TALLINN UNIVERSITY OF TECHNOLOGY
School of Information Technologies

Marika Eik 203959IAAM

Design of Data-Driven Parking Service: Case Study of Ülemiste City

Master Thesis

Supervisor
Juri Belikov
Ph.D
Co-Supervisor
Alari Krist
M.Sc

Tallinn 2023

TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond

Marika Eik 203959IAAM

Andmepõhise parkimisteenuse disain: juhtumianalüüs Ülemiste City näitel

Magistritöö

Juhendaja
Juri Belikov
Ph.D
Kaasjuhendaja
Alari Krist
M.Sc

Tallinn 2023

Author's declaration of originality

I hereby certify that I am the sole author of this thesis. All the used materials, references to the literature and the work of others have been referred to. This thesis has not been presented for examination anywhere else.

Author: Marika Eik .....
(signature)

Date: May 18th, 2023

Abstract

The aim of the Master Thesis is to develop a data-driven service architecture of "Dynamic and Optimised parking prices at Ülemiste City parking lots" that will be consistent with the general Ülemiste City strategy. The present work creates an architectural vision of the information system enabling traffic data analyisis, calculation and optimisation of dynamic parking prices and digital display of the prices on the screen before entering the parking lot. The modelling of the enterprise architecture was guided based on the value stream and capability analysis.

To achieve the target of the work, the Author performs the analysis of enterprise strategy and architecture by mapping the strategic goals, problems, priority programs and core capabilities. The outcomes of the Master Thesis include the actualisation of the Ülemiste City strategy together with the strategic model and business architecture. The priority programs, such as: Mobility, Data Strategy and Architecture, as well as Data Analytics were highlighted. The goals and capability analysis of mobility program were developed. The shortcomings and the problems of the existing business process at parking lots, business requirements, functional and non-functional requirements of the information system under development were investigated and considered. The detailed system analysis of the parking service with dynamic and optimised prices was implemented. The architectural vision of the information system providing the parking service with dynamic and optimised prices was developed. Based on the outcomes described, the problem of the thesis was solved and the objective was achieved.

The thesis was written in English and contains text on 93 pages, 5 chapters, 35 figures, 11 tables.

Annotatsioon

Magistritöö eesmärk on arendada andmepõhine teenuse arhitektuur: „Dünaamilised ja optimeeritud parkimistariifid Ülemiste City avatud parklates“, mis oleks joondatud Ülemiste City üldstrateegiaga.

Käesoleva magistritööga luuakse infosüsteemi arhitektuurne visioon, mis võimaldab liiklusandmete analüüsi, dünaamiliste parkimishindade arvutamist ja optimeerimist ning hindade digitaalset kuvamist ekraanil enne parklasse sisenemist. Ettevõtte arhitektuuri modelleerimisel lähtuti väärtusvoo ja võimekuste analüüsist.

Töö eesmärgi saavutamiseks teostab Autor ettevõtte strateegia ja arhitektuuri analüüsi, kaardistades strateegilised eesmärgid, probleemid, prioriteetsemad programmid ja põhivõimekused. Magistritöö tulemuste hulka kuulub Ülemiste City strateegia aktualiseerimine koos strateegilise mudeli ja äriarhitektuuriga. Esile tõsteti prioriteetsemaid programme, nagu mobiilsus, andmestrateegia ja arhitektuur ja andmeanalüüs. Töötati välja mobiilsusprogrammi eesmärgid ja võimekuste analüüs. Uuriti ja käsitleti olemasoleva äriprotsessi puudujääke ja probleeme Ülemiste City avatud parklates, ärinõudeid, arendatava infosüsteemi funktsionaalseid ja mittefunktsionaalseid nõudeid. Rakendati dünaamiliste ja optimeeritud hindadega parkimisteenuse detailne süsteemianalüüs. Töötati välja dünaamiliste ja optimeeritud hindadega parkimisteenust pakkuva infosüsteemi arhitektuurne visioon. Kirjeldatud tulemuste põhjal lahendati lõputöö probleem ja saavutati eesmärk.

Lõputöö on kirjutatud inglise keeles ja sisaldab teksti 93 leheküljel, 5 peatükki, 35 joonist, 11 tabelit.

List of abbreviations and terms

AS-IS	Current state, existing solution
AI	Artificial Intelligence
API	Application Programming Interface
ANN	Artificial Neural Network
BIZBOK®	A Guide to the Business Architecture Body of Knowledge®, Business Architecture Guild®
BM	Basic (naive) model
BPMN	Business Process Modelling & Notation, a graphical business process modeling language
BSC	Balanced Scorecard, tasakaalus tulemuskaart, method of strategic planning and management
DT	Decision Tree
DZ	Drizzle
ESG	Environmental, Social and Governance
EMHI	Estonian Meteorological and Hydrological Institute
FURPS+	Functionality, Usability, Reliability, Performance, Supportability, software requirements classification model, functional requirements, usability, reliability, supportability and design, development and interface requirements
GHG	Greenhouse Gas
GB	Gradient Boosting
GD	Gradient Descent
HR	Linear Regression with Huber loss
HTTP	Hypertext Transfer Protocol
ICT	Information and communication technology
IoT	Internet Of Things
KPI	Key Performance Indicator, performance measure
LSTM	Long Short-term Memory Neural Network

LR	Linear Regression
MoSCoW	Must have, Should have, Could have, Won't have, a method for prioritizing user requirements in agile software development
MVP	Minimum Viable Product, a minimum viable product with just enough features to put it into service as intended
ML	Machine Learning
MSG	Motorway Control System
MSE	Mean Squared Error
MLW	Machine Learning Workbench
MAE	Mean Absolute Error
NP	No Phenomena
RNN	Recurrent Neural Network
REST	Representational State Transfer
RA	Rain
RH	Relative Humidity
SWOT	Strengths, Weaknesses, Opportunities, Threats, analysis of internal strengths and weaknesses and external opportunities and threats, current situation analysis method in strategic planning
SSL	Suur-Sõjamäe-Lõõtsa
SGD	Stochastic Gradient Descent
SN	Snow
SHAP	SHapley Additive exPlanations
TO-BE	Future situation, desired solution
TOGAF®	The Open Group Architecture Framework, enterprise architecture framework
UML	Unified Modeling Language, a graphical modeling language used in an object-oriented approach
UTC	Coordinated Universal Time
XGB	Extreme Gradient Boosting

Table of Contents

List of Figures

List of Tables

Introduction

Over the past decades issues related to sustainable environment have become more acute due to rapid growth of urbanisation forms. Smart City concepts are targeted to support green environment through low carbon emissions technology. In the present work, the company under consideration is Mainor AS, who is a principal developer of Ülemiste City: the largest business campus in the Baltics. For the Ülemiste City (ÜC) following the Smart City best practices, the development of sustainable environment including the mobility is among top objectives.

One of the aims of the present thesis is the actualisation of strategic goals, the development of strategic view and business architecture of the Ülemiste City. The specification of strategic programs, identification of the objectives of mobility program are also under investigation. The detailed system analysis of one of the mobility programs: Development of dynamic and optimised parking prices, is under the main focus of the work. The target group of the parking service developed in this work is one-time/random visitors of ÜC. The current parking prices at open parking lots of ÜC make parking very attractive and accessible for random visitors. Thereby, ÜC does not receive the potential profit from the provided parking service in parking lots. On top of that, the current parking arrangement supports the increase of the number of vehicles moving within the city during the daytime hours, when people of the campus community prefer, for example, outdoor activities (walk meetings), resting in parks or designated outdoor areas.

The information system developed promotes the data analytics capability. It does not change the established management of parking at ÜC parking lots, which is beneficial as it does not require large investments. The parking pricing optimised by traffic data analysis will not affect the customers having monthly contract of parking, thus allowing to have control and management of regular campus customers. The expected results of promoting of traffic data analytics capability by the parking prices optimisation are: an increase of parking service

profit, increase of satisfaction and reputation of campus community members, as well as the decrease of the level of motorisation and CO2 emissions from the transportation.

Thereby, the study investigates the following questions:

- Which development areas of Ülemiste City the program (service) considered affects?
- What kind of problems related to Ülemiste City mobility the program developed solves?
- What are the shortcomings of the present solution of the equivalent business service?
- What indicators can measure the effectiveness of the program developed?
- What business requirements are needed to achieve the measurable outcomes?
- What new capabilities are needed to develop a desired data-driven service?
- How to design a business architecture and business process of a data-driven service, considering the strategic goals of Ülemiste City?

The Master Thesis consists of five chapters:

1. The first Chapter 1 gives an overview of Ülemiste City as organisation, addresses the main topic of the thesis. This Chapter defines the research problems, the goals, the relevance of the thesis topic, as well as the scope and expected results. It also includes the contribution of the Author.
2. The second Chapter 2 describes the main concepts of Smart City, gives an overview of urban forms indicators structurising the main development areas of Smart Cities. It also presents the main data sources in Smart Cities, addresses the role of artificial intelligence and machine learning in solving of problems related to urbanisation, traffic flow estimation.
3. The third Chapter 3 reflects the actualisation of Ülemiste City strategy utilising `TOGAF` enterprise and/or `BIZBOK` business architecture frameworks. This Chapter represents the Balanced Scorecard (BSC) Project demonstrating the strategic development areas of Ülemiste City with sub-domains and priority Programs supporting the achievement of measurable outcomes (KPI). The capabilities necessary for the development of Ülemiste City mobility program are also highlighted. `ArchiMate` modelling language is utilised to visualise the models created.
4. The fourth Chapter 4 covers the detailed system analysis of parking service at Ülemiste City parking lots with Dynamic and Optimised parking prices. In this Chapter the shortcomings of AS-IS parking management at Ülemiste City parking lots are evaluated. It also represents the strategic model of the service developed, makes value stream and

capability analysis, as well as demonstrates the architectural vision of the information system enabling traffic data analysis, calculation and optimisation of dynamic parking prices and digital display of these prices on the screen.

