennese plates, this group is excluded to gain a better view and a better scale of the chosen authorities. In total, more than one hundred different identification letters are present in the data for this month, but their number of occurrences is very low.

The number of entries of the authority on the last place of the presented list is only 327 for authority 'WN'. In comparison, the first place of the top twenty has a more than ten times higher value, which is 3697 entries for 'MD'. A special case of the depicted authorities is 'WD', which means Viennese diplomat. This group actually belongs to Vienna, but it is available as an own authority in the data. It is also the group with the lowest penalty ratio in the top twenty list. When 'WD' is excluded, the whole range of penalty ratio values lasts from 16.3 to 26.4, which makes a range of about 10 percent.

### One Week with Subset of Countries

In Figure 5.18 only the data of week 33 of district one is used to present a bar chart which is similar to the previous figure. It depicts the top ten countries based on the

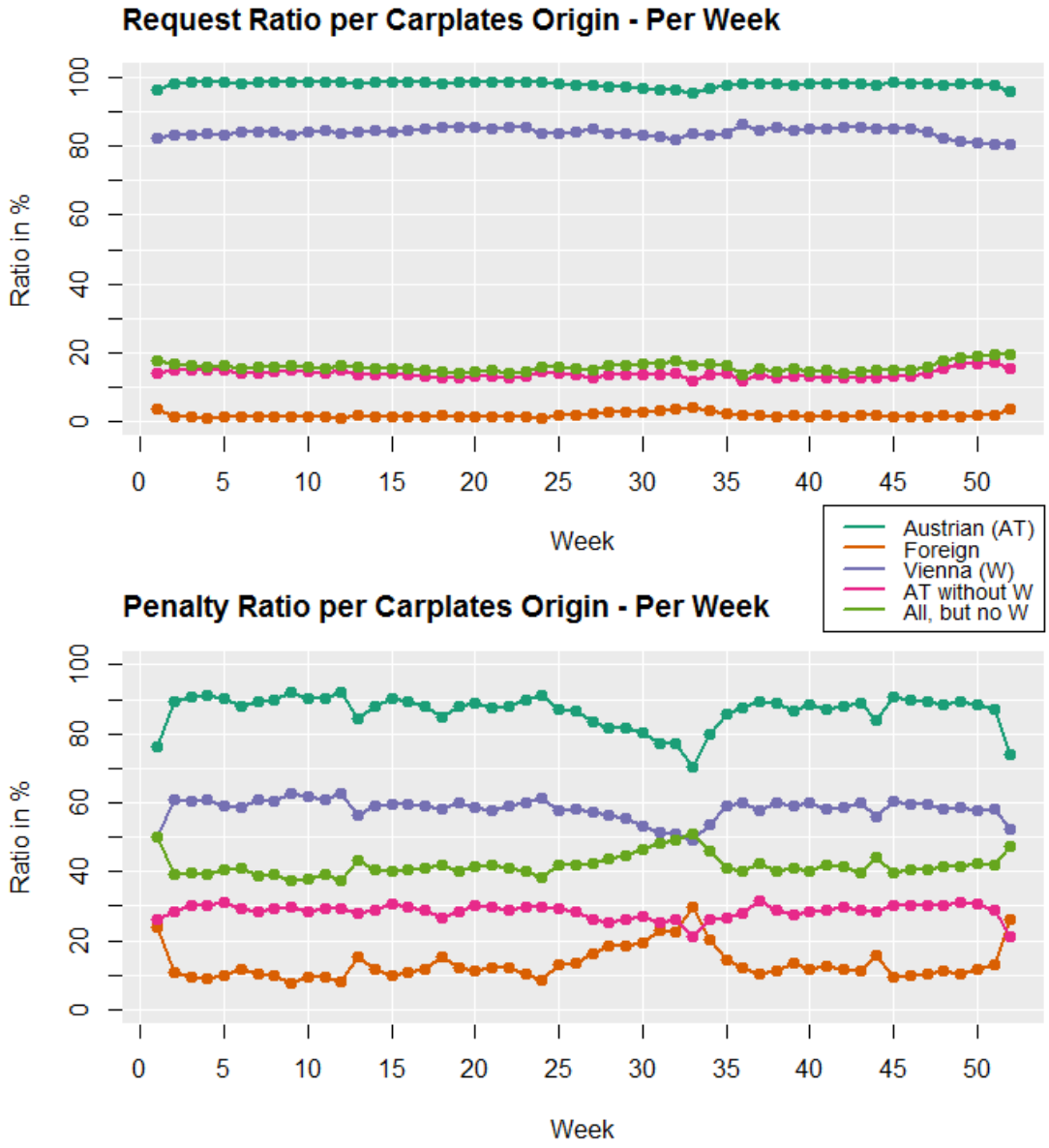


Figure 5.16: Comparison of ratio values of five carplate origin categories with weekly calculated values of the whole year of district one. Not all values are independent and therefore the sum per week is more than 100 percent.

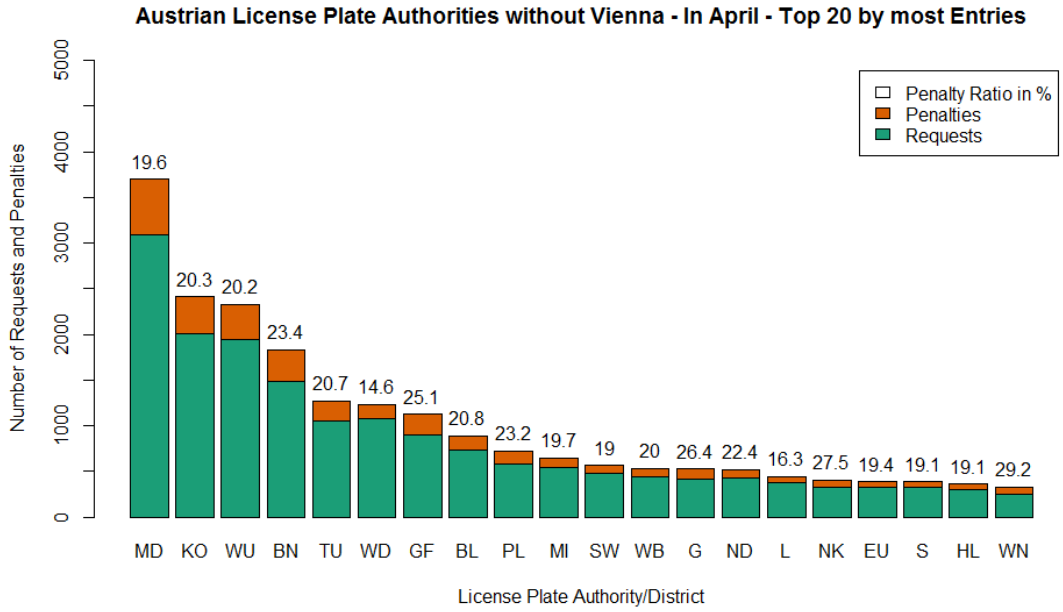


Figure 5.17: Bar chart of a selection of Austrian authorities of the April data of district one, showing the number of requests together with the number of penalties and the penalty ratio. The top twenty authorities by the number of entries are displayed, but plates with 'W' for Vienna are excluded.

number of entries, but Austrian license plates are excluded. Week 33 is chosen because it has the highest penalty value variation for the category of foreign license plates in Figure 5.16 and it is also the week with the 2nd most entries with 1609 to 1712 in week one. As the data is limited to the period of one week, the amount of entries is low. In total, about 30 different countries are present in the entries of this week, but also the wildcard value 'XXX', which can be chosen by the warden if the country is unknown, has the 12th highest amount of entries in the ordered list.

