h_da
HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES
fbi
FACHBEREICH INFORMATIK

h_da
HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES
fbmn
FACHBEREICH MATHEMATIK
UND NATURWISSENSCHAFTEN

Darmstadt University of Applied Sciences

Faculties Computer Science & Mathematics and Sciences

Forecasting Traffic Congestion States based on Motorway Grid Cells using Floating Car Data

Thesis for the acquisition of the academic degree
Master of Science (M. Sc.)
in the degree programme Data Science

submitted by Karen Schulz

Supervisor: Prof. Dr. Eva Brucherseifer, Faculty of Computer Science
Co-Supervisor: Prof. Dr. Antje Jahn, Faculty of Mathematics and Sciences

Date of Issue: 01.09.2020
Date of Submission: 01.03.2021

Selbstständigkeitserklärung

Ich versichere hiermit, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die im Literaturverzeichnis angegebenen Quellen benutzt habe. Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten oder noch nicht veröffentlichten Quellen entnommen sind, sind als solche kenntlich gemacht. Die Zeichnungen oder Abbildungen in dieser Arbeit sind von mir selbst erstellt worden oder mit einem entsprechenden Quellennachweis versehen. Diese Arbeit ist in gleicher oder ähnlicher Form noch bei keiner anderen Prüfungsbehörde eingereicht worden.

Datum Ort Unterschrift

Abstract (English)

Congestion is a major and increasingly limiting factor for mobility on motorway networks. Using floating car data, congestion states were usually predicted for road segments that were identified by an additional map-matching tool. Not using a map-matcher, 12 billion floating car data observations were cell-wise grouped into two direction classes representing two directions of a motorway. To the best knowledge of the author, the direction distinctive grid-based approach for assigning floating car data to motorway segments is proposed for the first time in this study. The well-known random forest classification algorithm was utilised for developing and forecasting models for single segments and the segment's collective of 1,000 motorway segments and 45 million observations. Evaluation was based on the metrics F_1-score, misclassification rate, and Bookmaker Informedness. Heuristics based on the average velocity of all motorists at specific points on the motorways served as ground truth for forecasting a segment's congestion state into one of the two classes: *free-flowing* and *congestion*. Whole grid forecasting models delivered better results in comparison to single segment models for four highly congested motorway segments. Major influential factors for the five-minute forecast of the segment collective were features regarding the velocity and the traffic count. Whole grid models are seemingly capable of adding value to the congestion state forecast in 5, 10, 20, 30, and 60 minutes in the future in the whole of North Rhine-Westphalia by considerably exceeding the F_1 and Bookmaker Informedness baseline scores for the 1,000 sample segments. The computational effort was more than 30% lower when using the direction distinctive grid-based approach in comparison to a map-matcher approach for assigning road segments to 2.6 million floating car data observations.

Keywords: traffic congestion, FCD, grid, direction-distinctive, segment, ITS, RF

Abstract (German)

Stau ist ein starker und limitierender Faktor für die Mobilität auf Autobahnen. Bei der Nutzung von floating car data werden üblicherweise Stauvorhersagen auf Straßensegmenten gemacht, welche durch einen sogenannten map-matcher mit den floating car data-Punkten verknüpft werden. Ein map-matcher wurde in dieser Arbeit nicht genutzt, sondern die 12 Billionen gruppierte floating car data-Punkte wurden richtungsunterscheidenden Zellen in einem Netz zugeordnet, welche zwei Richtungen einer Autobahn darstellen. Nach bestem Wissen der Autorin wurde der Ansatz zum ersten Mal in dieser Arbeit vorgeschlagen. Der bekannte random forest Klassifizierungsalgorithmus wurde angewandt, um Vorhersagemodelle für einzelne Segmente und für alle 1000 Segmente im Verbund mit 45 Millionen Beobachtungen zu entwickeln. Die Evaluationsmetriken F_1-score, misclassification rate und Bookmaker Informedness wurden zur Evaluierung herangezogen. Auf der mittleren Geschwindigkeit aller Fahrzeuge an spezifischen Punkten basierende Heuristiken dienten als ground truth Stau-Labels mit den Klassen *frei-fließender Verkehr* und *Stau*. Für vier einzelne Segmente mit viel Stau führten auf 700 verschiedenen Segmenten basierende Vorhersagemodelle zu besseren Ergebnissen als Modelle für einzelne Autobahn-Segmente. Im Modell mit den vielen verschiedenen Segmenten waren Geschwindigkeits- und Fahrzeuganzahl-Variablen die stärksten Einflussfaktoren. Solche Modelle scheinen einen zusätzlichen Wert für Stauvorhersagemodelle für bis zu 60 Minuten in der Zukunft generieren zu können, da sie für den betrachteten Datenkorpus erheblich besser waren als baseline-Modelle des F_1-scores und der Bookmaker Informednesss. Der Rechenaufwand war bei dem Ansatz von richtungsunterscheidenden Zellen in einem Netz zur Generierung von Segmenten deutlich geringer als bei dem genutzten map-matcher Ansatz bei 2,6 Millionen floating car data Beobachtungen.

Schlagwörter: Stau, FCD, Netz, richtungsunterscheidend, Segment, ITS, RF

Acknowledgements

I would like to thank my supervisor Prof. Dr. Eva Brucherseifer for constantly discussing my work with me during the time of creation. She inspired my thought processes and uncovered obscurities the work had during its development.

This master thesis was developed in close cooperation with the data analytics team of the company urban mobility innovations [1]. Urban mobility innovations has great expertise in the mobility sector and offers various mobility solutions to customers. A big thank you goes to the whole data analytics team that always had an open ear for me and supported me in the analytical process. I would like to express my great thanks to Dr. Stefan Radomski in particular who majorly supported me and also doing so even when he was occupied with a lot of other work as well.

The dataset Verkehrsinformationszentrale Nordrhein-Westfalen utilized for extracting ground truth data and details regarding it were gladly received from Volker Gronau and Tilo Voigt of Landesbetrieb Straßenbau North Rhine-Westphalia [2].

I thank Dr. Charla Schutte who proofread my work regarding the English language. She helped me develop my English writing skills in this work.

Last but not least, I thank my companions Betül Çetinkaya, Daniela Di Schiena, Anna Giron, Franziska Schmidt, and Björn Severitt from the data science field for discussing the work with me and giving me valuable hints at formerly unclear points in the work.

Contents

List of abbreviations

ANN artificial neural network

ARIMA auto-regressive integrated moving average

bagging bootstrap aggregating

BM Bookmaker Informedness

CNN convolutional neural network

CSV comma-separated values

DT decision tree

FCD floating car data

FCM fuzzy C-means clustering

FN false negatives

FP false positives

fpr false positive rate

GIS geographic information system

GPS global positioning system

HA historical average

HDFS hadoop distributed file system

HMM hidden Markov model

IG information gain

IQR interquartile range

ITS intelligent transportation system

K-NN k-nearest neighbours

LR logistic regression

LSTM long short-term memory

MAE mean absolute error

mcr misclassification rate

ML machine learning

MLlib machine learning library

MLP multilayer perceptron

MRE mean relative error

MSE mean squared error

NRW North Rhine-Westphalia

RF random forest

RMSE root mean squared error

SARIMA seasonal auto-regressive moving average

SVM support vector machine

TN true negatives

tnr true negative rate

TP true positives

tpr true positive rate

VIZ.NRW Verkehrsinformationszentrale Nordrhein-Westfalen

VPR structured vehicle passage record

XML extensible markup language

List of Figures

List of Tables

1. *Introduction*

The introductory part of this work explains why the topic of forecasting traffic congestion states is relevant, why the grid-based approach was examined and what questions can be answered using which methods. A short thesis outline is presented as well.

1.1 Motivation

Almost every motorist has encountered a traffic-congested motorway that is time-consuming and disruptive to plans. The buildup of traffic congestion is due to a higher demand on available road capacity [5]. It can also lead to enlarged CO_2 emissions. Obviously, motorists would like to avoid getting stuck on a traffic-congested road. Drivers can be warned of traffic congestion via radio or a navigation system. Such systems gather congestion information from, for example, a machine learning (ML) congestion state classification algorithm. Identifying and forecasting traffic congestion is one of the aims of intelligent transportation systems (ITSs). ITSs comprise topics improving transportation efficiency through state-of-the-art methods and technologies, e.g. ML algorithms. Forecasting traffic congestion is a non-trivial problem since it is influenced by many parameters and has a spatiotemporal nature. Moreover, motorists in the same traffic situation might even classify the congestion state differently due to their different interpretation of the traffic situation.

This work focuses on the motorway network, which is a congestion-prone road network with generally the most throughput. The North Rhine-Westphalia (NRW) state is used as an experimental area for this work. It is the most congestion-troubled state in Germany - in 2019, 36 % of Germany's traffic congestion and slow-moving traffic events were located in NRW [6].

Floating car data (FCD) is a traffic information data source and floating car data (FCD) ob-

servations are gathered through global positioning system (GPS). Many existing congestion state determination methodologies use a map-matcher. Map-matchers need a geographic information system (GIS) map that divides the road network into road segments. A road segment represents a carriageway part on the motorway in this work since the work focuses on motorways. A map-matcher is generally used for mapping FCD points onto road network segments. ML models can be subsequently utilized to forecast congestion states for each motorway segment.

A disadvantage of the map-matcher approach is that the additional map-matching tool needs to be provided and maintained. Moreover, the GIS map can be outdated and that might then well lead to wrongly assigned road segments. Furthermore, using predefined road segments of the road network can lead to not enough FCD on road segments for generating feature values. Additionally, the assignment process itself uses a complex ML algorithm.

1.2 Objective

This study evaluates the potential of forecasting traffic congestion states for carriageway segments with a scalable ML model based on motorway grid cells. Motorway carriageway segment boundaries were defined by the boundaries of grid cells. The motorway carriageways inside a cell, or in other words, the segments, were identified through a direction distinctive grid-based approach. To the best knowledge of the author, the motorway direction distinctive grid-based approach is proposed for the first time to forecast congestion states on motorways in this study.

A scalable ML model is developed using a high FCD volume. Segments are assigned to FCD through the direction distinctive grid-based approach. The potential is evaluated based on model performances with a different number of regarded motorway segments. Forecasting is performed for the following future time periods: 5, 10, 20, 30, and 60 minutes. The computational effort of assigning segments to FCD points is compared between the motorway direction distinctive grid-based approach and the map-matcher approach. The evaluated segments are distributed on the whole NRWs area. This leads to the validity of the results for a diverse spatial spectrum of motorways. Heuristics based on the average vehicle velocity, gathered by a traffic detector on a carriageway point, determine the ground truth

congestion state for the according motorway segment. Segment boundaries are determined by cell boundaries. The evaluation results are intended to be compared with congestion state prediction methodologies using the map-matcher for assigning FCD to segments and using other ML algorithms.

Grid cells are modelled direction distinctive in this work to relate to the carriageways of a motorway. In other words, two direction classes are generated for each grid cell that have opposite motorway directions. Forecasting congestion states based on segments gathered through motorway direction distinctive grid cells is a map-matcher alternative which resolves shortcomings as described in the section above. Furthermore, the motorway direction distinctive grid-based approach relies on a much simpler basic structure. Features are derived from FCD grouped by segment. The additional challenges of the motorway direction distinctive grid-based approach lie mainly in the distinction of different motorway segments inside a cell and on the impact on feature values of additional FCD points that are not recorded on a motorway. Therefore, robust features must be generated for forecasts.
In this study, congestion states are separated into two classes: *free-flowing* and *congestion*. The reason for this is that only one data source was available for gathering ground truth labels that made the distinction between the two classes. Moreover, the assumption of a major FCD volume from motorways inside cells is made in this work. It is further assumed that traffic detectors, which do not record congestion, observe free-flowing traffic.

1.2.1 Research questions

The research questions accompanying this work are:

- How can FCD be related to carriageway segments of motorways in a grid-based setting?
- What are the differences of the five-minute forecasting performance varied between the single segment and whole grid setting using the motorway direction distinctive grid-based approach for highly congested motorway segments?
- Which engineered features are seemingly valuable for forecasting congestion states using the motorway direction distinctive grid-based approach in the whole grid setting?

- Which congestion-state forecasting time periods seem to be valuable in the whole grid setting using the motorway direction distinctive grid-based approach?
- Can the motorway direction distinctive grid-based approach compete with map-matcher-based approaches regarding the forecasting performance?
- Is the computational effort of the motorway direction distinctive grid-based segment gathering approach lower than the effort of the map-matcher approach?

1.2.2 Contributions

A data processing pipeline was established utilizing FCD and ground truth data to form a huge data corpus. ML models were developed and evaluated based on the data corpus. The experimental steps are outlined below.

Experimental setup:

- Gathering of ground truth congestion state data set
- Development of a data preparation pipeline based on a direction distinctive grid-based segment gathering
- Development of features based on FCD aggregations

Experimental results:

- Description of important model features through
  - Visualisations
  - Expert knowledge
- Testing of model performances for the
  - Single segment model
  - Whole grid model
- Comparison of the computational time between segment assignments for the direction distinctive grid-based and the map-matcher approach

1.2.3 Research boundaries

The following research boundaries were set for this work:

- Differentiation of congestion states into free-flowing and congestion class
- Utilisation of FCD set to forecast congestion states
- No benchmark data set available for FCD data

1.3 Thesis outline

The fundamentals of the thesis are outlined in Chapter 2. They comprise selected automated traffic data acquisition methods, the FCD structure and FCD aggregation methods, the machine learning induction methodology for a binary classification task, and evaluation metrics for a binary classification task.
A literature review can be found in Chapter 3. Key aspects regarding ML congestion state estimation for road segments are identified, condensed, and thoroughly described. Several field studies are described in detail to show the diverse spectrum of experimental setups and to interpret the results according to their setups. Traffic congestion detection based on grid cells is addressed by selected case examples. A differentiation of the grid-based and map-matcher approach is made, and a discussion on further developing the grid-based approach for referring to motorway segments is presented.
Chapter 4 describes the utilized methodology in detail. Key aspects are the general model development framework, grid-based specifics including the direction distinctive grid cell approach as well as the ML methodology.
In Chapter 5, the experimental setup is outlined including a brief profile of the computer infrastructure. The raw and processed data are presented for receiving a thorough overview of the utilized database in the ML models.
Experimental results in Chapter 6 are divided into descriptive insights, forecasting of traffic congestion states, and a comparison of the computational effort using the motorway direction distinctive grid-based and map-matcher approach.
Chapter 7 comprises the conclusion of the thesis. A summary as well as a discussion and suggested future work are provided.

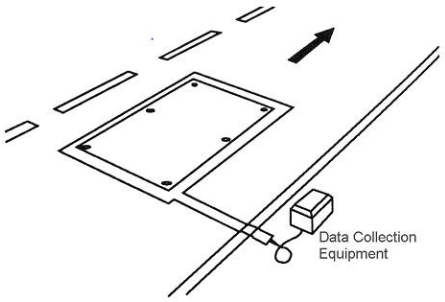
2. *Fundamentals*

The fundamentals of this work are presented in this chapter. They comprise selected automated data acquisition methods that are needed to gather traffic information, to form features as well as the target variable. Features are aggregated based on grid cells in this work and the basic grid-based approach is hence described. The alternative map-matcher approach is outlined as well. Utilized ML algorithms are presented afterwards to understand the modelling setup. Model evaluation is as important as the ML method itself and utilized evaluation metrics are therefore outlined.

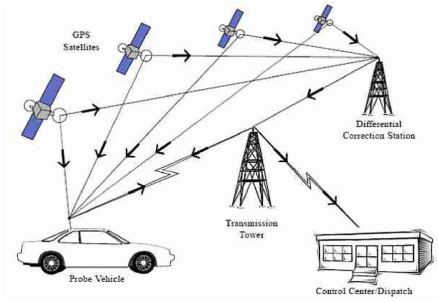
2.1 Selected automated traffic data acquisition methods

Automatic traffic data acquisition forms the foundation of automatically determining traffic congestion states with data-driven approaches. Amongst many data acquisition methods, a popular traditional as well as a popular modern method are introduced.

On the whole, traditional traffic information data sources use detectors located along the roadside [7]. According to Leduc [7], the induction loop technology is a conventional method from this class and basically consists of a data recorder and a sensor placed on or in the road. Leduc further illustrates that induction loops are embedded in roadways through a square formation that generates a magnetic field as can be seen in Figure 2-1a. The information is then transmitted to a counting device placed on the side of the road as can be seen in the figure as well. Two induction loops in a row, so-called double induction loops, are utilized to measure the vehicle's velocity since they have lower measurement error margins compared to single induction loops [8].
Compared to traditional data gathering methods such as induction loop detectors, floating car data (FCD) provides a sample of road user data. In contrast, traffic detectors are able



(a) Single induction loop [9]



(b) FCD gathering through GPS [7]

Figure 2-1: Selected automated traffic data acquisition methods

to capture every vehicle at specific points. Nevertheless, FCD has several advantages such as lower cost, wider coverage, and higher mobility [10]. It can furthermore locate a vehicle across its entire route [7]. FCD is collected through the GPS signal of road users from e.g. mobile devices as can be seen in Figure 2-1b. The position as well as the velocity and the heading, or in other words the compass direction, can be determined through GPS.

2.2 Floating car data

FCD is the main data structure used in this work and therefore described thoroughly in this section. An FCD sample with the typical columns *id*, *latitude*, *longitude*, *datetime exact*, *velocity*, and *heading* can be seen in Table 2.1. Different motorists can be distinguished through their *id* values. The *latitude* and *longitude* determine the location and the *heading* refers to the movement's compass direction in the interval of [0, 360) degrees. The first observation in Table 2.1 belongs to a motorist with *id* 1, driving 80 km/h at 2019-08-01 00:00:23 with

Id	***Datetime exact***	***Latitude***	***Longitude***	***Velocity***	***Heading***
1	2019-08-01 00:00:23	51.404660	7.473666	80	225
1	2019-08-01 00:01:15	51.402817	7.467361	91	226
1	2019-08-01 00:02:01	51.401985	7.465269	94	221
		...			
2	2019-08-01 00:00:58	51.654835	7.035406	107	93
		...			

Table 2.1: Sample floating car data set

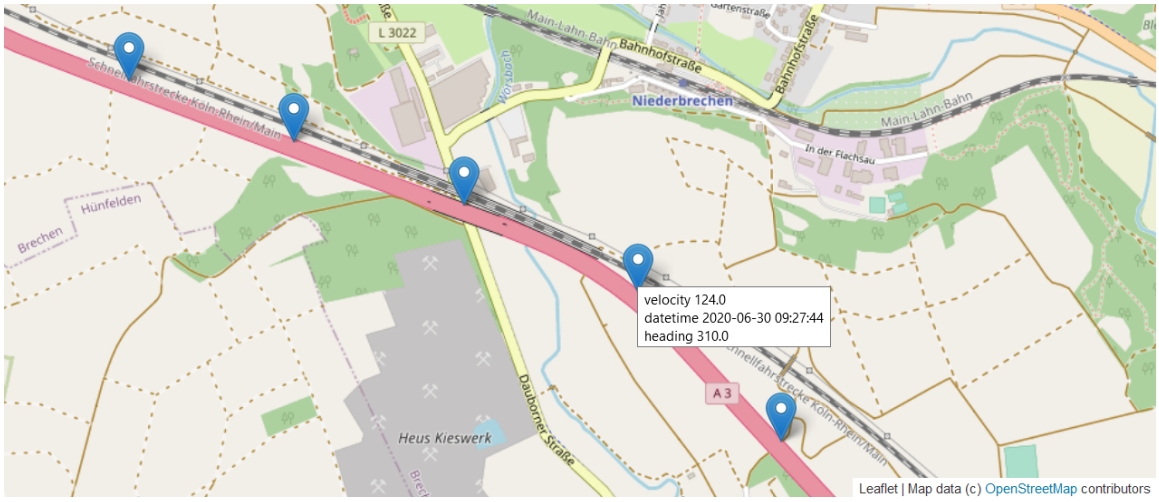


Figure 2-2: Excerpt of a floating car data *id* trace

the GPS coordinates 51.404660 and 7.473666 and a heading of 225 degrees, corresponding to the southwest travelling direction. A second and third observation from the motorist with *id* 1 is shown and the motorist can have an arbitrary number of subsequent observations, represented by the three dots. The next line shows the first observation belonging to a motorist with *id* 2. The final line represents numerous observations from various motorists. Figure 2-2 follows a trace of one *id* or, in other words, of one motorist on the A3 motorway. The blue markers represent the corresponding *latitude* and *longitude* values of the FCD points. Near to the fourth marker from the top is a text box showing the remaining FCD feature values. One can see that the vehicle had a velocity of 124 km/h, drove there in June 2020 in the heading direction of 310 degrees, approximately corresponding with the northwest direction.

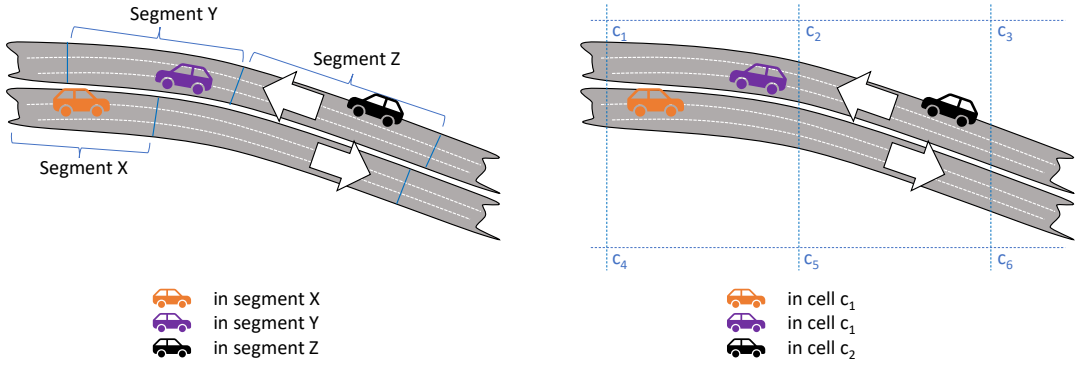
2.3 Floating car data aggregation approaches

Two approaches for assigning FCD points to spatial areas are presented. FCD points can be grouped based on the spatial areas as exemplarily shown in Section 2.3.2.