5. The fifth Chapter 5 reflects the traffic data management and analysis, as well as the building of machine learning models forecasting the amount of cars entering the Ülemiste City through the Suur-Sõjamäe-Lõõtsa road-intersection during the daytime hours. This Chapter also addresses the study of the effect of weather properties on the models accuracy. Linear Regression (LR), LR with Huber Loss (HR) and XGBoost (XGB) models were tested on uni-variate time series, which were first transformed to supervised learning problem by sliding window method. The Basic (naive) Model (BM) considered the present day vehicles and forecasted the same amount for the next day.

The Author would like to cordially thank her supervisors Juri Belikov and Alari Krist for their support, fruitful discussions and professionalism. The Author highly appreciates the knowledge and skills obtained during the discussions with the Thesis Advisor Margarita Matson. The Author would like to express her gratitude to Alexander Jung from Aalto Technical University who enabled her a real opportunity to dive into the world of data science and machine learning and to achieve positive results. The Author is thankful to Paul Leis for the discussions, useful comments and objectivity in the evaluation of her thesis. The Author would like to express her gratitude to Kadi Pärnits, who is one of the leaders and successors of Ülo Pärnits ideas, for the interest to Author's work, supportive attitude and useful discussions. The support from R8 Technologies during the Thesis preparation is also highly appreciated. The Author cannot forget the help and good attitude of all administrative staff and her group members at TalTech.

Last but not least, the Author would like to express her appreciation to her family, friends and Mainor AS team, who encouraged and followed the progress of the thesis.

1. Thesis objective

The first chapter gives an overview of the area of the Master Thesis and a company under consideration. This chapter defines the research problems, the goals, the relevance of the Thesis topic, as well as the scope and expected results. It also includes the contribution of the Author.

1.1 Ülemiste City project

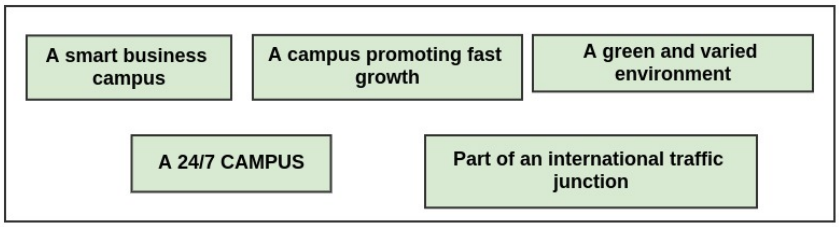
"Ülemiste City (ÜC) is the city of Future, where talents can grow, utilise their potential and succeed". This is the message and priority of the Ülemiste City. The creator of Ülemiste City, Ülo Pärnits, has imagined the development of this area as blooming and growing business campus with hundreds of companies and thousands of employees [2]. 'Education is capital!' was among most favourite sayings of Ülo Pärnits. Education was among higher priority values of Ülo Pärnits. He assumed that it is a thing that both the individual and the society should constantly invest in, throughout their lives.

The ideas and intentions of Ülo Pärnits were supported and implemented by him, his family and close friends [2]. Thereby in 2005, Ülo Pärnits started with an exciting project called Ülemiste City– an intellectual environment for active young Estonian talents who would not have to leave their country, but could impact the world from here; to be a smart business city and an engine of the Estonian economy. Mainor AS is the developer of Ülemiste City campus.

1.2 Ülemiste City: Organisational goals

Ülemiste City has established five strategic goals: A smart business campus, A 24/7 CAMPUS, A campus promoting fast growth, A green and varied environment, Part of an international traffic junction, Figure 1a. The strategic goals are related to development areas of ÜC: Economy, Knowledge, Community, Services and Environment. Each of these areas includes three sub-domains forming a "Radar", which is represented in Figure 1b. The present work covers Green City and B2C services within the Environment and Services development areas,

Figure 1b. A successful economy, an attractive campus, happy talents and data strategy are the principles that support ÜC to achieve its goals and concrete outcomes.



(a) Five strategic goals (author created).

(b) "Radar" [3].

Figure 1. Five strategic goals established by Ülemiste City. "Radar" describing five development areas in Ülemiste City with 15 dimensions. Green City, B2C services are shaded as the fields covered by this study.

1.2.1 Ülemiste City Mobility Program. Thematic scope of the thesis

Ülemiste City mobility program is defined based on the requirements and principles of Environmental and Services development areas. The mobility program addresses the objectives

represented in Table 1. The listed range of mobility objectives can be evaluated, supported and solved based on data. Data-driven transportation problem solving, involves the analysis of transportation related data using methodologies, such as big data analytics, machine learning.

In the Table Table 1 the objectives highlighted in the present work are marked by "YES" and determine the thematic scope of the present study. The financial and profitability analysis of advanced parking management implemented by dynamic and optimised parking prices are not within the scope of the thesis.

Table 1. Objectives of mobility program of ÜC. The objectives highlighted in the present work are marked by "YES" and determine the thematic scope of the present study

Objective	Covered by this study
Advanced parking management (Parking Lots/Houses)	YES
Reduce the number of cars/vans within the ÜC area	YES
Increase the number of non-motorised vehicles	NO
Improve accessibility to ÜC by public transportation	NO
Increase vehicle sharing services	NO

The present study supports the development of Ülemiste City from the mobility perspective, i.e. it analyses, forecasts and gives knowledge for the design of traffic flow within the ÜC area. In more detail, the present work is focused to develop a data-driven service architecture models along with the capabilities necessary to promote dynamic parking pricing, Figure 2.

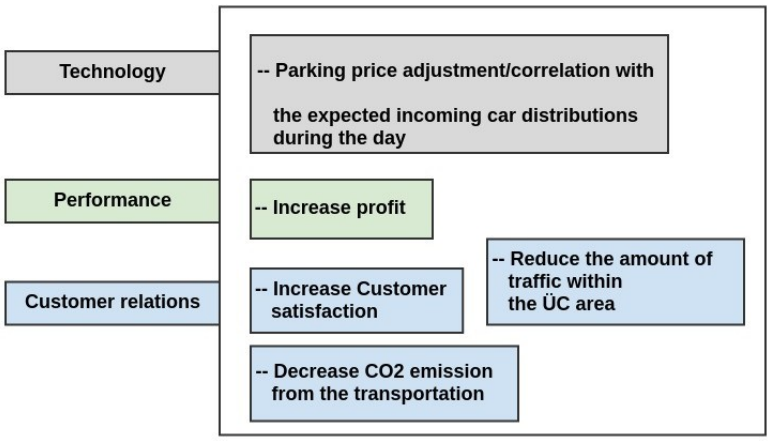


Figure 2. Goals for motivation model of parking service at parking lot with dynamic and optimised pricing (author created).

1.2.2 Motorisation in ÜC area

The problems of motorisation and parking are important not only for large cities, but also for industrial areas, such as, for example, Ülemiste City. At the moment, Ülemiste City has grown into the largest and fastest growing smart business centre in the Baltic States, prioritising its talents and education. The estimated amount of employees coming every day to ÜC is between 12000–13000. Most of them use personal cars to come to work. The increase of traffic amount, associated with urbanisation, directly leads to an increase in CO2 emissions from the transportation. The management of large traffic flows requires an improved approach that can support and facilitate the development of more effective public transportation options and sharing services, as well as the planning of real estate sector. Nowadays, machine learning (ML) methods are widely used to support the solution of the problems in various engineering and technological fields. Recent studies show that ML models can be successfully used to support and improve the development of urbanisation forms. For example, to estimate the level of CO2 emissions from the vehicles, it is necessary to know and forecast with a certain degree of probability the number of vehicles passing a certain area and the average distance travelled by them.

The available parking services of ÜC include parking houses, as well as the open parking areas on the territory of ÜC. ÜC is striving to become green city and offer sustainable environment to its customers, meaning the reduction of the number of cars and vans actively moving within the City during the working hours. One of the solutions for the increased motorisation level within the ÜC area can be the development of dynamic and optimised pricing for open parking lots, which would correlate with the forecasted distribution of cars and vans entering the ÜC territory during the working hours. The high motorisation level at ÜC at present, may be related to current parking pricing at the parking lots, which is very attractive for one-time visitors and may significantly increase the amount of traffic during the day.

1.2.3 Thesis topic relevance and novelty

For the Ülemiste City supporting and striving to follow the Smart City best practices, the development of sustainable environment is among top objectives. Advanced parking management can be partly addressed by dynamic and optimised pricing for parking lots, supported by the results of traffic data analysis. This approach assumes the development of an information system, enabling the optimisation of parking prices at ÜC parking lots.

The ability of tracking and forecasting the traffic volumes and employing this data in the CO2 emission calculation, gives the Ülemiste City a realistic estimate, which can be used in cooperation with the city of Tallinn in a ways, such as:

- The development of real estate sector;
- Planning of the public transportation;
- Address customers awareness and education.

The relevance of management and quantification of traffic flow within the ÜC is also related to Taxonomy compass [4] and environmental, social and governance (ESG) reporting [5]. The traffic management and forecast allow to monitor and control the level of CO2 emissions from the transportation, which may encourage the use of light traffic means, such as bicycle and e-scooter, better management and organisation of public transportation and overall "greener" environment in the Ülemiste City.