Generally, the number of entries is low for each country, only German license plates with 'D' do have nearly 500 entries. Then there are four other countries with 100 to 200 entries and all other categories are below one hundred. It can already be seen in Table 5.7 that the penalty ratio of foreign car plates is about 40 percent in the whole year. In the chosen week, the penalty ratio is especially high, which is also shown by the top ten categories. The lowest value is 57.9 for 'GB' and the highest value is 92 percent for license plates with 'I' as the country. Six out of ten values are above 85 percent.

Both diagrams, Figure 5.17 and Figure 5.18 are examples for visualizing this type of data for different origins of the cars' categories. The chosen categories and presented values can be completely different for other combinations of districts, months and weeks. The two categories, license plate authority and country are also not limited to a certain type

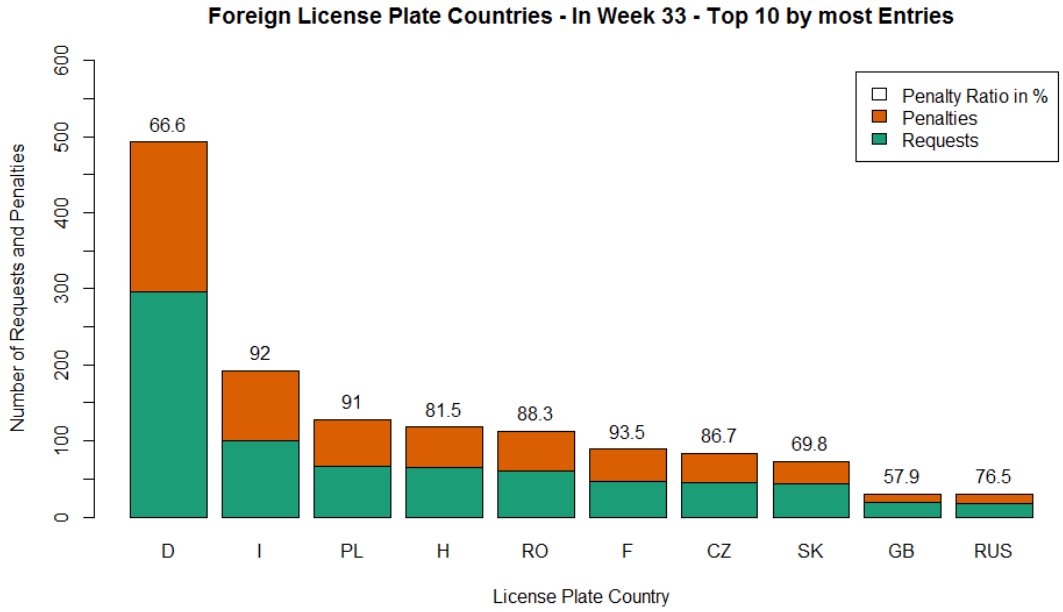


Figure 5.18: Bar chart of a selection of foreign countries of the week 33 data of district one, showing the number of requests together with the number of penalties and the penalty ratio. The top ten countries by the number of entries are displayed, but plates from Austria are excluded.

of time period or district.

### Year by Coordinate Points

For the following visualization, the data of the whole year of district six is used and for the dimension of place, the coordinate points are considered. Beforehand, a map of the chosen district is presented in Figure 5.19, for better understanding of the upcoming visualization. It is the same map as it is used in Chapter 4, but it is printed here again for easier handling. The source of the map is again the website of the City of Vienna[7] and the map is provided by Stadt Wien - ViennaGIS[8]. The chosen square area fits to the borders of the district and therefore also areas outside of the district are included in the following diagram. Two very special cases are the top-left and the bottom-right corner, because there are big areas which do not belong to the district.

The available data of district six with respect to the coordinate points is already listed in Table 5.5. Figure 5.20 shows two tile maps based on the whole year data of district six. Each tile depicts the ratio of foreign license plates to Austrian license plates in this area

[7] https://www.wien.gv.at/stadtplan/, Accessed 2019-08-10
[8] http://www.wien.gv.at/viennagis/, Accessed 2019-08-10

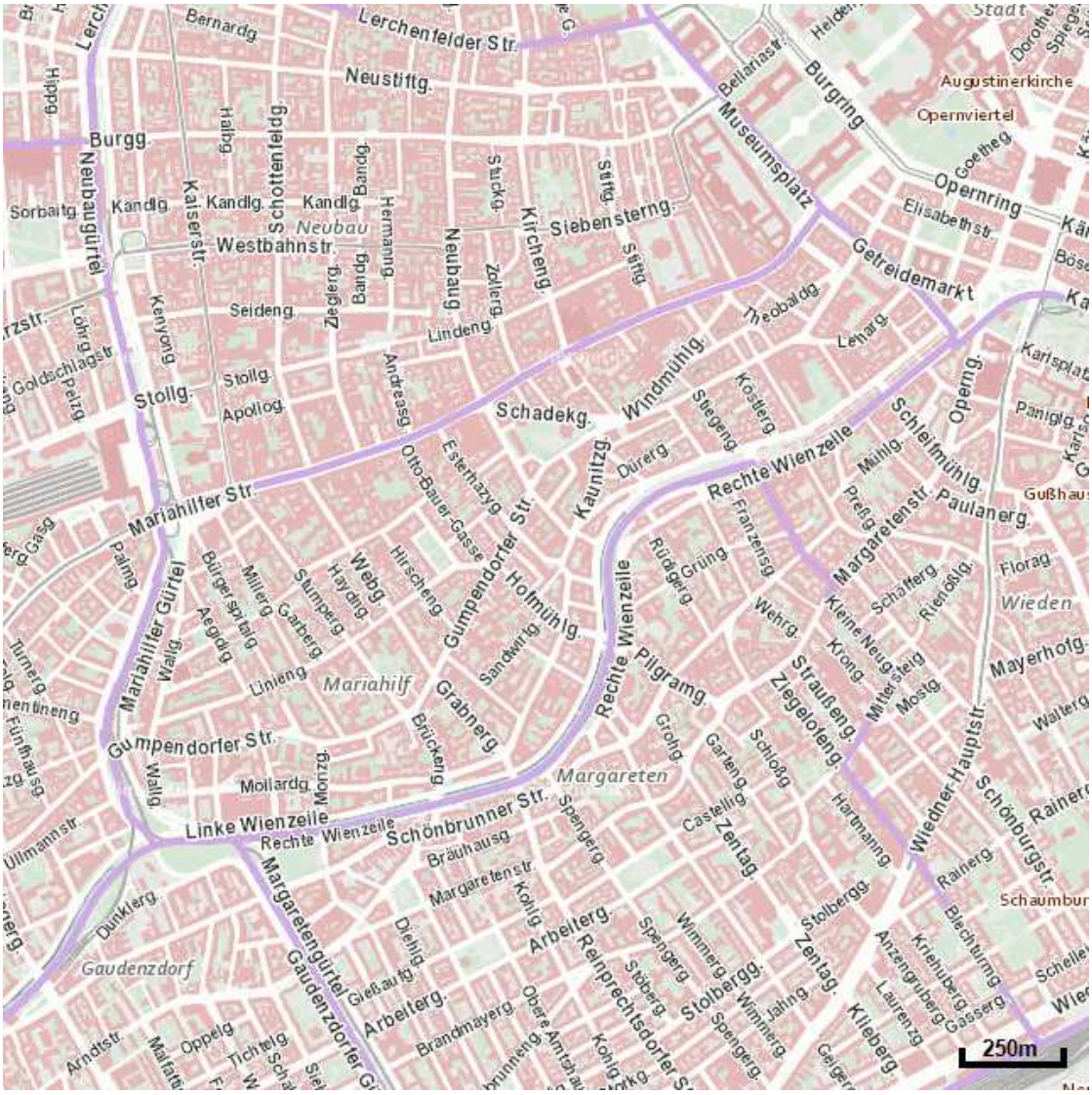


Figure 5.19: A map of district six in Vienna with border lines included.

