University of
Zurich UZH

GEO 511
Master′s Thesis
31 January 2020

Parking occupancy prediction based on historical parking data and geographic data

Michael BALMER
14-709-018

Supervised by:
Dr. Haosheng HUANG

Faculty representative:
Prof. Dr. Robert WEIBEL

Geographic Information Systems
Department of Geography
University of Zurich

Abstract

In light of a growing population and increased urbanization, car parking is critical for cities to manage. As a result, the prediction of parking occupancy has received attention in recent years. In spite of the fact that many external data sources have been considered in the prediction models, the underlying geographic context has mostly been ignored. To analyse the contribution of spatial information to parking occupancy prediction models, centrality, land use and Point Of Interest (POI) data were incorporated in Random Forest (RF) and Artificial Neural Network (ANN) prediction models in this thesis. Model performances were compared to a baseline, only including historical and temporal data input. Moreover, the influence of the amount of training data, the prediction horizon and the spatial variation of the prediction were explored. The inclusion of spatial information could be attributed to a performance improvement of up to 25% compared to the baseline. Moreover, as the prediction horizon expanded and predictions became less reliable, the relevance of spatial input increased. In general, land use and POI data proved to be more beneficial than centrality. Among all geographic features, the amount of office space in the vicinity of the respective parking segments had most predictive relevance. The amount of training data input did not have a significant influence on the performance of the RF model and should be subject to further research. The ANN model, conversely, achieved optimal result on a training input of 5 days. Likely attributed to varying occupancy patterns, prediction performance disparities could be identified for different parking districts and segments. Generally, the RF model outperformed the ANN model on all predictions. A possible explanation is its robustness to overfitting and its simplicity compared to the ANN.

Key words: Parking occupancy, prediction, geography, spatial information, machine learning

Acknowledgements

Firstly, I would like to thank Dr. Prof. Robert Weibel for offering me the opportunity to conduct this research. Also, I would like to express great appreciation to my supervisor, Dr. Haosheng Huang, who guided my through each stage of the process. Lastly, I wish to acknowledge the support from my friends, colleagues and family.

Contents

List of Figures

List of Tables

List of Abbreviations

ANN	**A**rtificial **N**eural **N**etwork
ARIMA	**A**uto **R**egressive **I**ntegrated **M**oving **A**verage
DOW	**D**ay **O**f the **W**eek
FFNN	**F**eed **F**orward **N**eural **N**etwork
IncMSE	**Inc**rease in **M**ean **S**quared **E**rror
MAE	**M**ean **A**bsolute **E**rror
MSE	**M**ean **S**quared **E**rror
ML	**M**achine **L**earning
OSM	**O**pen **S**treet **M**ap
POI	**P**oint **O**f **I**nterest
RF	**R**andom **F**orest
RNN	**R**ecurrent **N**eural **N**etwork
SVR	**S**upport **V**ector **R**egression
TOD	**T**ime **O**f the **D**ay

Chapter 1

Introduction

A growing population and increased urbanization are challenges our world is facing over the course of this century. It is predicted that approximately 70% of the world's population will be living in urban areas by the year 2050 (Bellissent, 2010). Efficiently managing the infrastructure and services is hence an increasingly important issue. Car parking facilities and traffic are among the most critical services cities need to manage.

Finding a free parking space in urban areas can be a very challenging task, especially for drivers who are unfamiliar with the area. Traffic congestion, coupled with scarcity of parking locations as well as changing parking rules and restrictions can make parking frustrating and stressful (Rajabioun and Ioannou, 2015). Worse yet, vehicles roaming around looking for free parking spaces, so-called parking search traffic, has far-reaching socio-economic consequences. A significant amount of fuel is wasted, intensifying environmental problems. Moreover, car drivers usually go at lower speeds when looking for an on-street spot which can lead to further congestion (Hampshire et al., 2016). Additionally, an increased number of traffic accidents is linked to the search for parking (Bush and Chavis, 2017).

Shoup (2006) examined the problem of parking search traffic in certain study areas of several US cities. He found that, on average, it amounts to as much as 30% of total traffic whereas average search time is more than 8 minutes. These findings, however, are difficult to generalize, as the experiments were conducted in problematic areas where long cruising times are expected. In a small business district in Los Angeles with 470 parking spaces, it has been estimated that parking search traffic accounts for 950,000 additional miles driven in a year. This adds up to an emission of approximately 730 tons of CO_2 (Shoup, 2007). Another study assessed random car-trips in the Netherlands and concluded that 30% of all trips ended in parking search cruising (van Ommeren et al., 2012). Average cruising time, however, was only 36 seconds, making up a few percent of total travel time. Moreover, they found that parking search traffic was more prevalent in large cities.

A potential solution to help mitigate parking search is the provision of parking-related data. If drivers are notified about the parking availability situation at their intended destination, the traffic congestion can be controlled and potentially mitigated. For off-street parking, signposts have signalled parking locations and a count of vacant parking spaces for many years already (Axhausen et al., 1994). Generally, these systems are referred to as parking guidance and information systems. Incoming and departing vehicles are monitored at the barriers and the number of empty spaces is derived. For on-street parking, however, the acquisition of parking availability is more complex. The deployment of intelligent sensors in parking lots has proved to be a practical method for monitoring parking occupancy as well as getting an insight from the vast amounts of data collected (Zheng et al., 2015). Ideally, this information can be transmitted to drivers' navigation systems in search of a parking space in the form of dynamic parking maps. As these systems are associated with high installation and maintenance costs (Xu et al., 2013), only a few cities have deployed them. Wireless in-ground parking sensors, for instance, have been installed in select street blocks of the central business district of Melbourne, Australia (Intelligent Sensors, Sensor Networks and Information Processing, 2015) and in the 'smart' city of Santander, Spain (SmartSantander, 2015). The city of San Francisco has made real-time parking availability data open to the public, encouraging people to make their own decisions regarding parking (Zheng et al., 2015).

In an ultimate attempt to develop more adaptive traffic management and traveller information systems, the prediction of parking occupancy has received significant attention in recent years. As an integral part of a comprehensive parking guidance and information system, it is of interest to traffic engineers, given the growing need for the development of more adaptive traffic management and traveller information systems (Ermagun and Levinson, 2018). In order to make use of real-time parking related data, parking occupancy prediction is particularly useful for drivers. The predictive element is essential, as it provides occupancy information for the near future. For instance, drivers are notified whether there are any free parking spaces at their planned destination and time of arrival. In such a way, people can plan their trips ahead of time, allowing them to customize the destination, departure time and even mode of transportation. Once close to the destination, drivers are guided directly to a vacant parking space. In the last decade or so, numerous automated parking prediction mechanisms have been discussed in the literature, making use of information and communications technology and machine learning (ML) (see Chapter 2). It has the potential to greatly reduce uncertainty, since at the time of decision making, parking spots cannot be guaranteed to be free (Rajabioun et al., 2013).

In the expectation to make predictions more reliable, various external sources of information have been incorporated in parking occupancy prediction models. The underlying geographic context of the study area, however, has not been considered. Although recent literature suggests that spatial information may be relevant for parking prediction problems (Rajabioun and Ioannou, 2015; Bock, 2018; Richter et al., 2014), the incorporation of geography in prediction models has not received due attention. Particularly, the inclusion of spatial information such as land use, POIs and the spatial configuration of the street network poses promising potential to be integrated. Within this framework, the main objective of this thesis is to assess the contribution of spatial information to parking occupancy prediction. Specifically, the following four research questions will be addressed:

1. To what extent does spatial information help improve the performance of on-street parking occupancy prediction models?
2. How do results change under varying amounts of training data and temporal prediction horizons?
3. How do predictions vary spatially?
4. How does the predictive performance of two popular ML algorithms (RF and ANN) compare?

The remainder of this thesis is structured as follows: Chapter 2 presents the state of the art with regard to parking occupancy prediction and points out research gaps. The study area and data are described in chapter 3, followed by an illustration of the methodology in chapter 4. Subsequently, the main findings are presented in chapter 5 and discussed in chapter 6. At last, chapter 7 draws a conclusion and provides an outlook to future work.

Chapter 2

Related research

The following chapter provides an outline of the literature with regard to parking occupancy prediction. Firstly, there will be a broad overview of the existing literature, defining parking occupancy and the context in which it is applied. Secondly, the three main prediction approaches in the literature, namely stochastic, statistical and ML, are discussed along with their underlying models. Then, the input data is examined, with a focus on geographically related features. Finally, a brief summary of the chapter is provided and research gaps are identified. A comprehensive overview of relevant literature is shown in table 2.1 at the end of the chapter.

2.1 Parking occupancy prediction problem

In the literature, parking occupancy prediction has been defined as the estimation of occupancy for a specific parking facility at a given time in the future based on parking-related information (e.g. Zheng et al., 2015; Li et al., 2018; Bock, 2018). Initial studies focused mainly on the occupancy prediction of off-street parking facilities, such as garages and parking lots. This is due to the fact that they are ubiquitous in many cities, the problem is simpler and the data are more accessible (Monteiro and Ioannou, 2018). In recent years, however, focus has increasingly shifted to on-street parking prediction. It is more challenging, due to the absence of lot entrances and significant changes of the occupancy rate of a parking street segment following every parking/leaving event (Bock, 2018). The literature distinguishes between the derivation of occupancy rates for regression problems and occupancy classes for classification problems. The majority of literature considered the former, whereas it is pointed out that the latter are usually easier to interpret for users (Richter et al., 2014).

Most of the studies focused on short-term (less than 1 hour) and medium-term (less than 12 hours) prediction (e.g. Ji et al., 2015; Klappenecker et al., 2014; Li et al., 2018). For short-term predictions, the use of real-time data in combination with historical data has been proven to be efficient, not only to keep precision and integrity of the predicted data (Sun et al., 2018) but also to capture unusual deviations of parking demand (Hössinger et al., 2014). As the number of steps and overall length of the prediction time-span increases, the prediction error increases exponentially, losing the forecasting meaning altogether (Fang et al., 2018; Sun et al., 2018). Hence, a large number of papers rely on real-time parking information, while considering past occupancy data (e.g. Liu et al., 2018; Li et al., 2018; Rajabioun and Ioannou, 2015). When exploiting average daily trends for long-term predictions, the sole use of historical occupancy data can suffice to point out generalized trends over time (Peng and Li, 2016; Richter et al., 2014).

2.2 Prediction algorithms

With regard to parking occupancy prediction algorithms, three approaches can be identified in the literature (Xiao et al., 2018; Mei et al., 2019). Firstly, model-based approaches have been developed to estimate occupancy by considering the stochastic arrival and departure of a parking site. A parking lot model is established through queue theory and Markov chain. Secondly, parametric statistics-based approaches are implemented using mathematical methods such as exponential smoothing, availability mean and Autoregressive Integrated Moving Average (ARIMA). Thirdly, non-parametric ML approaches, including Artificial Neural Networks (ANNs) and regression, are developed using artificial intelligence methodologies. In the following sections, the above mentioned approaches and their implementations in the literature are presented.

2.2.1 Model-based

A widely-used prediction approach involves the establishment of an underlying model for the parking process, where model parameters are estimated to make parking occupancy predictions. Stochastic arrival and departure processes are usually explicitly employed for this approach. It is mostly applied to off-street parking facilities and is based on the presumption that vehicles arrive to parking spaces following a Poisson distribution. It is usually employed to describe counts of events that are assumed to occur randomly in time (Gart, 1975). A number of studies (Caliskan et al., 2007; Klappenecker et al., 2014; Peng and Li, 2016; Wu et al., 2014) made parking occupancy predictions using a continuous-time Markov Chain. It considers the fact that vehicles can park or leave the facility at any time. The arrival rate of vehicles at a parking facility can be modelled with a maximum likelihood approach, derived from occupancy data while the departure rate is given with an autoregressive Gaussian process. Mostly, relatively short prediction horizons are considered ranging from less than an hour to a few hours. Besides the model parameters, the current time is used as the main predictor. Caliskan et al. (2007) based their parking availability model on vehicular ad hoc networks, which are used to exchange information among vehicles. Similarly, Lu et al. (2009) relied on advanced technologies for the provision of arrival and departure rates directly. Caicedo et al. (2012) suggested to make parking availability predictions based on requests allocations. Xiao et al. (2018) proposed a prediction framework based on parameter estimation methods and a predictive framework that benefits from time-dependent analytical properties. To take into account real-world conditions, they proposed to handle inter- and intra-day variations of arrival and departure patterns, as well as special events.

2.2.2 Parametric statistical

A second approach applies statistical methods to derive future availability from observed data directly. A model based on the naïve assumption that the occupancy rate stays at a constant average level was developed by Hössinger et al. (2014), applied on parking availability data in Vienna, Austria. Due to the strongly recurrent pattern of parking availability, they additionally developed a slightly more sophisticated model, the average day curve model. It is based on the assumption that occupancy follows an average daily curve. They found that real-time data, however, are indispensable to predict unusual deviations. Such deviations can occur due to events or exceptional circumstances. Similarly, Monteiro and Ioannou (2018) implemented parking occupancy prediction using availability mean and variation and normally distributed availability methods. Furthermore, with the addition of real-time data, they modelled the availability variation as a normal random variable. Only using historical parking occupancy data, Richter et al. (2014) focused on a prediction approach exploiting general recurring daily patterns by averaging all values within a timeslot of the day. To refine the model, the distinction between weekday and weekend, as well as each day of the week (DOW) was made, significantly improving the prediction performance.

Statistical time series methods, in contrast to naïve statistical aggregate functions, tend to be more powerful and have been a popular approach for making predictions in transportation problems (Karlaftis and Vlahogianni, 2011). In doing so, the evolution of a system is considered, with historical observations indexed by time. This approach is especially useful when little knowledge is available on the data or when the prediction variable cannot be related to explanatory variables (Zhang, 2003). Examples of methods include autoregression, moving average, or the combination realized as ARIMA. Rajabioun and Ioannou (2015) implemented a multivariate autoregression model that can predict parking availability for both on-street and off-street parking. Liu et al. (2018) similarly proposed an autoregression model with real-time sensing data, motivated by the fact that there is a linear association between lagged observations and current observations.

Implementations using ARIMA for parking occupancy prediction have been abundant in the literature. In time series analysis and forecasting, ARIMA is a well-known and widely used model, not least because of its statistical properties and its flexibility of representing different types of time series. Moreover, it is relatively easy to produce and straightforward (Vlahogianni et al., 2016). Dias et al. (2015) suggested an ARIMA prediction approach for occupancy status of public bicycle stations in Barcelona, Spain. Status are classified into 5 classes from full to empty.

In similar fashion, Badii et al. (2018) found that an ARIMA approach can make satisfactory predictions, under the condition that the training was recomputed every hour. This, however, is a considerable cost, especially if there is a large number of car park facilities. Yu et al. (2015) established an ARIMA model to forecast the remaining spaces of a central mall parking lot in real-time by constantly updating the data. Time series analysis, model parameter estimation and model adaptive testing was carried out to establish the model. Chen (2014) investigated the relationship between prediction error and aggregation level. Taking into account the geo-location, the aggregation smoothed the pattern, making it more predictable. Among others, he implemented an ARIMA model, demonstrating that the prediction error is negatively correlated with increased aggregation.

2.2.3 Non-parametric machine learning

Several ML algorithms such as ANNs, RF, Support Vector Regression (SVR) and K-means have been implemented to predict parking occupancy. In the following section, the application of each technique in the literature is reviewed.

Amongst various ML techniques, particularly ANNs have been implemented in a large number of scientific papers as a means to establish parking occupancy prediction models. Not only have they proven to deliver satisfactory prediction results for the most part, but also can they be considered state of the art nowadays (Xiao et al., 2018). While being computationally more expensive than statistical models, ANNs have been shown to relax constraints such as stationarity and linearity. Moreover they seem to be more adaptable to sudden shifts in the data in short-term forecasting models (Vlahogianni et al., 2016).

Feedforward Neural Networks (FFNNs), the simplest type of ANN models, have been advocated to make parking availability predictions. Yu et al. (2015) and Pengzi et al. (2017) used this type of model to make short-term predictions with time and recent parking occupancy observations for parking occupancy in Nanjing, China and Xi'an, China, respectively. In a similar manner, Zheng et al. (2015) and Chen (2014) made parking occupany predictions in San Francisco, with longer prediction horizons. Badii et al. (2018) made use of a FFNN to predict the occupancy rate for parking spaces in Florence, Italy. In doing so, they implemented external factors such as weather and traffic as data input. A wavelet neural network, a combination of wavelet transform and ANN, has been proposed by Ji et al. (2015) and Fang et al. (2018) as an alternative to a conventional ANN.

Recurrent Neural Networks (RNNs), more complex ANNs with loops, have been proposed in parking occupancy prediction schemes due to their strength in solving problems that are sequential and time-varying (Qolomany et al., 2017). Camero et al. (2019) implemented a short-term RNN for the occupancy prediction of several car parks in Birmingham, UK. Similarly, Vlahogianni et al. (2016) suggested a real-time time series occupancy scheme based on RNNs in Santander, Spain. Further, the usage of a long short term memory network, an extension of a traditional RNN, was proposed in the literature. Li et al. (2018), Shao et al. (2019) and Sun et al. (2018) argued that it is especially suitable for the problem of parking occupancy prediction due to the fact that it is able to overcome the problem of long-term dependencies.

Another ML technique that has been implemented for parking occupancy prediction is RF, an ensemble learning method consisting of a multitude of decision trees. Emphasising their robustness and competitiveness, Bock (2018) an RF model to predict occupancy status of parking segments from crowdsensed data in San Francisco, USA. Dias et al. (2015), conversely, used an RF model to make long-term occupancy predictions for a public bicycle sharing programme in Barcelona, Spain.

In addition, regression models have been deployed in the literature. Both a linear regression and an SVR model were implemented by Leu and Zhu (2015) to predict the number of available parking spaces for bicycle stations in Taipei, Taiwan. Additionally, they considered weather information and occupancy status of nearby stations. An SVR approach was also carried out by Zheng et al. (2015), using various combinations of feature inputs. Similarly, Badii et al. (2018) and Chen (2014) used SVR models for their prediction schemes on parking garages in the area of Florence, Italy and on-street parking facilities in San Francisco, USA, respectively.

A prediction strategy based on K-means clustering was implemented by Stolfi et al. (2017, 2019). It created groups of car parks and weekdays in different clusters. Ideally, sets of car parks that behave similarly can be described.

2.3 Input data

2.3.1 Temporal information

Due to the fact that the parking utilization rate follows recurrent within-day and day-to-day patterns (Chen, 2014; Xiao et al., 2018), temporal information such as the time of the day (TOD), the DOW and holidays have been implemented in many models to predict parking occupancy. The TOD is a very relevant factor to consider (e.g. Dias et al., 2015; Zheng et al., 2015; Richter et al., 2014). Either it is implemented in a time series (e.g. Vlahogianni et al., 2016; Liu et al., 2018) or as a feature in ML methods (e.g. Pflügler et al., 2016; Zheng et al., 2015; Badii et al., 2018). Furthermore, especially long-term predictions benefit from the distinction between DOWs (e.g. Richter et al., 2014; Vlahogianni et al., 2016; Rajabioun and Ioannou, 2015). The driver's parking behaviour also tends to be different on holidays. Li et al. (2018) incorporated the contrast between holidays and regular days as a feature in their model. Similarly, Wang et al. (2007) pointed out that days such as Labour Day, national day and Spring Festival influence parking demand.

2.3.2 Recent observations of parking occupancy

Recent parking occupancy information may be the most significant data input for future parking occupancy prediction. This stems from the fact that there is a strong temporal correlation for parking utilization (Rajabioun and Ioannou, 2015; Liu et al., 2018). While statistical time-series methods such as ARIMA explicitly rely on the parking occupancy status of previous time steps, the previous observations have been implemented as features in ML models in the literature. Bock (2018) found that the consideration of 5 previous observations is optimal for his model while others examined and implemented the observation information for 1 previous time step (Badii et al., 2018; Liu et al., 2018). Zheng et al. (2015) also experimented with the number of preceding time steps to include in the model.

2.3.3 Location

To a certain degree, spatial information has been taken into account for parking availability prediction models. Occupancy observations in adjacent roads were considered by Bock (2018). They were determined by predefined radii in beeline distance ranging from 100 to 800 metres and time intervals ranging from 0 to 60 minutes before the time at which occupancy was to be predicted. For the combination of these parameters, he computed the number and rate of empty parking spaces. Giuffrè et al. (2012) pointed out that the chance of finding a vacant parking spot generally varies depending on the location. Leu and Zhu (2015), on the other hand, used regional data (i.e. the occupancy status of the target station's neighbouring stations) as an important feature for their public bicycle occupancy prediction model. Similarly, Rajabioun and Ioannou (2015) pointed out that there is a correlation of parking usage between car parks that are at different distances from each other. As a result of the spatial correlation, they took into account the parking situation in neighbouring areas and incorporated the dependencies in their model. In contrast, spatial clustering was implemented by Richter et al. (2014). They examined the accuracy of predicting the parking occupancy within spatial regions and hypothesised that this should highlight spatial influences on the parking behaviour. The San Francisco on-street parking data was split into regions with three types of characteristics; commercial, touristic and residential. Implicitly, land use (more specifically functional zoning) was used as a predictor. Likewise, Chen (2014) aggregated parking spots based on their location, under the assumption that short walking distances are affordable for drivers. Different shape patterns were identified for the regions of the city, possibly representing different groups of travellers with distinct travel patterns.

2.3.4 Weather

Weather conditions (e.g. temperature, rainfall, humidity) are thought to influence the parking availability situation. Yang et al. (2003) and Greengard (2015) argued that weather information is of central importance, affecting the traffic behaviour and traffic flow intensity. Badii et al. (2018) showed that weather conditions 1 hour before the parking time have a significant impact on the parking behaviour. Similarly, Dias et al. (2015) found that the relative humidity plays an important role in making occupancy predictions for a public bicycle sharing system in Barcelona, Spain. Leu and Zhu (2015) also found that extreme weather conditions such as intense heat, rain and strong winds have an effect on bicycle rentals.

2.3.5 Events

Another factor that may influence parking availability are events like concerts or sport matches (Chen, 2014; Yang et al., 2003; Kimms et al., 2012). They can temporarily cause major increases in the volume of traffic, dramatically affecting the availability of parking spaces.

2.3.6 Traffic

Since traffic and parking are closely connected, it has been argued that the inclusion of traffic information is advantageous to predict parking occupancy. Especially traffic volume is an important factor, as high traffic volume makes it more difficult to find a vacant parking space (Shin and Jun, 2014; Yang et al., 2003; Hössinger et al., 2014). Moreover, Badii et al. (2018) suggested that apart from traffic volume, vehicle flow, concentration and average speed have high predictive relevancies.

2.4 Summary and research gaps

In the research literature, mainly three approaches have been implemented for parking occupancy prediction. These include model-based, parametric statistical and non-parametric ML. Predominantly for off-street parking facilities, the usage of model-based approaches has been popular. By doing so, stochastic arrival and departure processes are employed. Statistical methods, conversely, are based on aggregate functions and statistical time-series methods. They mainly take advantage of intra- and inter-daily parking occupancy patterns. Lastly, ML models have been found to be suitable to make parking occupancy predictions. Popular algorithms include ANN, RF, SVR and K-means.

A significant amount of the studies considered time and occupancy status of recent observations as their main data input in their model. Moreover, holidays, weather, traffic and events are additionally incorporated in an attempt to improve the model performance. Finally, a spatial component is included in some of the studies.

Given the large amount of research literature on parking availability prediction, however, the underlying spatial context of the study area has received little attention. The few studies that considered a spatial component in their prediction model only focused on occupancy status of adjacent roads, based on the assumption that there is a spatial correlation. The vast majority of the existing work, conversely, incorporated a data-driven approach. Implicitly, these models may assume that a relatively large amount of occupancy data takes into account spatio-temporal relationships between parking occupancy and the spatial environment. The explicit inclusion of information such as land use, POIs and the spatial configuration of the street network has, to the best of my knowledge, not been realized. Hence, its implementation in parking occupancy prediction models poses a great potential that has not yet been addressed.

TABLE 2.1: Summary of literature.