2.3.1 Map-matcher approach

A map-matcher matches FCD points to road segments. Zeidan et al. [11] stated that map matching is a key processing task in practically all analyses of urban location data as otherwise the findings cannot be related to urban infrastructure. The map-matcher approach

depends on a geographic information system (GIS) map recording motorways, main roads, and other road types. The GIS map represents road segments digitally through geometrical shapes. Figure 2-3a exemplarily shows motorway segments that are separated by blue lines. Segments are distinguished for both motorway carriageways on their own as throughout this work. The car symbols represent FCD points. The orange car drives on motorway segment X, the purple car drives on motorway segment Y, and the black car drives on motorway segment Z as shown in the diagram. A complex ML algorithm is used to assign FCD points to segments. A detailed description of map-matcher algorithms is beyond the scope of this work.



(a) Map-matcher approach

(b) Basic grid-based approach

Figure 2-3: Floating car data aggregation approaches

2.3.2 Grid-based approach

Spatial data such as data from induction loops and floating car data (FCD) can be grouped into disjoint grid cells based on latitude and longitude values. The grid's cell size can be statically allocated. A cell can have various geometric shapes such as rectangle, square, hexagon, or diamond.

The example provided in Figure 2-3b uses rectangular cells instead of segments for aggregating FCD points. The rectangular cells are represented by dashed blue lines and each cell has an index value in its top-left corner. FCD points, represented by the car symbols, are assigned to cells based on their rounded *latitude* and *longitude* value. The orange and the purple car are assigned to cell c_1. The black car is assigned to cell c_2. Generally, FCD points

Index		Features	
Datetime	***Cell middle***	***Avg velocity***	***%50 Velocity***
2019-08-01 00:00:00	51.40, 7.47	88	91
2019-08-01 00:05:00	51.40, 7.47	90	101
	...		
2019-08-01 00:00:00	51.65, 7.04	98	105
	...		

Table 2.2: Grid-based floating car data aggregation

can be located anywhere inside a cell's boundaries to be allocated to that cell.

FCD can be grouped based on its cell assignment. Furthermore, cell feature values can be extracted from temporal and cell-wise grouped FCD values. Table 2.2 presents temporal and grid-based aggregated features based on the composite *datetime* and *cell middle* index. The *datetime* is the five-minute window of the *datetime exact* feature from FCD as seen in Table 2.1. The *cell middle* variable describes the midpoint of a cell, attained by rounded *latitude* and *longitude* values from Table 2.1. The *avg velocity* and *%50 velocity*, the median velocity, are displayed as exemplary features. The feature values in the first line with the composite index 2019-08-01 00:00:00, 51.40, 7.47 were computed by only regarding the first three FCD observations from Table 2.1. The average velocity of the three FCD points was 88 km/h and the median velocity was 91 km/h. The second observation in the table shows feature values for the same cell in the following five-minute window. The other written-out observation displays feature values for the time window of the first observation but has a different cell index.

2.4 Machine learning induction methodology for binary classification

This section outlines ML models that can yield to forecasts for binary classification problems such as the binary *congestion state* determination. Binary ground truth labels for the *congestion state* determination formed a research boundary in this work since only ground truth data was available having two congestion states. The utilised data corpus for ML

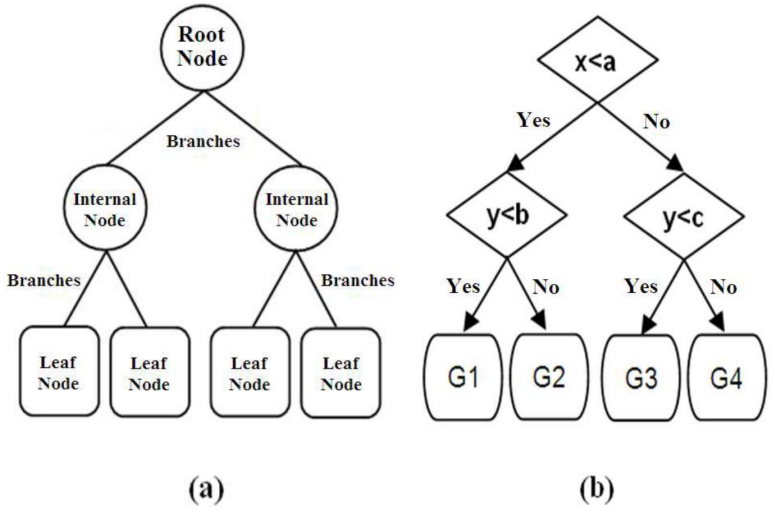


Figure 2-4: Decision tree [3]

models in this work was based on features and ground truth labels derived from FCD and induction loop records respectively. Decision trees (DTs) are presented first as the base for the subsequently described random forest (RF) modelling approach that was used for developing forecasts in this work.

2.4.1 Decision tree

A decision tree is a recursive machine learning (ML) model that can also be interpreted by non-expert users and serves as a solid induction method. The decision tree (DT) procedure was developed in the last century by several independent scientists from the statistical and machine learning field [12].

A decision tree is formed by a root node, internal nodes, leaf nodes and branches, as can be seen in Figure 2-4a. The root node of Figure 2-4b shows the split criterion $x < a$, which is the split leading to maximum information gain (IG). The maximum IG corresponds to the most homogeneous split subsets regarding the target variable. Branches represent a split criterion's values based on the observations. The majority ground truth class of observations from a leaf node is set as prediction class for these observations.

Several decision tree learning algorithms have been developed, such as the popular ID3, C4.5 and CART. DTs can be modelled in a distributed environment as well. The machine learning library (MLlib) of Spark 2.3.2, which is subsequently used, is therefore described in more

detail regarding DTs. MLlib supports the DT binary classification task using categorical and continuous features [13, 14]. It further uses a greedy algorithm for generating binary splits at each node through maximizing the IG. The IG is based on an impurity measure. For classification, the gini impurity is one option and defined in [14]. Training is performed distributedly on partitions of observations. The algorithm scales approximately linearly regarding several important parameters. Details can be found in [14].

Decision trees are capable of capturing non-linearities and feature interactions. Feature importances demonstrate the impact of the features on the target variable and can be interpreted especially by domain experts. The feature importance scores of a ML model are normalized to result in a sum of 1.

Hyperparameters of decision trees in *PySpark* are the maximal depth of the decision tree, minimal observations per node and the minimal IG. The impurity measure and binary classification threshold can be chosen as well.

2.4.2 Random forest

As the wording suggests, a random forest (RF) is an extension of the DT that utilizes several distinct DTs to gain predicting power in reducing the risk of overfitting. It is a bootstrap aggregating (bagging) ensemble model derived from DT individual models.

To form a RF, several decision trees are developed based on parts of the original data set, whereby each part, the tree-wise bootstrap sample, is chosen randomly. Furthermore, a random subset of the features is provided at each node of a tree. The single trees are not post-pruned and may be overfitted to their share of the data set. A combination of the single predictions is generally more robust than a DT prediction regarding overfitting. It can be seen analogous to human's knowledge of the crowd.

The RF algorithm of *Spark's* MLlib [15] is subsequently utilized and therefore shortly described. The algorithm supports both categorical and continuous features for the classification task. The decision tree implementation outlined in the section above is used. Different trees are trained in parallel.

Hyperparameters of DTs are used as hyperparameters for an RF as well. RFs in *PySpark* additionally have a hyperparameter specifying the number of trees.

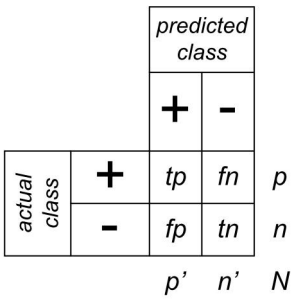


Figure 2-5: Confusion matrix [4]

2.5 Evaluation metrics for binary classification

Evaluation metrics used for evaluating binary classification problems such as the *congestion state* forecast with labels *free-flowing* and *congestion* are introduced. Binary classified data points can be assigned to one of four fields in the confusion matrix of Figure 2-5: true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN) (cf. whole section to [16]).

Selected metrics for binary classification based on the confusion matrix are displayed below. The misclassification rate and accuracy measures form the relation: accuracy $= 1 -$ misclassification rate. In the case of highly unbalanced classes, the misclassification rate (mcr) is very low when deterministically predicting the majority class. Both precision and recall do not regard TN and generally cannot be optimized at the same time due to their contradictory nature. The recall is equal to the true positive rate (tpr). When considering highly unbalanced target data, either the recall or the precision are generally on a very low level. A ML model generally leads to many FP if the negative class is the majority class. A model generally leads to many FN if the positive class is the majority class. The true negative rate (tnr) is generally quite small when the negative class is the majority class. The F-score combines precision and recall through their harmonic mean. Precision and recall are set to be equally important if the parameter $\beta \in [0, \infty)$ equals 1. If the F-score is utilized for highly unbalanced target data, the values lie on a relatively low level since either the recall or the precision measure has small values. The Bookmaker Informedness (BM) combines the tpr and the tnr and subtracts one. The BM evaluation values are generally not on a

higher or lower level when imbalanced target classes are faced.

$$
\begin{aligned}
\text{misclassification rate} &= \frac{\text{FP} + \text{FN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} && \in [0, 1] \\
\text{accuracy} &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} && \in [0, 1] \\
\text{precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}} && \in [0, 1] \\
\text{recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}} && \in [0, 1] \\
\text{tnr} &= \frac{\text{TN}}{\text{TN} + \text{FP}} && \in [0, 1] \\
F_\beta &= (1 + \beta^2) \frac{\text{precision} \times \text{recall}}{(\beta^2 \times \text{precision}) + \text{recall}} && \in [0, 1] \\
\text{BM} &= \frac{\text{TP}}{\text{TP} + \text{FN}} + \frac{\text{TN}}{\text{TN} + \text{FP}} - 1 && \in [-1, 1]
\end{aligned}
$$

3. *Literature review*

This review focuses on studies in the traffic congestion domain and excludes works from related fields such as traffic volume prediction. The review is separated into two main parts presenting the popular field of traffic congestion detection for road segments and the traffic congestion detection for grid cells as this approach was further developed in this work.

3.1 Traffic congestion detection for road segments

ML approaches towards traffic congestion detection for road segments are presented in this section. Only FCD from road segments was used for generating features on the road segments. Firstly, findings from literature are condensed and selected studies are presented in order to show their variety secondly. Thirdly, obstacles for comparing different studies are outlined.

3.1.1 Machine learning traffic congestion state determination characteristics

The determination of congestion states itself is a field of subliminal debate. Kerner [17] proposed three congestion states in his three-phase traffic theory. Studies on determining traffic congestion states differ in their number of congestion states as experiment settings in Section 3.2.2 show. To the best knowledge of the author, the number of congestion states in congestion state determination models range from two [18] up to five [19].

Modelling approaches can be divided into deterministic and non-deterministic/ stochastic techniques [20]. Deterministic methods comprise DT, RF, and logistic regression (LR) (cf. [21, Ch. 4.4]), whereas, for example, fuzzy techniques (cf. [22]) are non-deterministic methods. According to Jabbarpour et al. [23], non-deterministic congestion state modelling ap-

proaches outperform deterministic approaches by no vulnerability to the random nature of vehicle congestion but require higher computational capacity. Furthermore, no agreed-upon definition for congestion state assignment exists [24]. In Ma et al. [25], the assignment of ground truth traffic congestion states was based on a single indicator, speed, using a threshold of 20 km/hour. Fouladgar et al. [26] found the setting of a threshold for the whole network inappropriate since it could be different for different parts of the network according to him. Huang et al. [27] opposed using a single traffic indicator due to possibly different congestion states having the same traffic indicator value. Illustrating their line of thought, a mean velocity of 40 km/hour on a motorway segment could, for example, either refer to congestion or to free-flowing traffic in bad weather with storms and hail. The determination of congestion states is ordinary in our linguistic usage. Nevertheless, the human classification of a traffic situation in congestion states can differ, as described in [28].

The following paragraph examines the microscopic and macroscopic perspective of traffic data for ML models. Microscopic models are based on the knowledge of a single vehicle; macroscopic models use information from the entirety of vehicles [29]. According to Chen et al. [30], the traffic congestion state in a given time slot is the result of both global and local effects on the macroscale and microscale. They therefore argue that including both perspectives might improve evaluation performances. At the microscale, the variation of congestion states could be observed with precise details, while it would be hard to discover the global trend of large temporal scope according to them. In contrast, at the macroscale, the global trend of congestion levels could be easily revealed, while many details would be lost [30].

Characteristics of traffic data are presented in this paragraph. Traffic data is spatiotemporal data, whereby it implies the congestion state as well as other variables such as the mean velocity and the number of vehicles in an area. Guo et al. [19] stated that observations from nearby locations are strongly correlated with each other, referring to spatial correlation. Observations at adjacent time periods are strongly correlated and correlation diminishes when the temporal distance increases, referring to temporal autocorrelation [19] and local coherence [30] respectively. According to Chen et al. [30], traffic congestion levels also exhibit a temporal periodicity on workdays. Cheng et al. [32] showed experimentally that the

autocorrelation structure of the road network is dynamic and heterogeneous in both space and time by using both global and local autocorrelation measures.

Satisfying the traffic congestion data peculiarities described in the paragraphs above can be reached with different techniques for ML models. The techniques are presented for receiving an overview of the broad range of approaches for satisfying the spatiotemporal peculiarities. Firstly, the existing data set can be enriched through feature engineering. Fundamentally, ML models are able to use additionally generated features for modelling. Secondly, some models have supplementary opportunities for adapting to spatiotemporal data. For example, the filters as well as number of connections and nodes can be specified for artificial neural networks (ANNs) in order to illustrate the spatiotemporal characteristics. Thirdly, hybrid models that combine several models from different ML model classes, can be generated to better meet the spatiotemporal structure of the data.
A few examples of methodologies are presented addressing congestion state determination. Weighted exponential moving averages of consecutive time periods were used to address the temporal correlation at the macroscopic level in Pattara-Atikom et al. [28]. In generating features for a DT through the sliding window technique for consecutive time periods, the temporal correlation was addressed at the macroscopic level as well in Thianniwet et al. [33]. Chen et al. [24] generated the Moran quadrant feature that analysed spatiotemporal correlations. They further proposed a mixed forest model combining classification and regression trees. Chen et al. [30] used time series folding to address the temporal and seasonal correlation and multi-grained learning for capturing multiscale congestion patterns. They incorporated values from previous time periods through generating a 2-D matrix input called time series folding. For capturing temporal dependencies as well as macroscale and microscale congestion behaviour, they applied a series of convolutions on the input matrix.

Popular features for congestion state determination models are presented below. Zhang et al. [18] found that the flow, occupancy, velocity, and ramp flow features dominated the literature for capturing and predicting traffic conditions. Chen et al. [24] identified traffic flow parameters like flow rate, occupancy and velocity as usual evaluation indicators for traffic congestion. According to Huang et al. [27], vehicle velocity and traffic volume are among common features to determine traffic congestion.

Method, work	Now-casting	Rebalancing target ratio	# Cong. states	# Obs.	Measures	
					mcr	RMSE
DT [33]	Now	Equal proportions	3	448	9.71%	0.217
RF [34]	Now	- [1]	3	1124	12.50%	-
RF [24]	now	- [1]	5	2 M [2]	-	0.524
Fuzzy [35] [3]	Now	- [1]	3	440 [2]	11.21%	-
CNN [30]	10 Min.	- [1]	3	3 M	-	0.551
		...				
CNN [30]	60 Min.	- [1]	3	3 M	-	0.536

Table 3.1: Evaluation results of selected reviewed studies

Congestion states generally have an imbalanced class-distribution as, for example, the experimental data sets in [26] and [33]. Even though, rebalancing the data set was only described in a few studies such as in Thianniwet et al. [33] and Fouladgar et al. [26].

3.1.2 Selected machine learning case examples

This section exemplarily reviews several study designs to grasp their experimental setting in detail and classify their evaluation results. The reviewed studies were designed to predict (at the instantaneous time) or forecast (for future time periods) categorical traffic congestion states. They used numeric or categorical features derived from manual or automated data sources such as FCD or induction loops. Studies from different modelling perspectives were reviewed to demonstrate different methodological settings and the diverse spectrum of study setups. Table 3.1 briefly presents evaluation results of the studies. Results of [33], [34], and [35] should be regarded with special care due to small sample sets.

Evaluation results are described but due to a varying number of congestion states, often reported measures such as mean absolute error (MAE), mean squared error (MSE) and root mean squared error (RMSE) cannot be properly compared between all studies. It is hence additionally referred to the mcr as evaluation measure. Even so, this measure can be problematic considering an imbalanced class distribution. Examining the case, where

[1] not mentioned
[2] projected
[3] experiment B (reaches highest evaluation value)

only 1% of the observations belong to the congested state, a model always predicting free-flowing traffic still has an mcr of only 1%. Impedovo et al. [36] argued that it was difficult to identify and select the best algorithms to be adopted regarding traffic state estimation through videos. They argued that proposed systems had often been different at many stages and had adopted different data sets and testing conditions. Their argument also applies to the field of traffic congestion state prediction, as the usage of very different experimental setups in the exemplary studies shows. The studies are still enriching as insights into their experimental setup can be helpful for other studies.

Thianniwet et al. [33] proposed using a classification decision tree with three target categories for forecasting the congestion state. Data was collected through a GPS signal as well as through a video camera onboard a vehicle riding on several strongly congested roads in Bangkok, Thailand. The experiment was performed for approximately three hours which led to 448 observations in total after balancing the data set. Model features were the moving average velocity at time t, $t-1$, and $t-2$ with a sliding window of three and the moving average velocity at time t with a sliding window of five. The ground truth congestion states were generated with a majority voting of eleven car drivers. The evaluation value of the mcr was 9.71%.

A random forest classifier was used as a traffic congestion state forecasting model in Liu and Wu [34], based on approximately 1,000 data points from different road segments of the Shanghai traffic management information department in China. The random forest was based on the environmental features: Weather conditions, time period, special conditions of roads, road quality, and holiday. The target variable was the future traffic congestion state consisting of three classes. The classes were labelled through a threshold function of the congestion coefficient, $\frac{T-T_0}{T_0}$, with T indicating the actual travel time and T_0 the optimum travel time. The mcr was 12.5% regarding only 24 data points used as test data set.

Chen et al. [24] proposed a mixed forest model combining classification and regression trees to forecast the traffic congestion state, captured on five levels. The mixed classification forest consists of decision trees with a categorical target variable, congestion state, and continuous target variable, congestion coefficient. The target of the regression decision trees is classified and a mutual prediction of all trees is made through the mixed forest. The authors further proposed using the feature Moran quadrant, which analyses the spatiotemporal state of

urban road traffic. FCD data of taxis in the city of Chengdu in China were utilized for model training and evaluation. The study focused on 20 road segments around the Chengdu North railway station with one segment used for testing purposes. The Moran quadrant, a time factor, speed and congestion status at that time were the features for forecasting congestion states. The ground truth congestion states were generated through fuzzy C-means clustering (FCM) including the features average velocity and congestion coefficient. Between other measures, the false positive rate (fpr) of 4.17% was reported. The mixed forest was compared to the prediction models classified forest, post-classification forest, DT, bayesian algorithm, k-means, and support vector machine (SVM) and achieved the best result regarding the fpr.

Fuzzy techniques were proposed by Pongpaibool et al. [35] to forecast the traffic congestion state. The authors used tuned fuzzy logic and adaptive neuro-fuzzy techniques in order to emulate human expertise in determining the congestion state, which consisted of three levels. Data was collected through a video on a segment of a busy three-lane road in Bangkok, Thailand in a time period of almost three hours. Possible features were extracted from the video and comprised lane-wise vehicle volume and average speed reported every 30 seconds. The labels were produced through a majority vote of ten car drivers who watched the aforementioned video. The mcr was 11.21%.

Chen et al. [30] proposed using deep convolutional neural networks (CNNs) for short-term and long-term traffic congestion prediction (PCNN) in order to capture similar congestion patterns in neighbouring and seasonal time slots as well as exploiting multiscale properties of traffic congestion states. In the research, three congestion states were defined and the data was gathered through structured vehicle passage records (VPRs) from surveillance cameras containing vehicle ID, location, and timestamp. 614 road segments in Jinan, China were captured. The timespan was six weeks but due to preprocessing, some time periods were disregarded and the final number of observations was approximately 3 million. PCNN used congestion states as features and the target was the traffic congestion state forecast for 10, 15, 30, or 60 minutes. The ground truth congestion state was computed by thresholding the congestion coefficient and yielded to 64% *normal* traffic (first congestion state level). Unfortunately, only MAE, RMSE, and mean relative error (MRE) were provided lacking the mcr. The model's evaluation results were compared to and outperformed the results of historical average (HA), LR, auto-regressive integrated moving average (ARIMA), seasonal

auto-regressive moving average (SARIMA), k-nearest neighbours (K-NN), multilayer perceptron (MLP) and long short-term memory (LSTM).

In conclusion, the described studies vary in their feature data gathering approaches, ground truth gathering approaches, feature space, ML model, and number of congestion states.

3.1.3 Study comparison obstacles

Results of scientific works are generally compared to assess the scientific approaches. Obstacles when comparing results of different databases and study setups are addressed in this section.