for the whole year. As the data source, the first tile map uses all entries with calculated and estimated coordinate points, also called request points, while the second one uses all entries with closest coordinate points. If the minimum amount of ten foreign license plate entries for the first diagram and twenty entries for the second one are not fulfilled, the respective tile remains filled with grey color. The amount of entries of the second diagram is much higher and includes much more additional requests and therefore the ratio percentage values are lower.

The tiles in both maps show two certain characteristics, also by considering both maps together. There is one area on the left side of the district with a high ratio of foreign license plates. Secondly, the tiles on top of the maps are lighter than at the bottom, especially on the map which uses the closest coordinate point information. A possible explanation is, that those areas are more attractive for people with cars with foreign license plates, while darker areas, with lower ratio values are not.

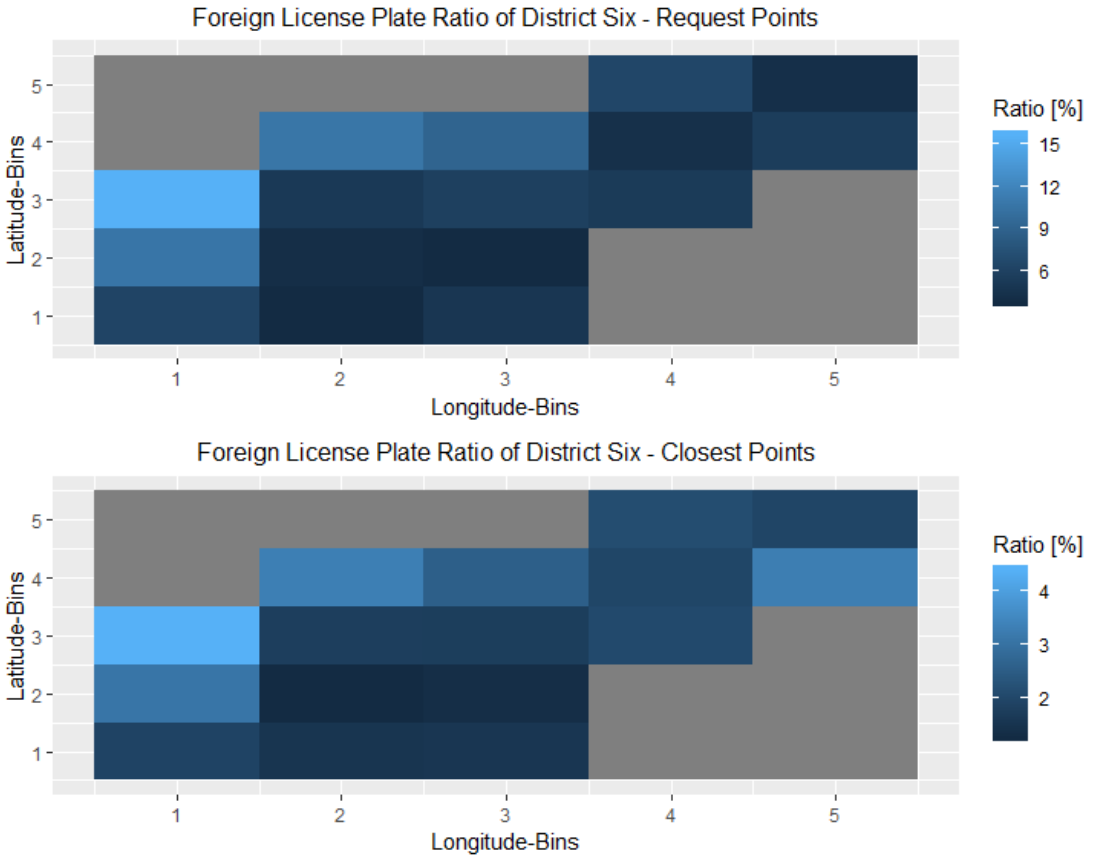


Figure 5.20: Tile maps of district six of the whole year data showing the ratio of foreign license plates to Austrian license plates. For the first tile map all entries with request coordinate points are used and for the second the closest coordinate points are used. All tiles with less than a certain amount of foreign entries are remained filled with grey color.

In Figure 5.20, both different coordinate point approaches are used. Each of the two approaches has its own disadvantage, which is the small amount of entries for the request coordinate points and the inaccurate position of the closest coordinate points. If the data shows the same characteristics for both types of coordinate points information, the significance of the outcome is increased.

## 5.6 Parking Density

In this section of the analysis and visualization part, it is the target to make some statements about the parking density in different levels of detail, based on the given source data. The main question is, if the data is suitable to make such statements and to which extent it can be made. The term "parking density" means the amount of parked cars in relation to the number of parking spaces. Therefore, it can also be an indicator for free parking spaces in an area. The first subsection describes a theoretical approach to calculate the parking density, while the later subsection clarifies why the data is not suitable for such an estimation.

### 5.6.1 Theoretical Application

This subsection describes the theoretical idea of how the enforcement data can be used to measure and describe the parking density. For calculating the parking density on a street for a certain time period, knowledge is needed about both affected indicators, namely the amount of parked cars and the number of parking spaces. If the parking enforcement is performed by a warden, the amount of parked cars can be easily obtained, because every entry in the protocol file directly represents a car on the street. It is assumed that for all entries the respective coordinate points can be achieved and thereby the position information is available. The second indicator, the number of parking spaces, is the difficult one, because this data is not given.

As an possible solution for the lack of knowledge about the number of parking spaces, Figure 5.21 shows an example street with three different situations, which are used to describe the process. Each line, A, B and C, depicts the same street X and all the parking spaces on different days, while the street including the four houses can be seen on the bottom. On the street, requests and penalties of cars are marked in the way how they are performed by a parking warden. Additionally, the timestamps are given for the first and the last penalty on this street. The main differences between the three cases are the amount of cars and the time period between both penalties, which also reflects the diverse number of requests and cars which are parked on the street. The first penalty is always performed in front of house number one, while the last penalty is always in front of the last house, which is number four.