1.3 Research Questions and Problem statement for a thesis

The present study contributes to the solution of the following problems related to ÜC Mobility program:

- Unreceived/Lost profit from the parking service at parking lots at ÜC;
- Increased motorisation level during the working hours;
- High CO2 emissions from the transportation;
- Unsafe area for pedestrians due to increased motorisation during the working hours.

The research questions for the thesis are defined as follows:

- What are the Strategic View and Business Architecture of the Ülemiste City?
- What kind of system analysis is necessary to implement the parking service with dynamic and optimised prices?
- What type of data analysis methods are useful for traffic flow forecasts?
- How traffic data management and analytics can support the implementation of parking service with dynamic and optimised prices?

The present work concentrates on the development of information system enabling the parking

at parking lots based on the dynamic and optimised prices. The optimised pricing is planned to be correlated with the expected incoming car distributions during the working hours. Thereby, the study promotes the capability related to traffic data analytics, meaning the selection of appropriate methods suitable for traffic forecast. The effect of weather properties on the vehicles forecasting accuracy is also under investigation. Weather properties can affect behavioural pattern of people. For example, the likelihood that on a rainy day a person will choose a car instead of bicycle is quite high. The counts of cars and vans determined by applying the external object detection ML model are going to be employed as time series data-sets. The inclusion of weather parameters can improve or degrade the accuracy of the models developed, demonstrating thus the degree of importance.

1.4 Thesis objective and expected results

The novelty of this study lies in the development of dynamic and optimised prices for parking lots at ÜC area according to the analysis of the flow of incoming cars. It is assumed that as a result of the employment of dynamic and optimised parking prices the revenue of the company will increase. The proposed dynamic and optimised parking prices will allow to control the number of one-time visitors arriving to Ülemiste City between the certain daytime hours. The one-time visitors are considered as those, which are not the members of the ÜC community. A demand on controlling the number of one-time visitors arriving to ÜC is related to ÜC mobility program, i.e. promotion of light traffic, such as bicycles and e-scooters.

The expected outcomes of the thesis are as follows:

- ÜC general strategy actualisation employing `TOGAF` enterprise and/or `BIZBOK` business architecture frameworks. The development of Balanced Scorecard (BSC) Project representing the strategic development areas of ÜC with sub-domains and priority Programs supporting the achievement of measurable outcomes (KPI). Design of ÜC Business Architecture, highlight of some priority programs, such as: Mobility, Data Strategy and Architecture, Data Analytics, as well as the capabilities necessary for the implementation of the programs mentioned.
- System analysis to implement the parking service at parking lots with dynamic and optimised prices including the strategic model, value stream and capability analysis, business information model, development of use cases, as well as the architectural vision of the information system.

- Methods development for traffic data management and analysis. Building the machine learning models enabling the forecast of incoming cars and vans during the working hours.

1.5 Author contribution

The present study enabled the Author to apply her knowledge of an IT analyst and business architect in practice. On top of that, the Author expanded her knowledge and experience in the field of data processing and analysis. During the thesis preparation, the Author acted in the following roles:

- Analyst;
- Business Architect;
- Data Scientist.

The main contributions of this study from the viewpoint of:

1. Design of information system enabling the data-driven dynamic and optimised parking pricing.
    - The distribution of the number of cars forecasted is employed as input for the development of optimised parking prices at parking lots of ÜC;
    - A highlight of relevant daytime hours based on feature importance evaluation, implemented as a part of data analytics, is used to support the dynamic price development.
      For incoming cars: until 7.30, at 9.30, 12.30, 13.30 and at 16.00;
      For incoming vans: 7.30–8.30, at 12.30 and after 16.00;
      For outgoing cars: 7.00, 8.00, 10.00, 13.30 and after 16.00;
      For outgoing vans: 6.00–7.00, after 18.00, 19.00.
    - Traffic flow amounts forecasted will be employed to evaluate the CO2 emission from the transportation in the ÜC area necessary for GHG reporting.
2. Traffic flow data analysis.
    - The accuracy of ML models developed is sufficient to forecast the counts of incoming/outgoing cars and vans between 6.00–19.00 with a step of 30 minutes on Tuesday, Wednesday and Thursdays;
    - ML models developed improved the accuracy of Basic (naive) Model (BM) by 26-

29% in average. XGBoost improved the accuracy of BM for incoming/outgoing cars the most;

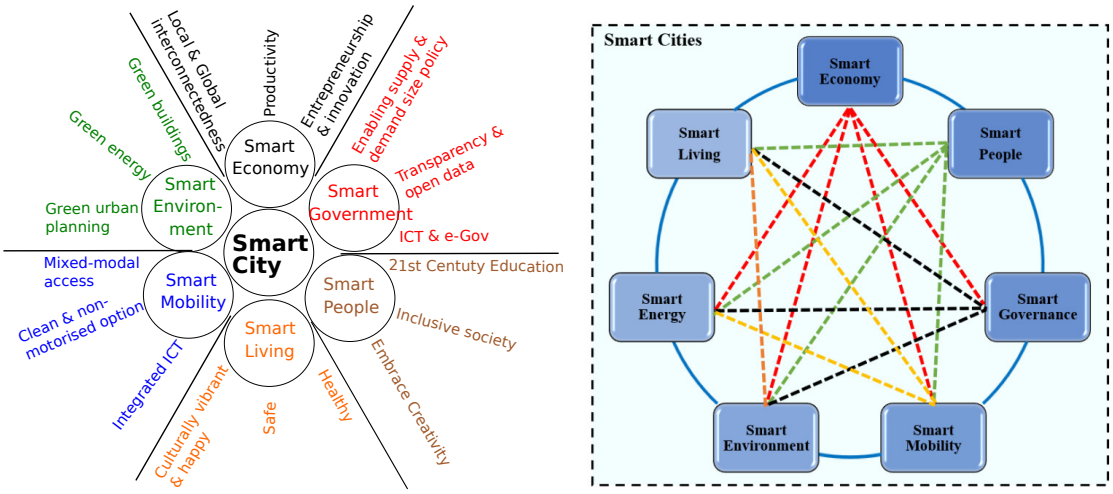
- Including weather features: improvement within 1-6%. Higher relevance: air temperature and weather phenomena;
- XGBoost: the effect of weather phenomena (snow or rain) for incoming cars. LR and HR: the effect of wind speed, air temperature, weather phenomena (no phen. or rain) for outgoing vans.

2. Smart City: state-of-the-art

This chapter describes the main concepts of Smart City, gives an overview of urban forms indicators structurising the main development areas of Smart Cities. It also presents the main data sources in Smart Cities, addresses the role of artificial intelligence and machine learning in solving of problems related to urbanisation, traffic flow estimation.

2.1 The concept of a Smart City

One of the main concepts of Smart City is to effectively manage the resources, optimise energy consumption, assure the green environment by adopting low carbon emission technologies. Among the goals of Smart Cities is to address upcoming challenges of conventional cities by offering integrated management systems with a combination of intelligent infrastructures [6]. Information and communication technology (ICT) plays an important role in Smart City by supporting the decision, implementation and ultimate productive services. The key components of Smart City include the areas given in Figure 3a and their interrelation is shown in Figure 3b.



(a) Key components of Smart City (author created). (b) Interrelation of the key components [6].

Figure 3. Smart City key components and their interrelation.

The study presented in [1] split the urban forms elements inherent to Smart Cities into the following sections: layout, landscape, infrastructure, density, housing/building type and land use. The parameters of these urban forms are given in Figure 4.

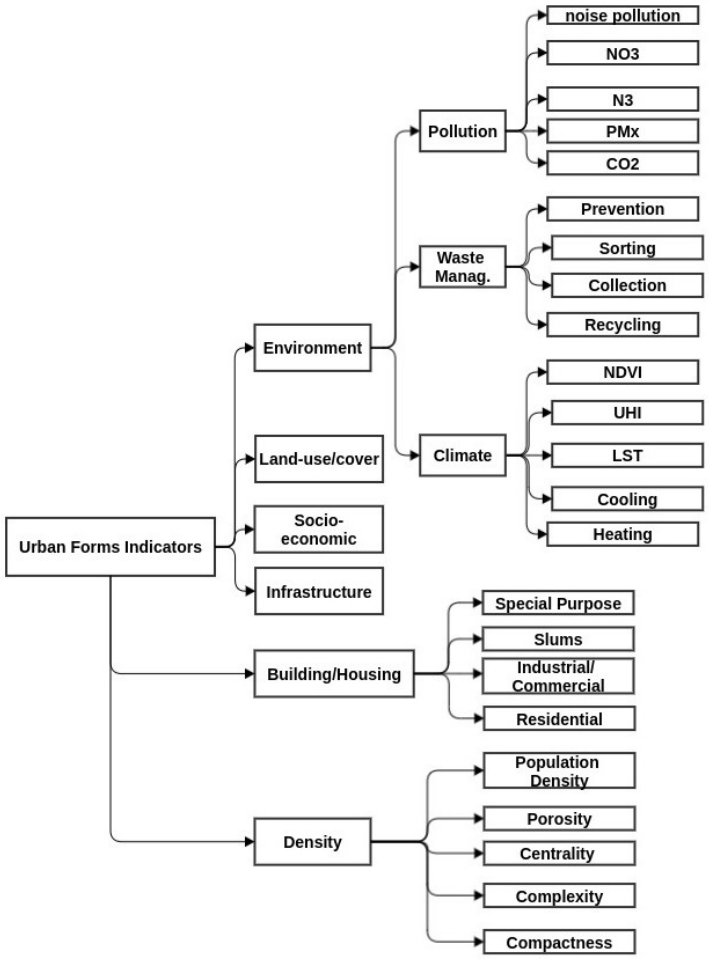
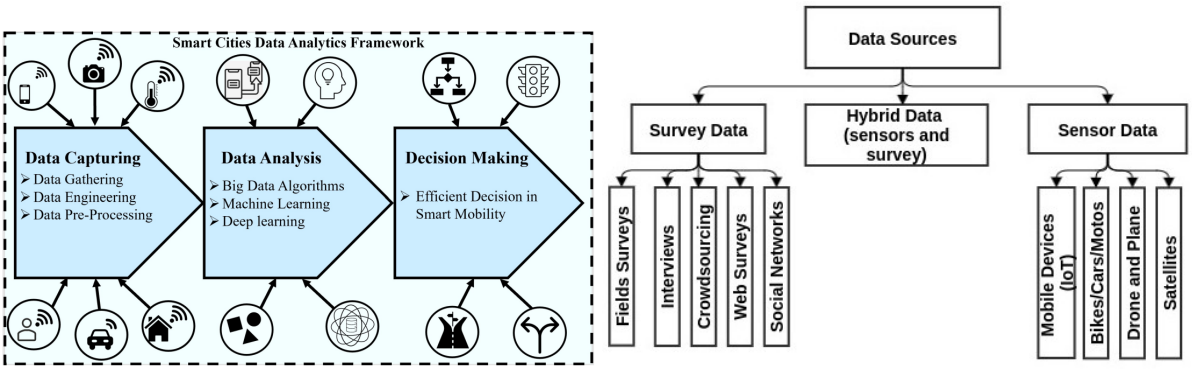