Research articles concerning traffic congestion state prediction and forecasting rely neither on benchmark data sets nor use published data predominantly, for example [24], [30], and [35]. Furthermore, researchers publish the utilized data very rarely. As a consequence, it is more difficult to compare the evaluation results of competing methodologies. One of the few open-source benchmark data sets proposed in a study is the Citywide Traffic Congestion Condition [37] by Fouladgar et al. [26]. It consists of information from stationary detectors. Loder et al. [38] constructed a large data set which is based on stationary detectors as well. An advantage of relying on stationary traffic detectors is the information gathering from the whole population of motorists on a point of the road. In contrast, an advantage of using FCD instead of stationary traffic detector data to forecast congestion states is the availability of data beyond stationary points. Forecasting models based on FCD can generally be utilised for much broader areas than models based on stationary traffic detector data. It was therefore relied on the FCD data source to generate features in the approach of this work. The data format as well as the information content differs between FCD and stationary detector data. It could therefore not be relied on the proposed benchmark data sets.
As addressed in the previous section, existing scientific work has a large variety in their study setups that was not compensated for by evaluation metrics. Therefore, no reliable comparison between studies could be made.

Publications use worldwide data sources for developing and evaluating modelling approaches, amongst them China [24], [19], [34], [29], Taiwan [39], and the USA [26]. Measures such as

the density of motorways per square kilometre of land area, the allowed vehicles and the recommended velocity can differ from country to country [40]. Transferring ML models and evaluation results throughout the globe has not yet been studied systematically.

3.2 Traffic congestion detection using grid cells

Traffic congestion detection methods using grid cells were seldomly developed in the past as only a few published studies in this field show. In the first section, studies using the grid-based approach are presented. In contrast, studies from Section 3.1 relied on road segments which are linked to FCD using a map-matcher. Differences between the grid-based and the map-matcher approach are therefore discussed secondly. Lastly, further developing the grid-based approach to determine congestion states for road segments is discussed.

3.2.1 Selected case examples

The case examples detecting traffic congestion in grid cells utilised the basic grid-based approach as presented in Section 2.3.2.

Liu et al. [10] categorized FCD from the core traffic area of Beijing into grid cells. They stated that selecting an appropriate cell size was crucial to overcome data density restrictions as well as to cover sufficiently small road segments. A traffic operation performance index of range $[0, 100]$, in which the value zero indicated free-flowing traffic and 100 indicated strong traffic congestion, was computed for each cell. The index was built through a normalization of the cell's free-flow speed divided by the current average speed. The results were visualised and an evaluation of the measure was not performed.

Zhao et al. [29] proposed a grid-based traffic flow influence concept and a corresponding traffic congestion diffusion model to characterise the congestion diffusion process in both spatial and temporal domains. They developed and evaluated a diffusion model to forecast the congestion state which was separated into congested and smooth classes. The ground truth labels were computed through thresholding the crowdedness value, which was based on the min-max scaled instantaneous velocity. The experimental area was the city of Shenzhen in China. The city's map was overlayed by a grid. Forecasting a grid cell's congestion state took the situation in the cell as well as in the neighbouring cells into account. The authors

Characteristics	Basic grid-based approach	Map-matcher approach
Dependencies	No	GIS map
Segment size	Manually definable	Pre-defined
Complexity of approach	Simple	Complex
FCD allocation to segments	-	<100%

Table 3.2: Differentiation between the grid-based and the map-matcher approach

found that the mcr converged to a upper threshold with the rising of the future time period. The model had a F_1-score of 89% in one district (an area including multiple cells) when forecasting the congestion state for the next time period. Unfortunately, a time period was not specified in terms of minutes and the class proportions of congestion and smooth classes were not mentioned.

3.2.2 Differentiation between the grid-based and the map-matcher approach

A map-matcher is regularly chosen to aggregate FCD for generating features of a data corpus used for forecasts. The grid-based approach is an alternative to the map-matcher processing procedure. Different characteristics leading to advantages and disadvantages of both approaches are outlined in Table 3.2.

Grid cells can be produced directly from the FCD features *latitude* and *longitude*. No further information is needed. The map-matcher approach has the disadvantage of needing a high-quality GIS map (cf. [10]) holding information about segments, which can be outdated.

The segments assigned by a map-matcher are pre-defined by the GIS map. However, the segment size of the grid-based approach generally refers to the grid's cell size and can be chosen manually. This can mitigate data sparsity issues when only a small amount of FCD observations is recorded in an area (cf. [29]). Derived from that, the computational effort decreases since fewer observations belong to the data corpus (cf. [29]).

The grid-based approach is fairly simple to understand and to implement. It can generally be relied on the infrastructure used to develop ML models. A map-matcher involves a complex estimation model, such as a hidden Markov model (HMM) (cf. [41]). The model can assign FCD points to road segments of the GIS map. It is a tool used in addition to the model

development framework.

FCD allocation to grid cells is 100% accurate, given that the GPS locations of FCD are entirely correct. But the grid cells do not correspond to motorway carriageway segments. That is a disadvantage if congestion states of roads of the motorway network is of interest. A map-matcher estimates assignments of FCD points to segments and is naturally less than 100% accurate. Even so, it is a sophisticated method for assigning FCD to road segments. Utilizing the advantages of the grid-based approach as outlined in this section and being able to assign FCD to motorway segments at the same time needed a further development of the grid-based approach.

3.2.3 Discussion

Having less dependencies and a simpler approach in addition to disposing the map-matcher tool, a closer look was taken at the grid-based approach. As stated in the previous section, FCD cannot be allocated to road segments in the grid-based approach. Figure 2-3b shows that once more since the orange and purple car are allocated to the same cell. Even so, the allocation of FCD to motorway segments is crucial when determining traffic congestion states for motorway carriageway segments, which is the objective of this work. Therefore, a method distinguishing between the two motorway carriageways inside a cell, a motorway direction distinctive grid-based approach is needed. For the majority of cells, two directions based on two carriageways of a motorway exist. Therefore, each FCD point in a cell should additionally be assigned to a direction class. Figure 3-1 displays the approach. The direction class is a binary variable with the values 0 and 1. The orange and purple car, symbolising FCD points, are still assigned to the same cell but differ in the assigned direction class. Every FCD point in a cell is either assigned to direction class 0 or direction class 1.

The transferability of results from scientific literature determining traffic congestion states with the map-matcher approach to the motorway direction distinctive grid-based approach is examined in this paragraph.

Several studies mentioned in Section 3.1 used FCD to extract features. The feature values were computed based on the composite index of predefined road segment and time window. The assignment of FCD to segments was achieved through a map-matcher. In contrast, the motorway direction distinctive grid-based approach uses observations of the according

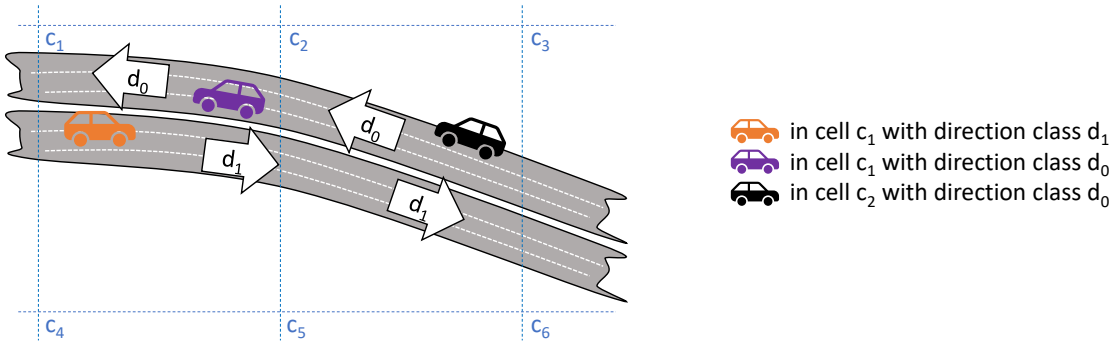


Figure 3-1: Motorway direction distinctive grid-based approach

direction class within the whole grid cell for computing feature values. This usually adds noise to the feature values.

Once a data corpus with features and ground truth is installed, the data structure type is the same between the motorway direction distinctive grid-based and the map-matcher approach except for a differing number of variables forming the composite index. The data corpus can be utilised for developing an ML model. Findings of studies that used predefined road segment can therefore be transferred to this work.

4. *Methodology*

This chapter describes and explains the utilized methodology. The ML pipeline including a data processing and a model development step is outlined. Grid-based specifics are important for this work and therefore presented in a separate section. Especially the segment assignment step with the help of direction distinctive grid cells is fundamental for the approach of this work. Performed descriptive analyses of the data corpus are described shortly. The choice of a basic inference model as well as its hyperparameter settings are justified, utilized techniques to generate model insights are described briefly and the choice of evaluation metrics is justified. Lastly, the methodology for comparing the computational effort of the segment assignment step between the direction distinctive grid-based and the map-matcher approach is described.

4.1 Model development framework

The ML pipeline used in this work is outlined in Figure 4-1 and therein utilized symbols rely on the ISO 5807 norm. The upper box shows the data processing flow including the input data. As can be seen in the figure, feature information was gathered from FCD comma-separated values (CSV) files and ground truth as well as a few features were extracted from extensible markup language (XML) files. The basic structure of FCD was described in Section 2.2, details regarding the utilized input data can be found in Section 5.2.1. Both input data types were processed and merged to form the data corpus. The segment assignment methodology for FCD is described in the following section due to its novelty and importance for the congestion state forecast. The data processing steps and the generated data corpus itself are described thoroughly in Section 5.2.2.

The data corpus was explored with the tools described in Section 4.3. Additional features were derived from analysing the data and ML model results. This feature engineering part is

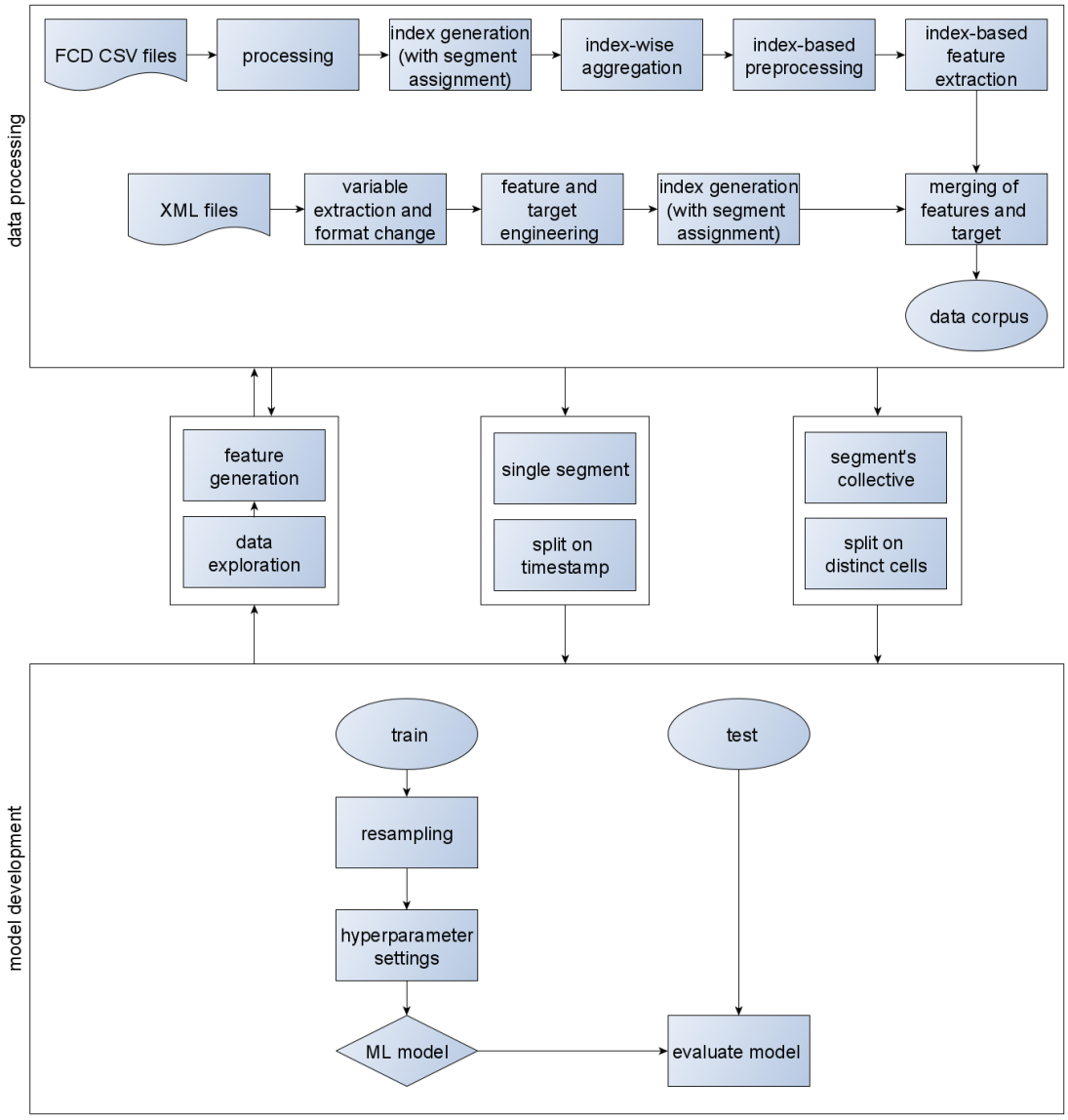


Figure 4-1: Machine learning pipeline diagram

shown through the left box in the middle section of Figure 4-1. ML models were developed for single motorway segments as well as for all segments having ground truth data as shown through the middle and right box in the middle section of the figure. In other words, models were built based on data of a single segment or all segments in a collective. Details regarding the train-test split in the two settings are described in section 4.2.3.

Model development and evaluation steps are shown in the lower box of Figure 4-1 and were performed for both cell settings. The train data set was resampled, otherwise the ML model would have learned to only forecast the majority *free-flowing* class. Model evaluation was based on the test data set with the original proportions of the two congestion state classes

since they represent the real conditions of the forecasting scenario. Details regarding the ML methodology can be found in Section 4.4.

4.2 Motorway direction distinctive grid-based specifics

The direction distinctive grid-based approach required some prerequisites, which are described in Section 4.2.1. The direction-distinctive grid cell modelling is a fundamental component of the grid-based approach for motorway segments and characterizes this work. The approach is presented and illustrated by an example in Section 4.2.2. The section 4.2.3 outlines the performed train-test split in two cell settings for the utilized spatiotemporal data corpus.

4.2.1 Prerequisites

Statical, disjoint rectangles were chosen as the geometrical shape of the uniform grid. This cell structure was selected since it is a basic shape which could be easily generated through rounding the *latitude* and *longitude* FCD variables. The size of each rectangle was set to approximately 110 m $\times$ 70 m first for being close to the 100 m $\times$ 100 m that worked best in the explorative analyses of Liu et al. [10]. Due to data sparsity issues it was switched to a larger size of approximately 1.1 km $\times$ 0.7 km. Implications from the wider cell size include congestion state forecasts at a lower granularity. Generally, the author expects better model performances with decreasing cell sizes that rely on data not being sparse.

It can generally be shown that most FCD points in a motorway-encompassing cell are recorded along the motorway. The FCD points on motorways thus dominate the observations inside a motorway-encompassing cell. A binary direction class variable was established additionally to the cell index to differentiate between motorway carriageway segments inside a cell. No distinction was made between observations from the motorway or other road types when accepting a minor share of noise from FCD points from other road types than the motorway. As a consequence, it was majorly relied on features that are robust to noisy data. For example, three observations were made inside a cell. Two observations belonged to vehicles on the motorway having velocities of 90, and 100 km/h respectively. One observation belonged to a vehicle riding on a small residential street having a velocity of 15 km/h. The mean velocity was then approximately 68 km/h and the median was 90 km/h.

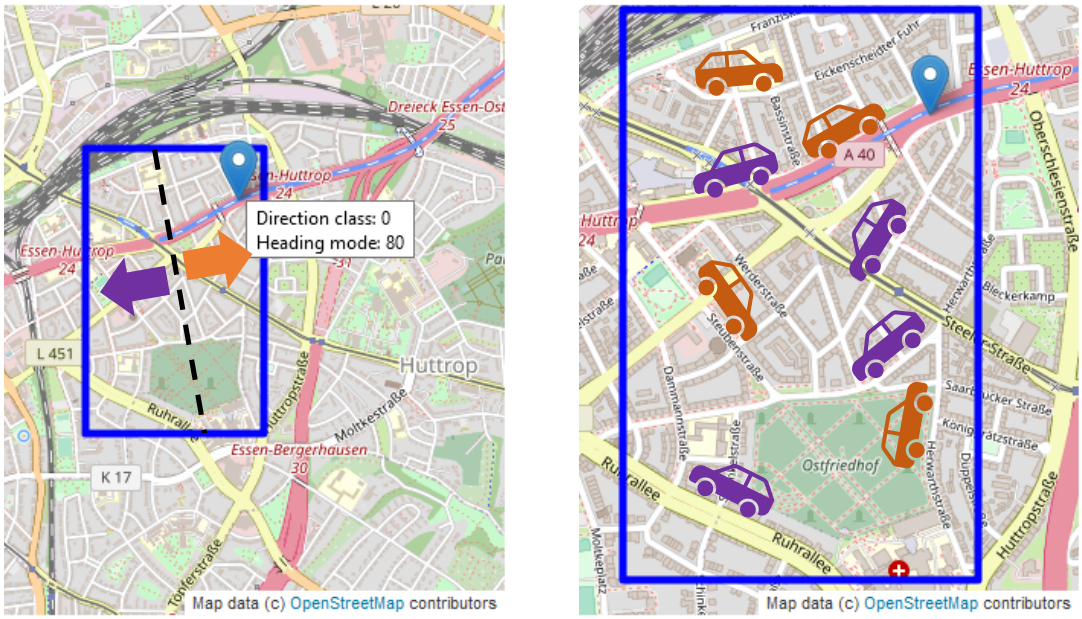
This brief example shows that the median is more robust to the noise from observations of streets other than motorways.

4.2.2 Direction distinctive grid cell modelling

Distinguishing between motorway directions inside grid cells is described due to its novelty and fundamental importance when forecasting traffic congestion states for motorway segments in grid cells. The following paragraphs describe the direction distinctive modelling procedure for grid cells and present an example of determining motorway directions for FCD points.

Firstly, FCD points were grouped cell-wise according to their *heading* values, or in other words according to their compass direction. Secondly, the mode of the *heading* values was taken as one of the two *direction classes* of the motorway part lying inside a cell. This implies the previously outlined assumption that most FCD points were recorded on the motorway. Thirdly, the other direction class was formed by the mode's counterpart of ± 180 degrees. Lastly, FCD was categorized into the two *direction classes* according to the minimal distance to the mode or to the mode's counterpart.

Figure 4-2a shows a direction class extraction example for the cell within the blue box. The blue marker represents the traffic detector location gathering ground truth labels and the white box near to the marker shows the allocated direction class and its heading mode. The orange arrow represents the mode with a heading of 80 degrees referring to direction class 0, which is the direction with ground labels. The purple arrow represents the mode's counterpart with a heading of 260 (80 + 180) degrees, referring to direction class 1. If the mode was 180 degree or higher, the mode's counterpart would have been associated with direction class 0. The dashed line illustrates the heading boundary between the two direction classes. In other words, an FCD point was assigned to direction class 0 if its heading value was in the interval [350, 360) or [0, 170). It was assigned to direction class 1 if its heading value was in the interval [170, 350).
Figure 4-2b illustrates the assignment of direction classes to several FCD points in the displayed cell. A vehicle symbol represents an FCD point and the vehicle's rotation shows the heading of the FCD point. Vehicles with orange colour were assigned to direction class 0



(a) Direction classes (0: orange, 1: purple) for one cell

(b) Direction classes for FCD sample

Figure 4-2: Assignment of grid cell motorway segments to floating car data

and purple vehicles were assigned to direction class 1. Table 4.1 presents the FCD points with assigned direction classes. The table has the column *cell middle*, which is the same for all observations since all regarded observations belong to the same cell. The column *heading mode* is the mode of all heading observations within the cell. The mode cannot be derived from the sample observations in the table since it relied on a larger dataset. The binary

Id	...	***Heading***	***Cell middle***	***Heading mode***	***Direction class***
Orange, top	...	85	51.45, 7.03	80	0
Orange, 2nd from top	...	60	51.45, 7.03	80	0
Orange, 3rd from top	...	145	51.45, 7.03	80	0
Orange, last from top	...	10	51.45, 7.03	80	0
Purple, top	...	255	51.45, 7.03	80	1
Purple, 2nd from top	...	227	51.45, 7.03	80	1
Purple, 3rd from top	...	231	51.45, 7.03	80	1
Purple, last from top	...	296	51.45, 7.03	80	1

Table 4.1: Assignment of direction classes to a floating car data sample

feature *direction class* was derived from the *heading* of an FCD point and the *heading mode*. The direction class assignment process of the first observation is described exemplarily. The first observation in the table belongs to the top orange vehicle symbol in Figure 4-2b and was assigned to direction class 0. The observation had a heading of 85 degrees and a distance of 5 degrees to the mode of 80 degrees as well as a distance of 175 degrees to the mode's counterpart of 260 degrees. The observation was assigned to the mode's direction class of 0 since the distance to the mode was smaller.

Once the according *direction class* is assigned to each FCD point, aggregation follows the example given in Section 2.3.2 with the composite index of *datetime*, *cell middle*, and additionally the *direction class*. Through using the direction class variable in the composite index, the index refers to time intervals of motorway segments instead of time intervals of grid cells.

4.2.3 Train-test split for different segment settings

A train-test split is a basic operation in the ML field. This work used spatiotemporal data and the spatial component was given through direction distinctive grid cells. The train-test split in the spatiotemporal context of this work is outlined in the following paragraph.