The knowledge of these three situations is now used to determine the maximum amount of parking spaces on street X and afterwards the parking density can be classified in a general way. Therefore, a main assumption is needed, that the sequence with the most

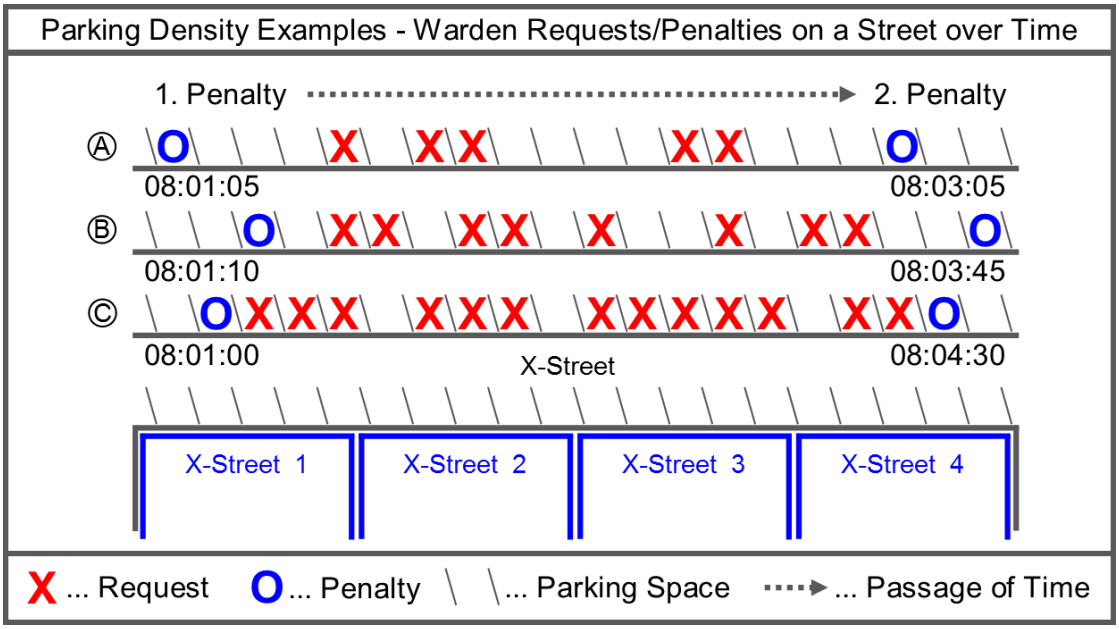


Figure 5.21: Three example situations A, B and C of requests and penalties performed by a warden on the same street with different timestamps for the first and the last penalty.

requests between two penalties is specifying the maximum amount of parking spaces, or at least is close to this value. As a consequence this assumptions make it possible to compare all other sequences with the maximum amount and to calculate a density value. As the maximum value is only an estimation, the calculated value is not exact, but at least it can be used to classify the parking density in general categories. Some examples for categories and their respective percentage values are listed in Table 5.9. For unequally distributed categories, the percentage ranges can be completely different.

The three presented situations from Figure 5.21, are calculated and classified as an example in Table 5.10. It shows the lines A, B, C and each one lists the number of cars as the sum of penalties and requests. The maximum parking spaces are set to the highest value of cars, which is 15 from line C. The parking density is calculated by the number of cars and the maximum amount of parking spaces. As an example classification, the unequally distributed categories, which are listed in Table 5.9 are chosen.

Table 5.10 in combination with Figure 5.21 now shows the lack of accuracy of this method. The amount of maximum parking spaces is calculated and defined as 15, but in reality there are 20 parking spaces available on the whole street. If the one or two are excluded because they are before and after the penalties, the number of spaces would still be 18. A possible way to prevent this wrong assumptions is to increase the maximum number of parking spaces before calculating the ratio. An example would be a static factor of 1.2, to increase the number by 20 percent. Obviously this can lead to situations where the maximum number of parking spaces is higher than the real amount, because when the

| **Two Levels - Equally distributed** | |
|---|---|
| Low | 0-50 % |
| High | 50-100 % |
| **Three Levels - Equally distributed** | |
| Low | 0-34 % |
| Medium | 34-67 % |
| High | 67-100 % |
| **Three Levels - Unequally distributed** | |
| Low | 0-50 % |
| Medium | 50-80 % |
| High | 80-100 % |

Table 5.9: Example category classifications for the calculated parking density.

| **Example parking density calculation** | | | | | |
|---|---|---|---|---|---|
| | Period of time | Pen+Req = Cars | Max. spaces | Ratio (≈ %) | Category |
| A | 8:01:05 - 8:03:05 | 2+5 = 7 | 15 | 46.7 | Low |
| B | 8:01:10 - 8:03:45 | 2+8 = 10 | 15 | 66.7 | Medium |
| C | 8:01:00 - 8:04:30 | 2+13 = 15 | 15 | 100 | High |

Table 5.10: Example parking density calculation of the presented three different sequences.

parking enforcement is performed, also all parking spaces can be occupied by cars and the increased value would then be too high. The usage of an static factor can help in some situations, but it is wrong for other cases.

In general, a rough estimation of the parking density is possible by this approach, but the accuracy is limited. Some improvements can also be made, for example by an additional consideration of the elapsed time during the first and the last penalty. This would also make it possible to allow a mixture of multiple requests and penalties, if for each type of entry an average amount of seconds is used. It can even go that far, that a gap of more seconds than normal is considered as a free parking space, but this approach is really dangerous because an amount of other circumstances can influence the behavior of a parking warden.

The previous example has also shown some additional problems when interpreting the penalty density. As it is shown, the third sequence, in combination with the other two, makes it possible to calculate the parking density for this specific street, at least in relation to other situations on the investigated street. However, the considered parking density is only relevant for this particular snapshot when the parking warden walks on the

street for three minutes. It can additionally be assumed that with these three sequences the outcome can also be estimated for some minutes before and afterwards, but here the problems already start. Is it relevant for 10, 20 or 30 minutes, or even for the whole morning time? This answer cannot be given by an estimation which is only based on three sequences of the same time period from different days. Therefore the whole process has many weak spots.

### 5.6.2 Limitations of the Data

During the description of the theoretical approach of calculating the parking density in Subsection 5.6.1, example situations are used to explain the whole concept. Some problem situations and questions have already been stated, but this section provides a sublist of sample circumstances and facts, why the given data is actually not suitable to calculate the parking density and why it even can not give an at least moderate estimation:

- **Similar Entry Sequences of Many Wardens Needed** For being able to estimate a certain segment of a street, many parking warden entry sequences are needed with very similar structure. Besides the already stated example with two penalties on two addresses, also requests with calculated coordinate points can be on the addresses. Nevertheless, by considering also requests, the required calculation power for finding all the segments and the sequences with the same reference points would be very big. The approach can also be extended by allowing neighbor addresses for the same segments, but than an additional handling of the obviously different amount of possible parking spaces has to be some how considered.
- **Relevant Time Period of an Estimation** As already mentioned in Subsection 5.6.1, if an estimation for the parking density is made, there occurs the question, what the relevant period of time for this estimation is? A really interesting statement would be the parking density on a street segment in the morning of a certain weekday, but a relevant estimation can only be made if enough values are available for each hour between 8 and 12 o'clock. The amount of required parking density values for this time period is totally random. Among others, it depends on the fluctuation of the cars for each point of time and the length of the investigated street segment. The main problem is that the described estimation is only a snapshot of the current parking density, but does not give much information about the minutes before and after this point of time.
- **Maximum Number of Parking Spaces** This point has already been mentioned at the end of Subsection 5.6.1, and it means the lack of knowledge about the maximum number of parking spaces. If all parking spaces of a street are never occupied at the same time, or at least in these situations are never checked by a parking warden, then it is impossible to estimate the maximum number of parking spaces and to calculate the parking density correctly. Also the possible way of

adding some additional percentage to the estimated maximum number of parking spaces is leading to other wrong values for streets where all parking spaces are often occupied.