		Method					
Study	**Location**	**MB**	**ST**	**ML**	**Data input**	**Prediction horizon**	**Step (min)**
Camero et al. (2019)	Birmingham, UK			✓	T,D,R	medium	30
Mei et al. (2019)	Hangzhou, China			✓	T,D,R	long	5,15,30
Shao et al. (2019)	Melbourne, Australia			✓	T,R	short	1
Stolfi et al. (2019)	UK		✓	✓	T,D,R	medium	15-30
Badii et al. (2018)	Florence, Italy		✓	✓	T,D,R,TR,W	long	15
Bock (2018)	San Francisco, USA			✓	T,D,R,L	short	5
Fan et al. (2018)	China			✓	T,D,R	medium	10
Fang et al. (2018)	China	✓		✓	T,R	long	15
Li et al. (2018)	Beijing, China			✓	T,D,H,R,E,W	medium	3
Liu et al. (2018)	Melbourne, Australia		✓		T,D,R	medium	5
Monteiro and Ioannou (2018)	Los Angeles, USA	✓	✓		T,D,TR	medium	5
Sun et al. (2018)	Shenzhen, China			✓	T,R	short	2
Xiao et al. (2018)	San Francisco, USA	✓			T,D,E	short	2
Pengzi et al. (2017)	Xi'an, China			✓	T,R	short	1
Stolfi et al. (2017)	Birmingham, UK		✓	✓	T,D,R	medium	30
Peng and Li (2016)	Shenzhen, China	✓			T,D	medium	20
Pflügler et al. (2016)	Munich, Germany			✓	T,D,H,E,TR,W	short	
Vlahogianni et al. (2016)	Santander, Spain			✓	T,D,R	short	1
Dias et al. (2015)	Barcelona, Spain		✓	✓	T,D,H,W	long	15
Ji et al. (2015)	Newcastle, UK			✓	T,D,R	short	5
Leu and Zhu (2015)	Taipei, Taiwan			✓	T,L,W	short	1
Rajabioun and Ioannou (2015)	San Francisco, USA		✓		T,D,R,L,TR	short	1
Tiedemann et al. (2015)	Berlin, Germany			✓	T,D,H,R	medium	15
Yu et al. (2015)	Nanjing, China		✓	✓	T,R	short	15
Zheng et al. (2015)	San Francisco, USA			✓	T,D,R	medium	15
Chen (2014)	San Francisco, USA		✓	✓	T,D,R,E	medium	60
Hössinger et al. (2014)	Vienna, Austria		✓		T,D,H,L	medium	30
Klappenecker et al. (2014)		✓			T	short	1
Richter et al. (2014)	San Francisco, USA		✓		T,D,L	long	5
Wu et al. (2014)	Taiwan	✓			T,D,R	medium	5
Rajabioun et al. (2013)	San Francisco, USA	✓			T,R	medium	5
Caliskan et al. (2007)	Brunswick, Germany	✓			T	medium	1
Wang et al. (2007)	Beijing, China		✓		T,D,H	long	

Note: MB: Model-based; ST: Statistical; ML: Machine Learning; T: TOD; D: DOW; H: Holidays; R: Recent occupancy observations; L: Location; E: Events; TR: Traffic; W: Weather; short: < 1h; 1h $\leq$ medium $\leq$ 12h; long: $>$12h.

Chapter 3

Study area and data

3.1 Study area

The overall study area for this work is the city of San Francisco. It is located on the west coast of the USA at approximately 37°47′N, 122°25′W and encompasses an area of some 120 km^2. The parking segments of interest are arranged in 9 clusters, mainly in the north-eastern part of the city.

3.2 Data

3.2.1 Parking occupancy data

The historical parking occupancy data is provided by the San Francisco Municipal Transportation Agency as part of SF*park*, a large-scale smart parking project. The overall goal was to increase parking efficiency and drivers' experience as well as to evaluate demand-responsive pricing (San Francisco Municipal Transportation Agency, 2014). In order to monitor parking occupancy, on-street parking spaces were equipped with sensors and recorded occupancy continuously from April 2011 to July 2013. In total, more than 8000 spaces were selected and sensors were installed in three phases starting from the formal launch in 2011. Sensors were distributed across 10 parking management districts in the city and cover more than 400 street segments in business districts, residential districts and a touristic area. Earlier than expected, some sensor batteries started to fail in 2012, rendering the data record incomplete for some of the locations.

Each parking space exhibits at least 1 sensor, an in-ground and self-powered wireless device that detects the presence of a vehicle primarily with a magnetometer. The signals are then transmitted via pole-mounted repeaters and gateways to the data warehouse. Occupancy status are derived from parking events, sent by the sensors. Namely, they register the start and the end of a parking session (i.e. when a vehicle arrives in or departs from the parking space) and the operational status (i.e. whether the sensor is functional). Once a new event is received, the status of the space changes. Occupancy rates, thereafter, are calculated on the basis of occupancy status durations and aggregated to whole hour increments. Each data record contains a timestamp, ID and name of the parking segment, the number of occupied parking spaces and the current parking capacity. An exemplary occupancy time series over three days is displayed in figure 3.2.

For this thesis, data from the initial phase of the project from April 2011 to June 2011 were considered. Totally, this data covers hourly occupancy status from 312 on-street parking segments distributed across 9 parking districts. A map of the distribution of the parking districts and their corresponding parking segments is shown in figure 3.1.

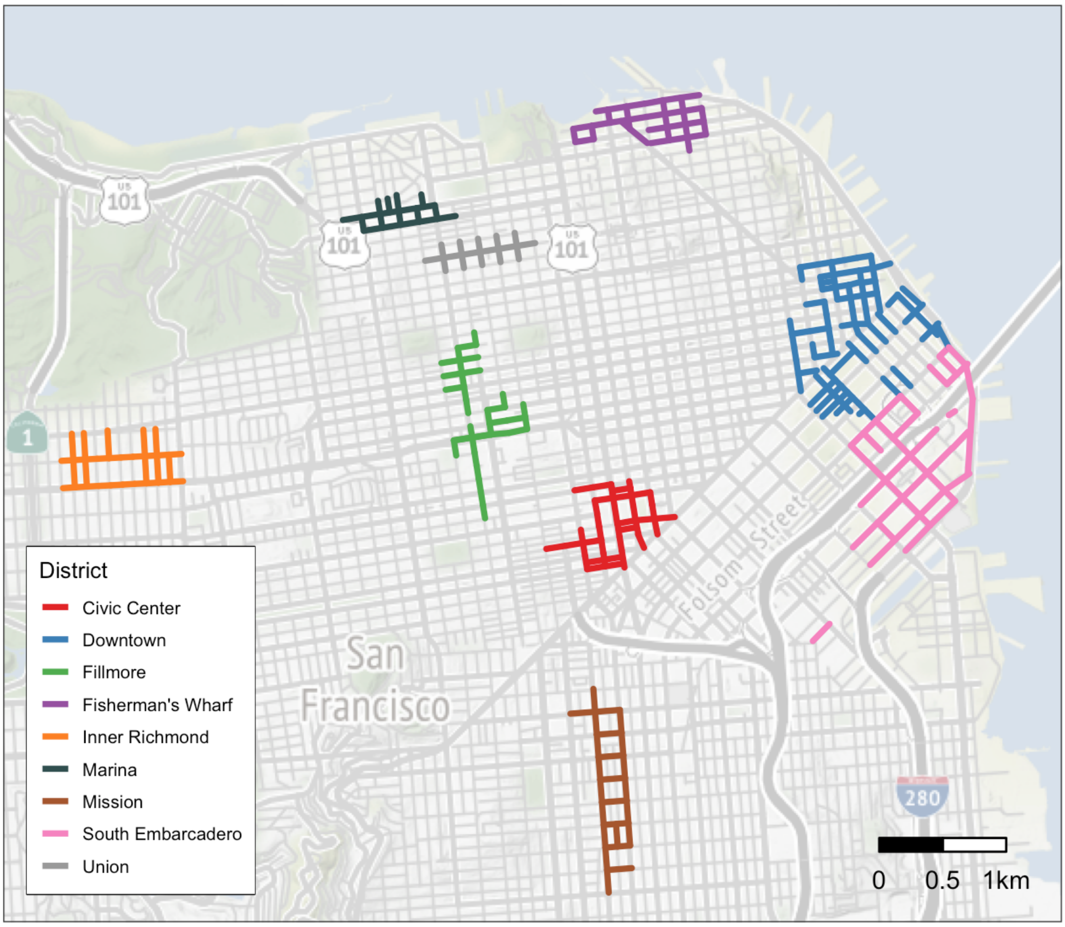


FIGURE 3.1: Parking district locations in San Francisco.

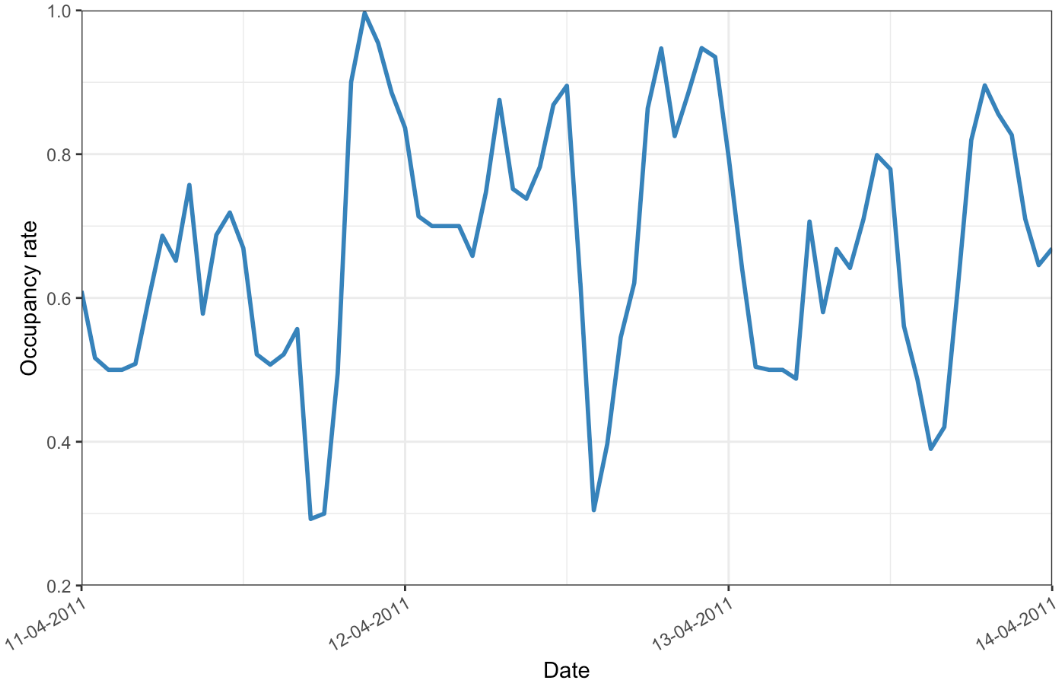


FIGURE 3.2: Example for the occupancy time series of a parking segment over three days.

3.2.2 Spatial data

Primarily OpenStreetMap (OSM) data from early 2019 (OpenStreetMap, 2019a) was utilized to map the street network in this thesis. It consists of a very detailed street and pedestrian network. Road class keys as shown in table 3.1 were considered.

TABLE 3.1: OpenStreetMap road class keys, in descending order.

Value	**Comment**	**Type**
Motorway	Restricted access major divided highway	Roads
Trunk	Most important roads that are not motorways	
Primary	Main roads in cities	
Secondary	Medium-large roads in cities	
Tertiary	Small roads in cities	
Unclassified	Least important through roads	
Residential	Serve as an access to housing	
Service	Access roads to or within private area	
Footway	Footpaths for pedestrians	Paths
Steps	For flights of steps (stairs) on footways	

Additionally, street segment data was used, provided by the City and County of San Francisco (City and County of San Francisco, 2014b). It includes information on the street centreline for each street segment (i.e. the street section between two intersections) including street name, address numbers and coordinates.

The land use data is a mixture of the San Francisco Assessor's table and a commercial business dataset and is freely available, provided by the City and County of San Francisco (City and County of San Francisco, 2016). The data includes land use categories for every parcel. For this thesis, 3 categories were considered: Industrial (production, distribution and repair), office (management, information, professional services) and residential. Additionally, information regarding the number of residential units and the square footage of each parcel is available. Land use categories are assigned if the square footage of any non-residential use is 80% or more of its total uses, otherwise it becomes *mixed*.

POIs were selected and three categories were defined. Business comprises the locations of all registered businesses which are provided by the City and County of San Francisco (City and County of San Francisco, 2014a). Public transport includes the locations of all train stations and stops within the city, derived from early 2019 OSM data (OpenStreetMap, 2019b). Touristic comprises the 20 most popular tourist attractions, according to Tripadvisor (as of 01/07/2019) (TripAdvisor, 2019). A description of each POI class is shown in table 3.2. Figure 3.3 visualizes the locations public transport and touristic POIs.

TABLE 3.2: Overview of POI classes.

Class	**Description**	**Number of points**
Business	Registered businesses	60216
Public transport	Railway stations and stops	71
Touristic	Top 20 tourist destinations	20

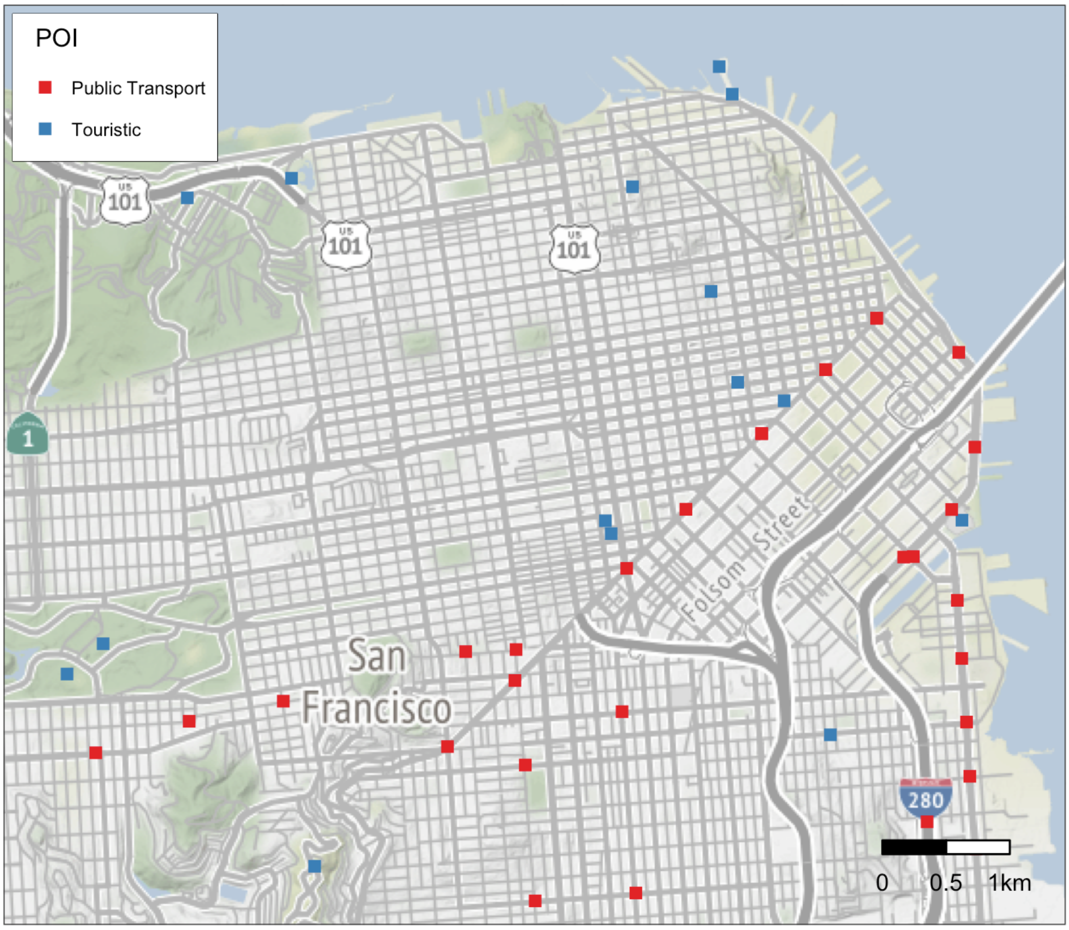


FIGURE 3.3: Public transport and touristic POI locations in San Francisco.

Chapter 4

Methodology

4.1 Overview

An overview of the workflow is shown in figure 4.1. In a first step, the street data and the parking occupancy data had to be preprocessed. The spatial data was then quantified according to the methods specified in section 4.4, based on which the geographic features were derived. Moreover temporal and historical occupancy features were defined. Given the input data, the prediction models were trained and validated which finally resulted in the occupancy prediction.

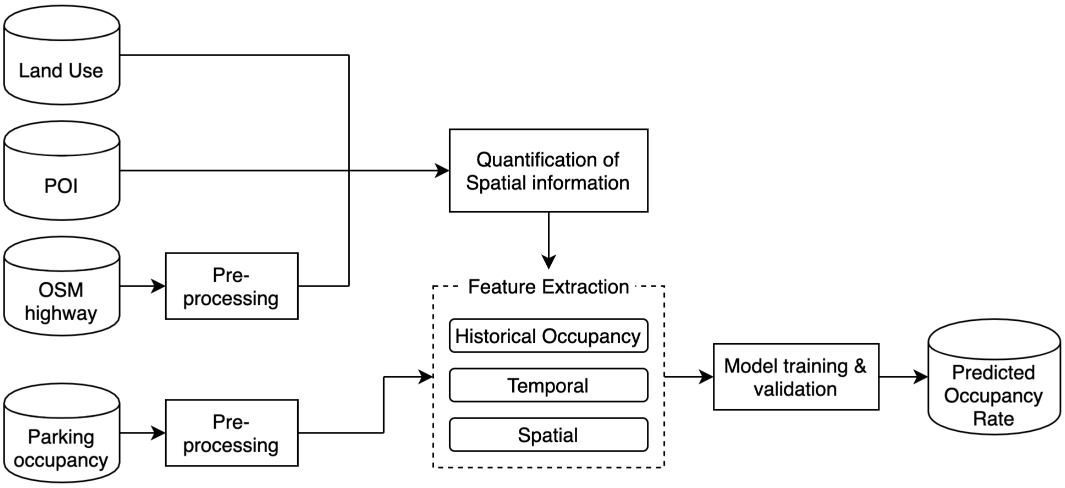


FIGURE 4.1: Overview of workflow.

4.2 Tools

All methods in this thesis were implemented in the software environment R 3.6.1 (R Core Team, 2019) and the spatial information system ArcGIS 10.6.1 (Esri, 2019). ArcGIS was used for spatial analysis, whereas R was used for data-preprocessing, spatial analysis, prediction model implementation and data visualization. The most relevant R packages for this thesis are listed in table 4.1.

TABLE 4.1: Most relevant R packages.

Package	Author(s)	Application
caret	Kuhn et al. (2019)	RF model tuning
dplyr	Wickham et al. (2019)	Data pre-precessing
ggmap	Kahle et al. (2019)	Spatial data visualization
ggplot2	Wickham et al. (2019)	Data visualization
igraph	Csardi (2019)	Centrality calculation
keras	Falbel et al. (2019)	ANN model implementation
randomForest	Liaw et al. (2018)	RF model implementation
raster	Hijmans et al. (2019)	Spatial data analysis
tensorflow	Falbel et al. (2019)	ANN model implementation

4.3 Data preprocessing

In a first step, the parking occupancy data had to be prepared. A total of 11 weeks of data records was selected, from 11.04.2011 - 26.06.2011, starting just after the launch of the SF*park* project. In that initial phase, sensors of all parking segments were fully functional with the exception of the parking district West Portal. Consequently, records from that district were excluded. In a next step, occupancy rates of all records was derived as follows:

$$Occupancy\ rate = \frac{Total\ occupied\ seconds}{Total\ vacant\ seconds + Total\ occupied\ seconds} \tag{4.1}$$

Further, all parking segment locations had to be georeferenced, from the street block numbers provided in the data. By doing so, coordinates from the street segment dataset were joined with the parking segment locations. Additionally, some missing parking segment coordinates were manually added.

Moreover, OSM street data had to be converted into network datasets. Two types of networks were considered: A road network including all roads, as well as a pedestrian network including all paths and roads with the exception of Motorway and Trunk (see section 3.2.2). Impedence was defined as the distance in metres. Whereas the road network has restrictions regarding turns and driving directions, no restrictions were defined for the pedestrian network.

4.4 Quantification of spatial information

In order to include spatial information as data input of the prediction models, it needed to be quantified in a meaningful way. By doing so, geographic predictors were created by assigning the values to each of the parking segment's location. In the following, the 3 main concepts for the derivation of the geographic features in this thesis are described.

4.4.1 Centrality

Centrality is a fundamental concept in network analysis and has been used in various fields to investigate territorial relationships (Wilson, 2000). Especially in urban areas, centrality has been studied by transforming the street network into a relational graph, representing urban street patterns as spatial networks. By doing so, streets are mapped onto graph nodes and intersections onto the edges between the nodes (Crucitti et al., 2006). In order to quantify the positional importance of an edge or a node, several centrality indices have been introduced. High centrality scores indicate a node's high importance within the network.

There is a multitude of centrality indices, some of which are conceptually similar. For this work, 3 indices with a low degree of similarity and correlation of values were computed for the entire street network of San Francisco. Those include betweenness, closeness and alpha centrality (see figure 4.2). For each edge representing a street segment, the mean values of the confining vertices were calculated. Betweenness centrality measures the number of times a node lies on the shortest path between two other nodes. It was first introduced by Freeman (1977) in the context of human communication in social networks. Closeness centrality (Bavelas, 1950) quantifies the mean length of shortest paths between a node and all other nodes in a network. The higher the score of a node, the closer it is to all other nodes. Accordingly, a high score indicates that a node is connected to other nodes exhibiting high scores. Finally, alpha centrality (Bonacich and Lloyd, 2001) is a variation of eigenvector centrality which is based on the concept that a node within a network is more important if it is linked to adjacent important nodes. Below are mathematical expressions for betweenness, closeness and alpha centrality in that order:

$$B(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}} \qquad C(v) = \frac{1}{\sum_w d(w,v)} \qquad x = (I - \alpha A)^{-1} e \tag{4.2}$$

Where:

σ_{st}: Number of shortest paths from node s to node t

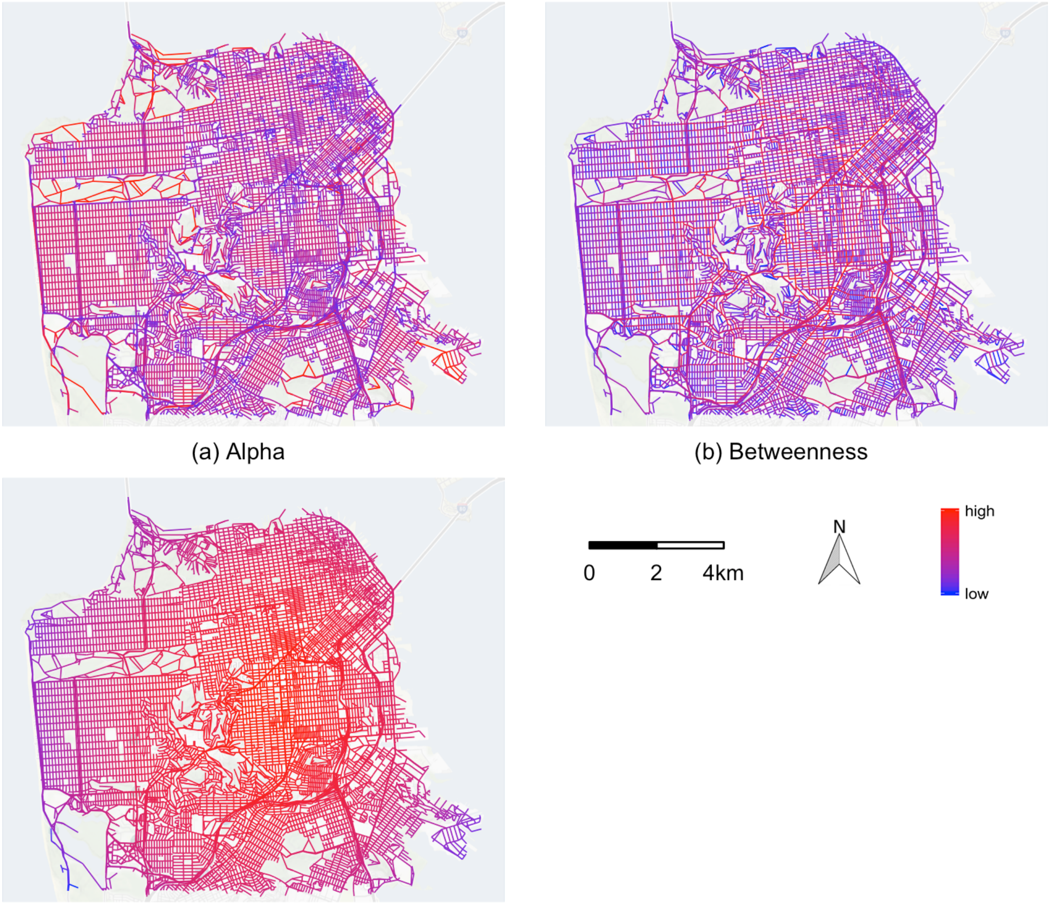
$\sigma_{st}(v)$: Number of shortest paths from node s to node t that pass through v

$d(w,v)$: Distance between vertices v and w

α: relative importance of endogenous versus exogenous factors in the determination of centrality

A: Adjacency matrix

e: effects of external status characteristics



(a) Alpha

(b) Betweenness

(c) Closeness

FIGURE 4.2: Street centrality San Francisco. Indicated as (a) Alpha, (b) Betweenness and (c) Closeness centrality.

4.4.2 Service area

Service areas are used to evaluate the accessibility of a facility, i.e. the service that is supplied along a traffic network (Talen and Anselin, 1998). In the context of this thesis, the associated buffering radius is drawn along the pedestrian network. The resulting service area is assumed to represent an area that covers a short walking distance from the parking space. This is based on the presumption that the driver does not necessarily find a parking space immediately at his end destination and is therefore prepared to walk for a short distance. Van der Waerden et al. (2015) found that the maximum distance drivers are willing to walk to their destination is mainly dependent on the trip purpose and the trip-related characteristics *visit duration* and *frequency of use*. It ranges from less than 50 metres to more than 500 metres. To cover the entirety of trip purposes in this work, the service area radius was set to 500 metres. Specifically, for each parking segment, a service area was created within which the respective land use category was quantified.

The land use features office, residential and industrial as well as business locations were quantified within the service areas of the parking segments. The first land use feature was measured according to the number of residential units whereas the second and third feature were assessed by the sum of their square footage. In doing so, land use parcels were considered that lie partially or completely within the service area. Similarly, the number of businesses within the service area was counted. Figure 4.3 below illustrates a street segment with its service area containing the parcels and their respective land use.