Models were developed for two different segment settings, single segments and the collective of segments, so-called whole grid setting. The approximate split proportions for both segment settings were: 70% training, and 30% test data.
Single segments were split into a train and a test data set based on its timestamp as can be seen on the left side of Figure 4-3. The blue rectangle represents a single cell with one direction class, referring to a motorway segment. Inside the cell, a timeline from August 2019 to February 2020 is displayed, characterising the examined time period. Since the congestion state proportions seem to differ monthly (cf. Section 6.1), cross-validation was performed for single segment models. Each month of data led to one cross-validation data subset. As presented in the figure, the first train data set consisted of data from September 2019 to February 2020 and the first test data set consisted of observations from August 2019. Each month was gone through this way and the final test data consisted of observations from February 2020.

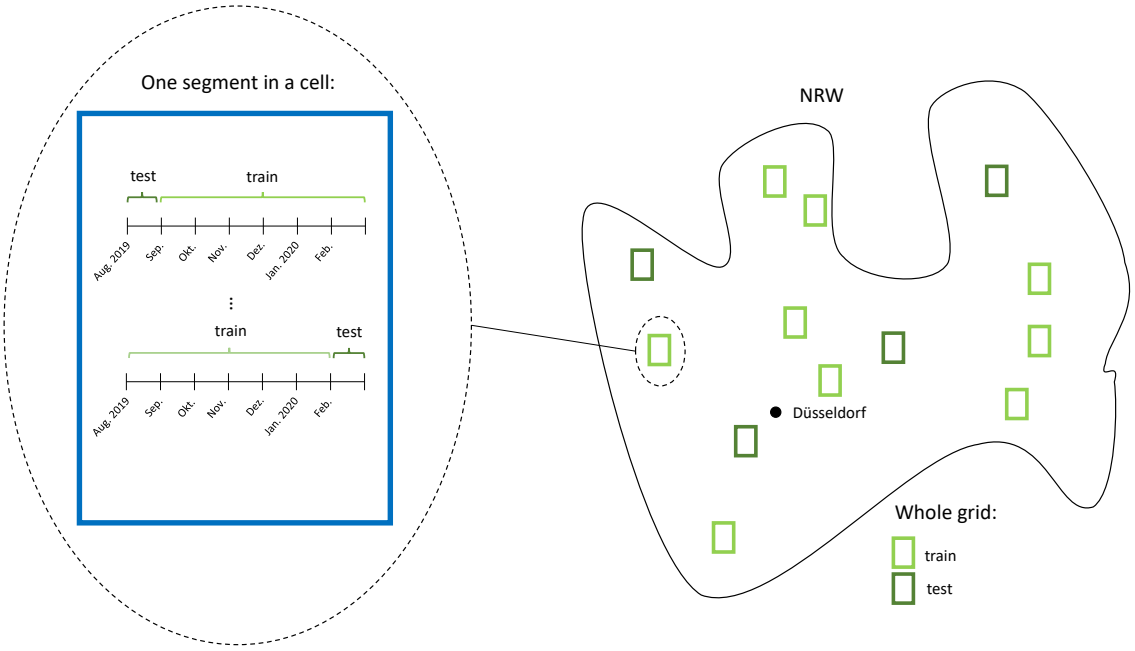


Figure 4-3: Train-test split in the single segment setting (left) and the whole grid setting (right)

The whole grid setting contained various cells and the train-test split was based on distinct cells, as shown on the right side of Figure 4-3. Most cells contained only one segment or, in other words, only one direction class. The area surrounded by black contours defines the state NRW. Rectangles inside the area represent exemplary cells. Cells with light green contours characterise cells belonging to the train data set and cells with dark green contours characterise cells belonging to the test data set. The cell that was displayed exemplarily in the single segment setting belongs to the train data set in the whole grid setting of the figure. The cells were randomly assigned to the train and test data set based on a seed.

4.3 Descriptive insights

The data corpus was explored regarding its spatiotemporal nature. The features and the target were examined through the following descriptive analyses:

- Descriptive plots of target regarding time and spatial component
- Statistical characteristics of features
- Correlation analysis between two numeric features respectively

- Connection analyses of each feature with the target
- Histograms of selected numeric and categorical features with respect to the two target classes

These analyses were selected due to a wide coverage of describing the spatiotemporal nature of the target variable, of demonstrating basic characteristics as well as associations between the features, and of describing associations between features and the target.

4.4 Machine learning methodology

The utilized ML inference model is justified along with the pursued feature engineering strategy in this section. The data resampling method and model hyperparameter settings are presented and vindicated, and strategies to generate model insights are outlined. Quality criteria for the encountered binary classification problem are justified lastly.

4.4.1 Machine learning inference modelling

The number of congestion states in this work was predefined through the labels *free-flowing* and *congestion* of the ground truth data, leading to a binary classification problem. Ground truth labels were gathered with the help of deterministic heuristics. The ground truth congestion state labels have therefore a deterministic nature. It was hence relied on a deterministic congestion state determination algorithm. The random forest (RF) modelling approach, which is suitable for binary target data, was selected due to its light-weight and scalable nature and relatively well predictive power. A benefit of utilizing an RF is the good implementation ability for the big data infrastructure as described in Section 5.1. More complex algorithms such as neural networks would have been extremely intense regarding the computational effort since large data volumes were used in this work. Moreover, the RF is a basic and well-known ML inference model that was utilized by a few other studies in the traffic *congestion state* prediction domain as well (cf. Section 3.1.2).
The feature engineering strategy is described in the following paragraph. The existence of temporal autocorrelations of traffic data was stated by Guo et al. [19] and Chen et al. [30] in Section 3.1.1 of the literature review. The existence of temporal autocorrelations justifies the forecasting approach with instantaneous feature values as used in this work.

The data corpus contained features on the macroscopic level. Information of the single FCD points were condensed to information of its entirety in defined groups. Deciding for the macroscopic level instead of the microscopic level was based on the extremely smaller computational effort when utilising large volumes of FCD. Furthermore, as a potential analysis of the motorway direction distinctive grid-based approach, discovering global trends in the macroscopic perspective could reveal the usefulness of the approach. The basic features mentioned in 3.1.1 were generated and incorporated in the data corpus when possible. Temporal and spatial connections of the *congestion state* data as well its temporal periodicities were taken into account through generating model features representing these characteristics. Major influential features could be identified through their feature importance scores.

4.4.2 Data resampling and hyperparameter settings

Like in other scientific works (cf. Section 3.1), rebalancing the data corpus is considered. Resampling as a valid approach for imbalanced data was used in this work since the developed RF models would have disregarded the minority congestion class otherwise. The oversampling method was chosen to compensate for the imbalanced target. Observations with the congestion target class were oversampled for the train and the test data set respectively until the class proportions of 20% congestion and 80% free-flowing classes were reached. Undersampling would have led to very few observations caused by the strongly imbalanced target and was therefore disregarded.

Hyperparameter settings contribute to the predicting performance of an RF model. Utilized values of hyperparameters as described in Section 2.4 are therefore presented. Hyperparameter values were utilized as pointed out in Table 4.2. The default setting was selected for every hyperparameter value due to the huge dataset. A search for each hyperparameter separately or a grid-search would have been too much consuming regarding time and resources. Choosing the default hyperparameter setting seemed to be a reasonable choice due to no further knowledge about reasonable hyperparameter values.

Hyperparameter	Setting	Characteristics
Max. depth of DT	Default	5
Min. obs. per node of DT	Default	1
Min. IG of DT	Default	0
Impurity measure	Default	Gini
Number of trees for RF	Default	20
Binary classification threshold	Default	0.5

Table 4.2: Utilised hyperparameter values

4.4.3 Model characteristics

The variety of possible descriptive analyses of model results is large. A few meaningful analyses were relied on for a good model overview. The following steps were performed to attain model insights:

- Histograms of values from confusion matrix for selected features
- Descriptive plots of forecasting performance regarding spatial component

The histograms were selected due to a comparability with results from the ground truth data corpus description as well as interesting findings. The descriptive plots were selected to account for the discovered variety of performance values for different motorway segments.

4.4.4 Quality criteria for machine learning models

The quality of ML models was measured by their forecasting performance. It was only relied on the evaluation results of the test data set with original proportions of target classes. The reason for this is that the evaluation of the original test data set reflects the real conditions a ML model encounters. As a standard evaluation metric for classification tasks and often reported measure in investigated works from Section 3, the mcr was used for comparing results. The F-score was additionally relied on as a standard evaluation metric. Studies from Section 3.1 did not report this measure. It was assumed that neither wrongly predicting *congestion* nor wrongly predicting *free-flowing* was generally worse than the other. Therefore, the F_1-score was chosen as evaluation measure. The *congestion* class of interest was chosen and thus used as reference class. According to Luque et al. [42], both the mcr and

the F_1-score are biased when using imbalanced instead of balanced data. They suggested using, for example, the null-biased BM as evaluation metric. This measure was additionally relied on in this work to compare results especially for different target class shares.

Evaluation measure baseline values are presented to compare the predicting power of the RF models to a threshold of a naive model.
The naive baseline model leading to the F_1 baseline score always forecasted the *congestion* class. The score for the test data with original proportions (0.3% *congestion* class) was computed as follows:

$$F_1 = 2\,\frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} = 2\,\frac{0.003 \times 1}{0.003 + 1} = 0.006.$$

The baseline model with the mcr evaluation measure always forecasted the *free-flowing* class. The model's mcr is much better than the rate of a model that would always forecast the *congestion* class. The baseline model leads to an mcr of 0.003 (99.7% *free-flowing* class) for the test data.
The baseline model using the BM as evaluation measure always forecasted the *free-flowing* class and was computed as follows:

$$\text{BM} = \frac{\text{TP}}{\text{TP} + \text{FN}} + \frac{\text{TN}}{\text{TN} + \text{FP}} - 1 = 0 + 1 - 1 = 0.$$

Always forecasting the *congestion* class led to the same result, only the first and second summand were swapped. The baseline value was the same for the resampled data set and the data set with original class proportions due to the null bias characteristic of the BM regarding imbalanced target data.

The prediction of the current *congestion state* was performed additionally for a categorization of the results besides using the baseline models. The evaluation values of the instantaneous time serve as upper boundary for the developed forecasting models. The reason for this is that the feature values at the instantaneous time have a greater informative impact for the instantaneous congestion state determination than feature values at 5, 10, 20, 30, and 60 minutes before.

Experiments in the whole grid modelling setting were executed three times to account for model variations and the mean and standard error were reported. Experiments in the single segment setting relied on cross-validation results from seven subsets (cf. Section 4.2.3). When comparing the single segment and the whole grid setting, whole grid models were executed seven times for having an equal number of runs. When comparing the segment settings, cells of the single segments regarded in the single segment setting were deterministically put into the test data set of the whole grid setting and the remaining cells were assigned randomly to either the train or the test data set.

4.5 Methodology for measuring the computational effort of segment-assignment approaches

The computational intensity of the segment-assignment step was compared between the motorway direction distinctive grid-based approach and the map-matcher approach since the two approaches are very different. The differences of the basic grid-based approach and the map-matcher approach were presented in Section 3.2.2. These differences apply to the direction distinctive grid-based approach and the map-matcher approach as well, except that the direction distinctive grid-based approach leads to a differentiation of the two motorway carriageways in a cell. An FCD sample was relied on for measuring the computation time of the two segment-assignment approaches. The computational effort of the subsequent steps in the ML pipeline, which were described in Section 4.1, was not examined since exactly the same steps can be utilized after assigning segments to FCD.

Measuring the runtime is a valid method to capture the computational intensity. Experiments have been executed three times to account for variations and the mean and standard error were reported.

5. *Experimental setup*

The computer infrastructure is described in the first part of the experimental setup. The data processing step from initial data sets to the data corpus generation is presented in the second part.

5.1 Computer infrastructure

This work relied on the infrastructure provided by the cooperating company urban mobility innovations. The programming language PySpark, which was executed on a spark cluster using Spark 2.3.2, was utilized due to its scalability properties for big data. The huge input data and the generated data corpus were stored in a hadoop distributed file system (HDFS). For the data processing part, eight to 16 kernels with four to 12 GB memory each were used. One kernel had four executors. The ML part was less computationally intense and only two kernels with four GB memory each and four executors per core were utilized. Ipyleaflet [43] was used for geographic visualisations.

The computational intensity of the motorway direction distinctive grid-based and map-matcher segment-assignment step was measured on a system with the following allocated resources: virtual machine with an ubuntu operating system, AMD Ryzen 7 3700U processor with 8 threads and a base clock of 2.3GHz, and 6 GB memory. The open-source barefoot map-matcher [41] was used for measuring the computational intensity of the map-matcher segment-assignment step. The additional settings for the map-matcher server were ten server connections and eight threads.

5.2 Data processing

The data set for generating ground truth labels as well as two features is presented first to get an impression of the validity of the ground truth. FCD is shortly referred to as well since most features were derived from it, it was described thoroughly in Section 2.2. The data corpus is presented secondly as utilized for analyses in the following chapter.

5.2.1 Initial data sets

In the following paragraphs, the utilized target and feature data sets are presented. The ground truth target variable and two features were extracted from the Verkehrsinformationszentrale Nordrhein-Westfalen (VIZ.NRW) data set in the XML file format. Floating car data (FCD) was mainly utilised to extract features that were fed into the RF models. The characteristics of the two data sets are shown in Table 5.1. The NRW area was decided upon since this state was most troubled by congestion in Germany. In 2019, 36% of traffic congestion events were located in NRW [6]. Data from August 2019 to February 2020 was regarded in this work. Each minute, an XML file was generated containing information on the *congestion state* at traffic detector locations. Per day, 1,440 files were released with a volume of approximately 90 MB. The total number of regarded files was approximately 300 thousand and the total volume was approximately 19 GB. An FCD point was generated approximately every 15 seconds for a tracked vehicle. Approximately 58 million FCD observations were recorded per day with an approximate volume of 6 GB. Approximately twelve billion observations with an approximate volume of 1.2 TB were regarded in total.

Data set	**VIZ.NRW (XML files)**	**FCD**
Area	North Rhine-Westphalia	
Time span	August 2019 - February 2020	
Time interval	1 Minute	15 Seconds
Number per day	1,440 files	57,941,021 obs.
Volume per day	90 MB	5.6 GB
Total number	306,720 files	12,341,437,588 obs.
Total volume	19 GB	1.2 TB

Table 5.1: Characteristics of the utilized raw data sets

Data set of Verkehrsinformationszentrale North Rhine-Westphalia

The VIZ.NRW data set was chosen to serve as ground truth data set since it was the only one available for research purposes without a fee. Instantaneous traffic messages are available to the public through web scraping the VIZ.NRW [44] data set on the data platform openNRW. The *congestion state* ground truth target can be extracted from the XML files of VIZ.NRW as well as a feature containing information of crashes and a feature containing information of construction sites. The VIZ.NRW data set as used in this work was not web scraped but gladly received from Landesbetrieb Straßenbau NRW since the regarded time period would have been only up to three months otherwise. The extracted traffic *congestion states* were classified as *slow-moving* and *congestion*. Both congestion information types can generally be published by the police to VIZ.NRW. Only the *congestion* type can be obtained by observing the velocity of vehicles with vehicle detectors. The velocity is then received through two consecutive induction loops positioned inside the road. A congestion message is generated if the mean velocity falls below 30 kilometres per hour for more than a five-minute interval. Sending congestion messages discontinues if the mean velocity is above 45 kilometres per hour for more than a five-minute interval. The observed data revealed that only very few incidents belonged to the *slow-moving* class compared to the *congestion* class. This did not seem to represent reality and was most likely due to the fact that this traffic state was only captured by the police. Therefore the *slow-moving* class was disregarded in further analyses.

A traffic detector generates traffic incident messages at one road point of a motorway segment. The whole motorway segment was assigned to the gathered ground truth label since the behaviour is spatially autocorrelated as stated in [19]. Data analysis showed that there were approximately 1,000 traffic detectors recording traffic congestion incidents. In the absence of congestion incidents, the traffic state was assumed to be *free-flowing*. Therefore the target variable consisted of two traffic congestion states. The target variable was created in a one-minute time interval by VIZ.NRW. Only motorway segments with at least one congestion event in the observed time span were regarded. There were a few cases when multiple traffic detectors have been associated to the same direction class of a cell. Traffic congestion messages belonging to the traffic detector with minor congestion messages were

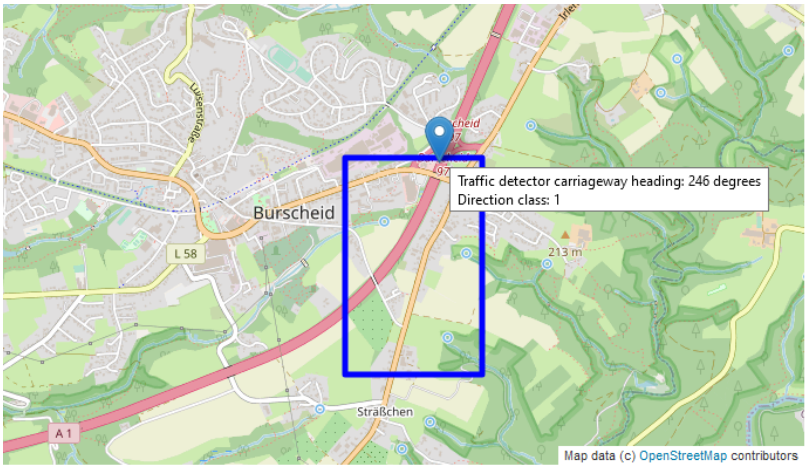


Figure 5-1: Cell (50.87, 6.97) and direction class 1

eliminated. This led to a reduction of approximately 4 thousand congestion messages. An exemplary traffic detector along with its according grid cell is shown in Figure 5-1. The figure shows a blue-marked traffic detector from which congestion incidents are reported on the A1 motorway. The traffic-detector-surrounding cell was illustrated by a blue box. The textbox shows that the traffic detector was assigned to *direction class* 1 with a heading of 246 degrees, corresponding to the west-southwest direction.

Floating car data set

Provided by the cooperating company urban mobility innovations, FCD used in this work was captured through the GPS signal of road users using a traffic-relevant app on their mobile devices. It therefore described a sample of road users. Every FCD point contained the following columns: *id*, *latitude*, *longitude*, *datetime*, *velocity*, and *heading*. An FCD sample was thoroughly described in Section 2.2.

FCD has been analyzed intensively by urban mobility innovations. Through this process, it is assumed that the utilized data is biased towards capturing heavy transport road users at a higher rate. It is not assumed to influence congestion state forecasts. The reason for this is that the biased FCD observations are still in the range of motorway velocities. It is further assumed that the coverage of road users is approximately 2%.

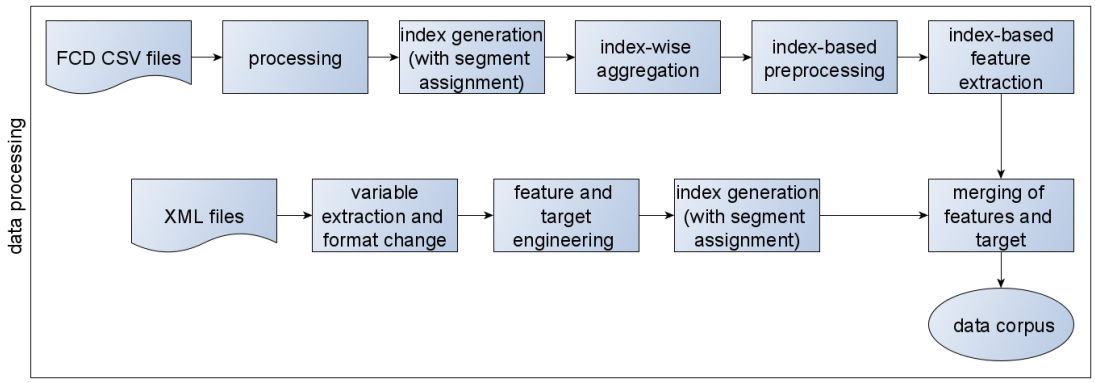


Figure 5-2: Data processing diagram

5.2.2 Data corpus generation

Figure 5-2 shows the processing pipeline's data flow diagram. The flow started at stored XML files of VIZ.NRW and CSV files of FCD data.

Amongst other variables, ground truth was extracted from congestion messages of VIZ.NRW's XML files in the *description* column. A sample of the generated data frame can be seen in Table 5.2. The table contains the *publicationTime*, the congestion incident *description*, and amongst others the location of the start and end of congestion through *latitude from*, *longitude from*, *latitude to*, and *longitude to*. The third displayed traffic message was a traffic congestion message. As part of the processing pipeline, observations containing congestion messages were selected and aggregated based on distinct five-minute time windows, called *datetime* parameter. The *heading* was engineered based on the location of congestion entry and removal point and was assigned to a direction class. Each cell midpoint, *cell middle*, was extracted through the according approximate *latitude* and *longitude* values. The index of the ground truth congestion states was established from the parameters *datetime*, *cell middle*, and *direction class*.

Approximately twelve billion FCD observations were considered for generating feature values that predict the congestion state. FCD was preprocessed by removing observations with implausible feature values such as negative velocity. The same index was established as for the ground truth data. The index parameters *datetime*, *cell middle*, and *direction class* were generated. Afterwards, features were generated for each five-minute interval of a segment through aggregating according FCD values. This led to approximately 1.4 billion observations. Several features were extracted and they are presented in Chapter 6.

PublicationTime	...	*Description*	*Latitude to*	...	*Longitude from*
2019-08-01 07:19:16	...	A4 Köln Richtung Olpe; Ausfahrt Kreuz Köln-Ost; Dauerbaustelle, Verbindungsfahrbahn gesperrt, bis 31.12.2020 Mitternacht	50.59458	...	6.92014
2019-08-01 07:19:16	...	A4 Heerlen/Aachen Richtung Köln; zwischen Köln-Klettenberg und Kreuz Köln-Süd; Gefahr durch Gegenstände auf der Fahrbahn; Pappe und Papier	51.40781	...	7.52615
		...			
2019-08-01 08:04:16	...	A1 Münster Richtung Dortmund; zwischen Kamen-Zentrum und Raststätte Lichtendorf Nord; **Stau**	51.51236	...	7.66717
		...			