- **Parking Spaces of Larger Areas** The number of parking spaces is not available for the mentioned street segments, but it is also not available for larger areas. If the amount would be available on a predefined sector level, it would give a chance to use the entries with coordinate point information to calculate the parking density for broader areas. Additionally, this approach cannot lead to a correct estimation because the total amount of entries in each area does also depend on the number of parking wardens who are working in the investigated area and period of time, which is often completely different.

- **Parking Time Limitations** Parking spaces exist in some areas where the time of day or the day itself completely influences the parking possibilities. There are certain hours or days, like lunchtime or school time, which are limiting the parking permissions street-wise. This does totally compromise all ways of calculating the maximum number of parking spaces in relation to the parked cars, because all available parking spaces could be occupied in a certain hour, while the algorithm thinks that there are much more spaces.

- **Walking Behavior of Parking Wardens** While for the coordinate point estimation, this point is not a big issue, it totally changes the possibilities for the parking density calculation. The path of a parking warden is not predefined and therefore it can also end up in special routes. This can lead to a high impact on the maximum number of parking spaces estimation, if the warden for example switches from one side of the street to the other exactly in the middle of an investigated street segment. As a consequence, the total amount of parking spaces is much higher in this case than for all other wardens who do not change the side of the street and instead check it again in the next sequence.

- **Short-term Parking Restrictions** Some of the least common, but as well existing circumstances are short-term parking restrictions. The period of time in which it is active can be completely different and lasts from single hours to whole months or even complete years. One of the most known examples are construction sites, which are often blocking all the parking spaces and make estimations of parking density nearly impossible for certain time periods. It influences the maximum number of parking spaces and also decreases the amount of entry lines because no cars are parking in the area during the active construction site.

- **Checking Amount of Data** If all the previous points are clarified, the last remaining topic is that there has to be enough data available for the calculation. Therefore it is necessary that the parking wardens often work in the investigated streets to generate enough entry lines, but this is only applicable for a very small amount of streets. Additionally, only entries with calculated coordinate points can

be used, because the closest coordinate point information is not accurate enough, which means that a big amount of data cannot be included. As a consequence, there is not enough data available to calculate the parking density for a broader context, but maybe only for very special streets and specific situations.

After all the listed circumstances are mentioned, a short check concerning the given entry lines is done to explain the lack of available data. For district one, the data has been checked for all entries with calculated coordinate points and closest coordinate points. Independently of the street length, the average amount of entries per day is calculated for each street and the minimum amount of necessarily available entries is set to 30 per day, as this would give a good chance for calculating an parking density estimation. As a result, there are only 15 streets left which fulfill the given criteria when using the calculated coordinate points and 60 streets for the entries with closest coordinate points information. For comparing this value, the whole district has about 280 streets. Basically, only entries with calculated coordinate points should be used for estimating the parking density as they are more accurate, but the ratio of 15 to 280 is very bad.

The whole topic of calculating the parking density is very complex and only applicable to specific streets under very special circumstances, because there is not enough data given to do this for bigger areas. Further investigations are not the topic of this thesis, but implementing and finding algorithms to improve this approach can be of high interest and should be part of the future work.

CHAPTER 6

# Critical Reflection

As a real world process, parking enforcement is an intermittently executed action in a totally inconstant environment and the main source for the basis data of this thesis. This chapter therefore describes some relevant problem situations, but also assumptions that have been necessary during all the previous chapters. Some of the issues have already been stated in the related chapters, but will be mentioned again in this chapter.

When working with data, it is crucial that it is correct and the processor can rely on all the given information. At least, this would be the ideal condition. Especially when working with data coming from a real world process, it always contains some kind of errors. This can be a total outage of an automatic system or a simple measurement error for a single instance of one particular product. In this context, the main element in the parking enforcement process is the parking warden, a human being operating its device. It easily happens that the warden is choosing the wrong street name or to make a typo while entering the house number. To some extent, it is possible to filter some kind of errors, but in the end the given data is as it is.

When a parking warden is checking the permissions of a car, besides the technical request, it is also possible that there is a paper parking ticket just behind the windscreen. If the warden is checking the paper parking ticket first and the car is not fined afterwards, the license plate will never show up in the protocol file. This means that some percentages of all cars cannot be considered in this whole thesis. The main assumption is, that the distribution of paper parking tickets is similar across all areas of the city and additionally, the usage of the technical solution is very high in Vienna. Nevertheless, some of the cars which are parking in the real world and have been checked by a warden do not show up in the data.

During the process of data preprocessing, many assumptions and also estimations are necessary and as a consequence, some inaccuracies can occur. One example is to choose the closest point on street from all the calculated points, which is basically only an

estimation of the position. All of the small emerged inaccuracies are the reason for choosing tile maps as the most detailed view, because directly showing the values on a street level would not be exact enough.

The whole process of gaining coordinate points is limited in the way, that a street is considered as a line and does not have a certain width. This is because the street graph does only provide the street details in this way for most of the cases. As a consequence, the opposite and frontage road flags are not considered in the whole thesis, but strictly speaking, this is possibly a lost of accuracy for some penalty entries.

The address service, which is used to request coordinate points for penalty entries, is relying on a combination of street name and house number. In the City of Vienna, a few street names do exist twice, which could be a possible problem when requesting coordinate points, especially if the district is not checked. For all sample tests, the second occurrence of a street has not been in the main area of the city where also parking enforcement is performed, for example in the outer districts. This circumstance can still be a problem for other applications in this area.

During the data transformation, also a parsing of the free text field for the penalty location is performed. This field is written manually by the parking wardens, hence it provides many different notations of the same meaning, which is already described in detail in the related chapter. The whole process of analyzing and interpreting written sentences or statements is an own research field and could at least fill a whole chapter in this thesis.

A circumstance which has only been stated marginally until now, is the fact that parking spaces along side a street are not permanently. Of course, a parking space mostly is out of concrete and stays as it is, but even the appearance of a street can change over time, which can influence yearly data about parking in a certain area. Two other examples are construction sites at the place of a parking space, also for longer times, and some parking spaces even change during the daytime from parking spaces to lanes of traffic.

The main source data of the parking enforcement are daily protocol files which include all actions performed by all the parking wardens in chronological order. This fact allows also situations in which a parking warden sends multiple requests of the same car consecutively. The license plates are anonymized during the import, which means that the recurring entries of the same license plate cannot be filtered afterwards. In contrast to multiple request entries for the same car, the amount of penalties remains normal, because a penalty is basically only performed once.

A parking warden has the possibility to cancel a fined penalty, which makes it possible to make a new penalty of the same license plate in a short period of time. As the cancellations are not given in the protocol file, some cars could be counted as two penalties even if only one parking fine notice exists in the end.

During the analysis, it is shown in one section, that a higher amount of requests leads to a lower ratio of requests to penalties. This value can be influenced by the fact, that a car,

which has already been fined, does not get fined again on the same day. As a result, each car, which is parked against the parking rules for a whole day, maybe only gets fined once, even if the parking wardens are passing by more often and send multiple requests for the other cars in the same street.