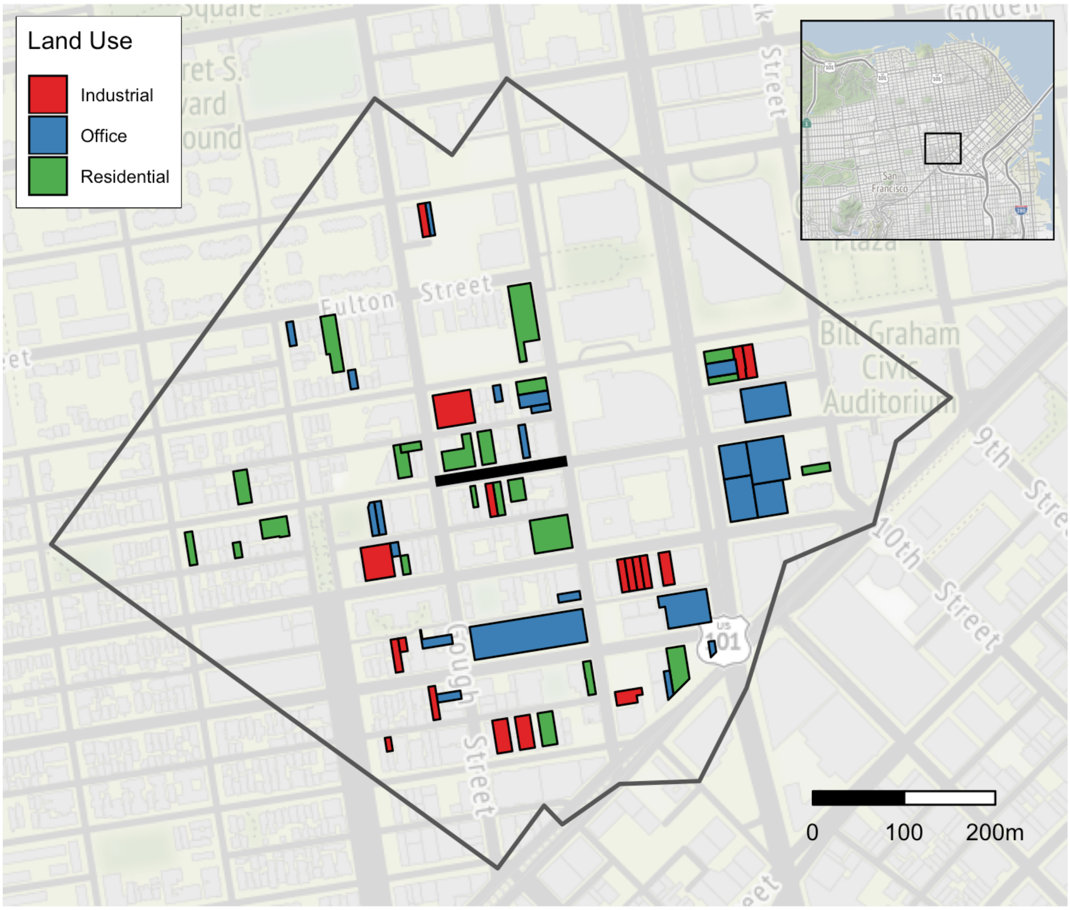


FIGURE 4.3: Exemplary parking segment and land use parcels within its service area of 500 metres. Service area boundaries are signified in dark grey.

4.4.3 Shortest path

As for the quantification of the touristic and public transport POIs in relation to the parking segments, shortest paths were computed. First introduced by Dijkstra (1959), the shortest path problem describes the problem of finding a path between two vertices in a network in such a way that the sum of the weights (or distance) of its intermediate edges is minimized. In the context of this thesis, single-source shortest paths were computed from the centre points of each parking segment to all POIs along the pedestrian network. As a result the mean distance (i.e. the average distance of all shortest paths from each parking segment to all points) was computed.

4.5 Geographic feature selection

Making use of the methods described above, nine geographic features were selected in total. Three features were allocated to each of the categories centrality, POI and land use. Table 4.2 shows all geographic features and their derivation.

TABLE 4.2: Overview of geographic features with their derivation.

Category	Feature	Description	Derivation
Centrality	Alpha	Parking segment alpha centrality	Centrality calculation
	Betweenness	Parking segment betweenness centrality	Centrality calculation
	Closeness	Parking segment closeness centrality	Centrality calculation
POI	Business	Businesses count	Service area
	Public transport	Mean distance to railway stations	Shortest paths
	Touristic	Mean distance to tourist attractions	Shortest paths
Land use	Industrial	Industrial floor area	Service area
	Office	Office floor area	Service area
	Residential	Number of residential units	Service area

4.6 Prediction approach

The aim of the prediction framework is to forecast the occupancy rate [0-1] of the parking segments for a specific time in the future. A prediction horizon ranging from 1 step to 10 steps is considered. One time step corresponds to 1 hour. Hence, occupancy values from 1 hour to 10 hours ahead were predicted. This was realized by using historical occupancy rates as data input. Accordingly, to make a 1 step ahead prediction, the occupancy rate 1 hour prior to the time at which occupancy is to be predicted was considered. Moreover, geographic features and the time were considered in the data input. X is defined as the data input or a feature vector and y corresponds to the prediction output. Generally, the prediction problem can therefore be described as follows:

$$X = \{t, O(t-k), GF_1, \cdots, GF_i\}, \quad y = O(t) \tag{4.3}$$

Where:

t: time

$O(t)$: Occupancy at time t

k: Number of steps ahead to be predicted

GF_i: Geographic feature i

In order to assess the effect of different predictors on the prediction models, 8 feature sets were defined, differing by their spatial information input. Feature set 1 solely includes temporal and historical occupancy information. It can be considered a benchmark, serving as a standard against which performances are compared. Feature sets 2 – 4 additionally include spatial information from the categories centrality, POI and land use, respectively. Feature sets 5 – 7 incorporate the unique combinations of the three categories, whereas feature set 8 combines all geographic features. Table 4.3 illustrates each feature set with their respective input categories.

TABLE 4.3: Overview of feature sets.

	Input categories				
Feature set	**Historical occupancy**	**Time**	**Centrality**	**POI**	**Land use**
1 (baseline)	✓	✓			
2	✓	✓	✓		
3	✓	✓		✓	
4	✓	✓			✓
5	✓	✓	✓	✓	
6	✓	✓		✓	✓
7	✓	✓	✓		✓
8	✓	✓	✓	✓	✓

4.7 Prediction models and parametrization

The prediction models were implemented with ML tools. A subset of artificial intelligence, they provide computer systems the ability to perform tasks without being explicitly programmed. Instead, a mathematical model is built based on sample data, so-called training data, to make predictions or decisions. Thereby, the primary goal is to allow the computer to learn automatically without human intervention (Bishop, 2006). In the case of supervised ML, the algorithm can use labelled examples to predict future events by applying what has been learned in the past to new data. Specifically, it learns a function that assigns an input object to an output value based on input-output pairs. In this thesis, two popular supervised ML algorithms were implemented to make parking occupancy predictions, namely RF and ANN. In the following, each algorithm is described.

4.7.1 Random forest

RF (Breiman, 2001) is an ensemble learning method that uses a collection of decision trees for regression and classification tasks. Explicitly, a randomly selected subset of predictors from the training data is used to build the trees. The algorithm subsequently outputs the mean prediction of the individual trees. RFs manage to correct the decision trees' tendency to overfit to their training set (Hastie et al., 2009) and have been used in real-world applications in various fields (Oshiro and Perez, 2012). Among supervised learning algorithms, RFs have provided competitive and robust results (Caruana and Niculescu-Mizil, 2006). Furthermore, compared to other popular algorithms, parameter tuning requires relatively low effort in order to achieve solid results (Hastie et al., 2009). Literature suggests that a higher number of trees is associated with an improvement in performance. Nevertheless, there is a threshold beyond which there is no significant gain in performance. Hence, the number of trees should be set in accordance with the computational environment (Oshiro and Perez, 2012). As a trade-off between required processing power and model performance, the number of trees was set to 200 for this thesis. Additionally, for each feature and training data set, two hyperparameters were tuned: the number of variables randomly selected at each split and the maximum number of terminal nodes each tree can have. Depending on the data input combination, the former was set to 2-5 whereas the latter was set to 100-1000. To determine optimal values, a grid search with cross validation was implemented. Other hyperparameters were set to their default values as stated in R's library randomForest (see `https://www.rdocumentation.org/packages/randomForest/versions/4.6-14/topics/randomForest`).

4.7.2 Artificial neural network

ANNs are connectionist systems, inspired by neural networks in the human brain. They have been implemented in various fields, for instance for engineering and technical problems. ANNs consist of neuron-like processing nodes, organized in layers. There are three types of layers, namely input, hidden and output. Every node is connected to all nodes in the previous layer. Notably, the number of nodes in the input layer corresponds to the number of input features, whereas the output layer consists of merely one node (for regression problems). In case of a FFNN, the information moves in one direction only, from the input layer through the hidden layers and finally to the output layer. ANNs with two or more hidden layers are commonly referred to as deep neural networks and are categorized as deep learning (Lecun et al., 2015). The data points, or inputs, are fed into the neurons in the input layer. Multiplying the input number with the neuron's weight results in the output of the neuron which is then transferred to the next layer. The conversion of the input signal of a node to an output signal is realized by the so-called activation function. Commonly, ANNs are trained with a backpropagation algorithm. It aims to minimize the error by altering the weights. This is effected by backpropagating information about the error in reverse through the network (Werbos, 1990). For this work, a FFNN with two hidden layers was built. Each of the hidden layers consists of 128 nodes which are activated by rectified linear unit activation function. A mini-batch gradient descent approach was implemented by setting the batch size to 32. The number of epochs was set to 100. The batch size defines the number of samples that are passed through the network before weights are updated whereas the number of epochs signifies the number of times the algorithm works through the entire training dataset. The architecture of the network is illustrated in figure 4.4.

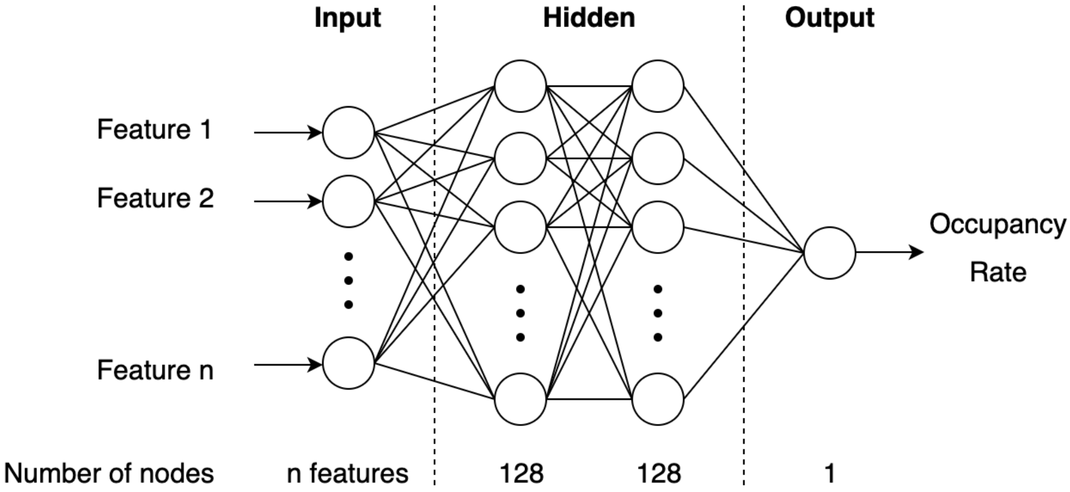


FIGURE 4.4: Artificial neural network architecture.

4.8 Performance assessment

To evaluate the performance of the prediction models in this thesis, 3 commonly used metrics were deployed. The Mean Absolute Error (MAE) measures the mean absolute differences between predicted and observed values. By contrast, the Mean Squared Error (MSE) indicates the standard deviation of the differences between predicted and observed values. Whereas the MAE gives equal weight to all errors, the MSE penalizes variance, weighing error with larger absolute values more than errors with smaller absolute values (Chai and Draxler, 2014). Finally, the coefficient of determination (R^2) indicates the level of explained variability in the dataset. Below, the mathematical expressions for the three metrics are given.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \qquad MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{4.4}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{4.5}$$

Where:

n:	number of data points
y_i:	observed values
$\hat{y}_i$:	predicted values
$\bar{y}_i$:	mean values

4.9 Experimental design

In order to address the research questions, experiments were performed on all 8 feature sets for the RF model and on feature sets 1 – 4 and 8 for the ANN model. To analyse different input scenarios, different train and test splits were applied. For each feature set, the amount of of training data was set to range from 1 day up to

10 days, in 1 day increments. Furthermore, to take into account weekday differences, the datasets were trained on the same weekdays. For instance, for a training data input of 3 days, 3 consecutive Mondays were considered. The testing dataset hereinafter consisted of records on the following Monday. Prediction performances among different weekdays were finally averaged. For all combinations, a ten-fold cross validation, a widely used and robust accuracy estimation method (Kohavi, 1995), was applied. Hence, training and validation subsets comprised 90% and 10% of the total reference data, respectively.

Moreover, for each data input scenario and feature set, different prediction horizons were considered (see section 4.6). Taking into account all feature sets, amounts of training data input, prediction horizons and both algorithms, a total number of 1300 runs was required, resulting in 1300 prediction outputs. Note that, in order to examine the influence of the prediction horizon on the performance, average values of all amounts of training data were considered. All experiments were performed using both RF and ANN algorithms and results were compared in terms of MAE, MSE and R^2. Moreover, of high interest was the evaluation of the contribution of the spatial information to the prediction models. This was achieved by comparing the prediction performance of the baseline (feature set 1) to those of all other feature sets for each training time period. Similarly, for each prediction horizon, the effect of the geographic input was examined. Moreover, in order to evaluate the predictive importance of each feature, the percent increase in MSE (%IncMSE) was derived. It indicates the increase in error as a consequence of a feature being permuted (values randomly shuffled). The higher the value, the more important is the feature.

In a last step, the variation of the model's predictive performance as a function of space was explored. Notably, model performances were evaluated on a quantitative and qualitative basis for the parking districts and segments.

Chapter 5

Results

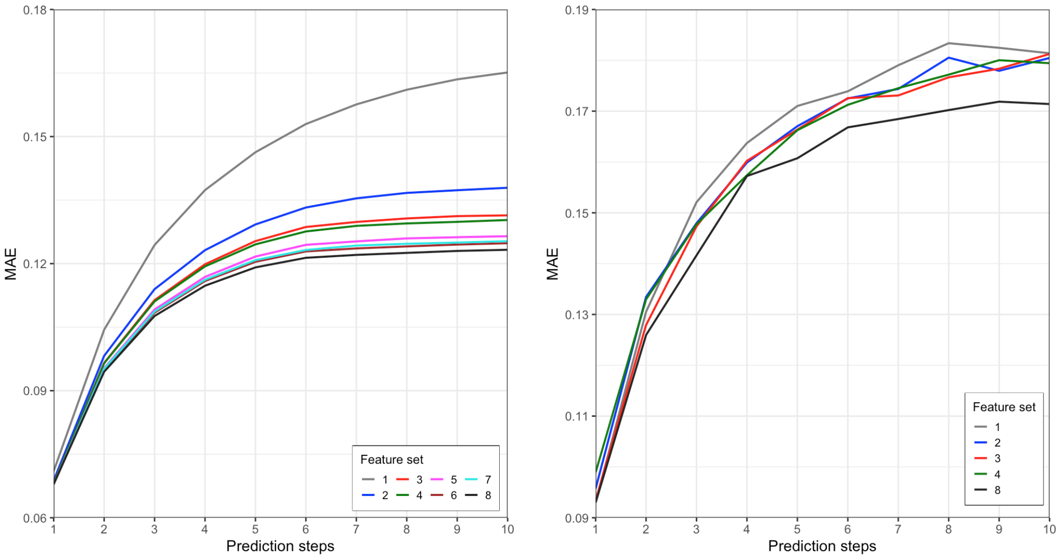
This chapter presents the experimental results. Addressing the second research question, sections 5.1 and 5.2 point out the influence of the temporal prediction horizon and the amount of training data on the predictive performance. Section 5.3 examines the contribution of spatial information on the prediction model, in reference to the first research question. Finally, section 5.4 examines the spatial variation of the prediction, pertaining to research question 3. Performance differences of the RF and ANN model are compared throughout the chapter, covering the fourth research question.

5.1 Prediction horizon

In this section, the effect of the prediction horizon on the model performance is evaluated. Figure 5.1 shows performances as a function of the prediction horizon of 10 steps in terms of MAE for both RF and ANN algorithms. Generally, it is apparent that an increased prediction horizon had a negative influence on the performance. This held true for all feature sets and both prediction models. Hence, the further in the future parking occupancy was predicted, the more error-prone was the result. One step ahead predictions achieved best results whereas 10 step ahead predictions performed worst. Notably, there was a steady decrease in performance as a function of the prediction horizon. For the RF model, a plateau was gradually reached for feature sets 2 – 8 (i.e. all except the baseline) after about 7 steps. Hence, the performance largely stopped deteriorating past that point. The performance of the prediction model using the baseline input, however, continued to be affected by more prediction steps. When using the ANN model, performance patterns were similar, regardless of the feature set that was used.

Evidently, a comparison of the feature set revealed that there was little difference for short-term predictions (i.e. few steps ahead predictions). For a 1 step ahead prediction, results were almost identical among feature sets. This applied to both RF and ANN algorithms. For instance, the mean MAE amounted to 0.071 and 0.068 for the worst and best performing feature set, respectively, for the RF algorithm. Conversely, for the ANN, MAE values of 0.099 and 0.093 were recorded for the worst and best performing feature set, respectively. Nevertheless, as the prediction horizon increased, differences were more prominent. On the maximum prediction horizon, the RF model achieved an MAE of 0.165 using feature set 1 whereas the usage of feature set 8 yielded an MAE of 0.123. By contrast, analogous results achieved by the ANN model amounted to 0.181 and 0.171. Hence, the difference was substantially smaller compared to the RF model.

Predictions on feature sets containing spatial information (feature sets 2 – 8) consistently outperformed the baseline (feature set 1) on all prediction steps when the RF model was considered. The same mostly applied to the ANN model. Table 5.1 provides an overview of model performances for all feature sets in terms of MAE, MSE and R^2. Results of prediction horizons of 1, 5 and 10 steps ahead are listed.



(a) RF

(b) ANN

FIGURE 5.1: Performance as a function of the prediction horizon. Comparison between (a) Random Forest (RF) and (b) Artificial Neural Network (ANN) algorithms.

TABLE 5.1: Random Forest (RF) and Artificial Neural Network (ANN) performance of all feature sets on a 1 step, 5 step and 10 step ahead prediction. Comparison in terms of Mean Absolute Error (MAE), Mean Squared Error (MSE) and coefficient of determination (R^2).

		1 step			5 steps			10 steps		
Feature set	**Algorithm**	**MAE**	**MSE**	**R^2**	**MAE**	**MSE**	**R^2**	**MAE**	**MSE**	**R^2**
1	RF	0.071	0.010	0.857	0.146	0.036	0.517	0.165	0.043	0.415
	ANN	0.093	0.017	0.811	0.171	0.048	0.402	0.181	0.051	0.329
2	RF	0.069	0.010	0.868	0.129	0.028	0.619	0.138	0.032	0.581
	ANN	0.096	0.018	0.797	0.167	0.045	0.420	0.180	0.051	0.332
3	RF	0.068	0.010	0.870	0.125	0.027	0.636	0.131	0.029	0.610
	ANN	0.093	0.017	0.803	0.166	0.045	0.419	0.181	0.051	0.341
4	RF	0.068	0.010	0.869	0.125	0.027	0.642	0.130	0.029	0.618
	ANN	0.099	0.019	0.787	0.166	0.045	0.420	0.179	0.051	0.349
5	RF	0.068	0.009	0.872	0.122	0.026	0.659	0.126	0.027	0.639
6	RF	0.068	0.009	0.873	0.120	0.025	0.665	0.125	0.027	0.646
7	RF	0.068	0.009	0.872	0.121	0.025	0.665	0.125	0.027	0.646
8	RF	0.068	0.009	0.874	0.119	0.025	0.673	0.123	0.026	0.655
	ANN	0.093	0.017	0.808	0.161	0.042	0.449	0.171	0.047	0.390

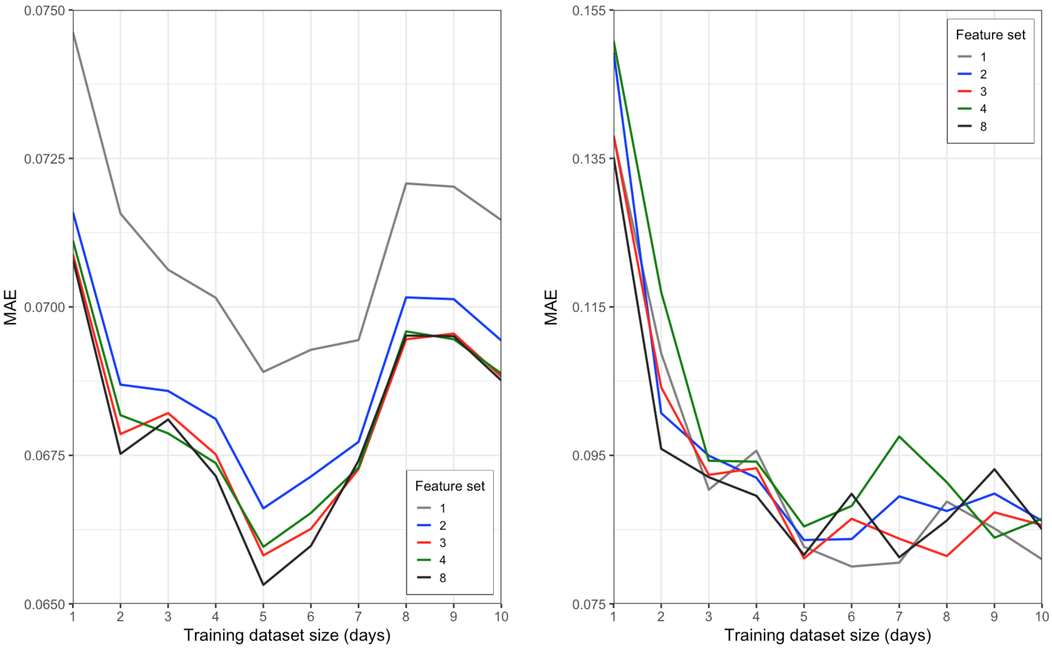
5.2 Training dataset size

In the following, the influence of the amount of training data on the model performance is examined. Training dataset sizes used for this experiment ranged from 1 day to 10 days in 1 day increments. Hence, for each feature set combination and prediction horizon, 10 values were recorded. Figures 5.2 and 5.3 illustrate the relationship between training dataset size and prediction performance for a 1 step ahead prediction and a 5 step ahead prediction, respectively. Figures for all other prediction horizons can be found in appendix A. Performances of feature sets 1 – 4 and 8 are shown.

When considering the RF model, an increased amount of data input was beneficial up to an amount of 5 days, when an optimum was reached. An input of additional training data (i.e. 6 – 10 days) was therefore not associated with model improvement. Rather, the model performance tended to deteriorate as more training data was considered. This applied to all prediction horizons. Moreover, patterns were very similar regardless of the prediction horizon. Also, performance differences as a function of the training dataset size were relatively small. For instance, there was a difference of approximately 0.005 in terms of MAE for feature set 8 when results of a training set of 5 days and 1 day were compared on a 1 step ahead prediction horizon. On a 5 step ahead prediction this difference amounted to approximately 0.009. It should be noted that the overall pattern was very similar for all feature sets and that the baseline was outperformed on all amounts of dataset input. Experiments using feature set 8 achieved best results consistently.

In comparison to the RF model, the impact of the training dataset size differed for the ANN model in terms of both pattern and magnitude. Although best results were generally also achieved using a training dataset size of about 5 days, the model benefited from an increased amount of training data to a much greater extent. This was especially apparent for shorter prediction horizons. An MAE difference of approximately 0.05 was recorded comparing feature set 8 on a data input size of 1 day and 5 days for a 1 step ahead prediction. Beyond an input of 5 days, the benefit became negligible. Similarly, on a 5 step ahead prediction horizon, an MAE difference of 0.03 was recorded. It is noteworthy that the baseline (feature set 1) was not outperformed by feature sets 1 – 4 and differences across the different feature sets were very small. Nonetheless, the usage of feature set 8 generally achieved best results.