Figure 4. Urban forms indicators [1].

Internet of Things (IoT) devices and applications are considered as the most important parts of Smart Cities, as they generate big data required to analyse and track the activities. Artificial Intelligence (AI) and machine learning (ML) methods have a vital role in the evolution of Smart Cities. ML is an area residing at the intersection of computer science and statistics, and its application in various fields, such as, for example, finance, civil engineering, healthcare, etc., is growing day by day. Recently, Smart Cities have been more and more included in these applications to address the needs of sustainable environment, optimal energy consumption. The modelling of indicators given in Figure 4 considering the input parameters, is an area for the application of ML models.

2.2 Data sources in Smart City

The authors of the study presented in [6] established a universal Smart City data analytics framework, which supports the decision making, Figure 5a. Urban data sources introduced in study [1] include: sensor data, hybrid data and survey data, Figure 5b. Converting the data from different sources into single unite measurements enables to apply ML models producing generalised/holistic outcomes.



(a) Universal Smart City data analytics framework [6]. (b) Data sources of Smart City [6].

Figure 5. Data management in Smart City.

The findings made in [6] report that the insufficient amount and heterogeneous training data may lead to over-fitting problem resulting in unreliable results. The accuracy of ML methods can be improved by increasing the training data. The authors found that the lack of the data can be minimised by meta-learning, enabling to build a model with a few samples only (zero shot, one shot, and a few shot learning). Pre-processing of data, such as, filling of missing information/values, select important and/or relevant features, also plays a vital role for the modelling.

Road monitors/sensors are typical collectors of traffic data, such as, vehicles speed, number of passing vehicles. The volume of the data is typically large, which usually requires pre-processing to make the further modelling and analysis effective.

2.2.1 Machine Learning methods for urban data

The study reported in [1] gave an overview of ML models employed for urban form applications. Based on this study the ML approaches can be aggregated, as it is shown in Figure 6.

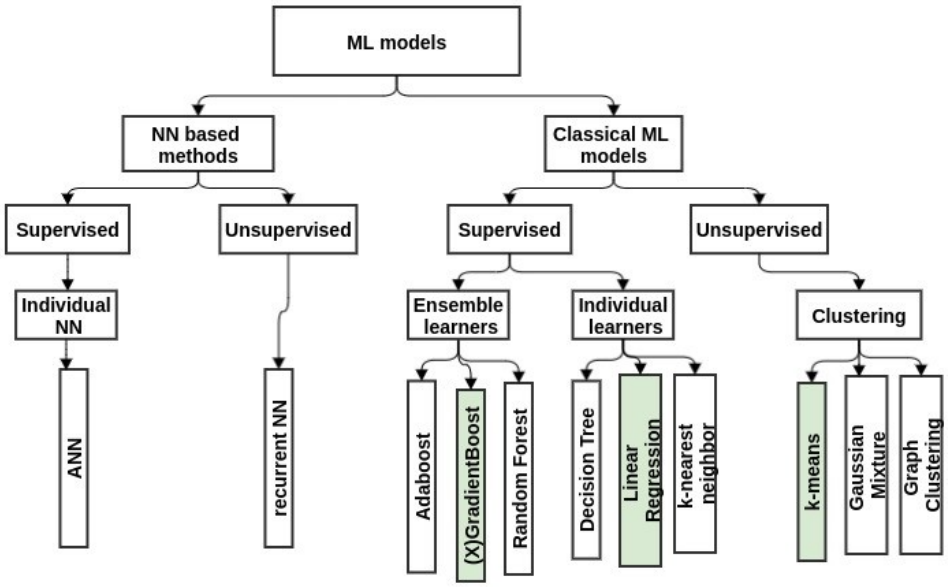


Figure 6. ML approaches for urban form applications [1].

In the study, the supervised and unsupervised ML models were compared based on input data: labelled/unlabelled, computing complexity, real-time: offline/real-time analysis, sub-domains: classification and regression, clustering and association, precision: accurate/moderate, number of classes: known/unknown. The authors gave clear and easy-to-use overview of ML methods considering the algorithmic principles, advantages/disadvantages and the opportunities/potentials to be applied in appropriate urban planning applications. The authors reported that among the classical ML methods, Figure 6, the Random Forest, Support Vector Machines (SVM), Linear Regression (LR) and Gradient Boosting were useful and mostly used in the studies of smart urban forms. On top of that, it was found that the efficiency of ML algorithm can be the low computational cost, as well as the ability to adapt and calibrate new data-sets. The explainability of results of ML models was also a relevant advantage.

The classical ML models applied in previous available studies on the smart urban forms depending on the target problem(s) are listed below. The data sources are given in brackets and the learning type, such as, classification [C], regression [R] and clustering [Cl] follows the

method and is given in square brackets.

Environment

- Pollution PM2.5 (sensors data): Gradient Boosting [C], SVM [C];
- Air quality (hybrid data): Linear Regression [R];
- Land use (sensors data): Random Forest [C], k-NN [C];
- Solid waste management (survey data): GB [R];
- Energy consumption (hybrid + sensors data): SVM [R], Clustering: k-means, GMM, DBSCAN [Cl];
- Environment and Environment quality (hybrid + sensors data): Decision Tree [C], SVM [C].

Infrastructure

- Traffic flow (survey data): Random Forest [R], Gradient Boosting [R];
- Travel behaviour (hybrid data): Association Rules.

Among the NN based methods applied for urban form modelling, Figure 6, were as follows:

Environment

- Solid waste management (survey data): ANN [R];
- Urban topology (hybrid data): ANN [C];
- Heating and Cooling (hybrid data): Recurrent NN [R].

Infrastructure

- Traffic flow (hybrid data): Recurrent NN [R].

Clustering. The clustering belongs to statistical processing and unsupervised learning. Clustering methods partition a data-set, consisting of m data points into a number of groups or "clusters", representing a subset of data points, which are more similar to each other than to data points in another cluster [7].

There are two fundamental requirements for data-set to be clustered: homogeneity and completeness. Homogeneity requires that all clustered data points should be of the same nature, i.e.

described by a similar set of characteristics. If cluster analysis is preceded by factor analysis (e.g. only quantitative values, no missing values, the min/max number of factors depending on the number of variables), then the sample does not need to be corrected/cleaned. Otherwise, the data-set should be checked for homogeneity and completeness.

The main goals of clustering analysis can be defined as following:

- Understanding the data by identifying cluster structure. Dividing the sample into groups of similar data points makes it possible to simplify further data processing and decision making, e.g. by applying its own analysis method to each cluster (e.g regression or classification);
- Data compression. If the initial sample is excessively large, then it can be reduced, leaving one of the most typical representatives from each cluster;
- Novelty detection. Outliers are selected that cannot be attached to any of the clusters.

Decision Tree (DT) is a simple non-parametric, easy-to-use transparent ML method. It is most favourable weak learner algorithm for ensemble combination. A straightforward algorithm of generating a DT is: * First select a feature to split on and place it at the root node; * Then repeat this process for all child nodes. The split is done on a feature which is most important. DT decides for the split that promises maximum information 'Gain'. Alternative to information 'Gain' is 'Gini Impurity'. Among the disadvantages are: requires large storage and the understanding of DT based methods is easy if few DTs are involved. Besides, DTs are not robust to data perturbations. Although, this can be solved by the aggregation of many DTs resulting in strong learning algorithms.

Support Vector Machines (SVM) is based on nonlinear transformations of features into a higher-dimensional feature space, where the classification problem becomes linear separable. SVM models are basically binary classifiers. With aggregation techniques, these can be made applicable to multi-class problems. Suitable for linear and nonlinear separable data. Has high precision and efficiency in large spaces. Less prone to over-fitting and stable, noise robustness and the problem of unbalanced classes. However, it is not suitable for large data-sets due to long training time.