For some visualizations during the analysis, the closest penalty point coordinate estimations from the data transformations are also used beside of the calculated coordinate points. As already mentioned, each of these estimations do have a distance value, which describes how many seconds it is away from the chosen position. In the applied analysis, these distance value has not been considered when making the tile maps. The weighting and influence of this value could be calculated by new approaches or algorithms, but this has not been done during this thesis.

A statistical analysis can only use fields and information which are provided by the source data or related external resources, but sometimes additional knowledge in the data would be great to investigate its influence. One example is the gender of a car driver who has parked a car, which is later checked by a parking warden. Of course, this example feature is not given in the data that is collected during the parking enforcement, but from a statistical point of view, such additional features and their impact on the parking behavior can be of high interest.

As the last statement for this chapter, it should be mentioned again, that the chosen visualizations during analysis only show the outcome and the results of specific combinations of levels of detail in the dimension of place and the dimension of time. For other districts or periods of time, the outcome is maybe different or even opposing to the described behavior. Nevertheless, the presented approaches provide the use cases for further investigations on other combinations of levels of detail.

CHAPTER 7

# Summary

This thesis has the aim of gaining a better understanding of the data which is collected during the parking enforcement process on the example of the City of Vienna. It should be shown what kind of information is available and how this could be analyzed and presented by statistical visualizations. The main source data are daily protocol files of the year 2018 generated on the backend servers. Each file consists of request and penalty entries sent by the parking wardens from their special working devices. For the usage and analysis of this data, which also contains highly personal information, strict conditions have to be fulfilled. As an example, the parking warden hash values, as well as the license plates have to be anonymized during the step of importing the data.

Five research questions have been formulated for achieving the aim of this thesis. Before answering those questions, a wide literature research has been made in various areas like coordinate points for localization, parking or statistical analysis and visualizations.

The first research question concentrates on the localization of all the available entry lines. As a result, it is possible to estimate the position of nearly each entry line by coordinate points, but in different quality. If a request is performed between multiple penalties, the chance increases that its position is estimated accurately, because only for penalties, the source data contains location information. The whole process of gaining coordinate points is performed during the preprocessing of the data, which is also used to bring it into the correct format for the analysis. Additionally, some open data, which are provided on different websites, are also used to enhance the amount of available information, like the address service provided by the City of Vienna.

For answering question two, the external environment is analyzed to achieve a list of possible levels of detail in the dimensions of time and place, by taking the results of the first question into account. While for the dimension of time, many different levels could be achieved, like yearly, monthly, weekly or daily views, the dimension of place is very limited due to different reasons.

The next research question concentrates on the possible indicators or features which are available by the data. Limited by the structure of the protocol files and the intermittently performed parking enforcement, two main features have been identified. These are the ratio between requests and penalties and different variations in the context of the origin of the license plates. By answering also question five, an analysis of these features for different combinations of levels of detail is performed. The outcome are multiple visualizations which provide approaches for achieving qualified statements and gaining additional knowledge about the parking behavior in a combination of the dimension of time and place. For example, one diagram shows the week, which has the highest percentage value of foreign license plates on all penalties for a specific district, which is week number 33 in district one.

The fourth research question is about a possible estimation of the parking density and its potential classification. The thesis provides a theoretical approach about how this could be achieved, but for the most part of the city the data does not include enough entries. Additionally, multiple barriers and problems are described, why this task is difficult by only relying on enforcement data, one example is the lack of knowledge about the total number of parking spaces.

As a final statement, the provided source data of a parking enforcement process is definitely suitable to gain additional information about the parking behavior in a city, but it is always necessary to consider the individual circumstances and the related approaches. An example is the amount of available data and the quality of the data, which is generally generated by human beings, the limiting factor in some situations. This thesis provides the possible approaches and strategies that can be used for a statistical analysis and visualization of the complete data of the City of Vienna for various combinations of the levels of detail.

## 7.1 Future Work

In this section, additional starting points for further topics concerning this thesis are described. Some of them are related directly and describe additional approaches or point of views, while others are possible improvements in specific situations.

As already mentioned in various situations of this thesis, the chosen combinations of the levels of detail in the dimensions of time and place are randomly chosen. These combinations could be extended a lot and easy by just using the same approaches for different levels, like other time periods or multiple districts. Especially in the context of using the coordinate points for the analysis and visualization, much more possibilities exist only by changing the size of each tile in a tile map.

The whole process of estimating the parking density is only described in theory and not followed up for those streets, which maybe provide enough data. This topic could fill up at least a whole chapter just for developing and investigating algorithms to use the

given data in a proper way for calculating the number of available parking spaces and afterwards for estimating the parking density.

During the preprocessing of the data, open data from external resources are used to enhance the available information. Two examples for additional extensions of this thesis are to include weather data, as well as the event calendar of the city. Both topics could totally influence the parking behavior and could especially describe some higher or lower numbers which have already been mentioned in this thesis.

Another way of enhancing the given information can be to include the traffic volume. This could lead to interesting dependencies between the amount of traffic and the parking behavior.

While this thesis only uses the enforcement data which is provided by the City of Vienna, there are also additional information collected in the context of parking. There are the electronically booked parking tickets via mobile parking that are sent to the servers by the car drivers. Additionally, a local resident can buy a long-term parking permission for being allowed to park in the base district without getting fined. If there would also be a possibility to include these two types of data into the already given source data, the knowledge of each single license plate would be increased a lot and additional analysis can be realized. Another requirement for realizing this would be to define additional conditions for using and getting the data at all.

In the area of statistics, other topics could be considered like generating models and working on forecasting. Furthermore, also the development of simulations of the process of parking or parking enforcement are possible.

During the realization of this thesis, which is working with the data of 2018, some changes of the protocol file are already in progress or have been done in the year 2019. For example, this includes entries of penalty cancellations or besides of just storing the district, also the field rayon should be used correctly to get a more detailed location description of all entry lines. An adaption and extension of the developed approaches of this thesis by these new features can also be a possible future work.

# List of Figures


3.1 An example timeline of two parking wardens and a part of the corresponding protocol. It depicts the mixture of requests and penalties of multiple wardens in the protocol file. . . . . 30
3.2 Overview of the data transformation steps and the included external resources, the street graph file and the address web service. . . . . 32
3.3 Simplified representation of three bounding boxes and the calculated points of Y-Street. The center point of a bounding box is calculated by the bottom-left and top-right point of a bounding box. The addresses lie on Y-Street and the green arrows mark the closest "point on street" of an address. . . . . 38
3.4 Example timeline of a warden, including illustration of gaining the estimated coordinate points for requests between two penalties of the same street. . . 41

4.1 A map of district six in Vienna with border lines included. . . . . 44
4.2 Plot of all calculated penalty coordinate points of district six. . . . . 45
4.3 Map of district six including coordinate points of all penalty entries. Additionally some special areas are marked and named with letters. . . . . 46