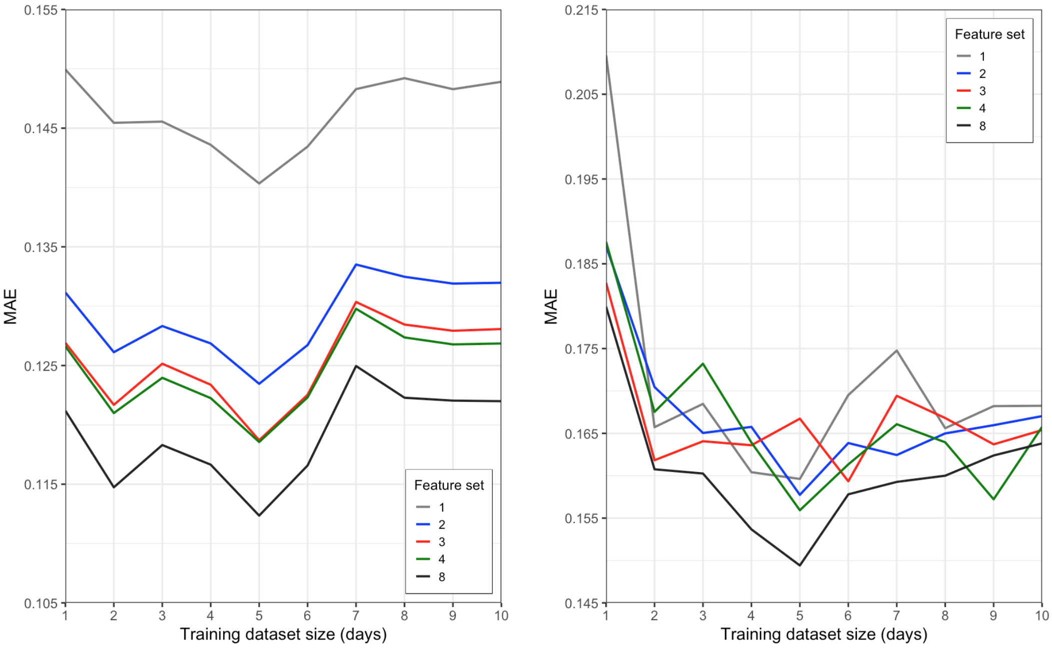
Table 5.2 shows RF model performances in terms of MAE, MSE and R^2 as a function of the amount of training data for feature sets 1 and 8. Accordingly, table 5.3 illustrates values for the ANN model. Both show results of a prediction horizon of 5 steps. For results of 1 and 10 step prediction horizons, refer to appendix B.



(a) RF

(b) ANN

FIGURE 5.2: Performance as a function of the training dataset size. Comparison between (a) Random Forest (RF) and (b) Artificial Neural Network (ANN) algorithms. 1 step ahead prediction. Note the small scale of the y-axis on the left subplot.



(a) RF

(b) ANN

FIGURE 5.3: Performance as a function of the training dataset size. Comparison between (a) Random Forest (RF) and (b) Artificial Neural Network (ANN) algorithms. 5 step ahead prediction.

TABLE 5.2: Random forest performance in terms of Mean Absolute Error (MAE), Mean Squared Error (MSE) and coefficient of determination (R^2) as a function of the training dataset size. Values for feature sets 1 and 8 on a 5 step ahead prediction.

	Feature set 1			Feature set 8		
Training dataset size (days)	**MAE**	**MSE**	**R^2**	**MAE**	**MSE**	**R^2**
1	0.150	0.038	0.493	0.121	0.026	0.652
2	0.145	0.035	0.514	0.115	0.023	0.693
3	0.146	0.036	0.501	0.118	0.025	0.657
4	0.144	0.035	0.511	0.117	0.024	0.633
5	0.140	0.033	0.521	0.112	0.022	0.688
6	0.143	0.034	0.518	0.117	0.023	0.678
7	0.148	0.037	0.524	0.125	0.028	0.650
8	0.149	0.037	0.525	0.122	0.025	0.683
9	0.148	0.036	0.538	0.122	0.025	0.690
10	0.149	0.037	0.523	0.122	0.026	0.679

TABLE 5.3: Artificial neural network performance in terms of Mean Absolute Error (MAE), Mean Squared Error (MSE) and coefficient of determination (R^2) as a function of the training dataset size. Values for feature sets 1 and 8 on a 5 step ahead prediction.

	Feature set 1			Feature set 8		
Training dataset size (days)	**MAE**	**MSE**	**R^2**	**MAE**	**MSE**	**R^2**
1	0.210	0.077	0.321	0.180	0.052	0.326
2	0.166	0.044	0.408	0.161	0.041	0.452
3	0.168	0.045	0.389	0.160	0.043	0.434
4	0.160	0.042	0.414	0.154	0.038	0.475
5	0.160	0.041	0.412	0.149	0.037	0.470
6	0.170	0.045	0.390	0.158	0.040	0.445
7	0.175	0.049	0.373	0.159	0.041	0.478
8	0.166	0.044	0.450	0.160	0.041	0.483
9	0.168	0.045	0.441	0.162	0.042	0.462
10	0.168	0.046	0.419	0.164	0.043	0.470

5.3 Geographic input contribution

5.3.1 Feature importance

In order to evaluate the contribution of individual features on the overall model performance, feature importances were determined for the RF model. Figure 5.4 gives insight into importances in terms of IncMSE. It should be noted that importances of all features are average values across all prediction horizons using feature set 8. Hence, the value of historical occupancy importance represents the mean importance value of all occupancy rates at 1 to 10 steps prior to the prediction target.

Figure 5.5 illustrates the importances of each individual occupancy rates 1 to 10 steps prior to the prediction target.

When average values were considered, the TOD exceeded an IncMSE of 0.05 and contributed most to the prediction model, followed by the historical occupancy. Nevertheless, parking occupancy rates at 1 and 2 steps prior to the prediction target showed most significance, implemented for 1 and 2 step ahead predictions, respectively. Parking occupancy rates at 3 or more steps prior to the prediction target gradually decreased in importance. Hence, the longer the prediction horizon, the less significant was the predictive significance of the historical occupancy input. In terms of geographic features, office showed the highest predictive importance, exceeding a value of 0.0125. The feature touristic contributed slightly less, followed by the feature business. The two most important geographic features contributed more to the model than historical occupancy rates at 7 or more steps prior to the prediction target. Furthermore, it is noteworthy that the land use features industrial and residential contributed significantly less than their counterpart office. Similarly, POI features business and public transport were less important than touristic. As an input category, centrality contributed least to the prediction model. Specifically, closeness was followed by betweenness and alpha centrality, in descending order.

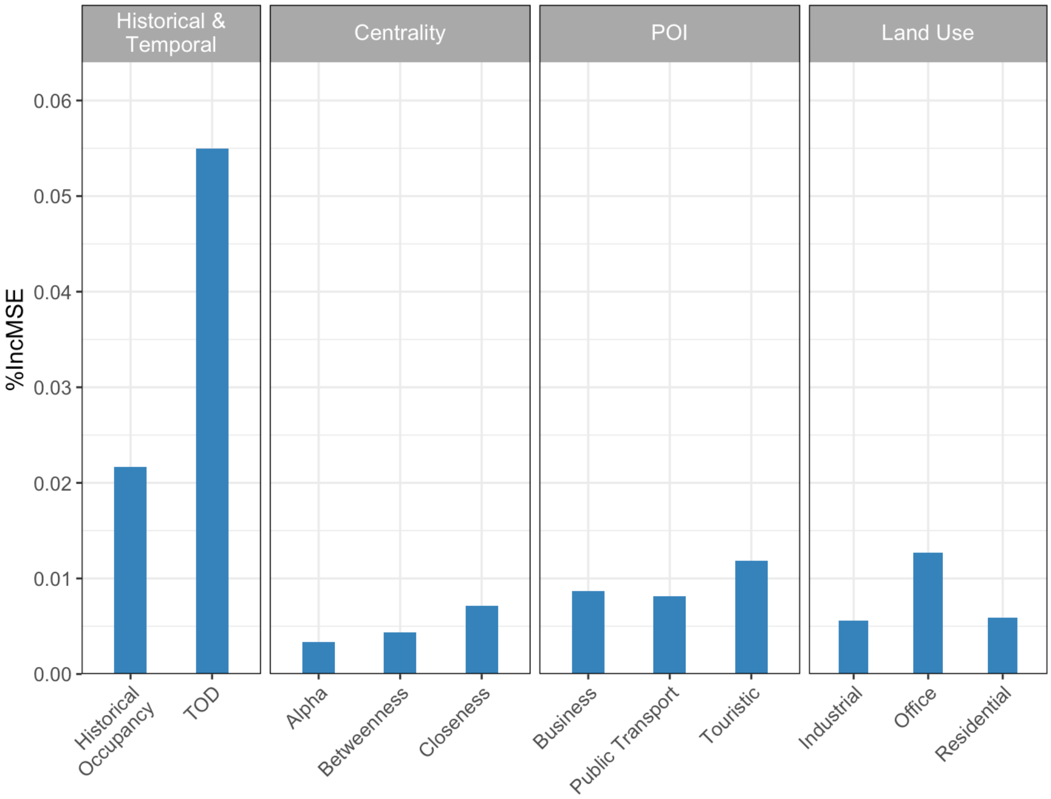


FIGURE 5.4: Random forest feature importances. Feature set 8. Mean values across all prediction horizons.

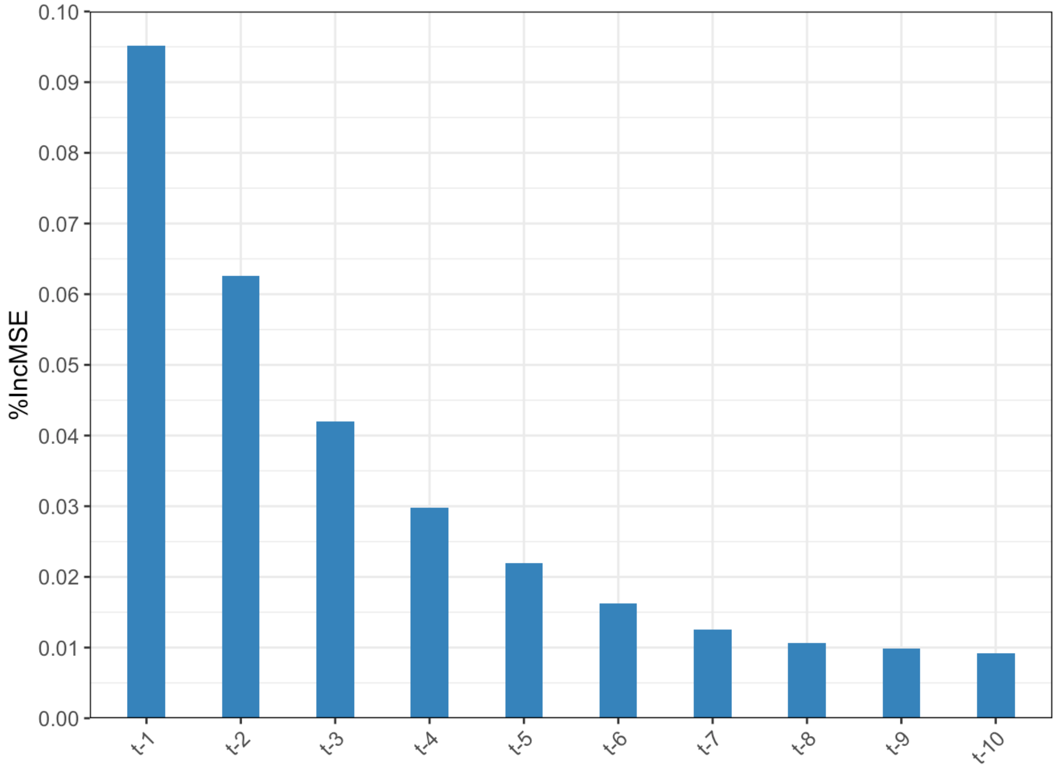
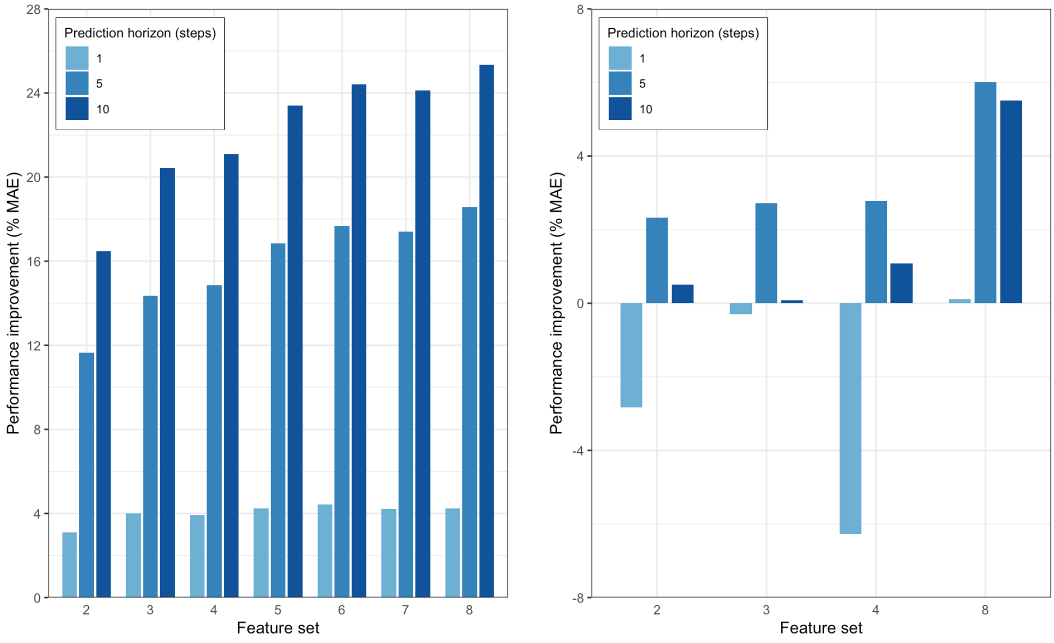


FIGURE 5.5: Random forest historical occupancy feature importances. Feature set 8.

5.3.2 Influence on prediction model performance

The following section illustrates the contribution of spatial information by comparing prediction results under usage of the different feature sets. In doing so, feature sets 2 – 8 were compared to the baseline. Figure 5.6 shows the relative performance improvement of each feature set compared to the baseline. Prediction horizons of 1, 5 and 10 steps were considered. The left and right subplots show results for the RF and the ANN model, respectively. The inclusion of spatial information yielded improvements of prediction results in any case for the RF model. On a 5 step and 10 step ahead prediction, the input of all spatial information (feature set 8) generated most improvement. On a 1 step ahead prediction, the combination of POI and land use improved most. Generally, the more data that was included, the better the model performed. Hence, feature sets 5, 6 and 7 (i.e. combinations of geographic categories) yielded better results than feature sets 2, 3 and 4. Moreover, the consideration of POI and land use information (feature sets 3, 4) lead to better results than the consideration of centrality information (feature set 2). Performance improvements became more significant with longer prediction horizons. On a 1 step ahead prediction, the inclusion of spatial information produced improvements of 3.1 – 4.2%, whereas on a 5 step ahead prediction and a 10 step ahead prediction these values amounted to 11.7 – 18.6% and 16.5 – 25.4%, respectively.

Performance improvements were not as significant for the ANN algorithm. Feature sets 2, 3 and 4 produced worse results compared to the baseline when occupancy rates were predicted 1 step ahead. The most significant deterioration was recorded for feature set 4, performing 6.3% worse in terms of MAE in comparison with baseline. A 5 step ahead prediction horizon, on the contrary, could be attributed to performance improvements ranging from 2.3 – 6.0% across all feature sets. When a 10 step ahead prediction was considered, performance changes were insignificant for feature sets 2, 3 and 4. Feature set 8 recorded an improvement of just under 6%.



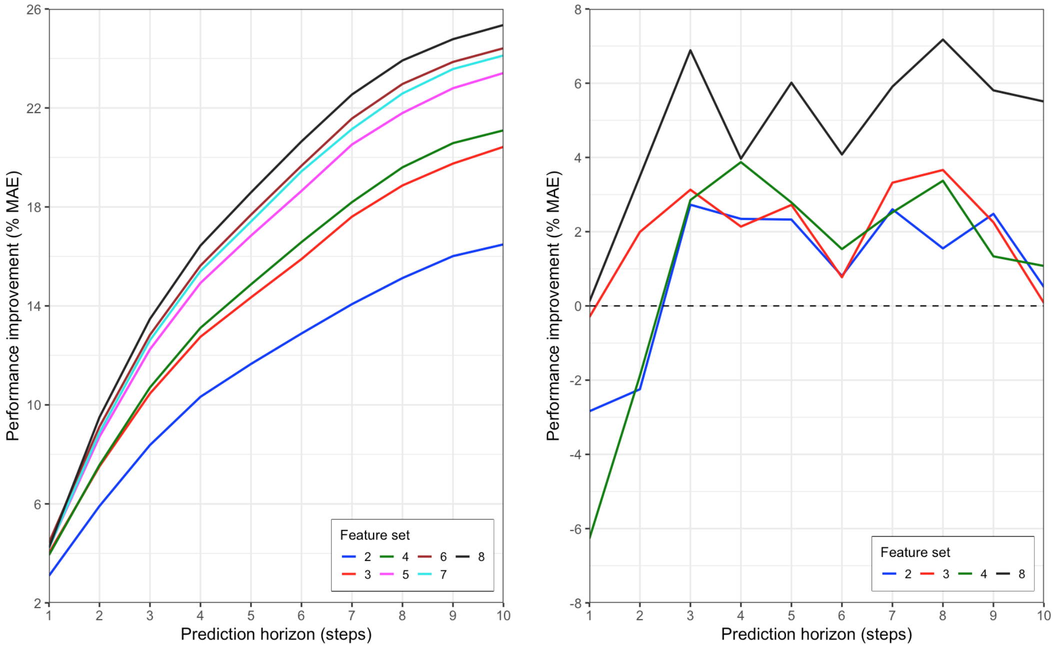
(a) RF

(b) ANN

FIGURE 5.6: Relative performance improvement using feature sets 2 – 8 compared to the baseline. Comparison of (a) Random Forest (RF) and (b) Artificial Neural Network (ANN) algorithms.

In similar fashion, figure 5.7 shows the performance improvement as a function of the prediction horizon. For the RF algorithm, the length of the prediction horizon was clearly correlated with a model performance improvement when results of feature sets 2 – 8 were compared to those of the baseline. The longer the prediction horizon, the more the values of each feature set diverged. Accordingly, on a short-term prediction horizon, values were very similar at a low level whereas differences were much higher on the maximum prediction horizon. Most significant improvements were recorded for feature set 8 across all prediction horizons.

Unlike the RF model, performance improvement for the ANN model did not steadily increase as a function of the prediction horizon. On a 1 and 2 step ahead prediction, feature sets 2 and 4 recorded prediction results that were worse than the baseline. Nevertheless, on longer prediction horizons, improvements were achieved. On average, feature set 8 showed most improvement across all prediction horizons, ranging from approximately 0 – 7%. Feature sets 2, 3 and 4, on the other hand, recorded lower values in the range of -6 – 4%. It should be noted that certain prediction horizons (e.g. 3 and 8) showed more improvement than others. However, no clear pattern could be recognized.



(a) RF

(b) ANN

FIGURE 5.7: Relative performance improvement using feature sets 2 – 8 compared to the baseline as a function of the prediction horizon. Comparison of (a) Random Forest (RF) and (b) Artificial Neural Network (ANN) algorithms.

The contribution of spatial information as a function of the amount of training data input is shown in figures 5.8 and 5.9 for a 1 step and a 5 step ahead prediction, respectively. Figures for all other prediction horizons can be found in appendix A.

Considering the RF model, rates of improvement varied across the amount of training data input. Generally, most improvement was achieved for a training dataset input of 2 days for all feature sets, whereas an input of 7 days yielded least improvement. This applied to all prediction horizons. Overall there is less contribution when an increased amount of training data was used. Nevertheless, this trend was relatively weak. On a 1 step ahead prediction, 4.0 – 5.8% improvement was recorded using 2 days as training dataset input. Conversely, a data input of 7 days yielded an improvement ranging from 2.5 – 3.2%. When 5 step ahead predictions were made, a data input of 2 days achieved an improvement of 13.3 – 21.1%, whereas an input of 7 days resulted in an improvement of 10.0 – 15.7%. Feature set 8 contributed most, considering all amounts of training data input.

By contrast, the inclusion of spatial information did not consistently improve the ANN model when different amounts of training data were considered. Moreover, it was not evident whether the amount of training data had an influence on the contribution of spatial information. On a 1 step ahead prediction, training inputs of 3, 6, 7 and 10 days resulted in worse performances compared to the baseline.

Contrarily, a prediction horizon of 5 steps entailed improvements for most amounts of training input. Although no clear pattern was evident, the consideration of feature set 8 tended to have best results.

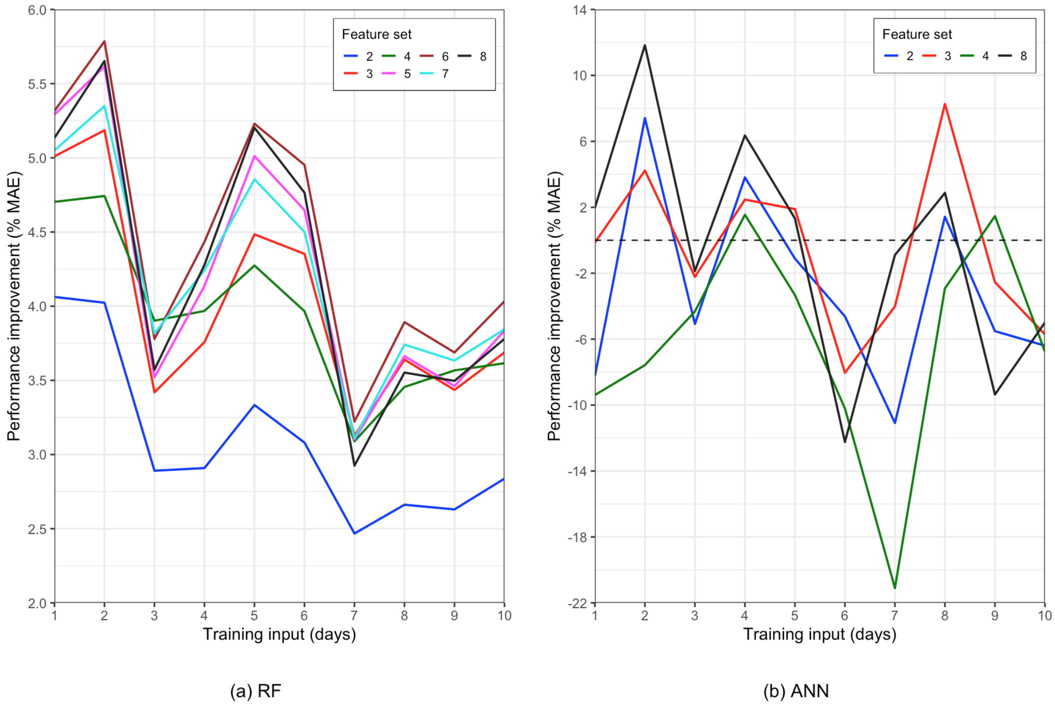
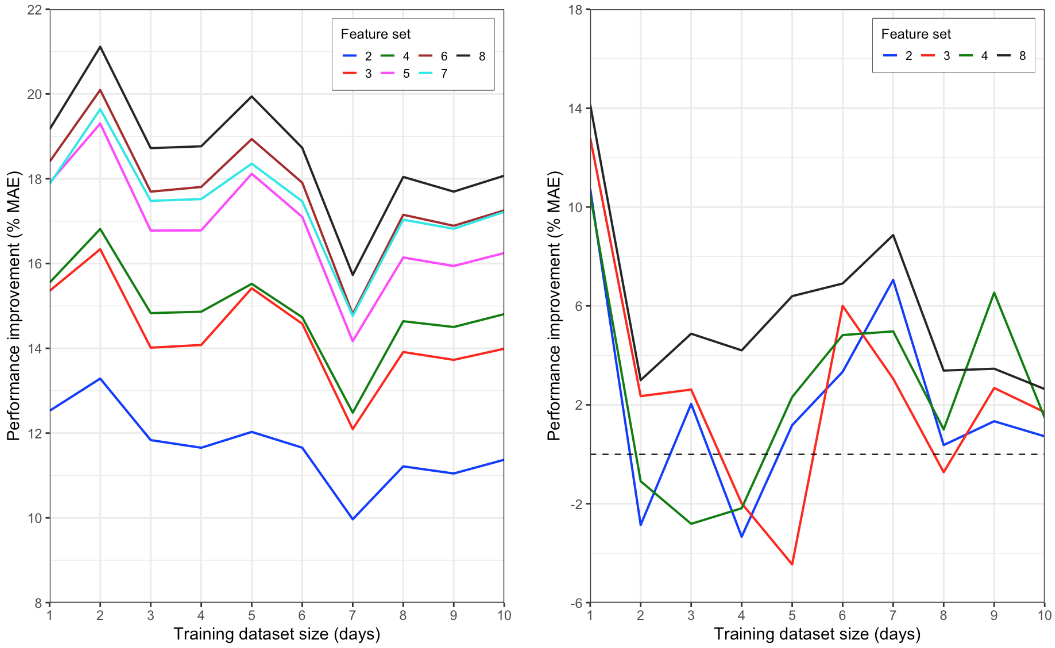


FIGURE 5.8: Relative performance improvement (feature sets 2 – 8 compared to the baseline) as a function of the training dataset size. 1 step ahead prediction. Comparison of (a) Random Forest (RF) and (b) Artificial Neural Network (ANN) algorithms.



(a) RF

(b) ANN

FIGURE 5.9: Relative performance improvement (feature sets 2 – 8 compared to the baseline) as a function of the training dataset size. 5 step ahead prediction. Comparison of (a) Random Forest (RF) and (b) Artificial Neural Network (ANN) algorithms.