Artificial Neural Network (ANN) is based on and schematically inspired by the functioning of the biological neurons. Among the advantages are high performance and computing power, efficiency for high dimensional problems, ability to work with complex characteristics, parallel processing capability and fault tolerance. The disadvantages are related to theoretical

complexity, requirement of careful adjustment and large amount of effective data.

2.3 Urban traffic flow estimation

The existing methods for analysing the flow of transportation networks are basically following two approaches, i.e. analytical and simulation-based [8]. One of the most widely used analytical approaches is the classical four-step model for urban transportation planning has been advocated for over 60 years. A new pathway is to use data-driven computational learning theory– a sub-field of Artificial Intelligence, for transport network flow analysis, for example, as it was addressed in the study by [9]. It is useful to keep in mind that only one data source cannot possess high positioning accuracy and high coverage simultaneously, which are both needed for traffic flow estimation. The volume of traffic can be adequately estimated based on data that includes both high positioning accuracy and coverage [8]. The typical traffic sensors include Closed-Circuit Television cameras, GPS devices, Vehicle Detectors. The information on traffic flow and density supports the analysis and extraction of traffic patterns to make the traffic conditions more effective.

Traffic prediction methods typically use only historical traffic data, although, the weather, temporal and spatial parameters also affect traffic conditions. The weather properties may affect the volume of traffic flow. The limitation of parking places on a sunny day may encourage the people to ride light traffic, such as, motorcycles, bicycles, while the rainy or snowy conditions may enforce to ride cars. Therefore considering the weather conditions may improve the traffic prediction accuracy.

The study presented in [10] focused on the building of ML model estimating the traffic flow based on the data from INRIX, which is a kind of mobile data collected from probe vehicles, where the speed is calculated as the average speed over a length of the road, and is called space mean speed. The work also used data from fixed measurements of the Motorway Control System (MCS), which controls the traffic flow. The motivation for the study was to create an intelligent model that automatically finds relationships between observed features, such as: speed, travel time, hour, weekday, location, and traffic flow. The relevance and usefulness of the features to improve the accuracy of the model were also evaluated.

3. Strategy and Enterprise Architecture

A target of the enterprise architecture is to support a company to achieve the concrete business outcomes through the strategy actualisation. The TOGAF [11] enterprise and/or BIZBOK [12] business architecture frameworks can be applied to map the strategy of a company, develop value stream(s) with the capabilities needed, and ArchiMate [13] modelling language can be utilised to visualise the models created.

3.1 Ülemiste City Strategy Analysis

Ülemiste City vision is defined, such as: Ülemiste City is the best well-known and desired business, living and studying environment for talents–smart experts, managers and companies who impact the (world) economy through their action. The mission includes: Developing an international, attractive, knowledge-based working, development and living environment that would increase the competitive ability of people and companies, bring talents home to implement their potential, and inspire the birth of new business models.

The established strategic goals of ÜC, related to development areas acting as drivers, for Balanced Scorecard (BSC) Project are given in Table 2.

Table 2. The established strategic goals of ÜC, related to development areas acting as as drivers, for Balanced Scorecard (BSC) Project.

Criteria for Defining Goals (Development areas)	**Define ÜC Goals**
Economy, Services	A smart business campus
Economy, Services	A 24/7 CAMPUS
Knowledge, Community, Services	A campus promoting fast growth
Environment, Community, Services	A green and varied environment
Environment, Community	Part of an international traffic junction

Detailed description of development areas/criteria (acting as drivers) for defining the strategic goals of ÜC:

Economy: increasing efficiency of economy and business operations;

Services: widen ÜC services allowing Talent to succeed;

Community: stronger community of campus talents;

Knowledge: widen education and research opportunities;

Environment: develop best location for business, work, life and study.

Based on the development areas and defined strategic goals represented in Table 2, the Balanced Scorecard (BSC) Project, summarising the objectives within each development areas, the respective measurements and some priority programs developed in the present study are given in Tables 3, 4.

Table 3. PART I: Balanced Scorecard (BSC) Project representing the strategic development areas of ÜC with sub-domains and priority programs supporting the achievement of measurable outcomes (KPI).

Devel. Area	Dimension, measurable (KPI)	Year2022	Year2025	Dimension Objective, measurable (KPI)	Year2022	Year2025	Program
Economy	**Reputation, satisfaction, [%]**		93	ÜC Employees satisfaction rate, [%]	84	90	P1, P2, P3
			100	ÜC Customer satisfaction rate, [%]	84	90	P1, P2, P3
	ÜC economic statistics, [%]		73	Employees in campus, [per]	12500	15000	
			73	Students in campus, [per]	1800	2000	
			73	Inhabitants in campus, [per]	900	2000	
			91	Corporate taxes: share of labor from the business sector, [%]	3.2	3.6	
			91	Corporate taxes: share of VAT from the business sector, [%]	2.2	2.4	
	ÜC development, [%]		88	Creation of new jobs, [%]	4.6	5.0	
			93	Efficiency of labour use: growth of additional value, [%]	12.6	13.0	
			93	Efficiency of labour use: Sales revenue per employee, [EUR]	151515	170000	
			86	Turnover and export: share of sales revenue in Estonia, [%]	2.7	3.5	P1, P2, P3
			86	Turnover and export: Export share of sales revenue, [%]	54.0	60.0	
			86	Turnover and export: Share of export in Estonia, [%]	4.14	4.5	
Knowledge	**R&D and innovation, [%]**		90	Technology hub:share of technology-driven production/services, [%]	47	52.2	
			85	R&D expenditures: financial volume (from turnover), [%]	11	13	
			77	Cooperation with universities and R&D: proportion of cooperating companies, [%]	50	65	
	Learning community, [%]		97	Employees with research degree: proportion of employees on campus, [%]	29	30	
			92	Number of students on campus, [per]	1830	2000	
	Start-up community, [%]						

Table 4. PART II (continuation): Balanced Scorecard (BSC) Project representing the strategic development areas of ÜC with sub-domains and priority programs supporting the achievement of measurable outcomes (KPI).

Devel. Area	Dimension, measurable (KPI)	Year2022	Year2025	Dimension Objective, measurable (KPI)	Year2022	Year2025	Program
Community	**International community, [%]**						
	Health promotion, [%]						
	Networks and events, [%]						
Services	**B2C services , [%]**		78	Customer satisfaction rate, [%]			P1, P2, P3
	B2B services, [%]		64				P1, P2, P3
	Accessibility, [%]		62	Employee/Customer satisfaction rate, [%]			
Environment	**Green City, [%]**		65	Eco-friendly buildings: LEED office buildings, [%]	87.4	100	
			65	Eco buildings: Elect. bike racks, [pcs]	157	200	
			65	Eco buildings: Elect. car chargers, [pcs]	15	50	
			96	Use of green areas: Green areas in ÜC, [m2]	44226	45000	
			96	Use of green areas: Green areas per person in ÜC, [m2/per]	3.7	4	
			91	Clean and healthy environment: Decrease of motorisation level, [%], CO2, [t]			P1, P2, P3
			91	Clean and healthy environment: Advanced parking mgt., [%]	81	85	P1, P2, P3
			54	Recycling: Waste sorting level on campus, [%]	11	60	
			54	Recycling: Waste per worker, [kg/per]	64	57	
	Work Environment, [%]		88	Employees satisfaction with env., [%]			
	Living Environment, [%]		74	Customer satisfaction with env., [%]			

Table 5. Programs (course of actions) linked to Ülemiste City strategic goals and outcomes (KPI's), Figure 7.

ID	Program description
P1	**Data Strategy and Architecture:** Governance, Master Data Mgt, Data warehousing, Data Quality, Data Architecture, Data Asset Planning and Inventory, Data Integration, Metadata Mgt.
P1.1	Increase ÜC digitalisation
P2	**Data Analytics**
P2.1	Data-driven services development, including B2C and B2B
P3	**Mobility**
P3.1	Parcel Delivery Service by robot Carriers (B2C, B2B)
P3.2	Dynamic pricing at Parking Lots (B2C)
P3.3	Increase the number of electric chargers in ÜC
P3.4	Develop light traffic opportunities, Encourage light traffic (bicycle, e-scooter)
P3.5	Encourage vehicles shearing services
P4	**Land-use**
P4.1	Increase/widen (better highlight) the means supporting the use of Green Areas
P5	**Waste management**
P5.1	Increase the level of ÜC community awareness on waste sorting practices
P5.2	Decrease the buildings CO2 footprint
P6	**Housing**
P6.1	Decrease the buildings CO2 footprint
P6.2	Develop different price level offices and services
P7	**Socio-Ecomonic**
P7.1	Widen the support of business networking within the ÜC area
P7.2	Attract/Develop 24/7 Campus by supporting the residential sector, facilities (as an example)
P7.3	Support and Encourage start-up infrastructure
P7.4	Increase Lifecycle services for companies(from incubator/testbed to startup hubs)

Table 6. Core capabilities necessary for the implementation of programs P3.2, P2.1, P1, Table 5.

CAPABILITIES	ASSESSMENT and/or SWOT based Weaknesses, Threats	PROGRAMS
– Traffic Data management, analysis – Forecast of incoming cars during the working hours – Financial and Business Analytics – IT development and administration	– Lost profit from Parking Service – High CO2 emissions from the transportation – Unsafe area for pedestrians due to high motorisation – Increased car traffic during the day	P3.2, P2.1, P1

3.1.1 KPI-s of Dynamic pricing at Parking Lots

The success of the processes and service under development will be estimated by the system of KPI-s presented in Table 7.

Table 7. The selected KPI-s to evaluate the success of the processes under development.