Table 5.2: Sample of ground truth data set

The ground truth labels and the features were merged in order to generate the data corpus used for modelling. The data corpus contained approximately 66 million observations but observations with missing feature values were eliminated from the data corpus. Characteristics of the data corpus can be seen in Table 5.3. The number of motorway segments, derived from the composite index variables *cell middle* and *direction class*, was approximately 1,000. Approximately 63,000 five-minute intervals existed in the time-period between August 2019 and February 2020. The mean number of five-minute time intervals without any missing feature values was approximately 42,000 and the total number of observations in the data corpus was approximately 45 million.

A sample of the data corpus is provided in Table 5.4. The corpus has the index parameters

Data set	**Data corpus**
Number of motorway segments	1,050
Mean number of five-minute intervals per motorway segment	42,479
Number of obs. in data corpus	45 M

Table 5.3: Characteristics of the data corpus

Index			**Features**			***Ground truth target***		
Datetime	***Cell middle***	***Dir. class***	***Avg velocity***	***%50 velocity***	...	**Current**	**5 Min.**	...
2019-08-01 08:00:00	51.51, 7.67	1	88	91	...	free-flowing	congestion	...
2019-08-01 08:05:00	51.51, 7.67	1	90	101	...	congestion	congestion	...
			...					
2019-08-01 08:00:00	51.65, 7.04	0	98	105	...	free-flowing	free-flowing	...
			...					

Table 5.4: Data corpus sample

datetime, *cell middle*, and *direction class*; it has the exemplary features *avg velocity* and *% 50 velocity*, and the exemplary current and five-minute near-future target time horizons. At 8:00 on 2019-08-01 in the cell with midpoint 51.51, 7.67 and direction class 1, the average velocity of FCD points was 88 km/h, the median velocity was 91 km/h, no instantaneous congestion was reported, and congestion was reported five minutes later. The congestion target class for five minutes in the future corresponds to the traffic congestion message seen in Table 5.2. Five minutes later in the same cell and direction class, congestion was reported at the instantaneous time as well as in five minutes in the future. The third observation in Table 5.4 belongs to 8:00 on 2019-08-01 in a different cell with the other direction class, and free-flowing traffic was found in the instantaneous time and the future five-minute interval.

6. *Experimental results*

In this chapter, the data corpus was described through analyses and *congestion states* were forecasted for the near future. The whole data corpus was used for generating descriptive insights and approximately 30% of the data corpus observations were used for the forecasting evaluation in the test data set. Furthermore, the computational effort of the developed direction distinctive grid-based approach was compared to the effort of a map-matcher. The results of this chapter were condensed lastly since the information density regarding the research questions is high in this chapter.

6.1 Descriptive insights

Descriptive and exploratory data analyses were performed to gain insights into the data corpus. Possible associations were evaluated as they can influence the ML model interpretation. For simplicity, descriptive insights rely on the five-minute future *congestion state* if not further defined.

6.1.1 Ground truth target

Several descriptive plots reveal patterns of the ground truth congestion behaviour in the observed time period of August 2019 to February 2020.

Figure 6-1 shows the monthly number of congestion events in thousands. To describe a data point, the number of observations in the data corpus having the ground truth *congestion* class was approximately 10 thousand in August 2019. One observation referred to a five-minute time interval of one segment, or more technical, of one direction class of a cell. A strong variation in the number of congestion events is apparent. Most congestion events were recorded in November 2019. The total number of congestion events in the regarded time period is 147 thousand.

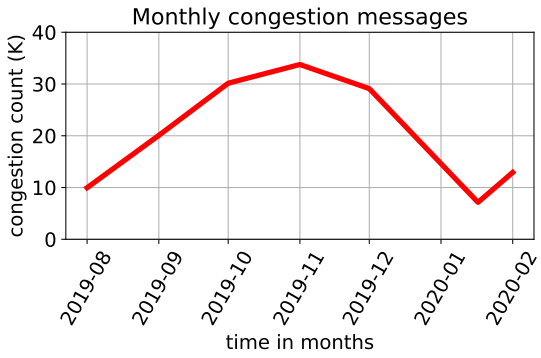


Figure 6-1: Observations with *congestion* target class in time course

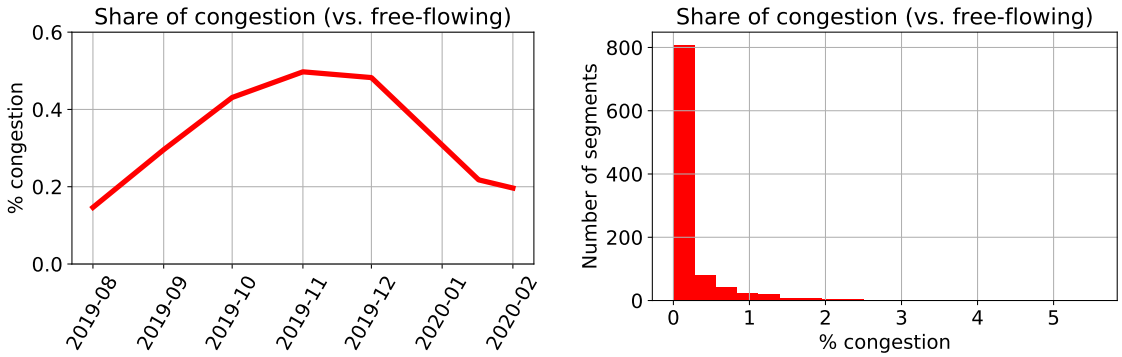


Figure 6-2: *Congestion* target class share in the temporal and spatial dimension

Figure 6-2 shows the share of the target label *congested* (vs. *free-flowing*) based on the time and spatial component. November 2019 recorded the highest share with a *congestion* rate of approximately 0.5%. By far the most motorway segments had a *congestion* rate of 0-0.25% during the whole time period, whereas only a few segments recorded 1% or more. The *congestion* rate based on all observations was 0.3%. It can therefore be stated that the target variable had extremely unbalanced classes. *Congestion* occurred much less often than *free-flowing* traffic, matching intuition.

6.1.2 Features

A feature overview is given first to get a better impression of the data. Thereafter, correlations between the metric features are examined to understand possible overlay effects in the developed ML models.

Feature	# Classes	Mode	Mode's frequency
Weekday	7	Friday	6.7 M
Hour	24	17	2.3 M
UnderConstruction	2	0	43.3 M
IsCrash	2	0	43.2 M

Table 6.1: Summary statistics of categorical features (N=45 M)

Overview of features

Tables 6.1 and 6.2 display statistical characteristics of the categorical and metric features in the data corpus. The feature values were extracted per observation of the index value, consisting of *datetime*, *cell*, and *direction class*.

Table 6.1 presents the number of classes of each categorical feature as well as the mode, outlining the most occurred class along with its occurrence rate.
The *weekday* feature incorporated seven classes. The most frequent class was *Friday*, which occurred approximately 6.7 million times in the data corpus.
The *hour* ranged from 0 to 23 having 24 classes. The mode was 17 and its frequency was approximately 2.3 million, implying that approximately 2.3 million observations were observed between 17:00 and 17:59.
The *underConstruction* feature with classes 0 and 1 refers to not under construction (class 0) or under construction (class 1). Most observations, approximately 43.3 million, were recorded from carriageway segments not under construction.
Class 0 of feature *isCrash* refers to no observed crash. On the contrary, class 1 refers to crashes. For most of the observations, approximately 43.2 million, no crash was reported.

In Table 6.2, the statistical key figures median, % 15 quantile and % 85 quantile are displayed for the metric features of the data corpus due to their robustness against outliers.
The quantile feature *% 20 velocity* is the upper threshold for the lowest 20% velocity values recorded per *datetime* interval, *cell*, and *direction class*. Its median was 79 km/hour, the 15% quantile for the 20% velocity quantile of an index was 23 km/hour and the 85% quantile was 114 km/hour. The assumption that most observations were recorded on a motorway can be maintained, since the minimum velocity is usually 80 km/hour on motorways. And

Feature	Unit of feature	15%	Median	85%
% 20 Velocity	km/h	23	79	114
% 50 Velocity	km/h	52	96	126
% 85 Velocity	km/h	83	116	144
Avg velocity	km/h	56	94	124
Stderr velocity	km/h	7	21	39
Traffic count		5	19	51
Traffic count distinct		2	4	10
Traffic count scaled robust		- 0.017	0.007	0.060
Traffic count scaled cell robust		- 0.275	0.029	0.531
Traffic count previous		5	19	51
Traffic count north		0	3	28
Traffic count east		0	7	31
Traffic count south		0	2	28
Traffic count west		0	8	31

Table 6.2: Summary statistics of metric features (N=45 M)

the slowest 20% of vehicles have a velocity of almost 80 km/hour as median. The statistical characteristics of the remaining velocity-related features can be interpreted analogue.

The feature *traffic count* refers to the number of grouped FCD observations with an according index of *datetime*, *cell*, and *direction class*. The 15% quantile *traffic count* value was 5, meaning that five observations were used to compute feature values. The median traffic count was 19 and the 85% quantile *traffic count* was 51.

Traffic count distinct reveals the number of observations from distinct *ids* of an index value. Distinct *ids* are generally linked to distinct vehicles. A *traffic count distinct* value is less or equal to the according *traffic count* value. Its median was 4, the 15% quantile was 2, and the 85% quantile was 10.

The *traffic count scaled robust* feature is the robustly scaled *traffic count* feature. The idea behind this feature is to capture the varying road infrastructure settings of differing cells. *Traffic count scaled robust* subtracts the median *traffic count* of the entire data corpus from the observation value and divides the result by the difference of the 90% *traffic count* quantile and the 10% *traffic count* quantile of the entire data corpus. The median value was 0.007, the 15% quantile was -0.017 and the 85% quantile was 0.06.

Traffic count scaled cell robust is the cell-wise robust scaled *traffic count* feature. It is supposed to describe deviations of the *traffic count* in differing time periods of a cell. The 15% quantile value was - 0.275 and the 85% quantile value was 0.531; the median was 0.029.
Traffic count previous measures the *traffic count* in the previous five-minute time interval of a motorway segment. The value had a 15% quantile value of 5 and an 85% quantile value of 51. The median *traffic count previous* was 19, as the median *traffic count*.
Traffic count north illustrates the *traffic count* of the motorway segment north to the considered motorway segment. The direction classes of the two neighbouring cells are the same. The minimal value is zero. It is observed when no FCD has been recorded for the northern motorway segment in the according time interval. The 15% quantile was 0, the median was 3, and the 85% quantile was 28.
The remaining traffic count values, from cells in the according compass directions of the regarded cell, can be interpreted just like the feature above.

Correlations between metric features

Several provided metric features are prone to reveal correlations due to causal dependencies. For example, the *traffic count* and *traffic count distinct* are causally connected. Strong correlations should be identified in order to interpret results correctly. Correlations between features are displayed through the heatmap of Figure 6-3. Medium ($0.6 < |corr| < 0.75$) and strong correlations ($0.75 < |corr|$) between features are discussed.

The features *% 20 velocity*, *% 50 velocity*, *% 80 velocity*, and *avg velocity* were medium to strongly positively correlated. This could have been expected from causal connections of the features. The feature values were derived from the same set of velocity data per index entry.
Traffic count, *traffic count distinct*, and *traffic count scaled robust* were medium to strongly positively correlated. This is well explainable since the *traffic count* measured every observation and the *traffic count distinct* measured observations of distinct *ids*.
The feature *traffic count scaled cell robust* was medium positively correlated to *traffic count* and *traffic count scaled robust*.
Traffic count previous was strongly positively correlated to *traffic count* and *traffic count scaled robust*, as well as medium positively correlated to *traffic count distinct*.

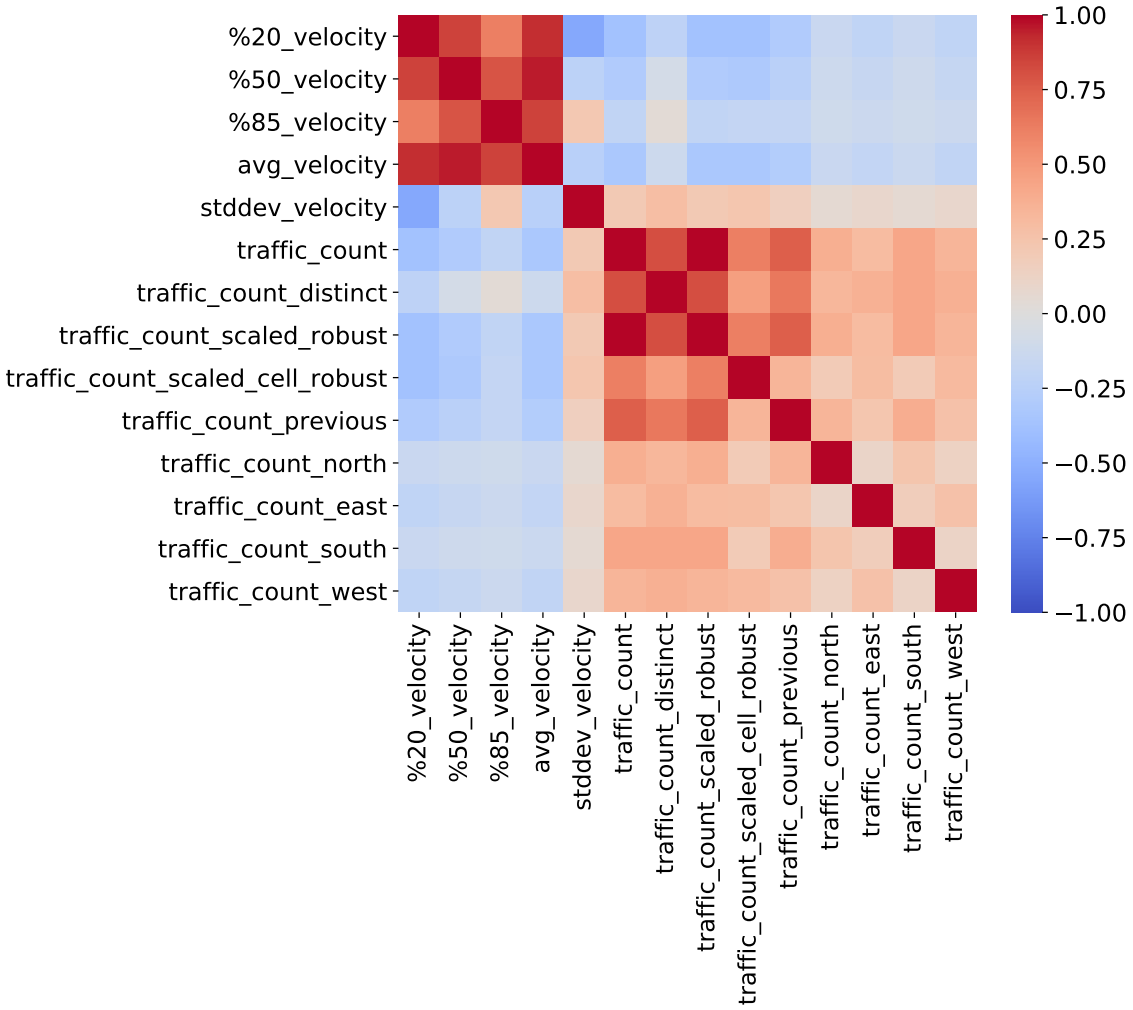


Figure 6-3: Correlations of metric features

6.1.3 Connections between feature and target

The connection between each feature and the current as well as five-minute future congestion state was examined through two different methods and is displayed in Table 6.3. The *congestion state* at that time was examined due to expected connections of velocity-related features and the target. The five-minute congestion state was investigated representatively for the forecasting target variables as before. The connection between categorical features and the *congestion state* was measured by a statistical independence test. The point biserial correlation coefficient (cc) was used for metric features and the *congestion state*.

Feature	**Type**	**Target current**	**Target 5 min. forecast**
Weekday	pv	<0.01	<0.01
Hour	pv	<0.01	<0.01
UnderConstruction	pv	<0.01	<0.01
IsCrash	pv	<0.01	<0.01
% 20 Velocity	cc	-0.05	-0.05
%50 Velocity	cc	-0.06	-0.06
% 85 Velocity	cc	-0.07	-0.07
Avg velocity	cc	-0.06	-0.07
Stderr velocity	cc	-0.01	-0.01
Traffic count	cc	0.09	0.08
Traffic count distinct	cc	0.05	0.04
Traffic count scaled robust	cc	0.09	0.08
Traffic count scaled cell robust	cc	0.09	0.08
Traffic count previous	cc	0.09	0.08
Traffic count north	cc	0.05	0.05
Traffic count east	cc	0.04	0.04
Traffic count south	cc	0.04	0.04
Traffic count west	cc	0.06	0.06

Table 6.3: Connections between each feature and the *congestion state*

Pearson's independence test, returning a p-value (pv), was performed for categorical features with the current congestion state at a significance level of 5%.
The null hypothesis of stochastic independency of the *weekday* and *congestion state* could not be accepted to the 5% significance level. Furthermore, there seems to be a stochastic dependency between the *hour* of day and the *congestion state*. The *underConstruction* feature and the *congestion state*, as well as the *isCrash* feature and the *congestion state* are seemingly interdependent.

The point biserial correlation coefficient $\in [-1, 1]$ was computed for a random sample of the data corpus (5%, 2.2 M obs.) in *scikit-learn*, since the functionality was not available in *PySpark*. The value zero was assigned to the *free-flowing* class, and the value one was

assigned to the *congestion* class. Moreover, a hypothesis test with an assumed correlation coefficient of zero was statistically significant to the 5% level for each combination.

Each metric feature was only weakly correlated with the *congestion state* at that time as well as the five-minute future *congestion state*. The results for the two target variables differ only marginally for each combination. The velocity-based features were barely negatively correlated to the *congestion state*. *Congestion* (label 1) was very scarcely linearly associated with declining velocities. A strong correlation of the *% 50 velocity* and the *congestion state* has been expected due to the ground truth gathering methodology. Ground truth *congestion states* have been computed through thresholding the average velocity of all vehicles at motorway traffic detectors. *Traffic count* related features were barely positively correlated to the *congestion state*. A positive correlation has been expected due to an increasing number of FCD points originating from a vehicle when spending more time inside a cell. More time spent in a cell segment is generally causally connected to less flowing traffic situations on a motorway. *Traffic count* features from neighbouring cells were slightly correlated to the *congestion state* as well. This is in alignment with expectation since the influence of neighbouring cells depends on the specific cell structure and cannot be generalized to all cells collectively.

6.1.4 Feature sets

Three feature sets were separated, as shown in Table 6.4. Strongly and medium correlated features were eliminated (one of them) in feature set 1 (*fs1*) leading to a better interpretation of feature importances. Every feature, presented in Section 6.1.2, stayed in feature set 2 (*fs2*) to gain additional prediction performance through the additional features. Even if the additional features were medium to strongly correlated to the base features, they could add some further information. Feature set 3 (*fs3*) excluded features measuring the traffic situation at neighbouring motorway segments.

In feature set 1, only one feature was decided on regarding the average velocity and quantile velocity features. The *% 50 velocity* feature was favoured over the *avg velocity* due to its robustness. Robustness is important since the ground truth average velocity relied only on vehicles on the motorway segment and the feature velocity values were computed by a small subset of vehicles which additionally did majorly, but not always travel on motorways but also on main roads, for example. The *traffic count* feature was selected as the representative

Feature	*Fs1*	*Fs2*	*Fs3*
Weekday	✓	✓	✓
Hour	✓	✓	✓
UnderConstruction	✓	✓	✓
IsCrash	✓	✓	✓
% 20 Velocity		✓	✓
% 50 Velocity	✓	✓	✓
% 85 Velocity		✓	✓
Avg velocity		✓	✓
Stderr velocity	✓	✓	✓
Traffic count	✓	✓	✓
Traffic count distinct		✓	✓
Traffic count scaled robust		✓	✓
Traffic count scaled cell robust		✓	✓
Traffic count previous		✓	✓
Traffic count north	✓	✓	
Traffic count east	✓	✓	
Traffic count south	✓	✓	
Traffic count west	✓	✓	

Table 6.4: Utilised feature sets

for the traffic count related features.

6.1.5 Visualisation of selected features

Selected *fs1* features from Table 6.4 are displayed graphically with respect to the ground truth target class of the five-minute forecasting time period.