5.4 Spatial variation

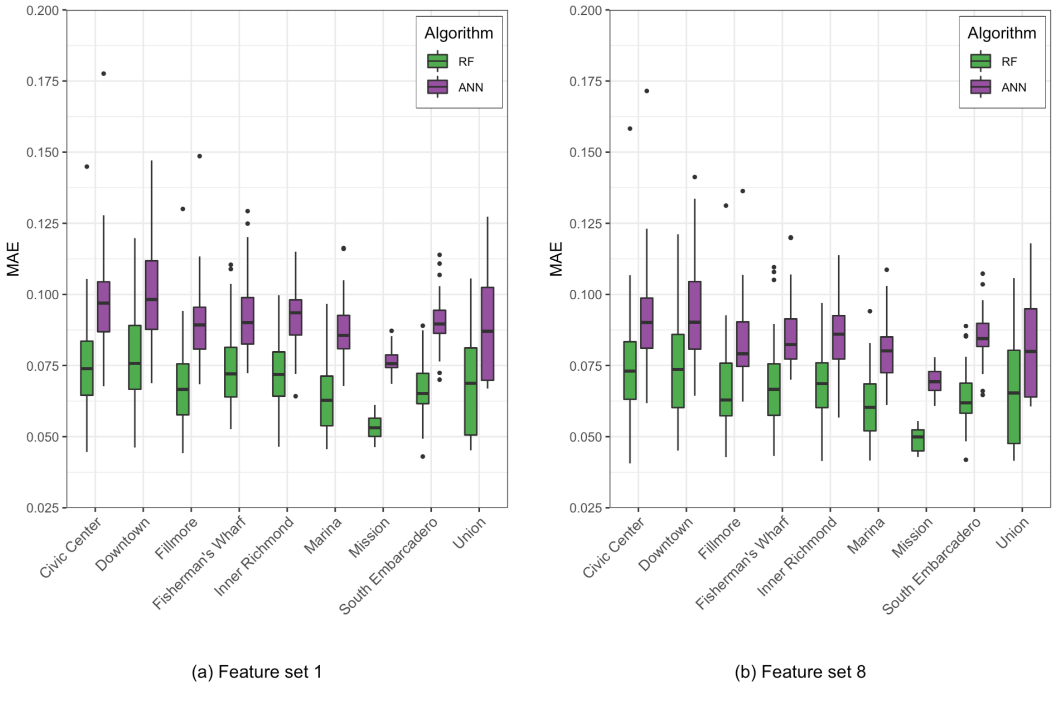
The following section focuses on the spatial variation of parking occupancy prediction. Specifically, the prediction model performance is compared and explored across space. As feature set 8 showed best overall performance across all experiments, performance results using feature set 8 and relative improvements compared to the baseline were utilized for the analysis in this section. Moreover, the focus lies on RF, the algorithm realizing superior results.

5.4.1 Quantitative assessment

In a first step, a quantitative assessment was carried out by comparing the prediction model performance among the parking districts (refer to figure 3.1 for the parking district's locations). The performances of each district on a 1 step and a 5 step ahead prediction are shown in figure 5.10 and 5.11, respectively. The left subplots signify performances of feature set 1 whereas the right subplots denote performances of feature set 8. Generally, there were performance disparities among the parking districts. The district Mission recorded best parking occupancy prediction results for its parking segments for both feature sets and algorithms. Accordingly, around 50% of segments in this district recorded an MAE lower than 0.05 on a 1 step ahead prediction using the RF model with feature set 8. In contrast, parking occupancy rates were most difficult to predict in the districts Civic Center and Downtown. The median MAE value in the former is just under 0.075 with above stated prediction set-up.

The introduction of spatial information (i.e. the usage of feature set 8) improved median occupancy prediction values for all districts on all prediction horizons for both RF and ANN algorithms. Moreover, performance differences between the algorithms were constant for each parking district, i.e. the RF model consistently outperformed the ANN model.

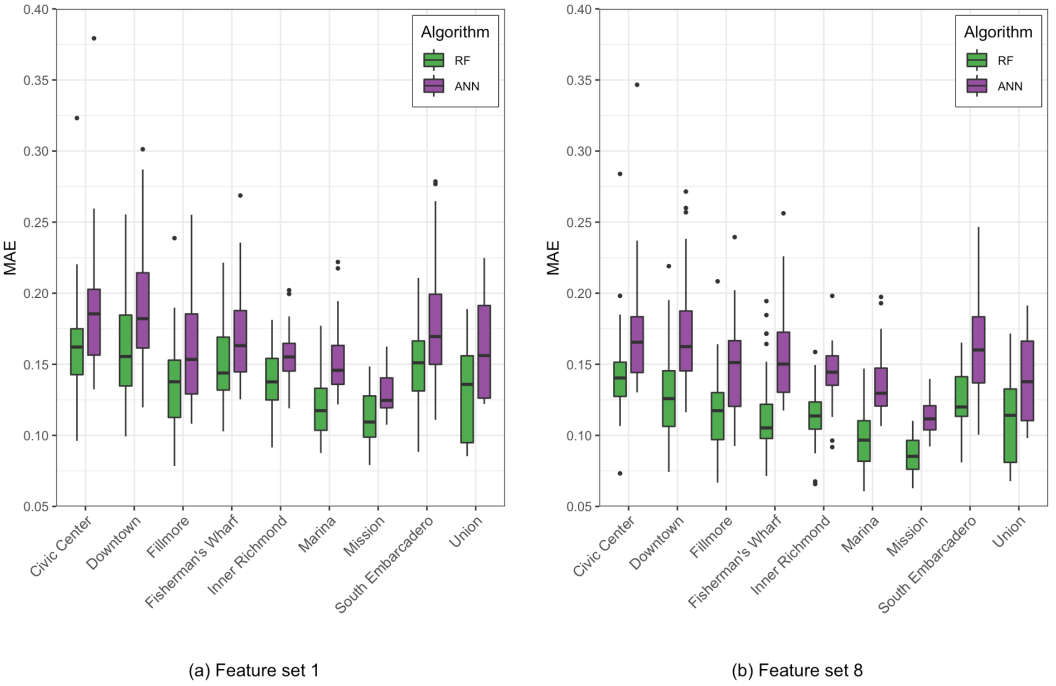
Noticeably, there were outliers that deviated greatly from median values, whereby most outliers lay in the upper range of the scale. Thus, both prediction algorithms recorded values exceeding 0.125 on a 1 step ahead prediction for parking segments in Civic Center, Downtown, Fillmore and Fisherman's Wharf. On a 5 step prediction horizon, some values exceeded an MAE of 0.3. Although the prediction models performed better using feature set 8 compared to feature set 1 overall, the inclusion of spatial information did not generally diminish the presence of outliers. Thus, occupancy rates of certain parking segments remained difficult to predict in spite of the addition of spatial information. On the contrary, in certain districts, there are parking segments whose occupancy rates were particularly easy to predict, recording values well below 0.05 and 0.075 on a 1 step and a 5 step ahead prediction, respectively.



(a) Feature set 1

(b) Feature set 8

FIGURE 5.10: Performance of parking segments aggregated by parking districts on a 1 step ahead prediction. Comparison of Random Forest (RF) and Artificial Neural Network (ANN) using (a) feature set 1 and (b) feature set 8.



(a) Feature set 1

(b) Feature set 8

FIGURE 5.11: Performance of parking segments aggregated by parking districts on a 5 step ahead prediction. Comparison of Random Forest (RF) and Artificial Neural Network (ANN) using (a) feature set 1 and (b) feature set 8.

Drawing on above-stated findings, relative model improvement for each district was derived by computing the relative differences in terms of MAE between feature set 8 and the baseline. As shown in figure 5.12, performance changes ranged from -10 – 30% on a 1 step ahead prediction. Prediction improvements deviated among the districts. Hence, the inclusion of spatial information did not entail equal improvement across space. Noticeably, occupancy rates of the vast majority of parking segments experienced a performance improvement. Most significant improvements were recorded in the districts Mission and Inner Richmond on a 1 step ahead prediction and in Mission and Fisherman's Wharf on a 5 hour ahead prediction. Median improvements were as high as 5 – 10% and around 25% for the former and the latter, respectively. Parking segments in Civic Center saw least prediction improvement with median values of 0% and 10% on a 1 step ahead prediction and a 5 step ahead prediction, respectively.

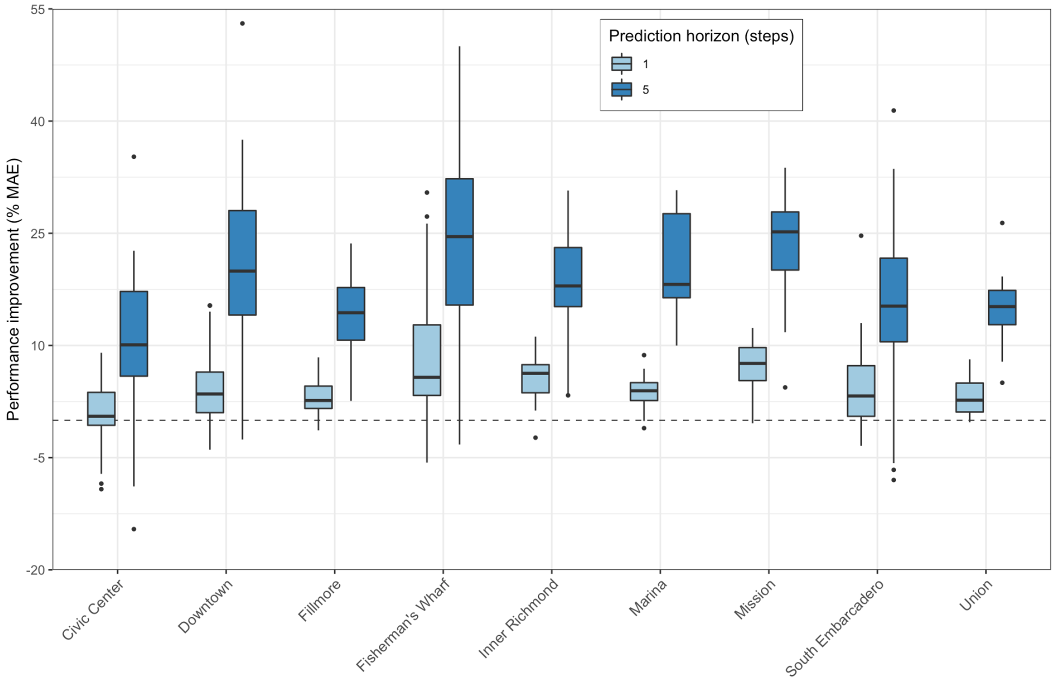


FIGURE 5.12: Relative performance improvement using feature set 8 compared to the baseline. Comparison of parking segments aggregated by parking districts on 1 step and 5 step prediction horizons. Random forest algorithm. The dashed line indicates no change.

Further, the distribution of parking segments in terms of performance improvement was explored, as shown in figure 5.13. When considering a 1 step ahead prediction, the distribution of performance improvement was relatively narrow. The bulk of parking segments recorded an improvement between 0 and 10%. Out of more than 300, occupancy rate predictions of 34 street segments saw an improvement of around 3% whereas the mean value of the distribution lay at 4.5%. On the contrary, the shape of the distribution was considerably wider on a 5 step ahead prediction. The mean value was approximately 18% improvement, whereas the highest number of street segments (21) recorded an improvement of around 17%. Although the inclusion of spatial information had more impact on a longer prediction horizon, there were more outliers to both sides of the scale.

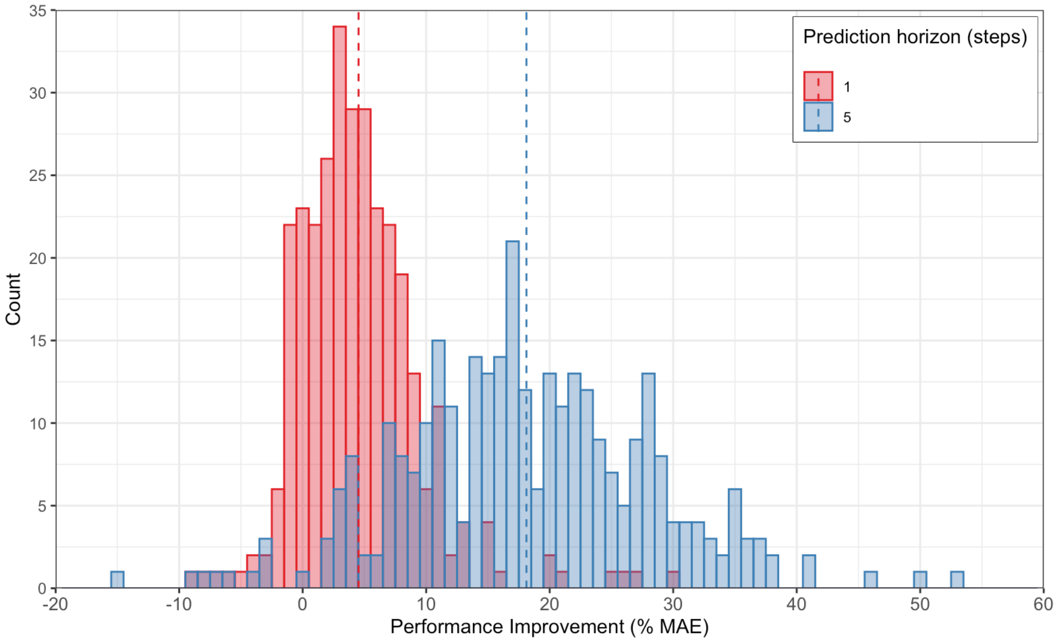


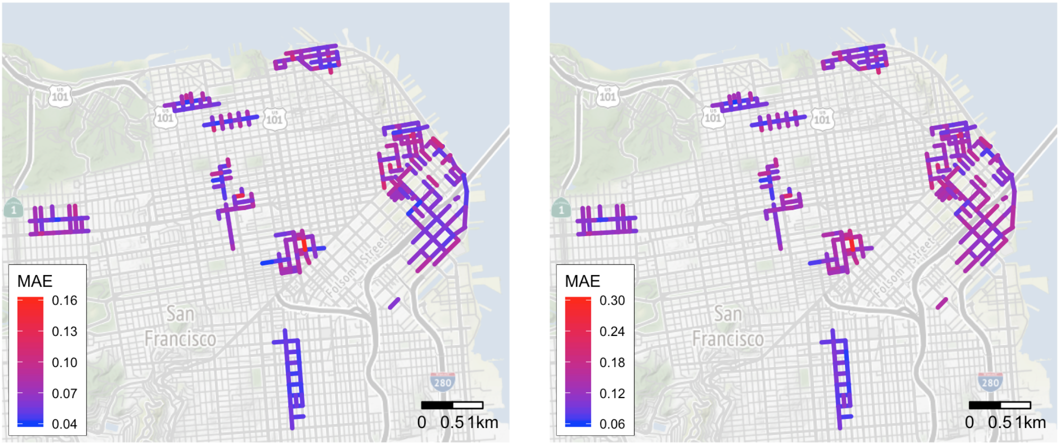
FIGURE 5.13: Distribution of parking segments in terms of relative performance improvement using feature set 8 compared to the baseline. Comparison 1 step and a 5 step ahead prediction horizons. Random forest algorithm. Dashed lines indicate average values.

5.4.2 Qualitative assessment

In the following, a qualitative assessment of the spatial variability is provided. The parking segments' occupancy prediction results using feature set 8 are visualized in figure 5.14, whereas the relative performance improvement of feature set 8 compared to the baseline is shown in figure 5.15. Results on a 1 step ahead prediction and a 5 step ahead prediction are shown on the left and right subplots, respectively. As for the performance of feature set 8, spatial patterns were very similar when comparing the 2 different prediction horizons. Hence, with a few exceptions, the model performance varied little on a spatial scale when different prediction horizons are utilized.

In general, no distinct pattern across space was identifiable. In many cases, parking segments whose occupancy rates were reliably predicted were next to or in the vicinity of others with relatively poor model performance. Especially in the northeastern parking districts of Fisherman's Wharf and Downtown as well as in the central districts of Civic Center and Fillmore, prediction results among parking segments were fairly heterogeneous. Nevertheless, in the southern parking district of Mission, parking occupancy was predicted very reliably for all segments.

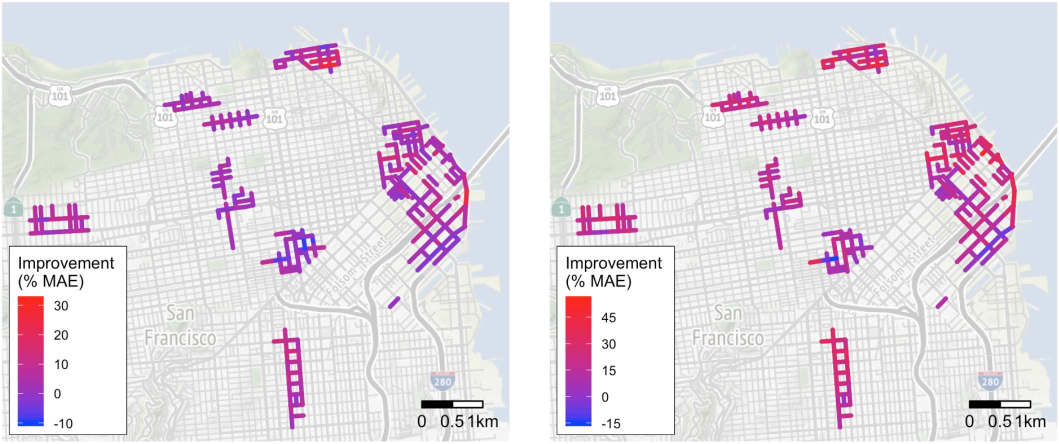
Similarly, there is no distinct pattern when the relative performance improvement was considered, regardless of the prediction horizon. Nevertheless, some parking segments with already difficult to predict occupancy rates did not benefit from the inclusion of spatial information, especially in the centrally located Civic Center district. Contrariwise, parking segments that achieved good prediction results using the baseline experienced further improvement by adding spatial information (e.g. Mission district).



(a) 1 step ahead

(b) 5 step ahead

FIGURE 5.14: Spatial distribution of performance using feature set 8. Random forest algorithm. (a) 1 step ahead prediction and (b) 5 step ahead prediction.



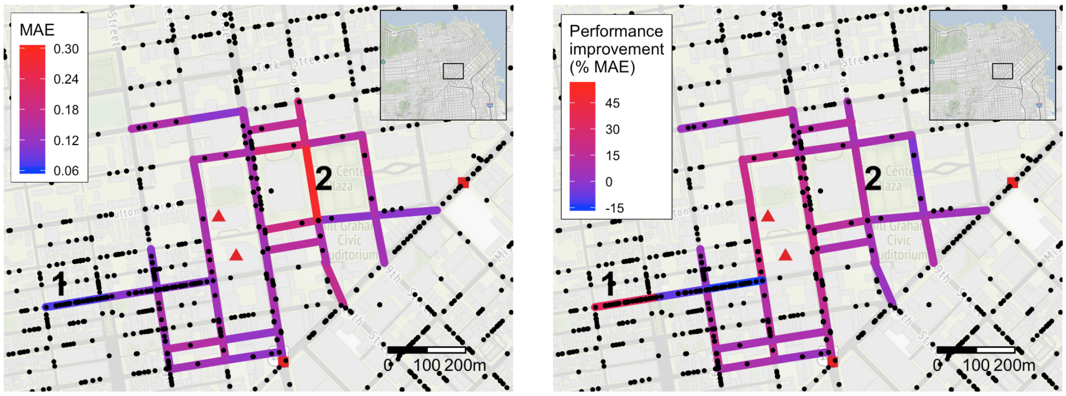
(a) 1 step ahead

(b) 5 step ahead

FIGURE 5.15: Spatial distribution of relative performance improvement using feature set 8 compared to the baseline. Random forest algorithm. (a) 1 step ahead prediction and (b) 5 step ahead prediction.

To further investigate the effect of space on parking occupancy prediction, 2 districts were selected for case studies, namely Civic Center and Fisherman's Wharf. Close-ups of each district are visualized in figures 5.16 and 5.17. The performance of feature set 8 and the relative performance improvement compared to the baseline on a 5 step ahead prediction are shown. Additionally, the locations of the POIs (business, touristic and public transport) are included. Evidently, there were 2 parking segments with especially distinct differences in terms of model performance in Civic Center (see 5.16. 200 Polk Street recorded an MAE of 0.28 whereas 500 Hayes Street recorded an MAE of 0.07. Moreover, the inclusion of spatial information improved the former and the latter by 12.5% and 35.2%, respectively. Other parking segments in the district recorded performance values of 0.11 – 0.20 and improvements of -14.6 – 22.7%. Furthermore, it is apparent that there are only few businesses in the vicinity of 200 Polk St, the parking segment that was most difficult to predict. On the contrary, a high number of businesses can be found at or near 500 Hayes St and other parking segments that were relatively easy to predict. Moreover, 2 tourist attractions and railway stops are located in the vicinity of the parking segments.

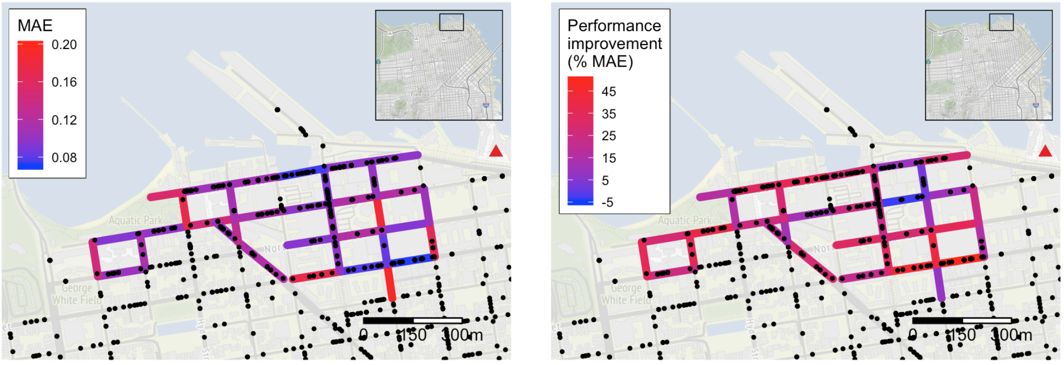
In the district Fisherman's Wharf (see figure 5.17, a significant number of parking segments recorded MAE values close to 0.2, whereas a few segments achieved results lower than 0.1. Similar to Civic Center, the locations of parking segments that were difficult to predict largely coincide with fewer business locations in the vicinity. Moreover, street segments whose occupancy were difficult to predict showed relatively weak improvement after the inclusion of spatial information, compared to street segments whose occupancy was easier to predict.



(a) Performance feature set 8

(b) Relative performance improvement

FIGURE 5.16: Spatial distribution of performance in parking district Civic Center using (a) feature set 8 and (b) relative performance improvement using feature set 8 compared to the baseline. Random forest algorithm on a 5 step ahead prediction. POIs are indicated as black circles (business), red squares (public transport) and red triangles (touristic). Labels show locations of street segments (1) 500 Hayes St and (2) 200 Polk St.



(a) Performance feature set 8

(b) Relative performance improvement

FIGURE 5.17: Spatial distribution of performance in parking district Fisherman's Wharf using (a) feature set 8 and (b) relative performance improvement using feature set 8 compared to the baseline. RF algorithm on a 5 step ahead prediction. POIs are indicated as black circles (business) and red triangles (touristic).

Another possible explanation of their diverging occupancy prediction results, occupancy rates for 500 Hayes St and 200 Polk St over a 2 week period are displayed in figure 5.18. Evidently, the occupancy rate for 200 Polk St fluctuated heavily, regularly reaching occupancy rates of 0 and 1. 500 Hayes St, on the contrary, recorded steady rates, mostly restricted between 0.6 and 0.8.

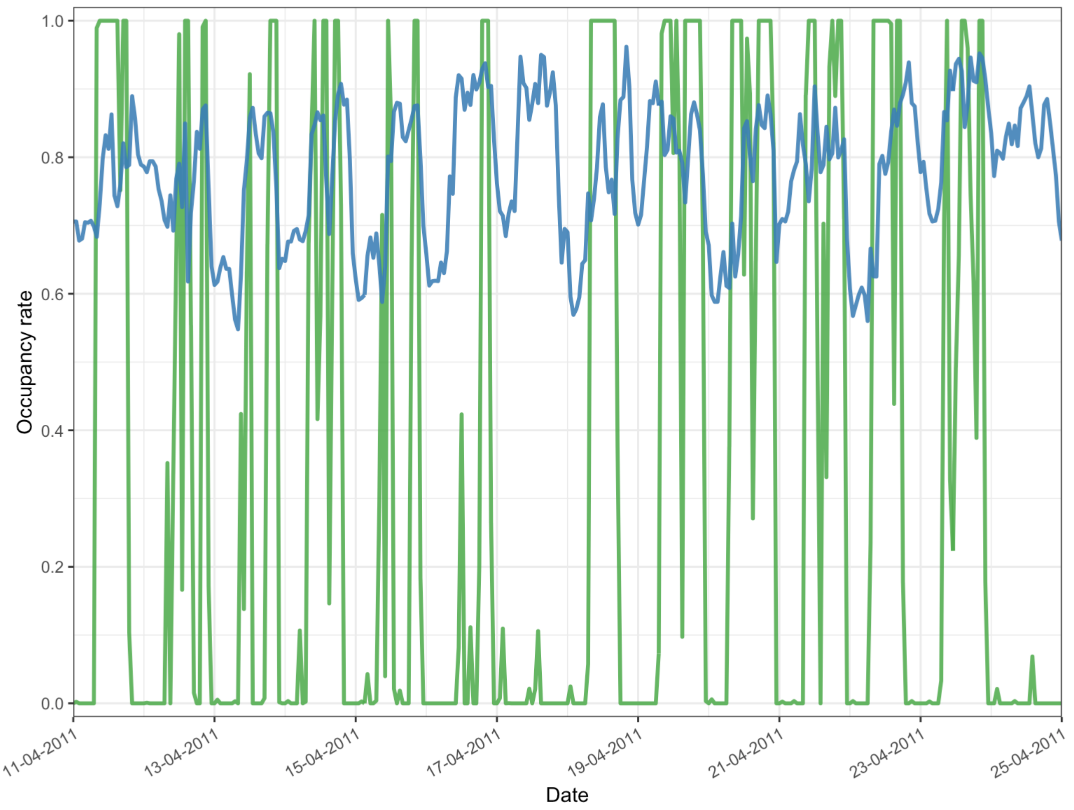


FIGURE 5.18: Occupancy rates of 500 Hayes St (blue) and 200 Polk St (green) parking segments over a 2 week period.