Process	Objective	KPI (measure)	Value (metric)
Application of dynamic and optimised pricing in Parking Lots	Increase profit from Parking Service during one year	% of increase	5-10%
	Decrease the number of cars during the working hours during one year	% of decrease	8-10%
	Decrease level of CO2 emission from the transportation during one year	tons	300-320 t

3.2 Ülemiste City Business Architecture

Ülemiste City generalised motivation model with the programs and capabilities forming the strategy model is represented in Figure 7. The capabilities necessary for realisation of the programs related to mobility objectives are represented in Figure 8. The capabilities that are developed in this study are marked with deep-red colour.

3.3 Ülemiste City central administration platform

Ülemiste City as a Smart City employs central administration platform, which is secure, scalable and versatile, and supports the IoT integration. Different cloud based platforms have entered the market within the recent years. The reasons for their wide use are connected to their ability to monitor and manage devices anywhere, regardless of the current location. On top of that, a cloud based ready platforms are able to address the information security issues. Cumulocity IoT platform was selected by the Ülemiste City as it supported easy IoT integrations with building management systems (BMS), enabled different architectures for data storage, as well as supported micro-service architecture giving flexibility in the selection of desired and necessary services. Cumulocity is a cloud based powerful platform, which was born in Silicon Valley, California 2010 and is constantly improving, flexible, and can save a

huge amount of software developers time. Cumulocity provides development framework in AngularJS, allows to build plugins and integrate them to the platform and/or to merge some other information system(s) using special API to Cumulocity platform. Various applications can be integrated and individual views for each user can be created utilising one platform. To access Cumulocity platform, there is a built-in REST API supporting the TCP/IP protocol [14]. It is possible to use both encrypted (HTTPS) and plain-text (HTTP) methods. On top of advantages described, the integration of devices is also made as easy as possible, i.e. by custom utilities for some popular development boards.

An example micro-service provided by Cumulocity platform is a Machine Learning Workbench (MLW), which can be accessed via MLW micro-service API. MLW is meant to facilitate the work of data scientists and machine learning practitioners by streamlining model training and evaluation activities. MLW application provides an integrated Jupyter Notebook environment enabling to perform data upload, pre-processing, modelling and visualisation. The data can be pulled to MLW via connectors to the data sources, such as Cumulocity IoT, or DataHub, where the data is collected from the devices connected to the Cumulocity IoT platform.

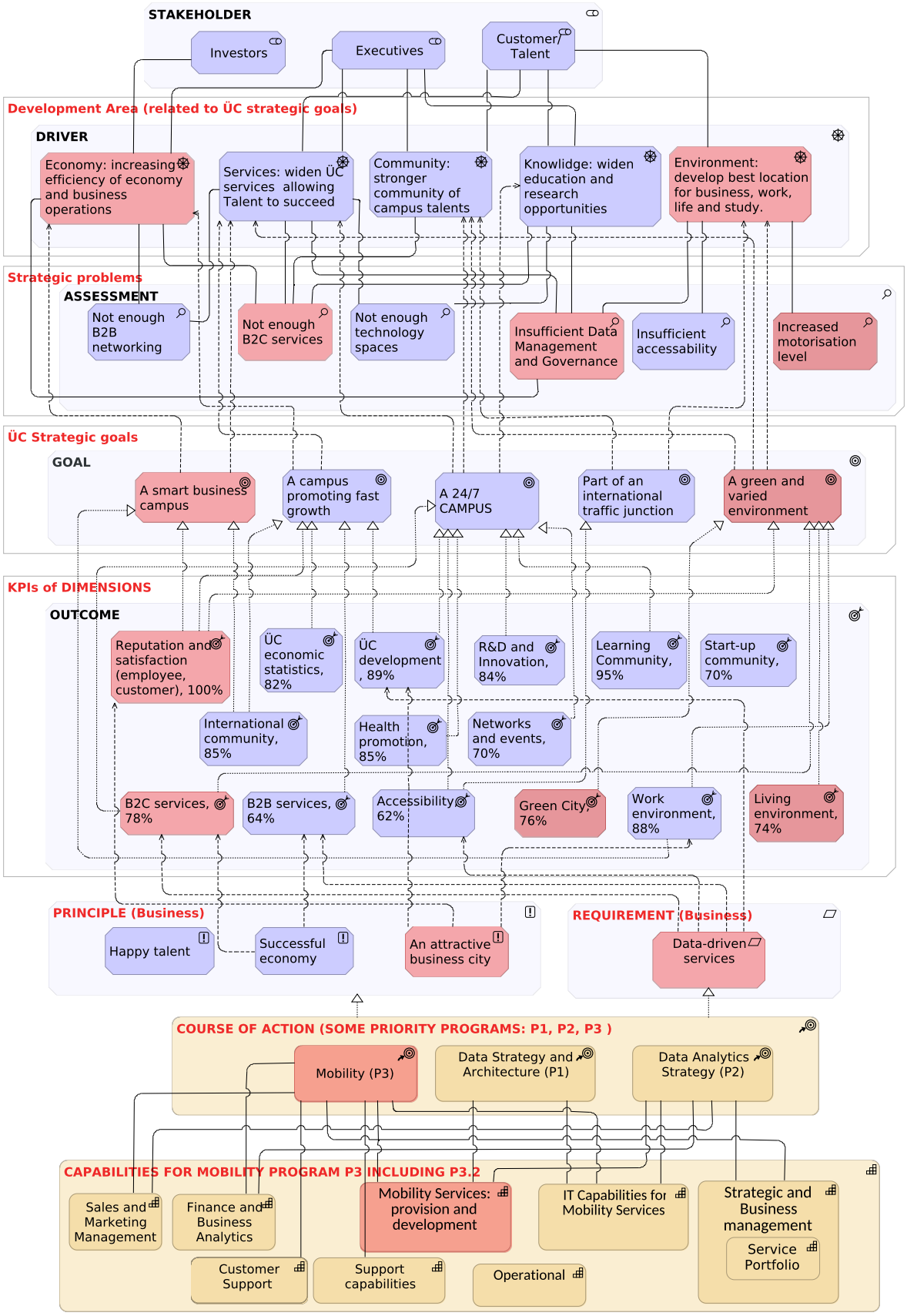


Figure 7. ÜC motivation and strategy models with some priority programs (author created). Entries marked by red colour are those affected by the service developed in this study.

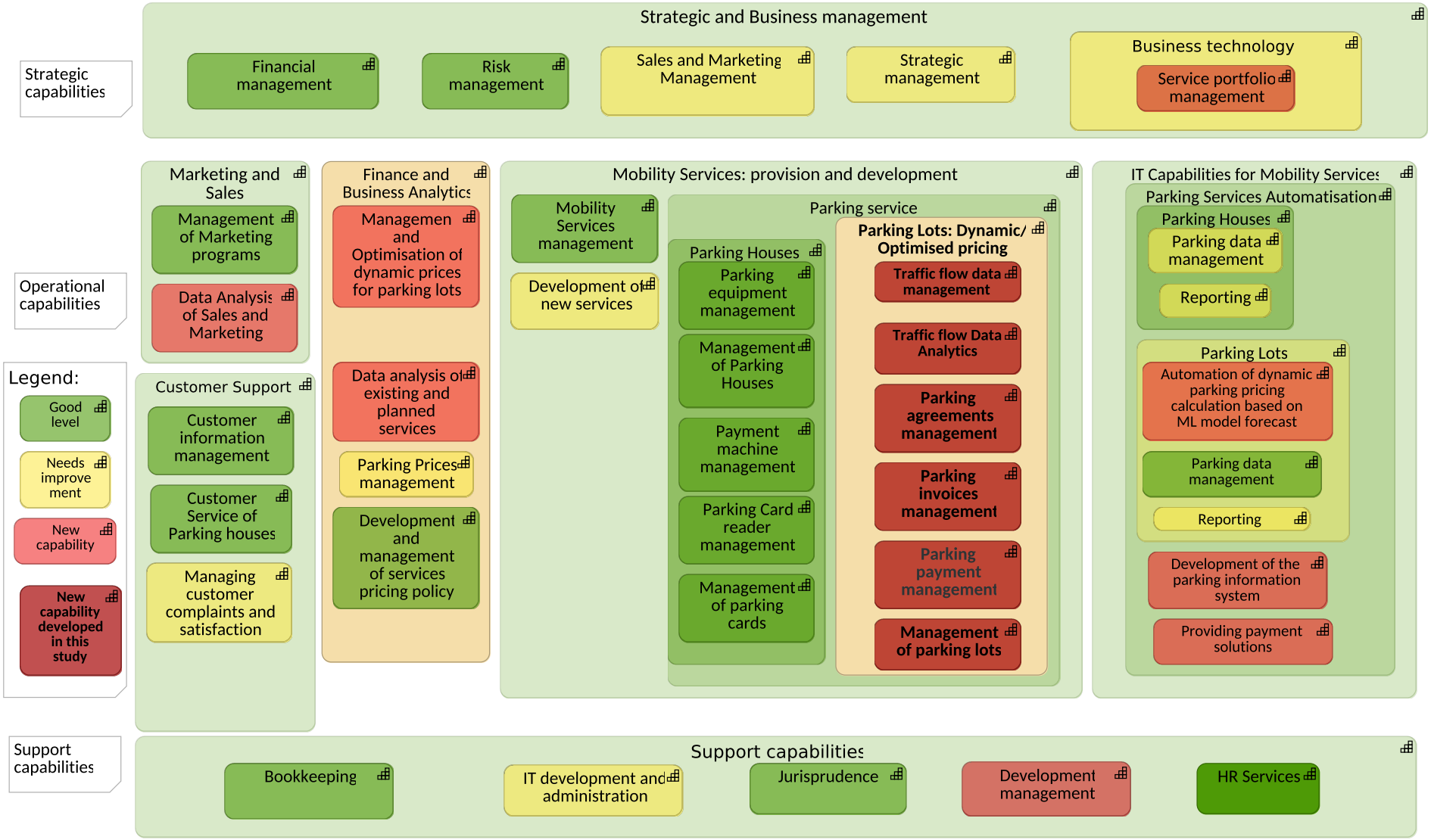


Figure 8. Strategic capabilities for Mobility program. The capabilities that are marked with deep-red colour correspond to the ones necessary for dynamic and optimised parking prices (author created).