Figure 6-4 shows histograms of the data corpus for features from the time-domain with respect to the *congestion* and *free-flowing* class respectively. The frequency of *free-flowing* traffic was lower on Sundays compared to the other weekdays. In contrast, *congestion* observations dropped strongly on Saturdays and Sundays. Tuesdays and Wednesdays were the most congestion-prone weekdays with the rest of the weekdays at a lower level. The

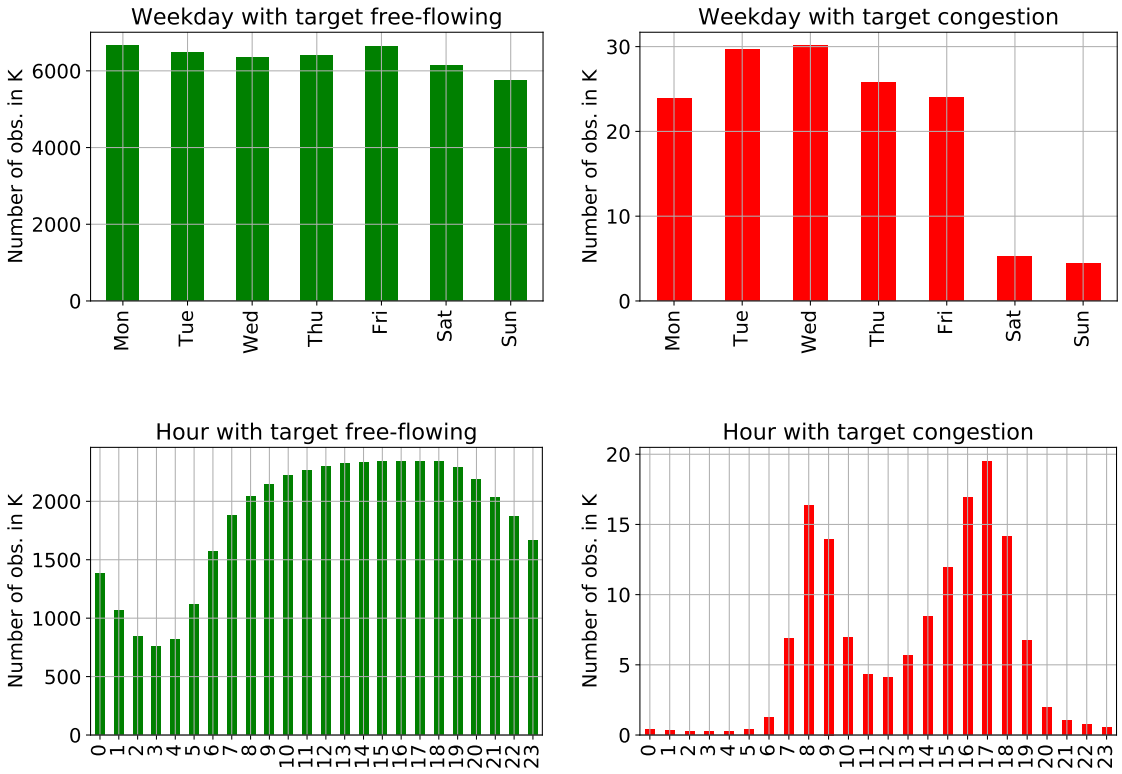


Figure 6-4: Histograms of time-domain features (N=45 M)

congestion behaviour in the data corpus was weekly periodic, as stated by scientific works from Section 3.1.1.

The *hour* of a day seems to lead to a differing FCD coverage which is derived from a differing frequency of the *free-flowing* congestion state. At nighttime, from midnight to 05:00, the frequency dropped strongly in the *free-flowing* class. *Congestion* occurred most frequently around peaks at 08:00 and 17:00. Only very few *congestion* observations were recorded at nighttime as well. The feature *hour* indicates a periodicity of *congestion*, which is congruent to findings in Section 3.1.1.

Figure 6-5 shows histograms of the data corpus for selected metric features with respect to the *congestion* and *free-flowing* class. The median velocity peaked between 87 and 105 km/hour for the *free-flowing* class, which is a decent travelling velocity on motorways. The low velocities could come from slow-moving traffic, temporarily major load on motorway surrounding streets, or badly arranged motorway segments. The median velocity for the

Figure 6-5: Histograms of selected metric features (N=45 M)

congestion class peaked at 23 km/hour and 81 km/hour. The first peak has been expected since congestion relates to slow velocities per ground truth definition. The second peak was not expected.

Traffic count histograms differ between the two *congestion states*. This was expected because a congested motorway segment results in more FCD observations, which are generally recorded in a specific time interval.

6.2 Forecasting traffic congestion states

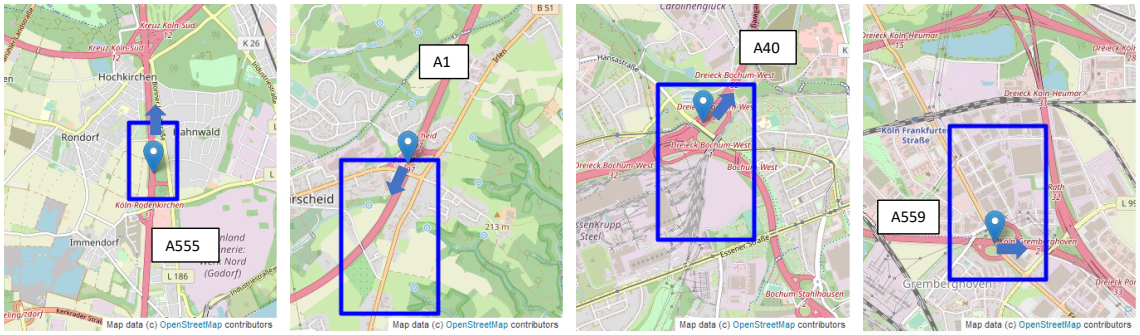
Traffic congestion states were forecasted based on the RF algorithm described in Section 4.4. Models that were developed and evaluated based on a single direction distinctive cell are presented firstly. The following sections rely on models developed based on the segment's collective. The feature importance scores are examined secondly. Evaluation results of the whole grid models are presented thirdly and interesting model characteristics are shown lastly.

6.2.1 Performance evaluation in the single segment setting

RF models developed on a train data set of one motorway segment are utilized to make forecasts on the same motorway segment in this section. Feature set 2 from Table 6.4 was utilized for model development if not mentioned differently.

Four segments are presented in their spatial context and evaluation results for single segment models of the segments are shown in three dimensions. The BM evaluation measure was relied on since this measure is robust against different proportions of the target classes in the four segments. The first evaluation dimension regards different time periods in the single segment setting and the second dimension examines the impact of the features from neighbouring cells in the single segment setting. In the third dimension, evaluation results of the single segment model and the whole grid model were compared. Four exemplary segments that had a *congestion* share of more than 1%, being among the top 30 segments with the largest share of the *congestion* class, were selected for these analyses. The idea behind relying only on segments with a large *congestion* share is the justification of the computationally more expensive single segment models only for segments with much congestion. Moreover, the concentration on a few segments has the advantage of introducing the direction distinctive cells separately. The analyses indicate the impact of the three dimensions for *congestion state* forecasts. All three figures show boxplots including the median, the interquartile range (IQR) and the whiskers. The BM baseline was displayed additionally.

Figure 6-6 shows the four exemplary segments within their cells in their spatial context. The markers represent the traffic detectors that led to ground truth *congestion states* and the arrows characterize the according carriageway directions. Cell (50.87, 6.97) in the blue rectangle as displayed in Figure 6-6a, incorporates a vertical motorway segment on A555. The traffic detector is located on the carriageway into the northern direction, corresponding to direction class 1 as symbolized with the blue arrow, and approximately 48,000 five-minute intervals belong to the segment. Its traffic arises from the neighbouring southern motorway segment. Figure 6-6b displays the motorway carriageway segment of A1 in cell (51.08, 7.13) in the southwest direction, corresponding to direction class 1, with approximately 53,000 observations. Traffic comes from the northern neighbouring cell. The federal highway B51 is located in parallel. The motorway segment of A40 in cell (51.48, 7.18) and direction class



(a) Cell (50.87, 6.97) and direction class 1 (N=48 K)

(b) Cell (51.08, 7.13) and direction class 1 (N=53 K)

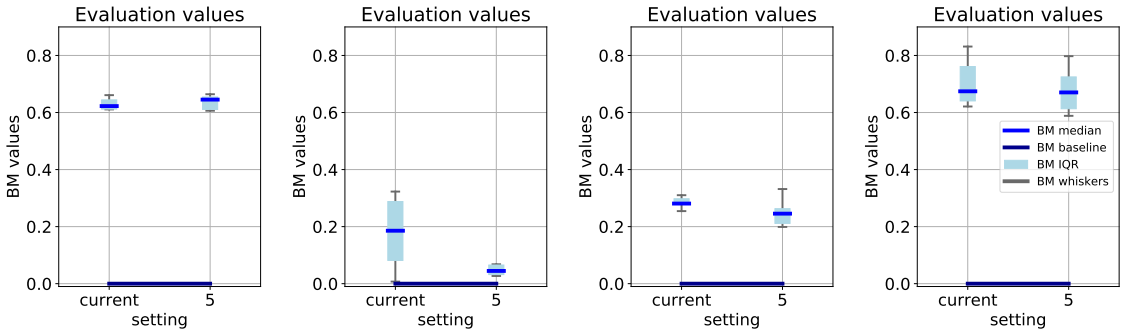
(c) Cell (51.48, 7.18) and direction class 0 (N=49 K)

(d) Cell (50.91, 7.06) and direction class 0 (N=53 K)

Figure 6-6: Exemplary motorway segments with an overproportional *congestion* share

0 into the northeast direction in Figure 6-6c had approximately 49,000 observations and obtains traffic from its western and eastern cell. Both neighbouring cells are responsible for forwarding vehicles since the regarded cell incorporates a motorway junction. The motorway segment of A559 in cell (50.91, 7.06) and direction class 0 of Figure 6-6c into the eastern direction and approximately 53,000 observations, has traffic coming from the western direction.

Figure 6-7 evaluates whether a performance drop can be observed between the instantaneous *congestion state* prediction and the five-minute forecast. The drop was expected due to autocorrelations in time (cf. Section 3.1.1). A comparable large drop from the instantaneous prediction to the forecast can only be seen in Figure 6-7b. It is noticeable that the IQR



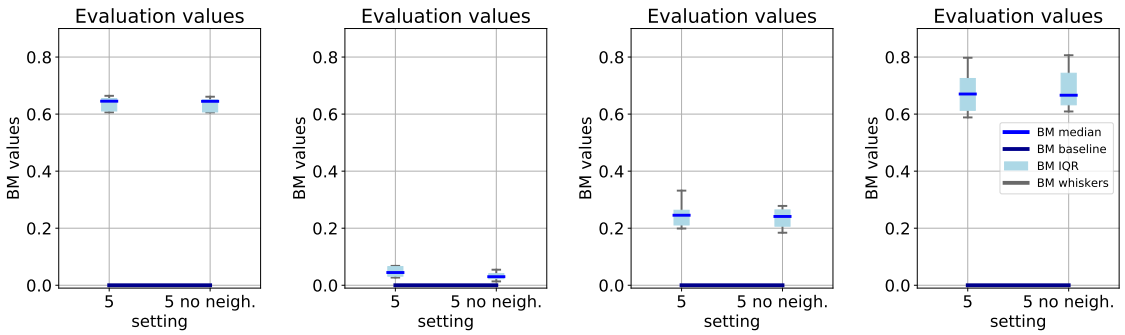
(a) Cell (50.87, 6.97) and direction class 1

(b) Cell (51.08, 7.13) and direction class 1

(c) Cell (51.48, 7.18) and direction class 0

(d) Cell (50.91, 7.06) and direction class 0

Figure 6-7: Random forest model performance evaluation diagram in the single segment setting for differing time periods



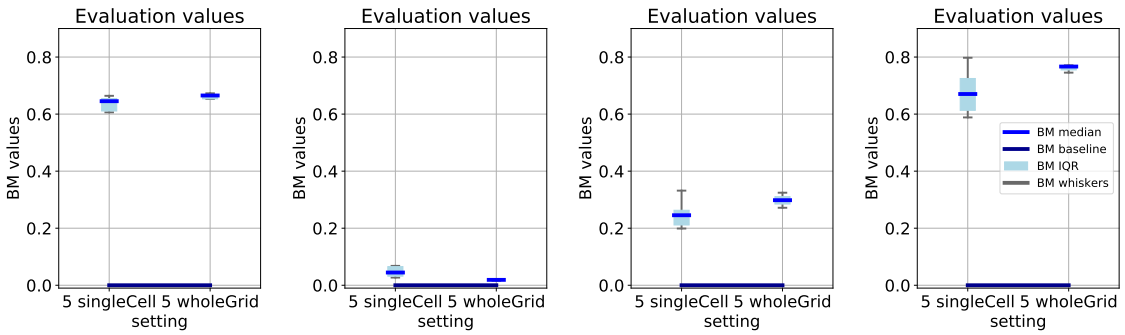
(a) Cell (50.87, 6.97) and direction class 1 (b) Cell (51.08, 7.13) and direction class 1 (c) Cell (51.48, 7.18) and direction class 0 (d) Cell (50.91, 7.06) and direction class 0

Figure 6-8: Random forest model performance evaluation diagram in the single segment setting for a differing feature space

in the figure was comparable large for the instantaneous prediction. Figures 6-7c and 6-7d show only a slight decrease in performance and figure 6-9a has even a slightly greater mean BM value for the five-minute forecast.

Figure 6-8 examines the impact of features from neighbouring segments for the five-minute forecast of *congestion states*. Feature set 3 from Table 6.4 was utilised in the right boxplot in each subfigure. Models for all four segments had a small drop in the mean performance when not taking the *traffic count* of neighbouring cells into account. Even so, the impact of the additional neighbouring segment features was quite small. The results indicate no impact of the features from neighbouring cells for models with highly congested motorway segments in NRW.

Figure 6-9 shows evaluation results for the five-minute near-time forecast in the single segment setting and in the whole grid setting. The figures show no continuously superior performance of the single segment models that was expected due to stronger adaption abilities to the regarded segment. The mean BM was lower in the single segment setting of Figures 6-9a, c, and d. Only Figure 6-9b shows superior performance of the single segment model but the segment was generally on a low-performance level as the values close to the baseline show. The evaluation results of a segment in the whole grid setting had only little variation showing that similar RF models were developed for different train-test splits. In contrast, the regarded single segment models had larger variations from different time windows of



(a) Cell (50.87, 6.97) and direction class 1 (b) Cell (51.08, 7.13) and direction class 1 (c) Cell (51.48, 7.18) and direction class 0 (d) Cell (50.91, 7.06) and direction class 0

Figure 6-9: Random forest model performance evaluation diagram in the single segment and whole grid setting

the test data sets. The results indicate no performance improvement by the single segment models in comparison to the whole grid models and the additional computationally effort cannot be justified for the four segments.

Generally, it can be seen that the evaluation values were in a quite large range between the four segments with an overproportional amount of *congestion* incidents. Motorway segments in cell (50.87, 6.97) and cell (50.91, 7.06) with direction classes 1 and 0 respectively had evaluation results on roughly the same high level. The segment in the cell with midpoint (51.08, 7.13) and direction class 1 had by far the worst performance that was only slightly better than the baseline for the five-minute forecast. The federal highway (Bundesstraße) B51 located parallel to the A1 segment might have badly influenced the results. The segment in cell (51.48, 7.18) and direction class 0 had a medium performance level. The medium performance could be due to quite noisy features arising from the second motorway at the motorway junction.

6.2.2 Feature importances in the whole grid setting

Feature importances in the whole grid setting are exemplarily shown for a five-minute forecasting model. Features of feature set 1 and feature set 2 of Table 6.4 were utilized to analyse feature importances. Feature importances are presented in order to dive deeper into the model decision-making process and for clarifying the model results from a causal perspective.

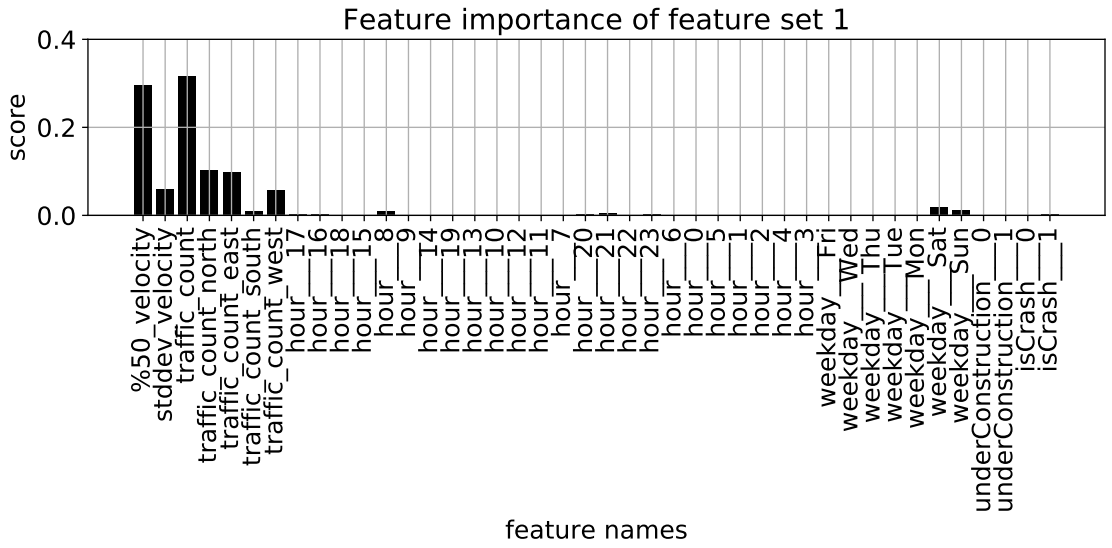


Figure 6-10: Feature importances for the five-minute forecast of feature set 1 based on the whole grid setting

Figure 6-10 shows feature importance scores of feature set 1. The *traffic count* seems to have a major impact on the *congestion state* forecast. A slightly smaller feature importance score had the feature *%50 velocity*. As shown in Section 6.1.3, this influence is seemingly not linear. The features describing *traffic counts* of neighbouring cells (*traffic count north*, *traffic count east*, *traffic count south*, *traffic count west*) seem to influence the *congestion state* as well. This was not expected in the whole grid setting since the impact of neighbouring cells depends highly on the directions of motorway segments from a causal perspective. The standard error of the velocity seems to have some influence on the *congestion state* determination as well. The congestion peak hour 08:00 and the hour after the second decrease of *congestion* incidents, 21:00 (cf. Figure 6-4), seem to influence the model to a small extent. Saturday and Sunday, in which congestion occurred much less frequently (cf. Figure 6-4), seem to have a small influence on forecasting the *congestion state*.

Figure 6-11 shows feature importance scores of features from feature set 2. Velocity-related features as well as traffic count-related features are seemingly of high importance for the congestion state forecast. Contrary to the feature importances of feature set 1, the scores are generally lower due to a spread to many correlated features. Traffic counts from seg-

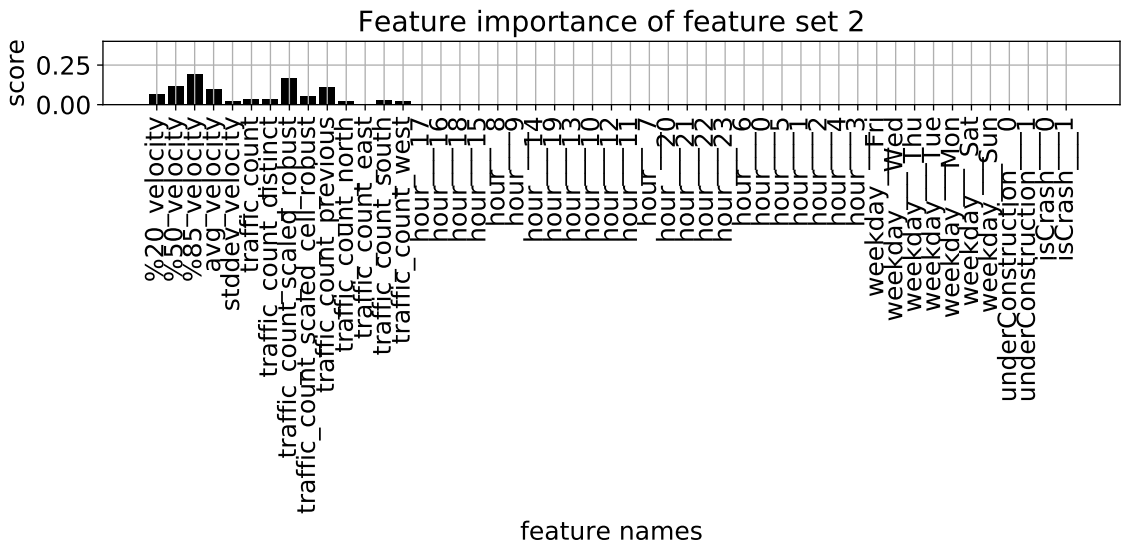


Figure 6-11: Feature importances for the five-minute forecast of feature set 2 based on the whole grid setting

ments of neighbouring cells do not seem to be important to determine the *congestion state*. This seems reasonable because the whole grid model forecast does not differentiate between different segments. The *weekday* and *hour* do not seem to be important for the *congestion state* forecast. Presumably, the timely periodicity was captured by the remaining features.

All in all, traffic count-related features and velocity-related features seem to be of major importance for forecasting the *congestion state*.

6.2.3 Performance evaluation in the whole grid setting

Evaluation results based on segments from the whole NRW grid are presented. The evaluation measures F_1-score, misclassification rate (mcr), and Bookmaker Informedness (BM) are displayed throughout for the test data set with original proportions of the target classes. Evaluation results of the resampled test data set can be seen in Appendix A. Feature set 2 from Table 6.4 was utilized for model development due to a greater source of information and thus at least equal model performance in comparison with the other feature sets.