5.5 Summary of results

In this section, the above stated results are summarized. The most relevant findings are as follows:

- Longer prediction horizons entailed less reliable predictions.
- An increased amount of training data did not necessarily improve the prediction model.
- The incorporation of spatial information helped improve prediction models. Improvements of up to 25% compared to the baseline were achieved on long-term prediction horizons.
- The inclusion of more spatial information was associated with better model performance. Land use and POI information were more beneficial than centrality.
- Among the spatial features, office showed most predictive relevance, followed by touristic and business.
- There were performance disparities among the parking districts. Nevertheless, no clear spatial pattern was evident.

- The introduction of spatial information into the model did not entail uniform improvement across space.

- The ANN model is outperformed by the RF model across all metrics. Moreover, the RF model benefited from spatial information to a much larger extent.

Chapter 6

Discussion

6.1 Benefits of the inclusion of spatial information

As part of the first research question, the benefit of adding spatial information to the parking occupancy prediction models was addressed. In order to do so, experiments were conducted which compared the relative performance improvement of data inputs containing spatial information (feature sets 2 – 8) with the baseline. Additionally, the importance of each individual feature was explored for the RF model using feature set 8. Overall, the RF model performance could be increased by 3.1 – 25.4% whereas the ANN model recorded performance changes of -6.3 – 7.2%, depending on the prediction horizon and feature set. Across all prediction horizons, feature set 8 performed best, recording improvements for both algorithms. Hence, in general, the more spatial information was used as data input, the better it performed. These findings are insofar relevant, as they point out that the incorporation of the underlying geographic context help improve parking occupancy prediction models. They imply that conventional models only taking into account temporal information can potentially be improved by incorporating spatial information. Hence, space matters when parking occupancy is predicted. Moreover, the results are novel, as, to the best of my knowledge, no study has previously considered the explicit inclusion of spatial information for parking occupancy prediction problems.

In other fields, however, geography has also been found beneficial as input for prediction problems. Krause and Zhang (2019) implemented POI/land use data for their traffic destination prediction model, with improved accuracy compared to the baseline. Also taking into account land use information, Luo (2010) made traffic demand predictions, potentially improving conventional models. Chan and Cooper (2019) used centrality information (specifically betweenness centrality) to predict bicycle mode share and flows, achieving comparable results to more complex models lacking spatial input. Similarly, Sarlas and Axhausen (2016) included network theory-based variables in their traffic volume prediction scheme, significantly enhancing the predictive accuracy.

Generally, the prediction models performed better using feature sets 3 and 4 compared to 2, while the usage of feature sets 6 and 7 entailed better results than the usage of feature set 5. This pattern was very clear for the RF model while it held partially true for the ANN model. Consequently, the spatial category land use was most valuable to the prediction model, closely followed by POI. Conversely, centrality contributed less to the prediction. Moreover, the combined usage of land use and centrality tended to be more beneficial than the combination of POI and centrality. It could be argued that the categories land use and POI are conceptually more similar and therefore had a similar impact on the prediction model. Moreover, the mere location of a parking segment with respect to the entire street network did not appear to play as influential a role as the configuration of space in its vicinity. The potential of the implementation of land use in a parking occupancy prediction scheme has been suggested by Richter et al. (2014). They split the sf*park* parking segments into regions with similar land use characteristics (i.e. commercial, residential and touristic) and showed regional similarities of parking occupancy predictions.

Results of overall contributions of geographic categories agreed with individual feature importances. Each POI feature contributed more to the model than centrality features. The land use feature office was considerably more beneficial than industrial and residential. This could partially be explained by the fact that many parking districts lack industrial areas. Especially the districts Inner Richmond, Fillmore and Marina, which are predominantly residential, provided little input to the model. The land use classes residential and office, on the other hand, are more evenly distributed with high concentrations in certain areas.

6.2 Training dataset size

The impact of varying amounts of data inputs on the model performance was investigated as part of the second research question. By doing so, the amount of input data to train the model was varied from 1 day to 10 days of occupancy data. Overall, little changes were recorded for the RF model. Although it achieved optimal results with a 5 day training input for all prediction horizons, an increased amount of input data past an amount of 5 days was associated with deterioration in model performance. The fact that an increased amount of training data did not lead to improved results is insofar unexpected as a larger amount of input data generally is associated with increased performance (Figueroa et al., 2012). In the field of parking occupancy prediction, Bock (2018) also pointed out that his RF model benefited from increased amounts of training data. Values larger than 20 days, however, only entailed small improvements. Similarly, Yang et al. (2003) found that smaller errors are associated with a larger sample size.

As the performance differences were relatively small, it could be argued that they are statistically insignificant and a small training input (e.g. 1 day) was sufficient for the model to learn. Since the recorded occupancy data was relatively uniform across time (i.e. there were no significant deviations), unexpected anomalies in the data can be excluded as a reason. Hence, further research is needed to investigate the influence of the training dataset size in the context of prediction model implemented with RF.

The ANN model, by contrast, benefited from an increased amount of training data significantly, especially on short-term predictions. It achieved best results when 5 or more days were considered as training input. After that, the learning curve likely reached a maximum. These findings are in agreement with Ji et al. (2015), who found that their model improved when more input data was added. This, however, was only the case up to a certain point, before the prediction error started to rise again due to overfitting of the training set. Similarly, Bock (2018) pointed out that his RF model benefited from increased amounts of training data.

It is noteworthy that changes in the amount of training data generally had little impact on the contribution of spatial information. Hence, the amount of training data did not substantially influence the geographic contribution.

6.3 Temporal prediction horizon

Further, the influence of the temporal prediction horizon on the model performance was examined. Prediction horizons of 1 to 10 steps were considered whereby 1 step corresponds to 1 hour. Hence, the prediction of occupancy rates of up to 10 hours in the future was considered. Unsurprisingly, as the prediction horizon increased, the less reliable prediction results became due to the fact that errors accumulate. This is in agreement with several studies that compared the length of the prediction horizon with the model performance Monteiro and Ioannou (2018); Zheng et al. (2015); Liu et al. (2018); Mei et al. (2019). Monteiro and Ioannou (2018) found that the usage of real-time data is only beneficial for a prediction horizon of up to 2 hours. After that point, using average occupancy values from previous days realized equal or better results. Similarly, Zheng et al. (2015) showed that their models incorporating historical occupancy data to make predictions for future values were useful for short-term predictions. After 15 steps (i.e. almost 4 hours), the incorporation of only temporal data showed similar results.

The fact that the length of the prediction horizon was correlated with the model performance was also reflected in the feature importances. The further the occupancy rates lay in the past, the less important they were to the model. As a result, past a certain prediction horizon, the temporal and geographic features were more important relatively. This is also consistent with the study by Zheng et al. (2015) who found that the inclusion of previous observations as a feature is very beneficial for short-term predictions. However, as the prediction horizon increased, other features became more relevant. In this thesis, this can be recognized by the fact that model performances using feature sets 2 – 8 increasingly diverged from the baseline as a function of the prediction horizon. Hence, spatial information showed more relevance with longer prediction horizons. Moreover, performances levelled off for long term horizons, suggesting that the usage of historical occupancy data is not beneficial for even longer prediction horizons.

Given the spatial expansion of the study area in this work, short-term predictions of up to 3 or 4 steps ahead are likely sufficient, as all destinations within the city are reached within a few hours at most.

6.4 Spatial prediction variation

The spatial variation of the prediction performance was evaluated as part of the third research question. A quantitative and a qualitative assessment investigated to what extent space influences parking occupancy prediction. Evidently, there were spatial disparities, manifested by prediction performances that were not uniform across space. Parking occupancy could be predicted more reliably in certain districts than others. However, as a whole, no clear pattern was recognizable. The same applied for the performance improvement after the inclusion of spatial information, i.e. the extent of improvement is not selective about space. Nevertheless, the model could be improved for many street segments' occupancy rates that were reliably predicted. Conversely, a considerable number of street segments recording unreliable predictions could not benefit to a great degree from the inclusion of spatial information. Hence, locations that were notoriously difficult to predict remained to be difficult to predict after the inclusion of spatial information in the model. Moreover, proximity of parking segments did not automatically suggest similar prediction performance, as suggested by Rajabioun and Ioannou (2015); Richter et al. (2014); Leu and Zhu (2015) who took into account parking situations of neighbouring segments, under the assumption that there is a spatial correlation of parking usage.

This was demonstrated by the parking segments 500 Hayes St and 200 Polk St (see figure 5.18), generating completely different results, despite their proximity. The fact that the spatial pattern is not very strong, would suggest that the location of a parking segment is not the driving factor for the prediction model performance. Rather, occupancy patterns could dictate the difficulty of prediction. The more the occupancy rate fluctuates through time, the more difficult its prediction becomes. This in turn, is likely related to the number of parking spots in a parking segment. Parking segments with fewer spots are more prone to higher fluctuations.

As seen in figures 5.16 and 5.17, the locations of POIs could also have an influence on the prediction performance. Evidently, the presence of businesses in the vicinity of parking segments appears to be associated with an increased prediction performance. Hence, the locations of businesses could either itself have an influence on the parking occupancy prediction or be a proxy for constant occupancy patterns. Touristic and public transport POI were few and far between and no conclusions could be drawn.

6.5 Machine learning algorithm comparison

Finally, the last research question aimed to discover performance differences of ML algorithms that were implemented in this thesis, namely RF and ANN. Most experiments discussed in the previous sections were performed using both algorithms. Therefore, there is a direct juxtaposition of each algorithm's performance. The RF's performance surpassed that of the ANN in every aspect. Under an optimal data input scenario on a 1 step ahead prediction using feature set 8, the RF model achieved an MAE of 0.065. By contrast, the ANN produced an MAE 0.082 under the same conditions. Moreover, the ANN model was much less receptive to the inclusion of spatial information. Averaged over all training dataset sizes, the inclusion of all spatial information (feature set 8) entailed an improvement of 4.2 - 25.4% compared to the baseline, whereas the ANN recorded an improvement of 0.1 - 7.2%.

In the literature, there is no consensus as to which algorithm is better suited to solve the problem of parking occupancy prediction. Both RF and ANN algorithms have been used to reliably make predictions. Bock (2018) implemented an RF model using a binary classification approach and reached an accuracy of 84.5% on a short-term prediction horizon. Similarly, Dias et al. (2015) aimed to predict occupancy status of a public bicycle service using 2 occupancy classes and achieved accuracies of just under 90%. A multitude of scientific literature, on the other hand, implemented ANNs. Zheng et al. (2015) made 15 minutes ahead predictions with their FFNN and recorded average MAE values of 0.089. Similarly, MAE values between just over 0.02 and 0.08 were achieved by RNN implementations realized by Vlahogianni et al. (2016), Shao et al. (2019) and Camero et al. (2019). It should be noted, however, that performance comparisons must be viewed with caution due to differences in the study area, input data and model architecture.

In recent years, the usage of ANNs saw a surge in parking occupancy prediction. Especially if the underlying relationships and features are unknown, ANNs are suitable for parking prediction problems (Pflügler et al., 2016). In particular, RNNs cater to the recurrent nature of parking occupancy time series. Conversely, RFs' robustness to overfitting has been pointed out as one of its main benefits compared to other algorithms, including ANNs (Hastie et al., 2009). Additionally, its relative simplicity and ease of tuning should be emphasised. Unlike the ANN, only few tuning parameters have to be considered. The fact that the RF model outperformed the ANN model considerably in this thesis could be explained by its above mentioned benefits. Nevertheless, the fact that no study has compared the prediction performance of RF and ANN in the realm of parking occupancy prediction directly does not allow a comparison to the literature. It should be stressed that more research is needed in terms of the incorporation of ANNs for the prediction of parking occupancy prediction models using spatial information.

6.6 Limitations of the work

The prediction models and its input data in this thesis exhibit limitations that need to be pointed out. Firstly, it is likely that the prediction model architecture and parametrization could be further optimized. Especially in case of the ANN, building a suitable model is challenging. Secondly, the methods by which the spatial data was quantified have their limitations. The land use categories that were utilized in this thesis (industrial, office and residential), were categorized as such if 80% of a parcel's floor area corresponds to said land use category. As a consequence, only parcels with a predominant industrial, office or residential use were considered for the analysis. A high number of parcels are classified as *mixed*, potentially also containing industrial, office or residential usage were therefore not considered. A potential distortion of the land use categories or an underestimation of certain categories could be consequences.

Chapter 7

Conclusion and future work

This thesis aimed to investigate the contribution of spatial information to the prediction of on-street parking occupancy. In order to do so, RF and ANN prediction models were implemented and spatial data regarding centrality, land use and POIs were used as training data input. To assess the contribution, the performance was compared to a baseline, comprising only historical and temporal training data. The experiments were conducted in light of the fact that the underlying geographic context has received little attention in parking occupancy prediction models. Moreover, the explicit inclusion of the spatial configuration of the street network, land use and POIs represents a novel concept. The key findings were as follows:

1. The inclusion of spatial information lead to a performance improvement of up to 3% and 25% on a short-term and a long-term prediction horizon, respectively. Hence, the incorporation of space adds value to parking occupancy prediction models. Generally, the consideration of a larger amount of geographic training features corresponded to increased performance. Consequently, the inclusion of a combination of centrality, land use and POI data achieved best prediction results. In particular, land use and POI information were more beneficial than centrality. In terms of features, office exhibited most predictive relevance, followed by touristic and business.

2. The amount of training data did not significantly impact the RF model's prediction performance. As occupancy data anomalies could be excluded, the reason is unknown and has to be investigated in further research. The ANN model achieved optimal results when 5 days of input data were trained. An increased amount of training data did not entail further improvement, likely attributed to the tendency of overfitting.

3. Generally, longer prediction horizons produced less reliable predictions. This can be explained by the fact that prediction errors accumulate as the prediction horizon increases. Moreover, as expected, the inclusion of spatial information showed more relevance on long-term predictions. This is attributed to the fact that, as historical occupancy becomes less important on longer prediction horizons, other features are relatively more significant.

4. There were prediction performance disparities across space. Moreover, no clear spatial pattern could be identified and proximity of parking segments did not necessarily signify similar prediction results. This could be explained by the fact that the performance is mainly dependent on occupancy patterns, which are not necessarily spatially correlated.

5. In terms of model performance, the RF model outperformed the ANN model in all respects. Moreover, the RF benefited from spatial information to a greater extent. A possible explanation for RF's superiority is its robustness to overfitting as well as its relative simplicity compared to the ANN.

In spite of the promising results, there are opportunities to further the research on parking occupancy prediction incorporating spatial information. Firstly, the consideration of more spatial sources is a potential future research could tap into. Information such as population distribution, the public transportation network and a wider variety of POIs could be taken into account. In addition, more land use categories and land cover information could prove to be beneficial for future work. Secondly, research could be extended in terms of prediction algorithms. Especially the usage of ANNs, considered state of the art in the field of transportation prediction, could be further exploited by considering different and more complex architectures. The linking of advanced statistical methods with ML would also be a conceivable option. Thirdly, alternative ways of parking occupancy data acquisition should be embedded in future parking occupancy prediction schemes. In such a way, the application of parking occupancy prediction could be extended to other places that currently do not have the means to monitor and manage parking occupancy. Crowdsensing is a viable option that could be implemented on large scales, integrating the population.

Bibliography

Axhausen, K. W., Polak, J. W., and Boltze, M. (1994). Effectiveness of the parking guidance information system in Frankfurt am Main. *Traffic engineering & control*, 35(5): 304–309.

Badii, C., Nesi, P., and Paoli, I. (2018). Predicting Available Parking Slots on Critical and Regular Services by Exploiting a Range of Open Data. *IEEE Access*, 6: 44059–44071.

Bavelas, A. (1950). Communication Patterns in Task-Oriented Groups. *Journal of the Acoustical Society of America*, 22(6): 725–730.

Bellissent, J. (2010). Getting Clever About Smart Cities: New Opportunities Require New Business Models. *Forrester Research, Inc.*

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, Berlin, Heidelberg.

Bock, U. F. (2018). *Dynamic Parking Maps from Vehicular Crowdsensing*. PhD thesis, Gottfried Willhelm Leibniz Universität Hannover.

Bonacich, P. and Lloyd, P. (2001). Eigenvector-like measures of centrality for asymmetric relations. *Social Networks*, 23(3): 191–201.

Breiman, L. (2001). Random forests. *Machine Learning*, 45: 5–32.

Bush, K. and Chavis, C. (2017). Safety Analysis of On-Street Parking on an Urban Principal Arterial. In *Transportation Research Board 96th Annual Meeting*, Washington, D.C., USA. Transportation Research Board.

Caicedo, F., Blazquez, C., and Miranda, P. (2012). Prediction of parking space availability in real time. *Expert Systems with Applications*, 39(8): 7281–7290.

Caliskan, M., Barthels, A., Scheuermann, B., and Mauve, M. (2007). Predicting parking lot occupancy in vehicular ad hoc networks. In *65th Vehicular Technology Conference*, pages 277–281, Dublin, Ireland. IEEE.

Camero, A., Toutouh, J., Stolfi, D. H., and Alba, E. (2019). Evolutionary Deep Learning for Car Park Occupancy Prediction in Smart Cities. In *12th International Conference on Learning and Intelligent Optimization*, pages 386–401, Kalamata, Greece. Springer.

Caruana, R. and Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In *23rd International Conference on Machine learning*, pages 161–168, Pittsburgh, USA. Association for Computing Machinery.

Chai, T. and Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, 7(3): 1247–1250.

Chan, E. Y. C. and Cooper, C. H. (2019). Using road class as a replacement for predicted motorized traffic flow in spatial network models of cycling. *Scientific Reports*, 9(1): 1–12.

Chen, X. (2014). Parking Occupancy Prediction and Pattern Analysis. Technical report, Department of Computer Science, Stanford University, Stanford, CA, USA.

City and County of San Francisco (2014a). Registered Business Locations - San Francisco. `https://data.sfgov.org/` (Accessed: 01/07/2019).

City and County of San Francisco (2014b). Street Segment and Intersection (CNN) Change Log. `https://data.sfgov.org/` (Accessed: 01/07/2019).

City and County of San Francisco (2016). Land Use. `https://data.sfgov.org/` (Accessed: 01/07/2019).

Crucitti, P., Latora, V., and Porta, S. (2006). Centrality measures in spatial networks of urban streets. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 73(3): 1–5.

Dias, G. M., Bellalta, B., and Oechsner, S. (2015). Predicting occupancy trends in Barcelona's bicycle service stations using open data. In *SAI Intelligent Systems Conference*, pages 439–445, London, UK. IEEE.

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

Ermagun, A. and Levinson, D. (2018). Spatiotemporal traffic forecasting: review and proposed directions. *Transport Reviews*, 38(6): 786–814.

Esri (2019). ArcGIS Desktop. `https://www.esri.com/en-us/arcgis/about-arcgis/overview`.

Fan, J., Hu, Q., and Tang, Z. (2018). Predicting vacant parking space availability: an SVR method with fruit fly optimisation. *IET Intelligent Transport Systems*, 12(10): 1414–1420.

Fang, X., Xiang, R., Peng, L., Li, H., and Sun, Y. (2018). SAW: A Hybrid Prediction Model for Parking Occupancy Under the Environment of Lacking Real-Time Data. In *44th Annual Conference of the IEEE Industrial Electronics Society*, pages 3134–3137, Washington, D.C., USA. IEEE.

Figueroa, R. L., Zeng-Treitler, Q., Kandula, S., and Ngo, L. H. (2012). Predicting sample size required for classification performance. *BMC Medical Informatics and Decision Making*, 12(8).

Freeman, L. C. (1977). A Set of Measures of Centrality Based on Betweenness. *Sociometry*, 40(1): 35–41.

Gart, J. J. (1975). *The Poisson Distribution: The Theory and Application of Some Conditional Tests*. Springer, Dordrecht.

Giuffrè, T., Siniscalchi, S. M., and Tesoriere, G. (2012). A Novel Architecture of Parking Management for Smart Cities. *Social and Behavioral Sciences*, 53: 16–28.

Greengard, S. (2015). Between the lines. *Communications of the ACM*, 58(6): 15–17.

Hampshire, R. C., Jordon, D., Akinbola, O., Richardson, K., Weinberger, R., Millard-Ball, A., and Karlin-Resnik, J. (2016). Analysis of Parking Search Behavior with Video from Naturalistic Driving. *Transportation Research Record*, 2543(1): 152–158.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York.

Hössinger, R., Widhalm, P., Ulm, M., Heimbuchner, K., Wolf, E., Apel, R., and Uhlmann, T. (2014). Development of a Real-Time Model of the Occupancy of Short-Term Parking Zones. *International Journal of Intelligent Transportation Systems Research*, 12(2): 37–47.

Intelligent Sensors, Sensor Networks and Information Processing (2015). IoT deployment in the City of Melbourne. `http://issnip.unimelb.edu.au/research_program/Internet_of_Things/iot_deployment` (Accessed: 01/07/2019).

Ji, Y., Blythe, P., Guo, W., Tang, D., and Wang, W. (2015). Short-term forecasting of available parking space using wavelet neural network model. *IET Intelligent Transport Systems*, 9(2): 202–209.

Karlaftis, M. G. and Vlahogianni, E. I. (2011). Statistical methods versus neural networks in transportation research: Differences, similarities and some insights. *Transportation Research Part C: Emerging Technologies*, 19(3): 387–399.

Kimms, A., Maassen, K. C., and Pottbäcker, S. (2012). Guiding traffic in the case of big events with spot checks on traffic and additional parking space requirements. *Central European Journal of Operations Research*, 20(4): 755–773.

Klappenecker, A., Lee, H., and Welch, J. L. (2014). Finding available parking spaces made easy. *Ad Hoc Networks*, 12(1): 243–249.

Kohavi, R. (1995). A Study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection. In *14th International Joint Conference on Artificial Intelligence*, pages 1137–1143, San Francisco, USA. Morgan Kaufmann Publishers Inc.

Krause, C. M. and Zhang, L. (2019). Short-term travel behavior prediction with GPS, land use, and point of interest data. *Transportation Research Part B: Methodological*, 123: 349–361.

Lecun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553): 436–444.

Leu, J.-S. and Zhu, Z.-Y. (2015). Regression-based parking space availability prediction for the Ubike system. *IET Intelligent Transport Systems*, 9(3): 323–332.

Li, J., Li, J., and Zhang, H. (2018). Deep learning based parking prediction on cloud platform. In *4th International Conference on Big Data Computing and Communications*, pages 132–137, Chicago, USA. IEEE.

Liu, K. S., Gao, J., Wu, X., and Lin, S. (2018). On-street parking guidance with real-time sensing data for smart cities. In *15th Annual IEEE International Conference on Sensing, Communication, and Networking*, pages 1–9, Hong Kong, China. IEEE.

Lu, R., Lin, X., Zhu, H., and Shen, X. (2009). SPARK: A new VANET-based smart parking scheme for large parking lots. In *Annual Joint Conference INFOCOM*, pages 1413–1421, Rio de Janeiro, Brazil. IEEE.

Luo, L. (2010). Research on the transportation demand prediction with land use model. In *International Conference of Logistics Engineering and Management*, pages 61–68, Chengdu, China. American Society of Civil Engineers.

Mei, Z., Zhang, W., Zhang, L., and Wang, D. (2019). Real-time multistep prediction of public parking spaces based on Fourier transform–least squares support vector regression. *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations*, pages 68–80.

Monteiro, F. V. and Ioannou, P. (2018). On-Street Parking Prediction Using Real-Time Data. In *21st International Conference on Intelligent Transportation Systems*, pages 2478–2483, Maui, USA. IEEE.

OpenStreetMap (2019a). Highway data, key:highway. `https://download.geofabrik.de/north-america/us/california/norcal.html` (Accessed: 14/05/2019).

OpenStreetMap (2019b). Railway data, key:railway. `https://download.geofabrik.de/north-america/us/california/norcal.html` (Accessed: 14/05/2019).

Oshiro, T. M. and Perez, P. S. (2012). How Many Trees in a Random Forest? In *8th International Conference on Machine Learning and Data Mining in Pattern Recognition*, pages 154–168, Berlin, Germany. Springer.

Peng, L. and Li, H. (2016). Searching parking spaces in urban environments based on non-stationary poisson process analysis. In *19th International Conference on Intelligent Transportation Systems*, pages 1951–1956, Rio de Janeiro, Brazil. IEEE.

Pengzi, C., Jingshuai, Y., Li, Z., Chong, G., and Jian, S. (2017). Service Data Analyze for the Available Parking Spaces in Different Car parks and Their Forecast Problem. In *International Conference on Management Engineering, Software Engineering and Service Sciences*, pages 85–89, Wuhan, China. Association for Computing Machinery.