4. System Analysis: Dynamic and Optimised parking prices

The target group of the information system under development is one-time/random visitors of ÜC. The system planned promotes the data analytics capability. It does not change the established management of parking in parking lots, which is beneficial as it does not require large investments. The parking pricing optimised by traffic data analysis will not affect the customers having monthly contract of parking, thus allowing to have control and management of regular campus customers. The expected results of promoting of traffic data analytics capability by the optimisation of parking prices for random visitors of ÜC are: an increase of profit from parking service by 5-10%, a decrease in the level of motorisation by 8-10% and CO2 emissions from the transportation by 300-320 tonnes during one year. In addition, it is expected that the satisfaction and reputation of campus community members will increase as well.

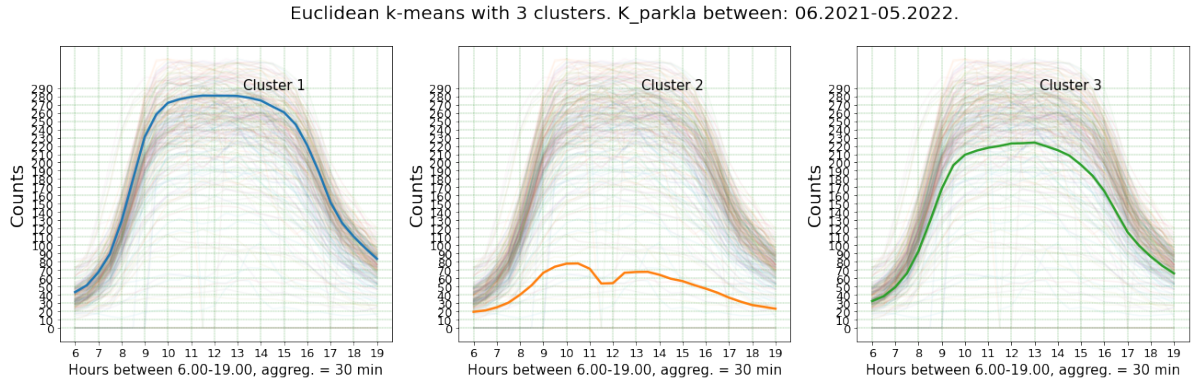
4.1 Parking management at ÜC parking lots: AS-IS process

In the present study the road-intersection at Suur-Sõjamäe-Lõõtsa was considered. Most probably the cars that are entering ÜC from this intersection are those who are willing to park their cars either at "K" parking lot or at "H", "L" parking houses. The data analysis of the parking lot "K", which is not a part of this work, showed that on most days it is completely parked from 9.30 to 15.30, i.e. the total number of parking spaces is 297 and more than 280 of them are occupied during the specified daytime period, Figure 9.

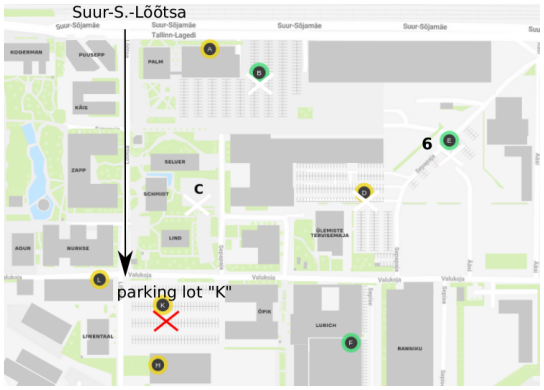
The configuration of the process flow given in Figure 10, demonstrates the current–AS-IS parking arrangement at ÜC parking lots.

4.1.1 Shortcomings of AS-IS Process

At the moment, parking in the ÜC parking lots is organised in a way that 2 hours of parking are free of charge and the following hours have a rate of 1 eur per every started 30 min. 24 hours of parking cost 6 eur. The established parking prices make parking very attractive and accessible for random visitors of ÜC. In other words, during working hours, parking for them becomes free or has a negligible cost. Thereby, ÜC does not receive the potential profit from the provided parking services in parking lots. On top of that, the existent parking arrangement increases the number of vehicles moving around the city during the daytime hours, when people of the campus community prefer, for example, outdoor activities (walk meetings), resting in parks or designated outdoor areas. The increased number of vehicles makes the outdoor activities for the campus community members unsafe. ÜC strives to develop an environment that allows both rapid economic growth and development of companies and start-ups, as well as sustainable and green environment within the ÜC area. The traffic within the ÜC area, increased due to random visitors during the daytime hours, affects the level of CO2 emissions from the transportation.



(a) "K" parking lot: Clustering (author created).



(b) "K" parking lot: Location [15].

Figure 9. Clustering results of "K" parking lot within the period from 06.2021-04.2022. The location of the "K" parking lot relative to Suur-Sõjamäe-Lõõtsa road intersection.

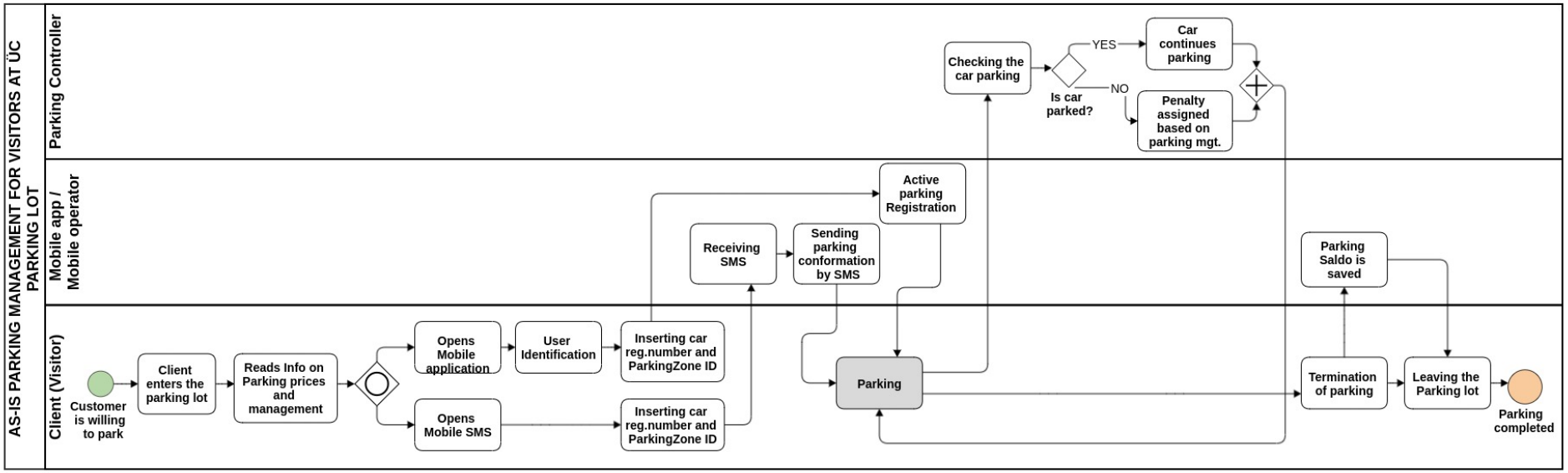


Figure 10. AS-IS process flow of parking management for visitors (non-agreement customers) at ÜC parking lot (author created).

4.2 Business requirements

Business requirements are the basis for the functional and non-functional requirements, as well as they may support the selection of the proper and useful software solutions. The Business requirements that are necessary to achieve the business goals determined for mobility program are given in Table 8. The detailed description of business requirements is as follows:

Traffic data request and processing (BR1). To optimise parking lot prices, it is necessary to have access to ÜC traffic data at intersections. The traffic data of ÜC are collected from Fyma database to Operational Data Store at Cumulocity platform. Fyma database includes the time series data of the counts of objects, such as, cars, vans, trucks, etc. Four main cameras in ÜC are capturing the vehicles that are coming in and out of the City;

Ability to perform traffic flow analysis and forecasting (BR2). Data analytics is necessary to estimate the incoming/outgoing traffic flow distributions during the daytime hours on every ÜC intersection. This information will support the optimisation of parking prices by specifying the traffic flow rush hours;

Ability to optimise the prices at open parking lots (BR3). The development and optimisation of prices for parking is planned to be carried out using financial and business analytics, which will be based on a forecast of incoming vehicles at a certain intersection of ÜC;

Digital display of dynamic parking prices (BR4). The dynamic and optimised parking prices will be displayed on a Screen before entering the parking lot.

Table 8. Business Goals and Requirements.