Figure 6-12 presents evaluation results of the RF model in the whole grid setting for five different time periods in the near future (5, 10, 20, 30, and 60 minutes in the future). The

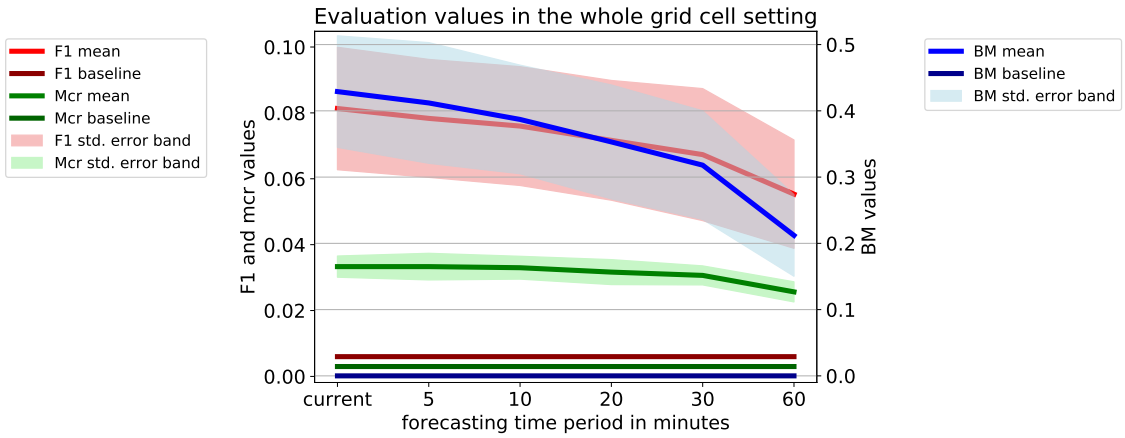


Figure 6-12: Random forest model performance evaluation diagram in the whole grid setting

forecasting time period is displayed on the figure's x-axis. The model's performances were measured with the F_1-score and mcr on the left y-axis and the BM metric on the right y-axis. The baseline values as presented in Section 4.4.4 served as lower thresholds for the RF model evaluation values. The mean and the standard error band are displayed for each evaluation metric as well. The evaluation values of the developed models can be additionally seen in Table A.1.

The mean and standard error band of the F_1-score were higher throughout, and therefore better than the baseline F_1-score for the test data set. This assigns a value to the forecasting models regarding the F_1-score. Even so, it is noticeable that the mean F_1-score decreased with proceeding forecasting time. The F_1 standard error was on a similar level for the different time periods. The mean F_1-score of the five-minute forecast was quite close to the prediction score of the instantaneous *congestion state*. In other words, only little performance was lost by forecasting five-minutes ahead on instantaneous information. A relatively large drop in performance can be seen between the 30-minute forecast and the 60-minute forecast. The decrease can be explained with the comparatively large margin of 30 minutes. The F_1-score generally ranged in a relatively low level for the developed models and the baseline respectively. This is due to the highly imbalanced target data that led to a small precision component of the F_1-score since the FP values in this scenario are generally high. The evaluation values of the mcr metric can be interpreted analogue to F_1-scores besides the fact that the smaller a value gets, the better it is. The mean mcr and its standard error band were higher throughout for the RF models in comparison to the baseline mcr. The forecasted *congestion state* in the five-minute whole grid model was misclassified more often

in comparison to the baseline. Two explanatory approaches are discussed and are based on the fact that the baseline mcr of 0,003 was very low. The first explanatory approach refers to the data quality of the data corpus. Usually, data is not entirely correct but several sources of possible errors exist. In this work, errors might have crept into the transmission of FCD or ground truth data directly. Noise might have sometimes been overrepresented in feature values, and errors might have sneaked into the segment assignment through direction distinctive cells. The mentioned data quality issues might have already led to a higher mcr than the baseline of 0,003. The second explanatory approach refers to the impact of wrong forecasts. Comparable big performance losses can occur through forecasting scenarios of wrongly forecasting the ground truth *congestion* class. Due to the highly imbalanced target, the sheer amount of FP is generally much higher than the amount of FN. In other words, when a model makes a mistake in its decision process, errors of forecasting ground truth *free-flowing* observations as *congestion* leads to more FP than FN would be generated when making a mistake in the opposite scenario. The model cannot be expected to avoid mistakes when forecasting ground truth *free-flowing* observations. Therefore, the results do not undermine the informative value of the models.
The mean BM and its standard error band were higher throughout than the baseline value. In other words, value can be assigned to the RF models regarding this metric. The mean BM decreased considerably with proceeding forecasting time. The standard error remained on a similar level for the differing forecasting time periods. The mean BM for the instantaneous time and the five-minute forecasting time were relatively close to each other. The model performance dropped only slightly when using feature information of the preceding five minutes for the *congestion state* determination. The mean BM dropped considerably from the 30-minute forecast to the 60-minute forecast which can be explained by the comparable large time interval.

As an overall performance evaluation of forecasting *congestion states* in the whole grid setting, the joint result of the evaluation metrics is considered. The mcr values were not taken into account due to reasons outlined above. It was therefore relied on the F_1-score and the BM measure. The mean and the according standard error band of both evaluation measures was considerably better in comparison to the baseline values. Hence, value can seemingly be generated by forecasting *congestion states* in the 5, 10, 20, 30, and 60 minutes near-future

time periods through utilizing an RF model in the whole grid setting. The forecasting performance seems to decrease with proceeding forecasting time. A decreased forecasting performance when time proceeds is reasonable when assuming temporal autocorrelations.

Comparing evaluation results from various studies is prone to biases due to the diverse spectrum of study setups that was shown in Section 3.1.2. A comparison of results from this work and works in Section 3.1.2 was made based on the mcr measure. Studies that used mcr for evaluation had a small sample size (< 2000) which could have led to quite considerable biases. The mcr for the test data set in the whole grid setting was lower throughout and therefore better than the mcr from other works (cf. Table 3.1). The mcr in this work might have been lower in comparison with other studies due to a smaller number of *congestion states* and a greater target class imbalance. Even so, the target data proportions were mainly not reported and it was not relied on unbiased evaluation metrics for imbalanced data. Amongst others, the differing number of congestion states could have furthermore influenced the evaluation results. A comparison to studies described in Section 3.2.1 was not made due to the differing target's spatial area. In this work, forecasts were developed for motorway segments.

6.2.4 Model characteristics in the whole grid setting

The five-minute forecasting model with seed one of the whole grid setting was selected for demonstrating characteristics and presenting insights regarding its forecasting performance.

The timely periodic behaviour seems to be captured by the model. Figure 6-13 shows TP and FN belonging to the left y-axis as well as FP relying on the right y-axis of the features *weekday* and *hour* in the two plots. Both plots show that the periodic *congestion* behaviour was captured. The proportions of the three measures TP, FN, and FP roughly correspond with each other and roughly match the proportions of the ground truth *congestion* observations in Figure 6-4 as well. Even so, many wrongly forecasted *congestion* observations can be seen. This point was already discussed in the previous section.

Figure 6-14 outlines the histograms of the *%50 velocity* and *traffic count* features. The TP and FN values correspond to the left y-axis and the FP values correspond to the right y-axis

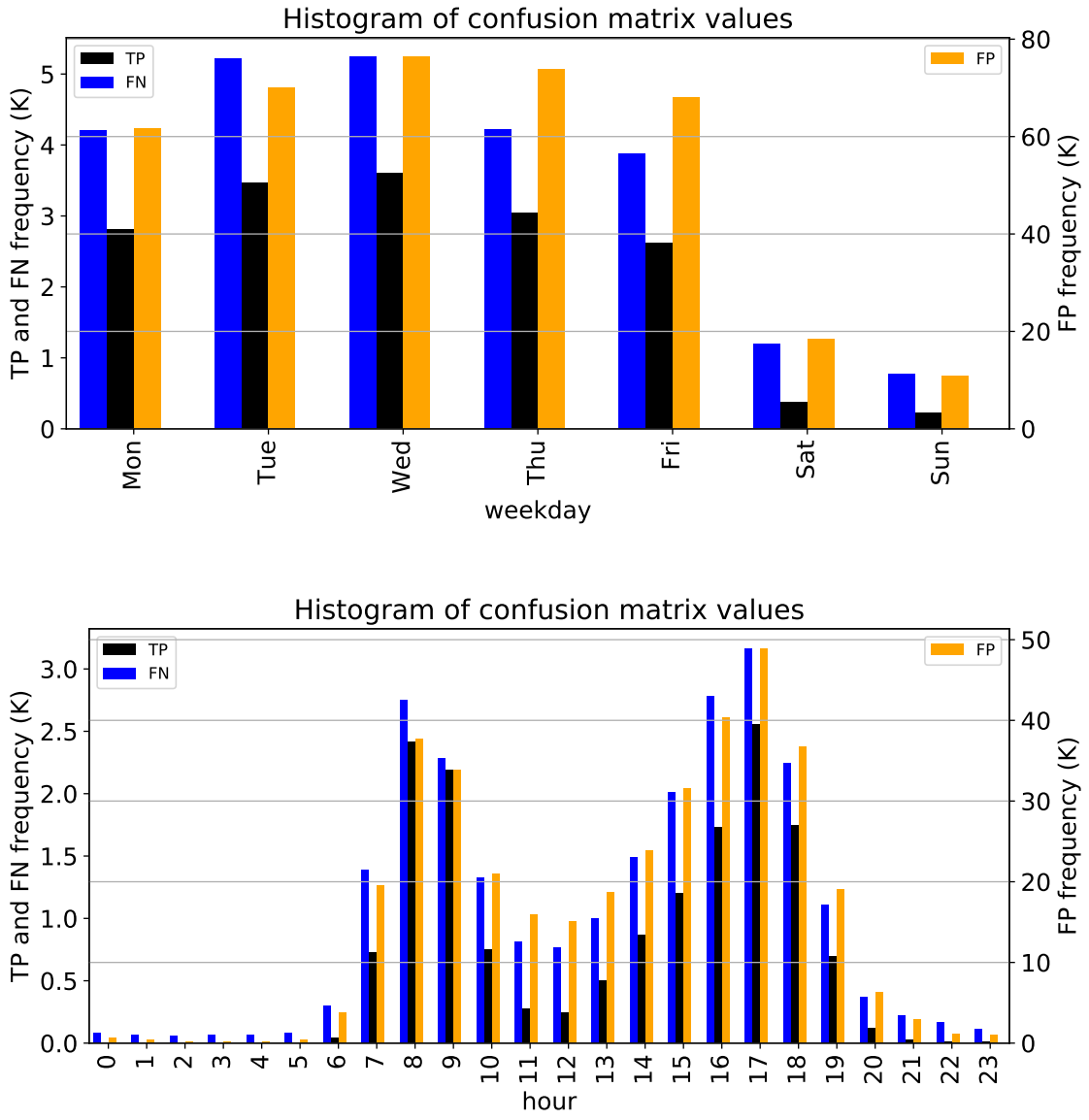


Figure 6-13: Histograms of confusion matrix characteristics from a five-minute forecast in the whole grid setting for time-domain features

again in both plots. The proportions of the TP and FP for the feature *%50 velocity* are roughly aligned. On the contrary, FN seem to have a different distribution which peaked at 90 km/h. Observations with a median velocity of 75 km/h or above were not assigned to the *congestion* class as can be seen in the figure. This raises the question of why observations with such high velocities were assigned to the ground truth *congestion* class in the first place. Possible causes are manifold, including overlapping effects of street segments from other motorways or roads. The congested motorway segment could also have a time period

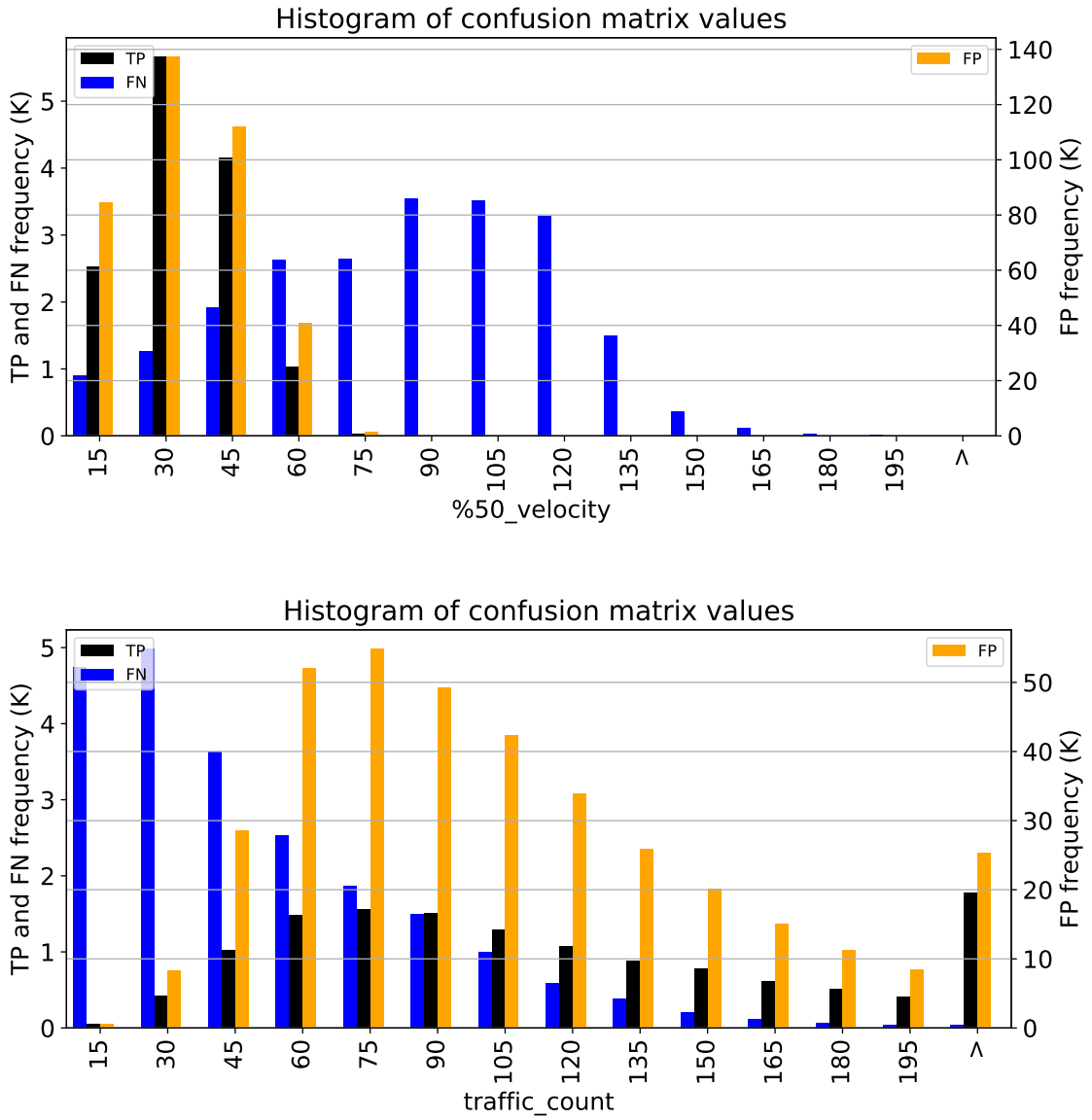


Figure 6-14: Barplots of confusion matrix characteristics from a five-minute forecast in the whole grid setting for selected metric features

of relief, which did not reach the boundary for assigning *free-flowing* traffic.

The *traffic count* feature, analogue to the *%50 velocity* feature, seems to have a different underlying distribution for FP than for TP and FN. The association of *congestion* and higher *traffic count* values is not as obvious as the relation of slow *velocities* and *congestion*. Traffic *congestion* generally leads to longer time periods on a motorway segment which corresponds to more FCD observations and hence to higher *traffic count* values.

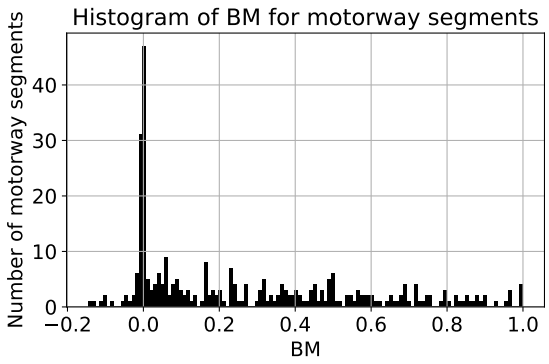


Figure 6-15: Evaluation values per motorway segment (N = 320) from a five-minute forecasting model in the whole grid setting

Furthermore, the BM value was computed separately for each motorway segment in the test data set of one run of the whole grid setting. It was only relied on this evaluation metric since it is the only robust one regarding different target class proportions. Different segments generally have different shares of the *congestion* and *free-flowing* target classes. Figure 6-15 shows the segment-wise evaluation measure of the 320 motorway segments. The plot shows a strong variation between evaluation values of distinct segments. The plot has a peak at the value zero coming from segments with very little *congestion* observations. To be precise, 56 segments had less than ten times *congestion* as ground truth target class, a tpr of zero and a BM value of less or equal to zero. The findings therefore do not question the model's benefit assessed in the previous section. The majority of segments (N=219) had BM values spread between zero and one. Therefore, the forecasting power for five minutes in the future seems to vary considerably between different segments in the grid. This could be explained causally by different carriageway segment sizes and spatial surroundings inside cells. A detailed inspection of strengths and weaknesses for various kinds of segments was beyond the scope of this work.

6.3 Computational effort comparison between different segment-assignment approaches

The difference in the computational time of the motorway direction distinctive grid-based and map-matcher approach was examined. As described in Section 3.2.2, the two approaches are quite distinct. In the motorway direction distinctive grid-based preprocessing step, FCD

Characteristics	Direction distinctive grid-based	Map-matcher based
Volume in GB	57 MB	
Number of obs.	2.6 M	
Area	Broad region of Düsseldorf (51.124375, 6.543125), (51.398238, 6.939885)	
Processing time in min.	13	19
Std. error in min.	<1	<1

Table 6.5: Segment assignment computational time for the motorway direction distinctive grid-based approach and the map-matcher approach

is assigned to road segments based on grid cells and direction classes. A map-matcher assigns FCD points to road segments through an algorithm based on a GIS map that incorporates road segments of the road network. Details of the utilized infrastructure can be found in Section 5.1. It was assumed that the more basic motorway direction distinctive grid-based approach led to less computational time.

The computational effort was benchmarked for an FCD sample data set of one day with a volume of approximately 57 MB, as can be seen in Table 6.5. It encompassed approximately 2.6 million observations in the broad region of Düsseldorf, NRW. The motorway direction distinctive grid-based preprocessing method led to a mean processing time of approximately 13 minutes and a standard error of less than a minute. The map-matcher approach had a mean processing time of approximately 19 minutes and also a standard error of less than a minute. As a result, the computational effort of the motorway direction distinctive grid-based segment-assignment step was more than a 30% less in comparison to the map-matcher approach. This conclusion is limited to the utilized infrastructure or in other words, results may differ when utilizing different amount of cores, servers, RAM, et cetera.

6.4 Subsumption of results

Results from the previous sections in this chapter are summarized at a higher level of abstraction in this part. Furthermore, the gathered results are condensed in the context of the research questions.

This paragraph explains why inductive reasoning was applied to the regarded data corpus that was called data sample below. Results that were found for the sample with approx-

imately 1,000 motorway segments in NRW were generalized throughout for the statistical population of motorway segments in the whole of NRW. The sample of 1,000 motorway segments incorporated a large variety of different segment sizes, of different motorways, and of surroundings of segments inside cells. It is therefore regarded as representative of the entirety of motorway segments in NRW. The model's reliability is furthermore drawn from the comparable small standard errors when using different segments in the train and test dataset.

Findings based on features and the target of ML models from the five-minute forecasting time interval were generalized for all regarded near-future time intervals due to temporal autocorrelations of traffic-related variables (cf. Section 3.1.1). Supplementary, generalizations are based on the received performance benefit of forecasting congestion states in even the largest 60-minute time interval and hence occurred autocorrelations between a 60-minute time interval of the traffic variables.

Results of single segment models for selected highly congested motorway segments and a five-minute forecast indicate that features of neighbouring cells add no value to single highly-congested segment models in the whole of NRW and a near-time forecast of up to 60 minutes. The results do not fit with the expected spatial correlation (cf. Section 3.1.1) and causal connection between instantaneous traffic from neighbouring motorway segments and traffic in the future five-minute interval of the regarded segment. A more sophisticated approach on single cells might be able to improve the evaluation performances. The results gave furthermore the indication that single highly-congested segment models should not be preferred over the more general so-called whole grid model for highly-congested cells in NRW and a forecasting interval of up to 60 minutes. The given indication is in-line with no found improvement from neighbouring cells.

Several features that were provided to ML models seem to capture valuable information to forecast congestion states up to 60 minutes on motorway segments in NRW based on their feature importance scores. Results for a five-minute whole grid model showed that velocity-related features and *traffic count* related features were of high importance. In other words, the approximated instantaneous velocity of motorists played an important role in determining the near-future *congestion state* in five minutes in the regarded data corpus.

The results are generalized to each forecasting time period due to autocorrelations in time of the features (cf. 3.1.1) and the segments assumed representativeness for motorway segments in the whole of NRW. Therefore, velocity-related features seem to be very important for forecasting *congestion states* on the motorway network of the whole of NRW and a forecasting interval of up to 60 minutes. This matches intuition since *congested* traffic is causally related to lower velocities. *Traffic count*-related features seem to be of high importance for forecasting *congestion states* on NRW motorway segments in the future time interval of up to 60 minutes as well. *Traffic count*-related features give an approximate measurement of the amount of vehicles on a motorway segment. The duration of vehicles on a motorway segment is causally important for determining traffic *congestion states* and the *traffic count* is assumed to represent the duration.

Whole grid ML models seem to be capable of adding value to *congestion state* forecasts in the whole of NRW for an arbitrary time interval of up to 60 minutes. The developed models appear to be capable of forecasting *congestion states* for the whole of NRW since the 1,000 regarded sample motorway segments were spread over the entire area. Developed models up to a forecasting time of 60 minutes in the future added value to the *congestion state* estimation. The closer the forecasting time period was to the instantaneous time, the better the *congestion state* forecast has been. The behaviour can be explained by temporal autocorrelations of features and the target.

A comparison of the evaluation results from this work and other studies is prone to biases due to the diverse spectrum of study setups (cf. Section 3.1.2). An additionally and widely utilized benchmark data set could have reduced the impact of differing study setups. As stated in Section 3.1.3, scientific work in the traffic congestion determination domain does not predominantly rely on benchmark data sets. Furthermore, no proper benchmark data set incorporating FCD existed until the time of writing this work. Results of introduced methodologies by other studies could therefore only be compared to results of this work based on different data corpora which strongly limited its informative value. Another barrier for comparing results was the unavailability of a standardized metrics utilization. Moreover, frequently presented metrics can only poorly adopt to a differing number of congestion states, to differing target class shares, and further differences in experimental setups.