Pflügler, C., Köhn, T., Schreieck, M., Wiesche, M., and Krcmar, H. (2016). Predicting the availability of parking spaces with publicly available data. In *46. Jahrestagung der Gesellschaft für Informatik*, pages 361–374, Klagenfurt, Austria. Gesellschaft für Informatik.

Qolomany, B., Al-Fuqaha, A., Benhaddou, D., and Gupta, A. (2017). Role of Deep LSTM Neural Networks and Wi-Fi Networks in Support of Occupancy Prediction in Smart Buildings. In *19th International Conference on High Performance Computing and Communications*, pages 50–57, Bangkok, Thailand. IEEE.

R Core Team (2019). R: A language and environment for statistical computing. `https://www.r-project.org/`.

Rajabioun, T., Foster, B., and Ioannou, P. (2013). Intelligent parking assist. In *21st Mediterranean Conference on Control and Automation*, pages 1156–1161, Platanias-Chania, Greece. IEEE.

Rajabioun, T. and Ioannou, P. (2015). On-Street and Off-Street Parking Availability Prediction Using Multivariate Spatiotemporal Models. *IEEE Transactions on Intelligent Transportation Systems*, 16(5): 2913–2924.

Richter, F., Martino, S. D., and Mattfeld, D. C. (2014). Temporal and Spatial Clustering for a Parking Prediction Service. In *26th International Conference on Tools with Artificial Intelligence*, pages 278–282, Limassol, Cyprus. IEEE.

San Francisco Municipal Transportation Agency (2014). sfPark. `https://www.sfmta.com/projects/sfpark-pilot-program` (Accessed: 01/07/2019).

Sarlas, G. and Axhausen, K. W. (2016). Research Collection. In *16th Swiss Transport Research Conference*, pages 1–23, Ascona, Switzerland. Swiss Transport Research Conference.

Shao, W., Zhang, Y., Guo, B., Qin, K., Chan, J., and Salim, F. D. (2019). Parking Availability Prediction with Long Short Term Memory Model. In *13th International Conference on Green, Pervasive, and Cloud Computing*, pages 124–137, Wuhan, China. Springer.

Shin, J.-H. and Jun, H.-B. (2014). A study on smart parking guidance algorithm. *Transportation Research Part C: Emerging Technologies*, 44: 299–317.

Shoup, D. C. (2006). Cruising for parking. *Transport Policy*, 13(6): 479–486.

Shoup, D. C. (2007). Cruising for Parking. *Access*, 1(30): 16–23.

SmartSantander (2015). Smart Santander. `http://www.smartsantander.eu/` (Accessed: 01/07/2019).

Stolfi, D. H., Alba, E., and Yao, X. (2017). Predicting car park occupancy rates in smart cities. In *2nd International Conference on Smart Cities*, pages 107–117, Malaga, Spain. Springer.

Stolfi, D. H., Alba, E., and Yao, X. (2019). Can I Park in the City Center? Predicting Car Park Occupancy Rates in Smart Cities. *Journal of Urban Technology*, pages 1–15.

Sun, M., Li, Z., Peng, L., Li, H., and Fang, X. (2018). FLOPS: An Efficient and High-precision Prediction on Available Parking Spaces in a Long Time-span. In *21st International Conference on Intelligent Transportation Systems*, pages 2937–2942, Maui, USA. IEEE.

Talen, E. and Anselin, L. (1998). Assessing spatial equity: An evaluation of measures of accessibility to public playgrounds. *Environment and Planning A*, 30(4): 595–613.

Tiedemann, T., Vögele, T., Krell, M. M., Metzen, J. H., and Kirchner, F. (2015). Concept of a Data Thread Based Parking Space Occupancy Prediction in a Berlin Pilot Region. In *Workshop on AI for Transportation*, pages 58–63, Austin, USA. AAAI.

TripAdvisor (2019). San Francisco Attractions. `https://www.tripadvisor.com/Attractions-g60713-Activities-a_allAttractions.true-San_Francisco_California.html` (Accessed: 01/07/2019).

van der Waerden, P., Timmermans, H., and de Bruin-Verhoeven, M. (2015). Car drivers' characteristics and the maximum walking distance between parking facility and final destination. *Journal of Transport and Land Use*, 10(1): 1–11.

van Ommeren, J. N., Wentink, D., and Rietveld, P. (2012). Empirical evidence on cruising for parking. *Transportation Research Part A: Policy and Practice*, 46(1): 123–130.

Vlahogianni, E. I., Kepaptsoglou, K., Tsetsos, V., and Karlaftis, M. G. (2016). A Real-Time Parking Prediction System for Smart Cities. *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations*, 20(2): 192–204.

Wang, Z., Yi, J., Liu, J., and Zhang, X. (2007). Study on the control strategy of parking guidance system. In *International Conference on Service Systems and Service Management*, pages 1–4, Chengdu, China. IEEE.

Werbos, P. J. (1990). Backpropagation Through Time: What It Does and How to Do It. *Proceedings of the IEEE*, 78(10): 1550 – 1560.

Wilson, A. G. (2000). *Complex Spatial Systems: The Modelling Foundations of Urban and Regional Analysis*. Routledge.

Wu, E. H.-K., Sahoo, J., Liu, C.-Y., Jin, M.-H., and Lin, S.-H. (2014). Agile Urban Parking Recommendation Service for Intelligent Vehicular Guiding System. *IEEE Intelligent Transportation Systems Magazine*, 6(1): 35–49.

Xiao, J., Lou, Y., and Frisby, J. (2018). How likely am I to find parking? – A practical model-based framework for predicting parking availability. *Transportation Research Part B: Methodological*, 112: 19–39.

Xu, B., Wolfson, O., Yang, J., Stenneth, L., Yu, P. S., and Nelson, P. C. (2013). Real-time street parking availability estimation. In *14th International Conference on Mobile Data Management*, pages 16–25, Milan, Italy. IEEE.

Yang, Z., Liu, H., and Wang, X. (2003). The research on the key technologies for improving efficiency of parking guidance system. In *International Conference on Intelligent Transportation Systems*, pages 1177–1182, Shanghai, China. IEEE.

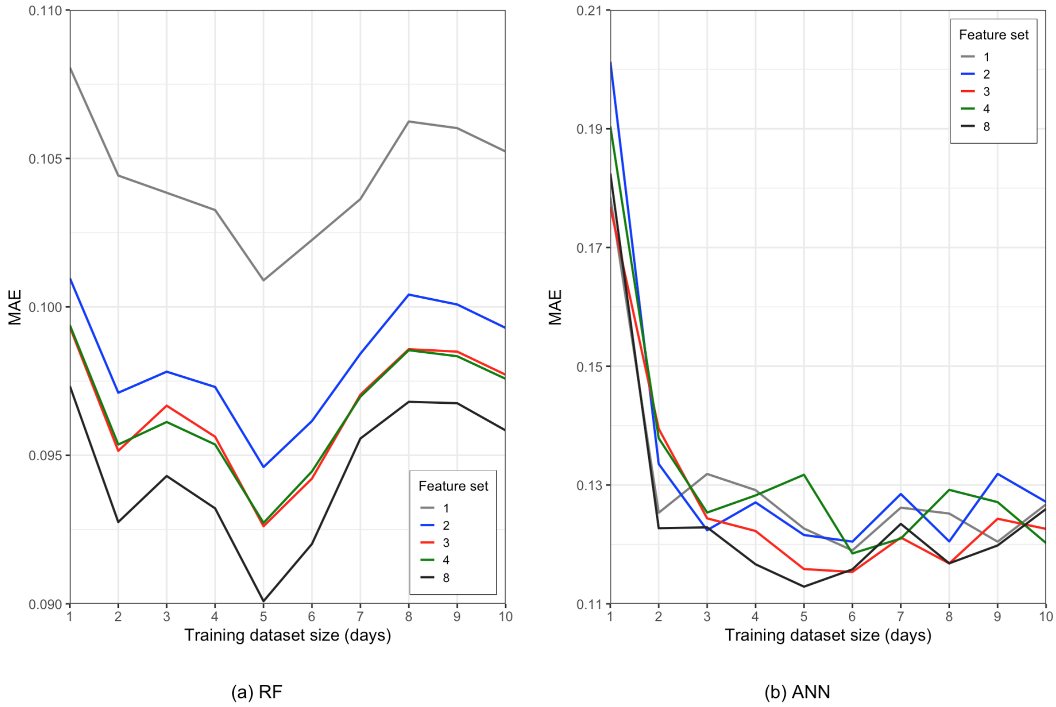
Yu, F., Guo, J., Zhu, X., and Shi, G. (2015). Real Time Prediction of Unoccupied Parking Space Using Time Series Model. In *3rd International Conference on Transportation Information and Safety*, pages 370–374, Wuhan, China. IEEE.

Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, 50: 159–175.

Zheng, Y., Rajasegarar, S., and Leckie, C. (2015). Parking availability prediction for sensor-enabled car parks in smart cities. In *10th International Conference on Intelligent Sensors, Sensor Networks and Information Processing*, pages 1–6, Singapore. IEEE.

Appendix A

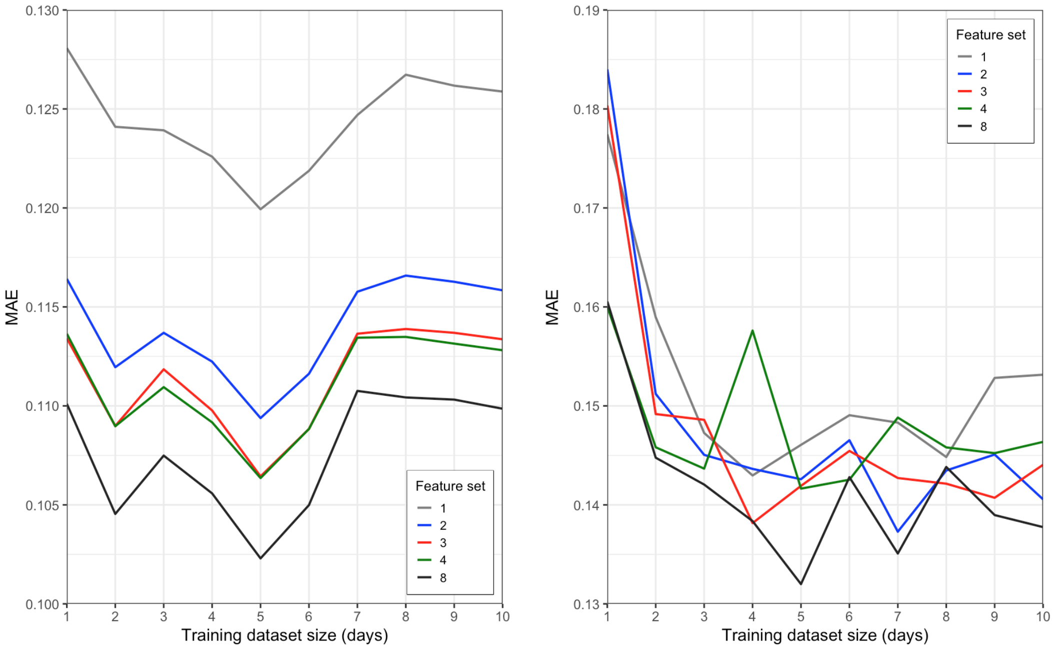
Figures



(a) RF

(b) ANN

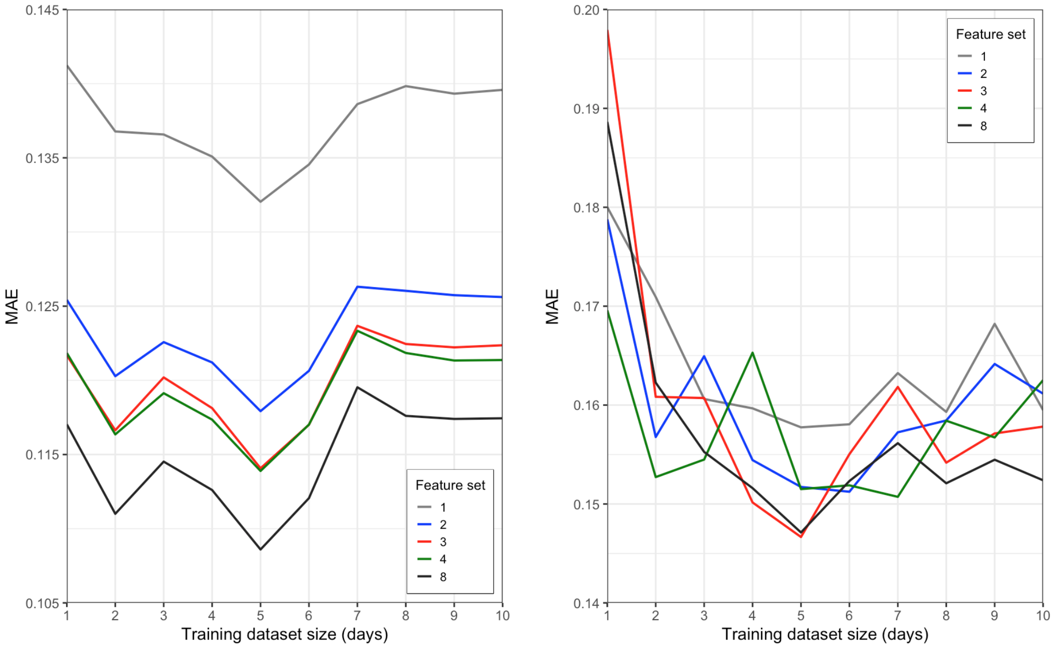
FIGURE A.1: Performance as a function of the training dataset size. Comparison between (a) Random Forest (RF) and (b) Artificial Neural Network (ANN) algorithms. 2 step ahead prediction.



(a) RF

(b) ANN

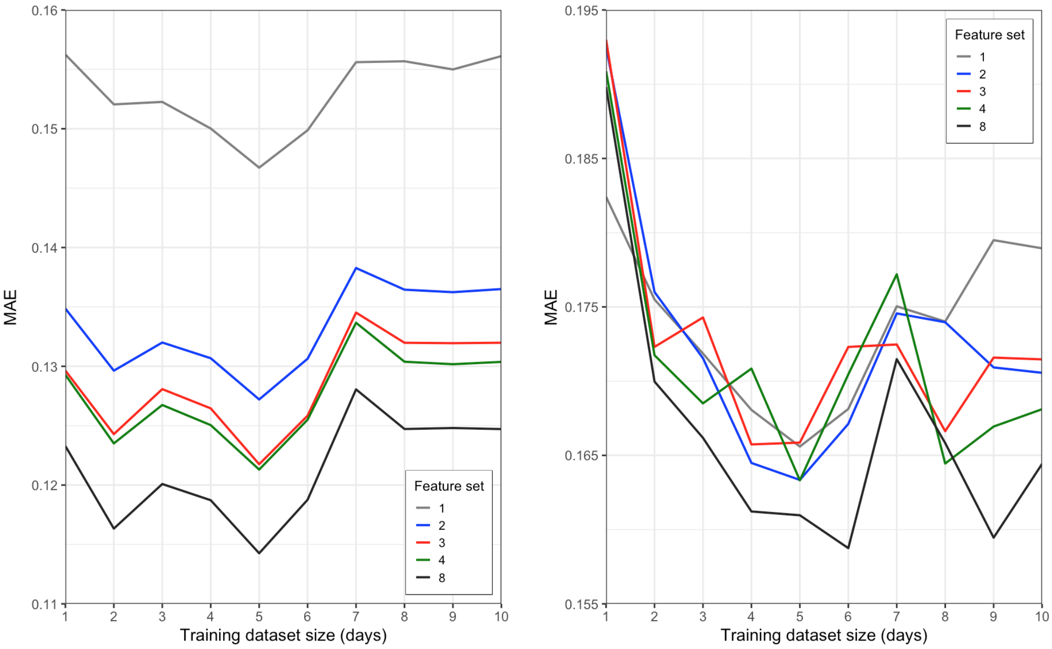
FIGURE A.2: Performance as a function of the training dataset size. Comparison between (a) Random Forest (RF) and (b) Artificial Neural Network (ANN) algorithms. 3 step ahead prediction.



(a) RF

(b) ANN

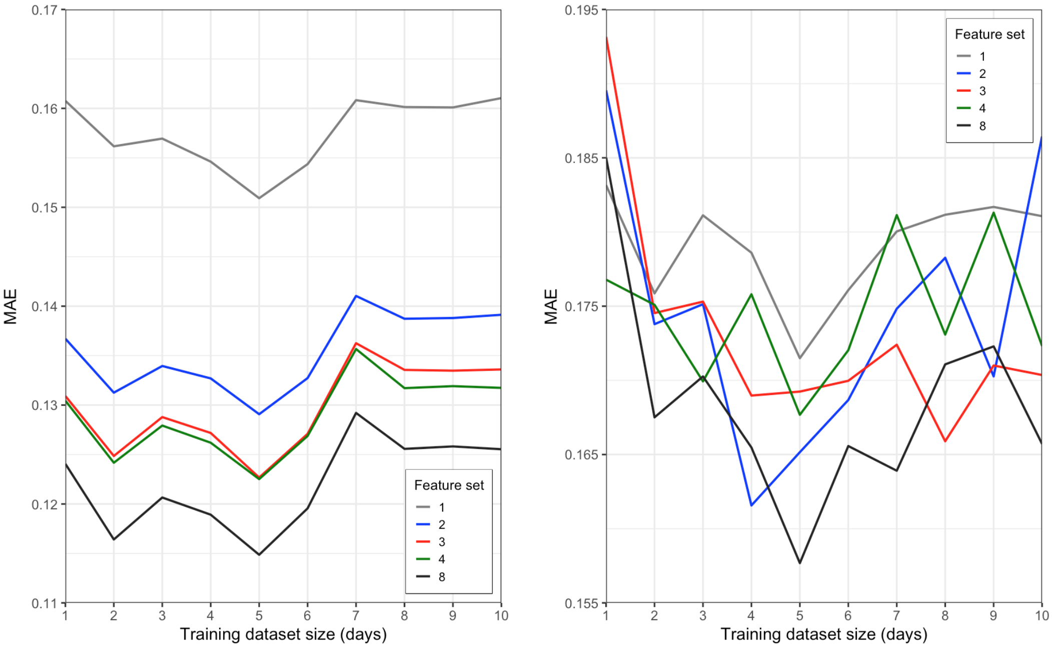
FIGURE A.3: Performance as a function of the training dataset size. Comparison between (a) Random Forest (RF) and (b) Artificial Neural Network (ANN) algorithms. 4 step ahead prediction.



(a) RF

(b) ANN

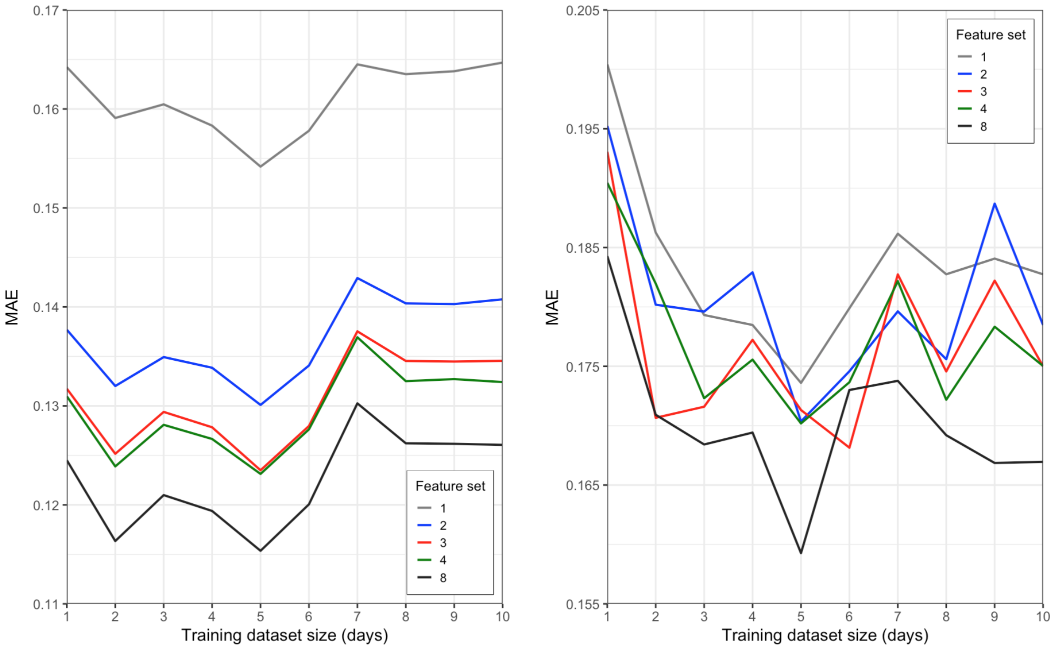
FIGURE A.4: Performance as a function of the training dataset size. Comparison between (a) Random Forest (RF) and (b) Artificial Neural Network (ANN) algorithms. 6 step ahead prediction.



(a) RF

(b) ANN

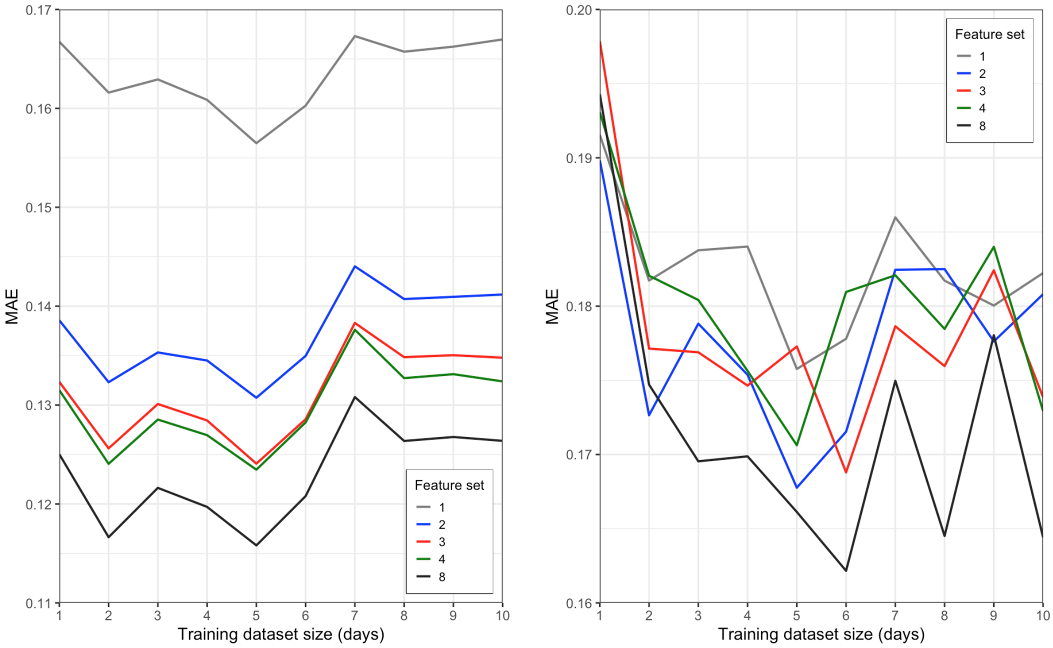
FIGURE A.5: Performance as a function of the training dataset size. Comparison between (a) Random Forest (RF) and (b) Artificial Neural Network (ANN) algorithms. 7 step ahead prediction.



(a) RF

(b) ANN

FIGURE A.6: Performance as a function of the training dataset size. Comparison between (a) Random Forest (RF) and (b) Artificial Neural Network (ANN) algorithms. 8 step ahead prediction.



(a) RF

(b) ANN

FIGURE A.7: Performance as a function of the training dataset size. Comparison between (a) Random Forest (RF) and (b) Artificial Neural Network (ANN) algorithms. 9 step ahead prediction.

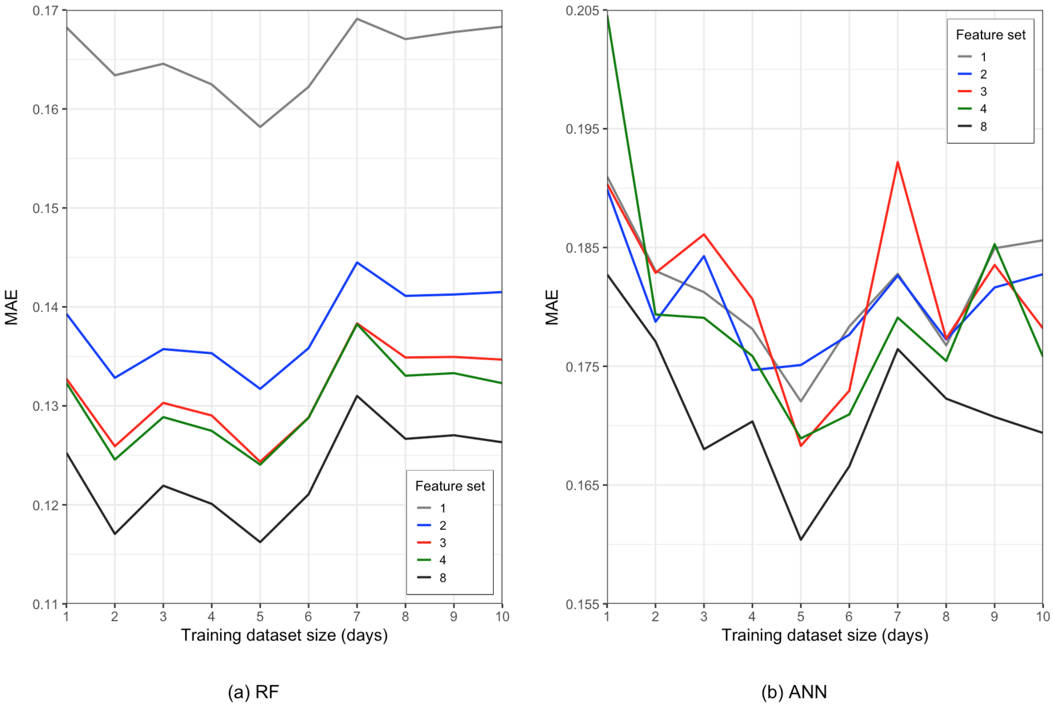


FIGURE A.8: Performance as a function of the training dataset size. Comparison between (a) Random Forest (RF) and (b) Artificial Neural Network (ANN) algorithms. 10 step ahead prediction.