5.1 Comparison of the six different half year splitting possibilities of district one of the total number of entry lines and the ratio of requests to penalties. MX labels the begin of the first half of the year, where the 'X' is the number of the month, i.e., M2 means that the first half of the year starts with February. 62
5.2 Comparison of the two different quarter splitting possibilities of district one, showing the four values of total number of entries, number of requests, number of penalties and the ratio of requests to penalties. 'Eco' is the economic approach where the first quarter starts with the first month of the year, while 'Meteo' is the meteorological approach where the first quarter starts with the first of March. . . . . 63
5.3 Comparison of the yearly data, based on monthly values for district one and six. The four diagrams are showing the number of entries, requests, penalties and the requests to penalties ratio. . . . . 64
5.4 Cross-Correlation functions of district one and six for showing the relation of the number of requests and the penalty ratio. . . . . 65

5.5 Comparison of the yearly data, based on weekly values for district one and six. The four diagrams are showing the number of entries, requests, penalties and the requests to penalties ratio. . . . . 66
5.6 Cross-Correlation functions of district one and six for showing the relation of the number of requests and the penalty ratio. . . . . 67
5.7 Histogram showing the number of requests per day of district one. All 15 days which have no performed request in the source data are excluded. . . 68
5.8 Two standard seasonal decomposition time series by Loess (STL) plots of district one, including data values, seasonal, trend and remainder part. The used standardizations are zero values on the left side and the mean on the right side. . . . . 69
5.9 Two boxplots depicting the penalty ratio of district one of all days grouped by the day of the week in the left plot and a workday to weekend day in comparison on the right side. The input data is not standardized, but y-axis values above 35 percent are distributed around this border value. . . . . 70
5.10 Two boxplots showing the number of requests and the penalty ratio of district one on every working day in certain time periods of the day. The input data is not standardized, but y-axis values above 25 percent in the right diagram are distributed around this border value. . . . . 71
5.11 A map of district one in Vienna with border lines included. . . . . 73
5.12 Tile maps of district one entries with calculated and estimated coordinate points with different amounts of bins in each row. The source is the complete year. The left side depicts the number of entries in a bin and the right side depicts the calculated penalty ratio. Ratio values above 100 are set to 100. 75
5.13 Tile maps of district one entries including closest penalty points information with different amounts of bins in each row. The source is the complete year. The left side depicts the number of entries in a bin and the right side depicts the calculated penalty ratio. Ratio values above 35 percent are set to 35. 77
5.14 Tile maps of district one entries including closest penalty point information with similar bin sizes showing the number of entries, but for different time levels. . . . . 79
5.15 Tile maps of district one entries including closest penalty point information with similar bin sizes showing the penalty ratio, but for different time levels. Ratio values above 40 percent are set to this value. . . . . 80
5.16 Comparison of ratio values of five carplate origin categories with weekly calculated values of the whole year of district one. Not all values are independent and therefore the sum per week is more than 100 percent. . . . . 86
5.17 Bar chart of a selection of Austrian authorities of the April data of district one, showing the number of requests together with the number of penalties and the penalty ratio. The top twenty authorities by the number of entries are displayed, but plates with 'W' for Vienna are excluded. . . . . 87

5.18 Bar chart of a selection of foreign countries of the week 33 data of district one, showing the number of requests together with the number of penalties and the penalty ratio. The top ten countries by the number of entries are displayed, but plates from Austria are excluded. . . . . 88
5.19 A map of district six in Vienna with border lines included. . . . . 89
5.20 Tile maps of district six of the whole year data showing the ratio of foreign license plates to Austrian license plates. For the first tile map all entries with request coordinate points are used and for the second the closest coordinate points are used. All tiles with less than a certain amount of foreign entries are remained filled with grey color. . . . . 90
5.21 Three example situations A, B and C of requests and penalties performed by a warden on the same street with different timestamps for the first and the last penalty. . . . . 92

# List of Tables


3.1 Examples of location free text values with positive determined house numbers and negative cases where no house number could be determined. . . . . 36

5.1 Amount of entry lines in comparison to the size of a district, sorted by the amount of entries per ha of the district size. . . . . 57
5.2 Amount of entry lines in comparison to the number of penalty lines, sorted by the ratio of entries to penalties. Listing only the districts with the highest entries to district size ratio. . . . . 57
5.3 Results of the data transformation process of the whole City of Vienna. Chosen data fields are in respect to the penalty ratio. . . . . 59
5.4 Results of the data transformation process of district one. Chosen data fields are in respect to the penalty ratio. . . . . 60
5.5 Results of the data transformation process of district six. Chosen data fields are in respect to the penalty ratio. . . . . 61
5.6 Results of the data transformation process of the whole City of Vienna. Chosen data fields are in respect to the origin of cars. . . . . 82
5.7 Results of the data transformation process of district one. Chosen data fields are in respect to the origin of cars. . . . . 83
5.8 Results of the data transformation process of district six. Chosen data fields are in respect to the origin of cars. . . . . 84
5.9 Example category classifications for the calculated parking density. . . . . 93
5.10 Example parking density calculation of the presented three different sequences. 93

# Bibliography


Lynd D Bacon. Graphical data analysis and visualization techniques for marketing research. In *The Seventh Annual Advanced Research Techniques Forum*. American Marketing Association, 1996.

Yoav Benjamini. Opening the box of a boxplot. *The American Statistician*, 42(4):257–262, 1988.

Howard Brody, Michael Russell Rip, Peter Vinten-Johansen, Nigel Paneth, and Stephen Rachman. Map-making and myth-making in broad street: the london cholera epidemic, 1854. *The Lancet*, 356(9223):64–68, 2000.

Stuart T Card, Jock D Mackinlay, and Ben Scheiderman. Readings in information visualization: using vision to think. Interactive Technologies, San Francisco, California, USA, 1999. Morgan Kaufmann Publishers. ISBN 978-1-55860-533-6.

Christine H. Chih and Douglass S. Parker. The persuasive phase of visualization. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 884–892, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-193-4. doi: 10.1145/1401890.1401996. URL `http://doi.acm.org/10.1145/1401890.1401996`.

Robert B Cleveland, William S Cleveland, Jean E McRae, and Irma Terpenning. Stl: A seasonal-trend decomposition. *Journal of Official Statistics*, 6(1):3–73, 1990.

Paul Cowpertwait and Andrew Metcalfe. *Introductory Time Series With R*. Springer Science & Business Media, Berlin, Germany, 01 2009. ISBN 9780387886978. doi: 10.1007/978-0-387-88698-5.

MJ Crawley. *The R Book*. John Wiley & Sons, Chichester, United Kingdom, 2007. ISBN 978-0470973929.

Chrystinne Oliveira Fernandes, Felipe Baldino Moreira, Simone Diniz Junqueira Barbosa, and Carlos José Pereira de Lucena. What your eeg wearable sensors can tell about you? In *Proceedings of the 1st International Conference on Internet of Things and Machine Learning*, IML '17, pages 2:1–2:10, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-5243-7. doi: 10.1145/3109761.3158389. URL `http://doi.acm.org/10.1145/3109761.3158389`.

David Freedman and Persi Diaconis. On the histogram as a density estimator: L 2 theory. *Probability Theory and Related Fields*, 57(4):453–476, 1981.

Michael Friendly. Visions and re-visions of Charles Joseph Minard. *Journal of Educational and Behavioral Statistics*, 27(1):31–51, 2002.

Salvador García, Sergio Ramírez-Gallego, Julián Luengo, José Manuel Benítez, and Francisco Herrera. Big data preprocessing: methods and prospects. *Big Data Analytics*, 1(1):9, 2016.