For these reasons, the question of whether the developed motorway direction distinctive grid-based models can compete with ML models based on map-matched street segments could not be answered. The study setups of scientific works including data source, experimental setup, and label gathering were too diverse for making reliable comparisons on the study results. Using an additional FCD processing framework incorporating a map-matcher would have been too consuming in terms of the computational effort and time since the FCD is of huge volume (cf. Table 5.1).

The computational time for the motorway direction distinctive grid-based segment-assignment step was more than 30% lower than the spent computational time when using the map-matcher approach given the prescribed settings and utilizing 2.6 million data points. Moreover, less computational effort leads to carbon emission saving and contributes to environmental protection. Excluding the computational time, a disadvantage of the map-matcher approach is the needed maintenance of a map-matching service. Furthermore, a map-matcher can have various adjustable parameters which can be cumbersome to tune.

7. *Conclusion*

This work evaluated the potential of forecasting traffic congestion states for motorway segments based on grid cells using a huge floating car data set. The motorway direction distinctive grid-based approach for linking floating car data to motorway segments is characteristic for this work. It was proposed for the first time in this study to the best knowledge of the author and utilized throughout for segment aggregations. Besides that, major areas of interest were differentiating between developing models for one and numerous motorway segments, generating feature insights, developing forecasting models for different near-future time periods, and comparing the computational effort for the motorway direction distinctive grid-based and map-matcher motorway segment assignment step.

7.1 Summary

This section summarizes the findings from other scientific works, the applied methodology and the utilized data sets. It furthermore emphasizes on the analytical findings of this work.

Several congestion state prediction studies exist in the machine learning field. Details of the summarized literature review can be found in Chapter 3. Studies using floating car data usually rely on a map-matcher for mapping floating car data points to road segments. The basic grid-based approach was pursued by only a few studies which mapped floating car data points to grid cells. Congestion state forecasts, not for entire grid cells but for motorway segments based on grid cells, are proposed for the first time in this study to the best knowledge of the author. The developed method was named motorway direction distinctive grid-based approach.

Ground truth labels were gathered for the verification of congestion state forecasting results.

The two ground truth congestion labels *congestion* and *free-flowing* were extracted from traffic congestion messages that were generated by thresholding mean velocities of traffic detector records. The labels originally referred to spatial points in North Rhine-Westphalia and were assigned accordingly to motorway segments. Details of the ground truth data can be regarded in Section 5.2.1.
Floating car data transformation led to features of the data corpus that were used to forecast congestion states. Each observation in the data corpus contained feature values at a given time and motorway segment, consisting of a cell and a binary motorway direction variable. Motorway directions were determined by taking the mode of the floating car data in each grid cell as one motorway direction (cf. Section 4.2.2). The opposite of the mode was used as opposite motorway direction in a grid cell. 12 billion floating car data observations were aggregated based on the time and segment (cell and direction) for generating feature values. Aggregations of floating car data and further computations formed features such as the median velocity. Ground truth traffic congestion states were merged to the feature corpus based on their indices for generating the data corpus of 45 million observations.

The data corpus, as can be exemplarily seen in Table 5.4, was split into a train and a test data set. So-called whole grid models utilized train and test data sets of disjoint cell groups incorporating approximately 700 and 300 motorway segments respectively. Models developed for single motorway segments used cross-validation data subsets based on the monthly timestamp. Congestion state forecasts were implemented for the near-future time periods: 5, 10, 20, 30, and 60 minutes. The forecasts relied on the random forest machine learning model. Due to highly imbalanced congestion state classes, the target class *congestion* (0.3% share) was oversampled in the train data set. Models were evaluated based on the F_1-score, the misclassification rate, and the Bookmaker Informedness.

Features from neighbouring motorway segments could not improve the evaluation performance of models developed and evaluated based on an exemplary motorway segment with a high congestion share for five-minutes in the future. Results hence indicate that a single segment model with a high congestion share does not improve its performance by relying on features from neighbouring cells (cf. Section 6.2.1). Furthermore, no superior performance of single segment models in contrast to whole grid models could be seen in the experiments.

This is an indication for preferring whole grid models over separate models for each highly-congested motorway segment in NRW and a forecasting time interval of up to 60 minutes.

The random forest marked several features as important for the five-minute congestion state forecast in the whole grid setting as analysed in Section 6.2.2. Through induction for the statistical population of the entirety of motorway segments in North Rhine-Westphalia and a forecasting time period of up to 60 minutes, the current velocity of motorists seems to play an important role in determining the future congestion state. Features in the traffic count feature domain refer to durations of vehicles driving on a motorway segment. This feature domain seems to be of great importance for forecasting congestion states in the whole of North Rhine-Westphalia as well. The importance of the mentioned feature domains for forecasting congestion states could have been expected from a causal point of view.

Random forest models for several forecasting time periods were developed to represent the whole of North Rhine-Westphalia's grid by relying on approximately 1,000 sample motorway segments with ground truth values. Models for the 5, 10, 20, 30, and 60 minutes near-future time periods were evaluated as can be seen in Section 6.2.3. The whole grid model for each described forecasting time period was able to add value to congestion state forecasts. The model is hence proposed for forecasts of up to 60 minutes on North Rhine-Westphalia's motorway segments.

The model evaluation results in this work are based on the motorway direction distinctive grid-based segment-assignment step. A comparison with evaluation results of models relying on segment assignments through a map-matcher could not be made in a reliable way (cf. Sections 3.1.2, 3.1.3). The underlying study setups differed strongly. It was therefore not possible to compare evaluation measures in a reliable way.

The computational time for assigning segments to 2.6 million FCD points can be reduced by more than 30% with the prescribed infrastructure (cf. Section 6.3).

7.2 Discussion

The relevance of the traffic congestion topic and of the proposed approach is stated in this section. Practical implications of the forecasting approach based on direction-distinctive grid cells are presented additionally. Limitations based on inductive reasoning and ground truth data is furthermore outlined.

Forecasts of traffic congestion can be implemented in navigation systems or intelligent transportation systems to warn users of upcoming traffic congestion. Road users could save time when not being caught in traffic congestion and CO_2 emissions could be diminished additionally. The motorway direction distinctive grid-based approach as used in this work is less complex, has less dependencies, can adopt better to data sparsity issues and needs maintenance of less tools as opposed to the widely utilized segmentation through a map-matcher. It is emphasized that this work proposes a machine learning model for forecasting congestion states in the area of whole North Rhine-Westphalia as justified in Section 6.4.
A few features were found to be of strong importance for forecasting congestion states. Noise seems to be harmless for these features since they still had a large impact on the congestion state forecast. The noise arises when using the motorway direction distinctive grid-based approach by an absence of only regarding data points gathered on motorways as the map-matcher approach does. The motorways are the road network of interest.

Using grid cells for determining motorway segments is an alternative to relying on predefined road segments determined by a geographic information system map. Even if the forecasting performance of the motorway direction distinctive grid-based approach was slightly worse than the performance when using a map-matcher, several advantages could compensate for that. The advantages of the motorway direction distinctive grid-based approach are a slim design and an apparently easy adaption to data sparsity issues. The slim design incorporates one less tool in only depending on the PySpark infrastructure and not additionally depending on a geographic information system map and a complex map-matcher. This could lead to less maintenance work when putting the machine learning model into production. Furthermore, the segment assignment through direction distinctive motorway cells would seemingly need less computational effort.

A caveat is the utilization of only seven from the twelve months of a year in this work. The single segment model performances with monthly cross-validated test data indicate a higher variation in performance results for unseen months. Therefore, applying the whole grid model between March and July could lead to a bit greater deviations from the performance results of Figure 6-12.
The methodology proposed in this work can generally be adopted to areas of sparse floating car data. In this case, models would have to be reevaluated for the customized cell size. This aspect was beyond the scope of this work.
The motorway direction distinctive grid-based congestion state forecast methodology was developed in a scalable infrastructure. The near-future forecasts of congestion states can enrich a navigation system or an intelligent transportation system. Scalability and a slim application design can become important aspects for forecastings congestion states in comparable large regions such as North Rhine-Westphalia.

Limitations of this research are the consideration of a data sample instead of the statistical population for generating feature values and the possibility that the sample is not representative. Furthermore, the true target values could differ from the data used as ground truth for this work since they only consider a deterministic threshold of one traffic indicator for determining only two congestion states.

7.3 Future work

Important possible fields of future work are outlined below. Opportunities for improving the data corpus and optimizing the model performance in the future are described. Possibilities for comparing models of different studies are pointed out. The determination of reasonable application areas for congestion state prediction models is proposed as well.

Distinguishing between two directions of motorway-encompassing cells seems to work well in many use cases. Deviations from the normal case might lead to biases in the feature values. Expanding the data processing step for a more detailed distinction of motorway directions could lead to additional direction classes in some cases. Using gaussian mixture models additionally could be an approach to determine the number of motorways and hence

the number of directions in a cell. Pursuing the direction distinctive grid-based approach, a more detailed distinction of directions is proposed as it could further improve the forecasts. Other scientific works ([19], [29]) as well as causal connections indicate performance improvements from the usage of traffic information from the nearby traffic network. Even so, this work did not show considerable improvement. Further developing the usage of features derived from cells in the neighbourhood should therefore be studied in the future.

Optimizing hyperparameters of the utilized machine learning model was beyond the scope of this work but it could improve the model's performance. Comparing various machine learning model types, including anomaly detection methods, based on a data corpus generated with the help of grid cells could also lead to improved modelling results. These two possibilities of adjusting the forecasting methodology in order to gain predicting power should be studied in the future.

General obstacles for relying on benchmark data sets are the use of non-open-source data sets in studies as well as heterogeneous data sources such as stationary detectors, FCD, and image data. Generating benchmark data sets for the diverse kind of data sources is recommended as it could lead to more validity and justification of proposed algorithms. Ideally, benchmark data sets for different data formats would rely on the same traffic situations. In this case, even models using different input formats could be compared. Metrics could be established that were reported standardly in scientific works in this field. It is further an open question if a metric can account for the various dimensions of different study setups. That would be a great help in comparing experimental results of different scientific works. Regarding this work specifically, the forecasting power of the motorway direction distinctive grid-based approach could not be related to the map-matcher forecasting approach of other studies. Linking the approaches is a missing piece that should be pursued.

Transferring ML methodologies and evaluation results throughout the globe has not yet been studied systematically, which might be partly due to lacking benchmark data sets. Research in this field could lead to a better base for comparing studies with data sources from a diverse spatial spectrum. It could hence lead to a wider applicability of ML models developed for specific areas as well.

Bibliography

[1] D. Heuser, "Urban Mobility Innovations," 16.10.2020. [Online]. Available: https://www.umi.city/

[2] "Landesbetrieb Straßenbau Nordrhein-Westfalen | Straßen.NRW," 16.10.2020. [Online]. Available: https://www.strassen.nrw.de/de/

[3] J. A. S. Sá, A. C. Almeida, B. R. P. Rocha, M. A. S. Mota, and L. M. Dentel, "Lightning Forecast Using Data Mining Techniques On Hourly Evolution Of The Convective Available Potential Energy," 2016, pp. 1–5. [Online]. Available: https://www.researchgate.net/publication/303773171_Lightning_Forecast_Using_Data_Mining_Techniques_On_Hourly_Evolution_Of_The_Convective_Available_Potential_Energy

[4] W. van der Aalst, *Process mining: Data science in action*, 2nd ed. Berlin and Heidelberg and New York and Dordrecht and London: Springer, 2016. [Online]. Available: http://lib.myilibrary.com?id=915730

[5] W. P. Lee, M. A. Osman, A. Talib, and A. I. M. Ismail, "Dynamic Traffic Simulation for Traffic Congestion Problem Using an Enhanced Algorithm," *World Academy of Science, Engineering and Technology, International Journal of Mathematical, Computational, Physical, Electrical and Computer Engineering*, vol. 2, no. 2, pp. 625–632, 2008.

[6] "Staubilanz 2019: NRW mit den meisten Staus | ADAC," 10.08.2020. [Online]. Available: https://www.adac.de/der-adac/regionalclubs/nrw/staubilanz-nrw/

[7] G. Leduc, "Road Traffic Data: Collection Methods and Applications: JRC-IPTS Working Papers," *European Commission Joint Research Centre Institute for Prospective Technological Studies: Seville, Spain*, 2008.

[8] F. Mocholí Belenguer, A. Mocholí Salcedo, A. Guill Ibañez, and V. Milián Sánchez, "Advantages offered by the double magnetic loops versus the conventional single ones," *PloS one*, vol. 14, no. 2, p. e0211626, 2019.

[9] "Chapter 2, Traffic Detector Handbook: Third Edition—Volume I - FHWA-HRT-06-108," 13.12.2020. [Online]. Available: https://www.fhwa.dot.gov/publications/research/operations/its/06108/02.cfm

[10] Y. Liu, X. Yan, Y. Wang, Z. Yang, and J. Wu, "Grid Mapping for Spatial Pattern Analyses of Recurrent Urban Traffic Congestion Based on Taxi GPS Sensing Data," *Sustainability*, vol. 9, no. 4, p. 533, 2017.

[11] A. Zeidan, E. Lagerspetz, K. Zhao, P. Nurmi, S. Tarkoma, and H. T. Vo, "GeoMatch: Efficient Large-Scale Map Matching on Apache Spark," in *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 10.12.2018 - 13.12.2018, pp. 384–391.

[12] J. Fürnkranz, "Decision Tree," in *Encyclopedia of Machine Learning*. Springer, Boston, MA, 2011, pp. 263–267. [Online]. Available: https://link.springer.com/referenceworkentry/10.1007%2F978-0-387-30164-8_204

[13] "Classification and regression - Spark 2.3.2 Documentation," 10.12.2018. [Online]. Available: https://spark.apache.org/docs/2.3.2/ml-classification-regression.html#decision-trees

[14] "Decision Trees - RDD-based API - Spark 2.3.2 Documentation," 10.12.2018. [Online]. Available: https://spark.apache.org/docs/2.3.2/mllib-decision-tree.html

[15] "Ensembles - RDD-based API - Spark 2.3.2 Documentation," 10.12.2018. [Online]. Available: https://spark.apache.org/docs/2.3.2/mllib-ensembles.html#random-forests

[16] M. Kubat, *An Introduction to Machine Learning*. Cham: Springer International Publishing, 2015.

[17] B. S. Kerner, *Introduction to Modern Traffic Flow Theory and Control*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009.

[18] X. Zhang, E. Onieva, A. Perallos, E. Osaba, and V. C. Lee, "Hierarchical fuzzy rule-based system optimized with genetic algorithms for short term traffic congestion prediction," *Transportation Research Part C: Emerging Technologies*, vol. 43, pp. 127–142, 2014.

[19] S. Guo, Y. Lin, S. Li, Z. Chen, and H. Wan, "Deep Spatial–Temporal 3D Convolutional Neural Networks for Traffic Data Forecasting," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 10, pp. 3913–3926, 2019.

[20] J. F. Zaki, A. Ali-Eldin, S. E. Hussein, S. F. Saraya, and F. F. Areed, "Traffic congestion prediction based on Hidden Markov Models and contrast measure," *Ain Shams Engineering Journal*, 2019.

[21] T. Hastie, R. Tibshirani, and J. H. Friedman, *The elements of statistical learning: Data mining, inference, and prediction*, 2nd ed., ser. Springer series in statistics. New York: Springer, 2009.

[22] C. Kahraman and Ö. Kabak, *Fuzzy Statistical Decision-Making*. Cham: Springer International Publishing, 2016, vol. 343.

[23] M. R. Jabbarpour, H. Zarrabi, R. H. Khokhar, S. Shamshirband, and K.-K. R. Choo, "Applications of computational intelligence in vehicle traffic congestion problem: a survey," *Soft Computing*, vol. 22, no. 7, pp. 2299–2320, 2018.

[24] Z. Chen, Y. Jiang, and D. Sun, "Discrimination and Prediction of Traffic Congestion States of Urban Road Network Based on Spatio-Temporal Correlation," *IEEE Access*, vol. 8, pp. 3330–3342, 2020.

[25] X. Ma, H. Yu, Y. Wang, and Y. Wang, "Large-scale transportation network congestion evolution prediction using deep learning theory," *PloS one*, vol. 10, no. 3, p. e0119044, 2015.

[26] M. Fouladgar, M. Parchami, R. Elmasri, and A. Ghaderi, "Scalable deep traffic flow neural networks for urban traffic congestion prediction," in *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 14.05.2017 - 19.05.2017, pp. 2251–2258.

[27] F.-R. Huang, C.-X. Wang, and C.-M. Chao, "Traffic Congestion Level Prediction Based on Recurrent Neural Networks," in *2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*. IEEE, 19.02.2020 - 21.02.2020, pp. 248–252.

[28] W. Pattara-atikom, P. Pongpaibool, and S. Thajchayapong, "Estimating Road Traffic Congestion using Vehicle Velocity," in *2006 6th International Conference on ITS Telecommunications*. IEEE, 21.06.2006 - 23.06.2006, pp. 1001–1004.

[29] B. Zhao, C. Xu, and S. Liu, "A data-driven congestion diffusion model for characterizing traffic in metrocity scales," in *2017 IEEE International Conference on Big Data (Big Data)*, 2017, pp. 1243–1252.

[30] M. Chen, X. Yu, and Y. Liu, "PCNN: Deep Convolutional Networks for Short-Term Traffic Congestion Prediction," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 11, pp. 3550–3559, 2018.

[31] L. Liu, J. Zhen, G. Li, G. Zhan, Z. He, B. Du, and L. Lin, "Dynamic Spatial-Temporal Representation Learning for Traffic Flow Prediction," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–15, 2020.

[32] T. Cheng, J. Haworth, and J. Wang, "Spatio-temporal autocorrelation of road network data," *Journal of Geographical Systems*, vol. 14, no. 4, pp. 389–413, 2012.

[33] T. Thianniwet, S. Phosaard, and W. Pattara-atikom, "Classification of Road Traffic Congestion Levels from GPS Data using a Decision Tree Algorithm and Sliding Windows," *Lecture Notes in Engineering and Computer Science*, vol. 2176, 2009.

[34] Y. Liu and H. Wu, "Prediction of Road Traffic Congestion Based on Random Forest," in *2017 10th International Symposium on Computational Intelligence and Design (ISCID)*. IEEE, 09.12.2017 - 10.12.2017, pp. 361–364.

[35] P. Pongpaibool, P. Tangamchit, and K. Noodwong, "Evaluation of road traffic congestion using fuzzy techniques," in *TENCON 2007 - 2007 IEEE Region 10 Conference*, 2007, pp. 1–4.

[36] D. Impedovo, F. Balducci, V. Dentamaro, and G. Pirlo, "Vehicular Traffic Congestion Classification by Visual Features and Deep Learning Approaches: A Comparison," *Sensors (Basel, Switzerland)*, vol. 19, no. 23, 2019.

[37] M. D. Laboratory, "Traffic dataset," 2016. [Online]. Available: https://www.dropbox.com/sh/uo634k3ybvmu1dc/AAAOxRpk-2Q_187fZ9tZmRABa?dl=0

[38] A. Loder, L. Ambühl, M. Menendez, and K. W. Axhausen, "Understanding traffic capacity of urban networks," *Scientific reports*, vol. 9, no. 1, p. 16283, 2019.

[39] F.-H. Tseng, J.-H. Hsueh, C.-W. Tseng, Y.-T. Yang, H.-C. Chao, and L.-D. Chou, "Congestion Prediction With Big Data for Real-Time Highway Traffic," *IEEE Access*, vol. 6, pp. 57 311–57 323, 2018.

[40] European Road Safety Observatory, "Motorways 2018," 2018.

[41] GitHub, "bmwcarit/barefoot," 01.01.2021. [Online]. Available: https://github.com/bmwcarit/barefoot

[42] A. Luque, A. Carrasco, A. Martín, and A. de las Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," *Pattern Recognition*, vol. 91, pp. 216–231, 2019.

[43] GitHub, "jupyter-widgets/ipyleaflet," 08.01.2021. [Online]. Available: https://github.com/jupyter-widgets/ipyleaflet

[44] "Open.NRW," 05.09.2020. [Online]. Available: https://open.nrw/dataset/verkehrsinformationen-der-viz-nrw-f-r-nordrhein-westfalen-1568186868

A. *Performance evaluation results*

Table A.1 presents evaluation results of the resampled test data set as well as the test data set with original class proportions in the whole grid setting.

Time period	**F_1-score (std. err.)**		**Misclass. rate (std. err.)**		**BM (std. err.)**	
	Resampled	Original	Resampled	Original	Resampled	Original
Baseline	0.33	0.006	0.20	0.003	0	0
Current time	0.57 (0.07)	0.081 (0.019)	0.13 (0.01)	0.033 (0.003)	0.43 (0.08)	0.43 (0.08)
5 Min.	0.56 (0.08)	0.078 (0.018)	0.14 (0.02)	0.033 (0.004)	0.41 (0.09)	0.41 (0.09)
10 Min.	0.54 (0.07)	0.076 (0.018)	0.14 (0.02)	0.033 (0.003)	0.39 (0.08)	0.39 (0.08)
20 Min.	0.50 (0.08)	0.072 (0.018)	0.15 (0.02)	0.032 (0.004)	0.35 (0.09)	0.35 (0.09)
30 Min.	0.47 (0.08)	0.067 (0.020)	0.15 (0.01)	0.031 (0.003)	0.32 (0.08)	0.32 (0.08)
60 Min.	0.35 (0.08)	0.055 (0.016)	0.17 (0.01)	0.026 (0.003)	0.21 (0.06)	0.21 (0.06)

Table A.1: Performance evaluation of the random forest whole grid model