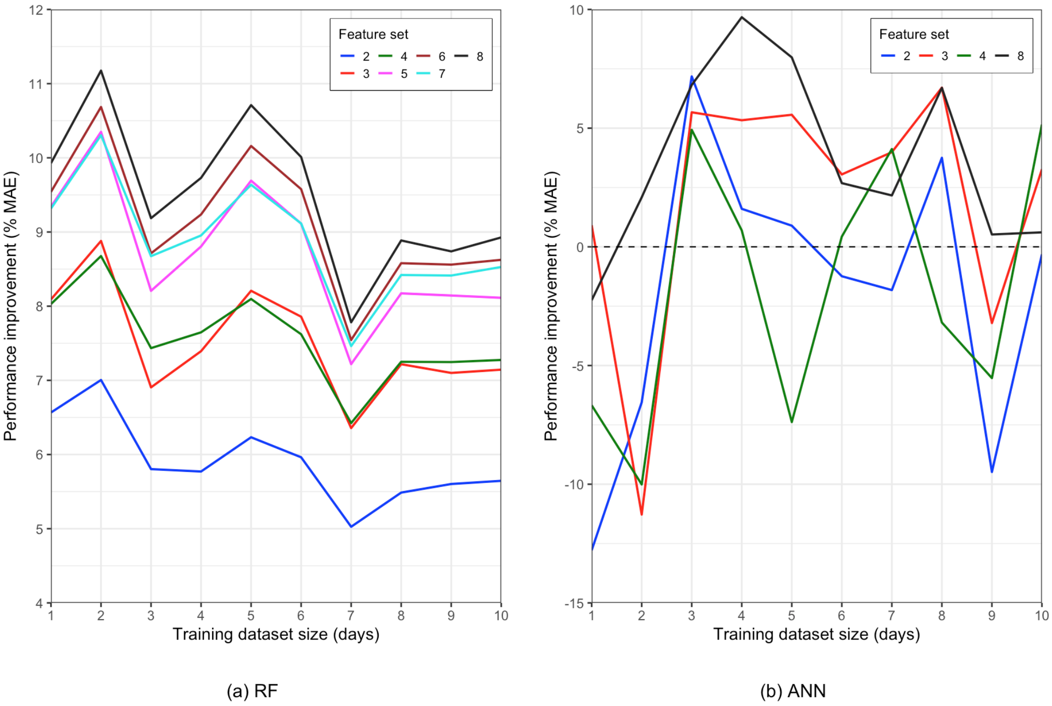
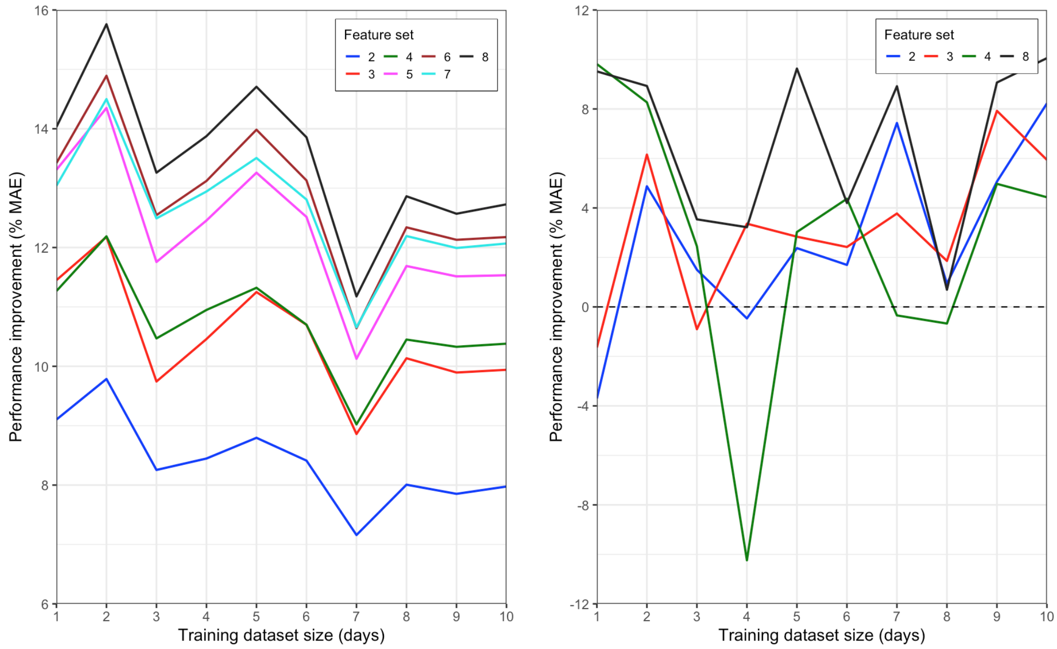


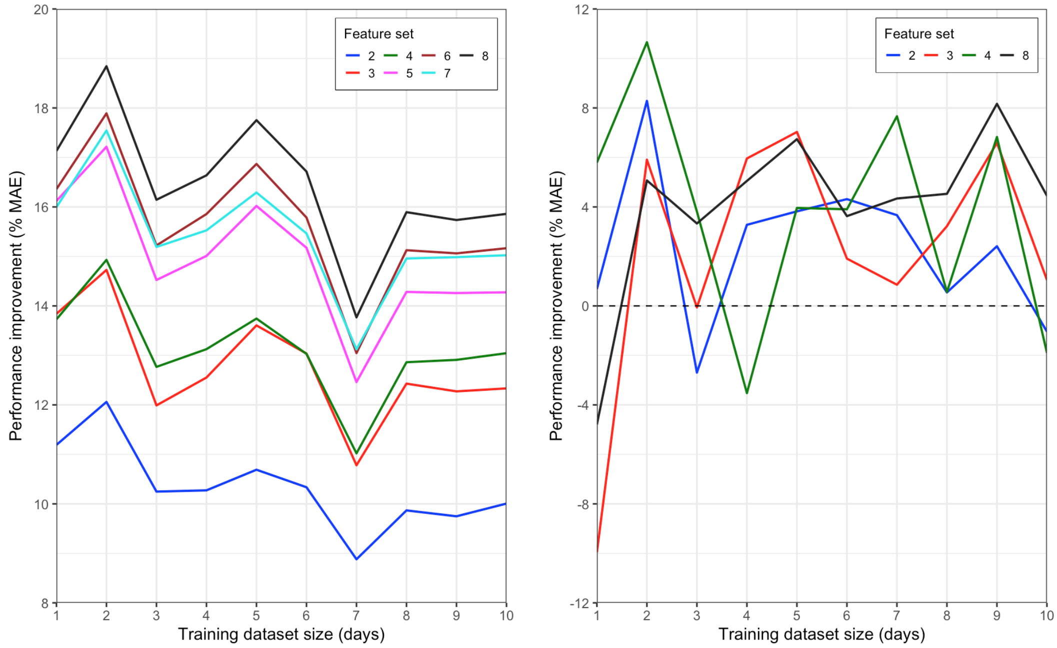
FIGURE A.9: Relative performance improvement (feature sets 2 – 8 compared to the baseline) as a function of the training dataset size. 2 step ahead prediction. Comparison of (a) Random Forest (RF) and (b) Artificial Neural Network (ANN) algorithms.



(a) RF

(b) ANN

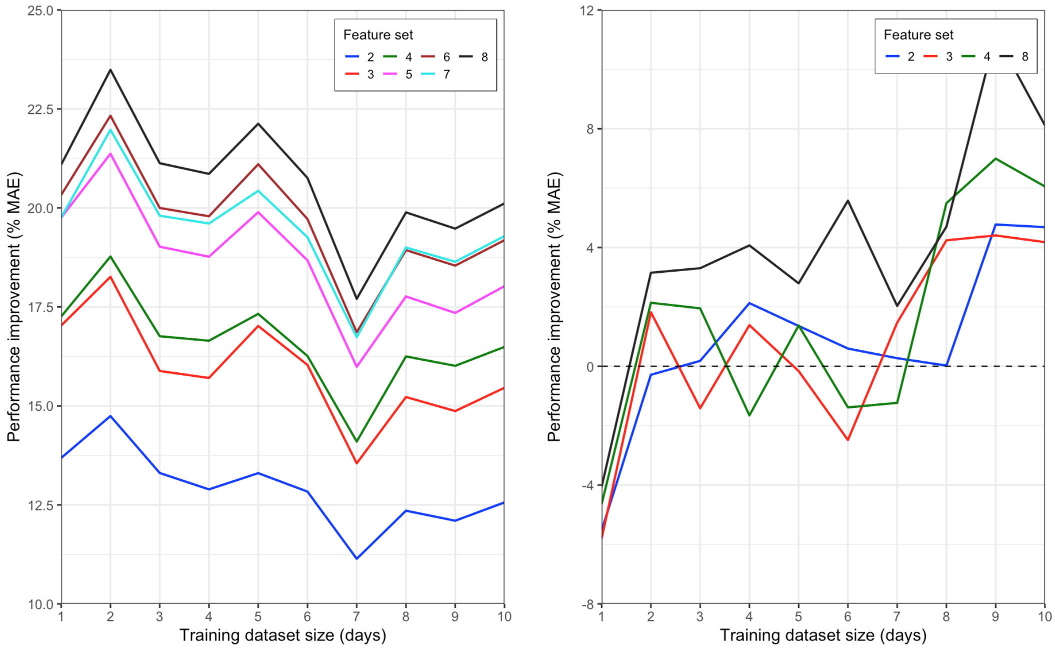
FIGURE A.10: Relative performance improvement (feature sets 2 – 8 compared to the baseline) as a function of the training dataset size. 3 step ahead prediction. Comparison of (a) Random Forest (RF) and (b) Artificial Neural Network (ANN) algorithms.



(a) RF

(b) ANN

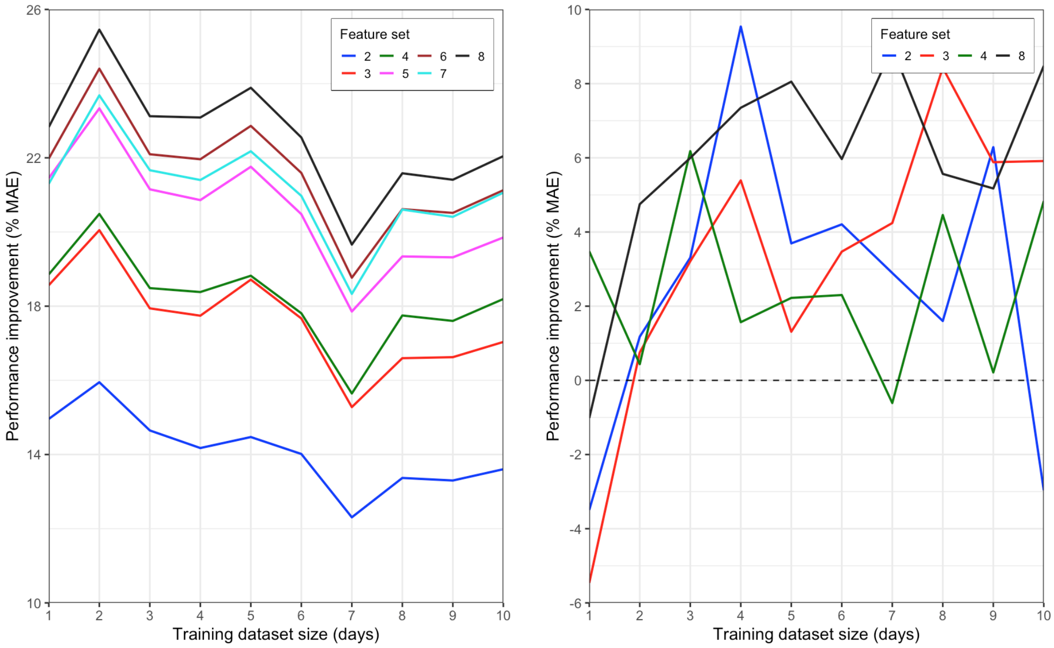
FIGURE A.11: Relative performance improvement (feature sets 2 – 8 compared to the baseline) as a function of the training dataset size. 4 step ahead prediction. Comparison of (a) Random Forest (RF) and (b) Artificial Neural Network (ANN) algorithms.



(a) RF

(b) ANN

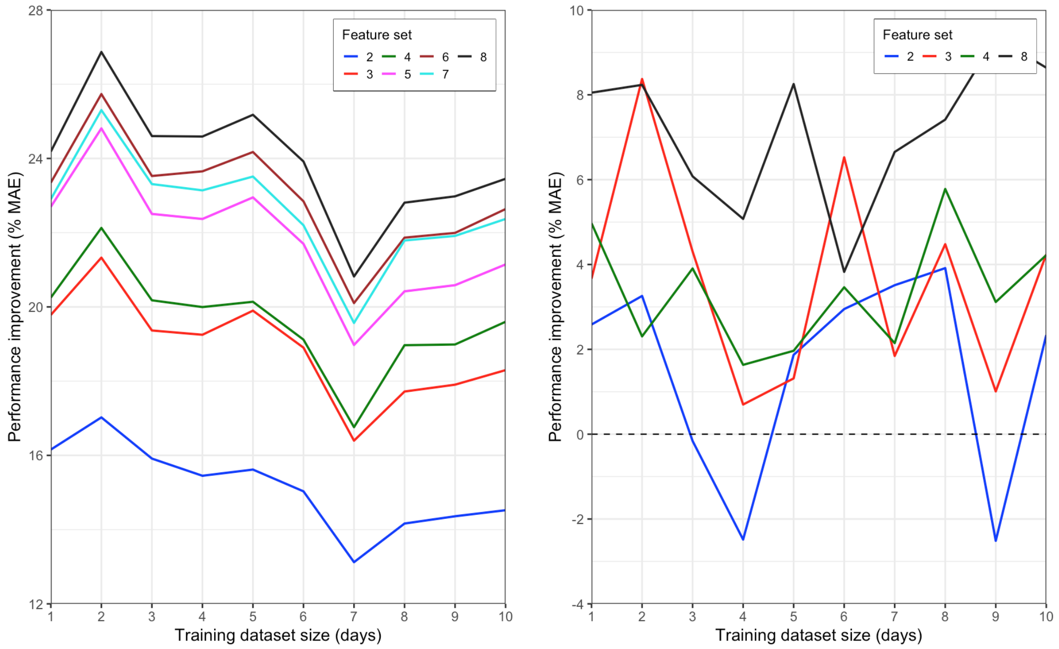
FIGURE A.12: Relative performance improvement (feature sets 2 – 8 compared to the baseline) as a function of the training dataset size. 6 step ahead prediction. Comparison of (a) Random Forest (RF) and (b) Artificial Neural Network (ANN) algorithms.



(a) RF

(b) ANN

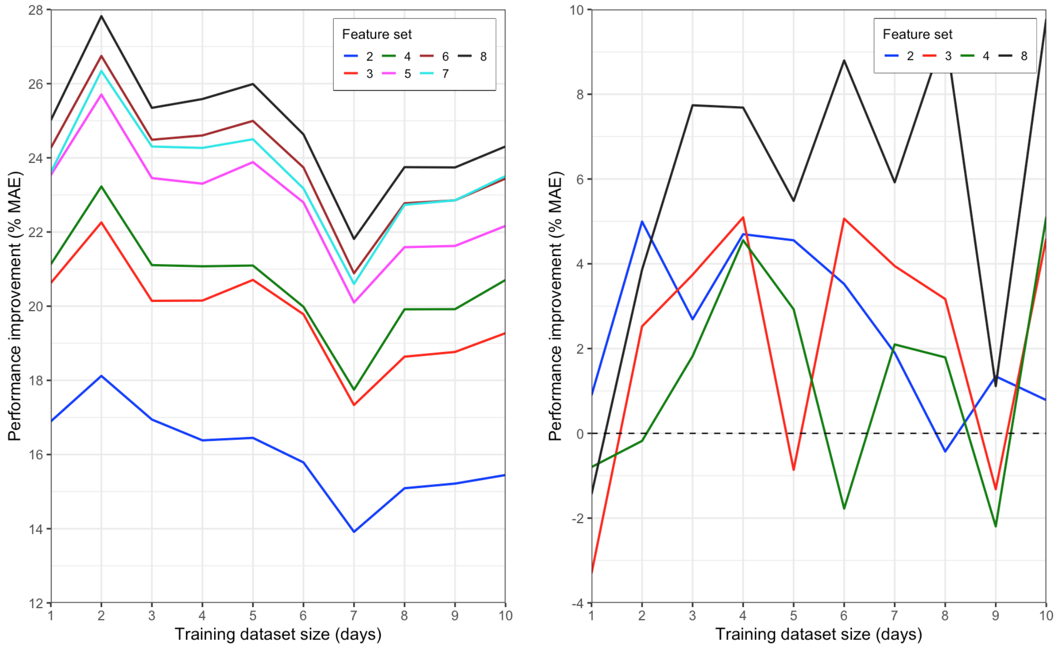
FIGURE A.13: Relative performance improvement (feature sets 2 – 8 compared to the baseline) as a function of the training dataset size. 7 step ahead prediction. Comparison of (a) Random Forest (RF) and (b) Artificial Neural Network (ANN) algorithms.



(a) RF

(b) ANN

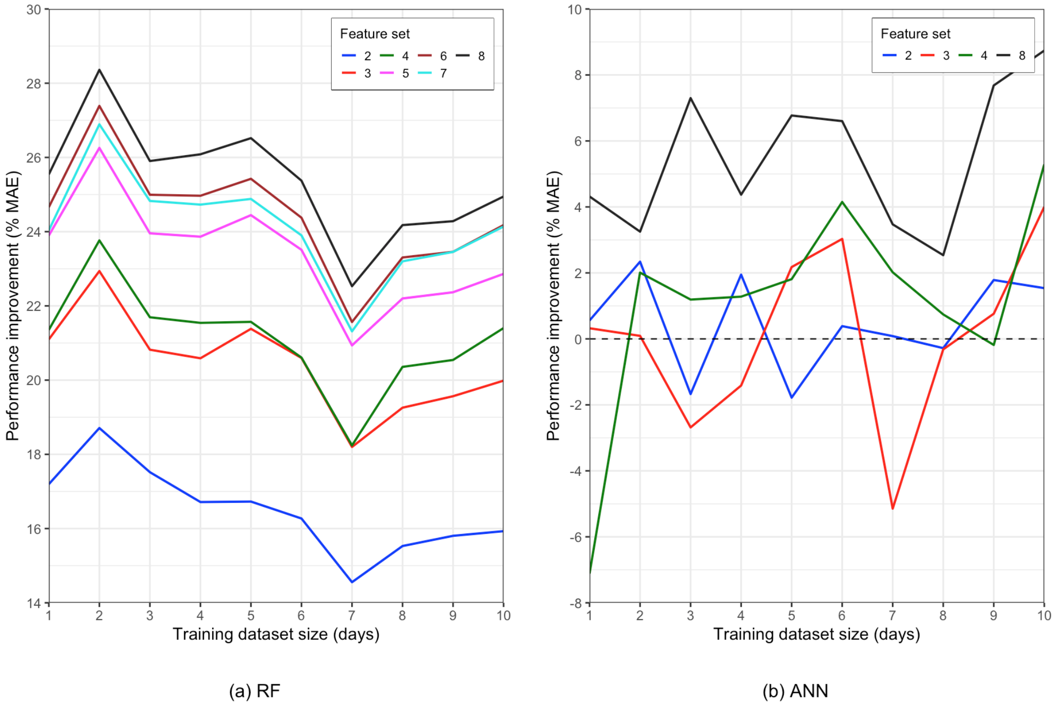
FIGURE A.14: Relative performance improvement (feature sets 2 – 8 compared to the baseline) as a function of the training dataset size. 8 step ahead prediction. Comparison of (a) Random Forest (RF) and (b) Artificial Neural Network (ANN) algorithms.



(a) RF

(b) ANN

FIGURE A.15: Relative performance improvement (feature sets 2 – 8 compared to the baseline) as a function of the training dataset size. 9 step ahead prediction. Comparison of (a) Random Forest (RF) and (b) Artificial Neural Network (ANN) algorithms.



(a) RF

(b) ANN

FIGURE A.16: Relative performance improvement (feature sets 2 – 8 compared to the baseline) as a function of the training dataset size. 10 step ahead prediction. Comparison of (a) Random Forest (RF) and (b) Artificial Neural Network (ANN) algorithms.

Appendix B

Tables

TABLE B.1: Random forest performance in terms of Mean Absolute Error (MAE), Mean Squared Error (MSE) and coefficient of determination (R^2) as a function of the training dataset size. Values for feature sets 1 and 8 on a 1 step ahead prediction.

	Feature set 1			Feature set 8		
Training dataset size (days)	**MAE**	**MSE**	**R^2**	**MAE**	**MSE**	**R^2**
1	0.075	0.012	0.842	0.071	0.010	0.862
2	0.072	0.011	0.852	0.068	0.009	0.872
3	0.071	0.010	0.854	0.068	0.009	0.870
4	0.070	0.010	0.853	0.067	0.009	0.870
5	0.069	0.010	0.857	0.065	0.009	0.875
6	0.069	0.010	0.859	0.066	0.009	0.876
7	0.069	0.010	0.869	0.067	0.009	0.882
8	0.072	0.011	0.861	0.070	0.010	0.876
9	0.072	0.011	0.863	0.070	0.010	0.876
10	0.071	0.011	0.861	0.069	0.010	0.876

TABLE B.2: Artificial neural network performance in terms of Mean Absolute Error (MAE), Mean Squared Error (MSE) and coefficient of determination (R^2) as a function of the training dataset size. Values for feature sets 1 and 8 on a 1 step ahead prediction.

	Feature set 1			Feature set 8		
Training dataset size (days)	**MAE**	**MSE**	**R^2**	**MAE**	**MSE**	**R^2**
1	0.138	0.033	0.700	0.135	0.034	0.578
2	0.109	0.021	0.768	0.096	0.017	0.816
3	0.090	0.015	0.823	0.092	0.015	0.828
4	0.096	0.017	0.810	0.090	0.015	0.823
5	0.083	0.013	0.832	0.082	0.013	0.836
6	0.080	0.012	0.829	0.090	0.015	0.828
7	0.081	0.013	0.847	0.081	0.013	0.854
8	0.089	0.015	0.830	0.086	0.014	0.837
9	0.085	0.014	0.833	0.093	0.016	0.836
10	0.081	0.013	0.841	0.085	0.014	0.844

TABLE B.3: Random forest performance in terms of Mean Absolute Error (MAE), Mean Squared Error (MSE) and coefficient of determination (R^2) as a function of the training dataset size. Values for feature sets 1 and 8 on a 10 step ahead prediction.

	Feature set 1			Feature set 8		
Training dataset size (days)	**MAE**	**MSE**	**R^2**	**MAE**	**MSE**	**R^2**
1	0.168	0.045	0.395	0.125	0.028	0.628
2	0.163	0.042	0.420	0.117	0.024	0.683
3	0.165	0.044	0.395	0.122	0.027	0.635
4	0.162	0.043	0.403	0.120	0.026	0.648
5	0.158	0.040	0.420	0.116	0.023	0.673
6	0.162	0.042	0.410	0.121	0.025	0.657
7	0.169	0.045	0.411	0.131	0.030	0.621
8	0.167	0.044	0.436	0.127	0.027	0.669
9	0.168	0.044	0.440	0.127	0.027	0.671
10	0.168	0.045	0.425	0.126	0.027	0.662

TABLE B.4: Artificial neural network performance in terms of Mean Absolute Error (MAE), Mean Squared Error (MSE) and coefficient of determination (R^2) as a function of the training dataset size. Values for feature sets 1 and 8 on a 10 step ahead prediction.

	Feature set 1			Feature set 8		
Training dataset size (days)	**MAE**	**MSE**	**R^2**	**MAE**	**MSE**	**R^2**
1	0.191	0.057	0.288	0.183	0.054	0.307
2	0.183	0.050	0.341	0.177	0.049	0.352
3	0.181	0.051	0.301	0.168	0.045	0.379
4	0.178	0.050	0.310	0.170	0.046	0.385
5	0.172	0.046	0.330	0.160	0.041	0.411
6	0.178	0.049	0.313	0.167	0.044	0.416
7	0.183	0.051	0.343	0.176	0.051	0.370
8	0.177	0.049	0.388	0.172	0.046	0.418
9	0.185	0.052	0.354	0.171	0.046	0.438
10	0.186	0.053	0.324	0.169	0.045	0.421

Appendix C

Scripts

All R code can be accessed on the GitHub repository under following link: `https://github.com/michaelbalmer/Master-thesis`.

Personal declaration

I hereby declare that the submitted thesis is the result of my own, independent work. All external sources are explicitly acknowledged in the thesis.

Signature:

Date: 31.01.2020