Sean Gillies, H Butler, M Daly, A Doyle, and T Schaub. The geojson format. *RFC 7946; The Internet Engineering Task Force*, 2016. URL `https://tools.ietf.org/html/rfc7946`.

Barbara Görres. Vom globalen Bezugssystem bis zur Umsetzung für die Praxis. Bonn, Germany, 01 2010 (accessed April 20, 2019). Institute of Geodesy and Geoinformation, University of Bonn. URL `https://www.gib.uni-bonn.de/team/lehrbeauftragte/bgoerres/papers/100715koeln_fuer_web.pdf`.

M. Graham and J. Kennedy. Using curves to enhance parallel coordinate visualisations. In *Proceedings on Seventh International Conference on Information Visualization, 2003. IV 2003.*, pages 10–16. IEEE, July 2003. doi: 10.1109/IV.2003.1217950.

Jerry L. Hintze and Ray D. Nelson. Violin plots: A box plot-density trace synergism. *The American Statistician*, 52(2):181–184, 1998. doi: 10.1080/00031305.1998.10480559. URL `https://www.tandfonline.com/doi/abs/10.1080/00031305.1998.10480559`.

Dang Hai Hoang, Thorsten Strufe, Quang Duc Le, Phong Thanh Bui, Thieu Nga Pham, Nguyet Thi Thai, Thuy Duong Le, and Immanuel Schweizer. Processing and visualizing traffic pollution data in hanoi city from a wireless sensor network. In *38th Annual IEEE Conference on Local Computer Networks-Workshops*, pages 48–55. IEEE, 2013.

Toby Dylan Hocking, Carson Sievert, T Tsai, and Susan VanderPlas. *Two new keywords for interactive, animated plot design: clickSelects and showSelected*, 2015, (accessed April 20, 2019). URL `https://raw.githubusercontent.com/tdhock/animint-paper/master/HOCKING-animint.pdf`.

Bernhard Hofmann-Wellenhof, Herbert Lichtenegger, and Elmar Wasle. *GNSS–global navigation satellite systems: GPS, GLONASS, Galileo, and more*. Springer Science & Business Media, Berlin, Germany, 2007. ISBN 978-3-211-73012-6.

Ernest Hovmöller. The trough-and-ridge diagram. *Tellus*, 1(2):62–66, 1949.

Alfred Inselberg and Bernard Dimsdale. Parallel coordinates: A tool for visualizing multidimensional geometry. In *Proceedings of the 1st Conference on Visualization '90*, VIS '90, pages 361–378, Los Alamitos, CA, USA, 1990. IEEE Computer Society Press. ISBN 0-8186-2083-8. URL `http://dl.acm.org/citation.cfm?id=949531.949588`.

HariNamSimran Kaur Khalsa. Effects of climate change upon birds in the mojave and sonoran deserts. Master's thesis, The University of New Mexico, Albuquerque, New Mexico, 2012. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.825.7525&rep=rep1&type=pdf.

Freddy Lécué, Simone Tallevi-Diotallevi, Jer Hayes, Robert Tucker, Veli Bicer, Marco Luca Sbodio, and Pierpaolo Tommasi. Star-city: semantic traffic analytics and reasoning for city. In *Proceedings of the 19th international conference on Intelligent User Interfaces*, pages 179–188. ACM, 2014.

Sang-Min Lee, Jinkwan Park, Daegeon Kwon, Bokuk Park, Hwan-Gue Cho, and DoHoon Lee. A new visualization method for pairwise time-series data with random walk plot. In *Proceedings of the 9th International Conference on Ubiquitous Information Management and Communication*, IMCOM '15, pages 64:1–64:8, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3377-1. doi: 10.1145/2701126.2701144. URL http://doi.acm.org/10.1145/2701126.2701144.

G Merkel. Positionsbestimmung mit GPS. *Mitteilungsblatt DVW-Bayern 4/1996*, 1996 (accessed April 20, 2019). URL https://www.dvw.de/sites/default/files/landesverein-bayern/VeroeffentlichungenMitteilungen1996_4/DVW_1996_4_Merkel_Positionsbestimmung_mit_GPS.pdf.

R Moll, RH Cameron, and M Schüssler. Vortices in simulations of solar surface convection. *Astronomy & Astrophysics*, 533:A126, 2011.

Edzer Pebesma et al. spacetime: Spatio-temporal data in R. *Journal of Statistical Software*, 51(7):1–30, 2012.

Kristin Potter, Hans Hagen, Andreas Kerren, and Peter Dannenmann. Methods for presenting statistical information: The box plot. *Visualization of large and unstructured data sets*, 4:97–106, 2006.

R Allan Reese. Boxplots. *Significance*, 2(3):134–135, 2005. doi: 10.1111/j.1740-9713.2005.00118.x. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1740-9713.2005.00118.x.

Thomas A Runkler. Data analytics. *Wiesbaden: Springer*, 10:978–3, 2012. doi: 10.1007/978-3-658-14075-5.

DAVID W. SCOTT. On optimal and data-based histograms. *Biometrika*, 66(3):605–610, 12 1979. ISSN 0006-3444. doi: 10.1093/biomet/66.3.605. URL https://doi.org/10.1093/biomet/66.3.605.

O. Shamir, I. Yacoby, and N. Paldor. The matsuno baroclinic wave test case. *Geoscientific Model Development Discussions*, 2018:1–16, 2018. doi: 10.5194/gmd-2018-260. URL https://www.geosci-model-dev-discuss.net/gmd-2018-260/.

Donald C Shoup. Cruising for parking. *Transport Policy*, 13(6):479–486, 2006. ISSN "0967-070X". doi: "https://doi.org/10.1016/j.tranpol.2006.05.005".

Robert H Shumway and David S Stoffer. *Time series analysis and its applications: with R examples*. Springer Science & Business Media, Berlin, Germany, 2017. ISBN 978-1441978646.

Swasti Singhal and Monika Jena. A study on weka tool for data preprocessing, classification and clustering. *International Journal of Innovative technology and exploring engineering (IJItee)*, 2(6):250–253, 2013.

Herbert A Sturges. The choice of a class interval. *Journal of the American Statistical Association*, 21(153):65–66, 1926.

Chi-Chia Sun, Shanq-Jang Ruan, Mon-Chau Shie, and Tun-Wen Pai. Dynamic contrast enhancement based on histogram specification. *IEEE Transactions on Consumer Electronics*, 51(4):1300–1305, 2005.

Franco Tecchia, Celine Loscos, and Yiorgos Chrysanthou. Visualizing crowds in real-time. In *Computer Graphics Forum*, volume 21, pages 753–765. Wiley Online Library, 2002.

Edward R Tufte. *The visual display of quantitative information*, volume 2. Graphics press, 2001.

Janet Vertesi. Mind the gap: The tube map as London's user interface. Ithaca, NY, USA, 2005 (accessed April 20, 2019). Science & Technology Studies Department, Cornell University. URL `https://pdfs.semanticscholar.org/75f8/acafd96663ca368442d1cd5f55dc125a1c6e.pdf`.

Robert Weber, Gerhard Walter, and Stefan Klotz. GPS-relevante Koordinatensysteme und deren Bezug zum österreichischen Festpunktfeld. *VGI-Österreichische Zeitschrift für Vermessung und Geoinformation*, 83:190–200, 1995.

MR Wigan. Parking models in transportation planing. *Traf. Eng. & Ctrl*, 16:488–489, 1975.