Notation	Business Goal (Mobility program)	Business Requirement (motivation model)
BR1	Decreased motorisation level/ CO2 emissions from the transportation	Traffic data request and processing
BR2	Decrease motorisation level/ CO2 emissions from the transportation	Ability to perform traffic flow analysis and forecasting
BR3	Increase profit from Parking service	Ability to optimise the prices at open parking lots
BR4	Increased Customer satisfaction	Digital display of dynamic parking prices

4.2.1 Functional and Non-Functional Requirements

All the requirements were prioritised using MoSCoW method, which is an easy-to-understand prioritisation technique for managing requirements. This method supports a common understanding of the system functionalities for the developers, analysts and the stakeholders (business side). Non-functional requirements were classified following the FURPS+ model utilised in software development. The abbreviations of MoSCoW and FURPS+ methods include, respectively:

- M - must have, must be, important and critical for MVP to function;
- S - should have, could be;
- C - could have, it would be good, expands user possibilities;
- W - won't have, MVP is out of scope.

- F - Functionality, functional requirements;
- U - Usability;
- R - Reliability;
- P - Performance;
- S - Supportability;
- "+" - other requirements, for example, for design, development, interfaces or physical infrastructure.

The functional requirements are described by Use Cases, represented in Section 4.6 Figure 15. The functionalities defined by use cases are in line with the minimum viable product (MVP) concept, which originates from the Lean Startup methodology in software development. The benefit of MVP approach is that it allows to validate a product/service without having to invest time and money in building the complete version.

The non-functional requirements were set considering the criteria typically used in the initial selection. The non-functional requirements chosen were, as follows:

- Usability: the system developed should be easily used and operated. For the user the operation should be smooth by providing better usability;
- Security: the system developed should be resilient for any type of malicious attacks;
- Data integrity: the system developed should enable integrity, consistency, and correct-

ness of the data in the application;

- Robustness: the system developed should be robust against invalid or erroneous inputs.

4.3 Parking management at ÜC parking lots: integration of dynamic and optimised pricing

In the present study, the number of incoming cars forecasted during the daytime hours between 6.00–19.00 is planned to be employed in use cases related to dynamic price development for parking.

The data-driven motivation model with capabilities necessary for the development of dynamic pricing of parking is represented in Figure 11. The capabilities that are developed in this study are marked with deep-red colour. The outcomes highlighted in Figure 11 are mostly targeted on the increase of company profit and well-being of ÜC customers and community. On top of that, it is expected that the service under development will decrease the motorisation level during the working hours and CO2 emission from the transportation.

The flow of cars is going to be distinguished between the incoming and outgoing lanes. Such division is necessary to estimate the incoming and outgoing amounts of vehicles, as well as to be able to regulate and control the traffic flow in the case of closure of one or both lanes.

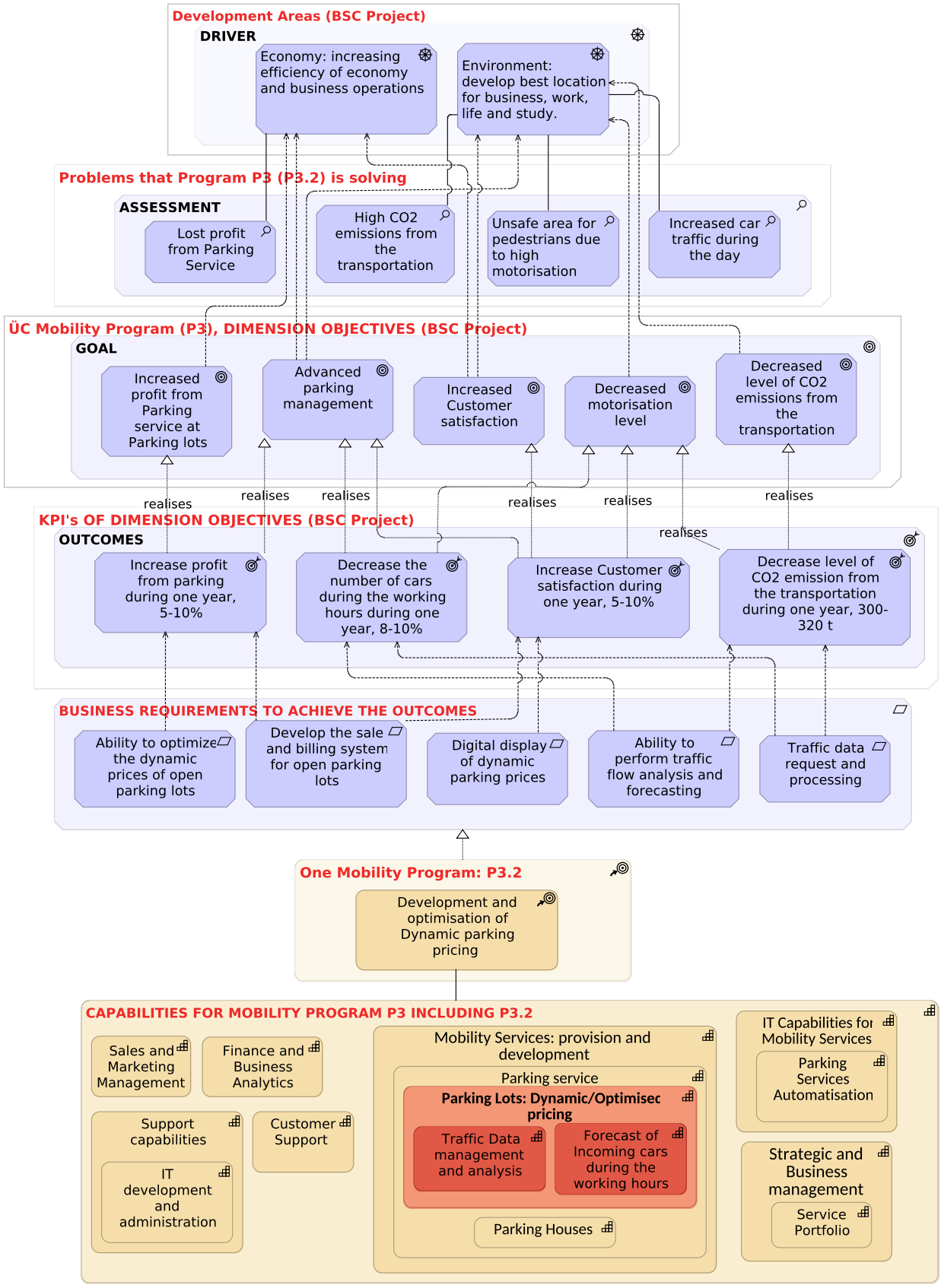


Figure 11. The data-driven motivation model with capabilities necessary for the development of dynamic pricing of parking. Capabilities marked by red colour are those implemented in this study (author created).

4.3.1 Value Stream and Capability Analysis

The value stream, integrating the capabilities necessary for the development of dynamic pricing of parking is represented in Figure 12.

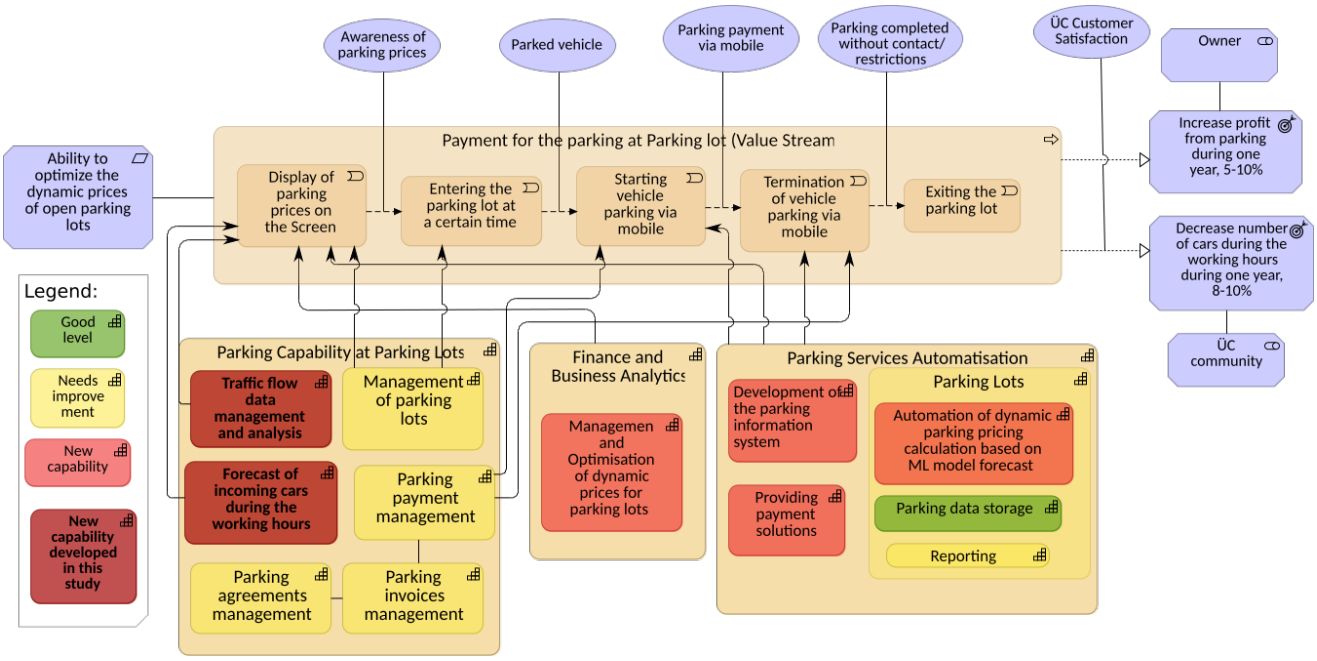


Figure 12. The value stream, integrating capabilities necessary for the development of optimised and dynamic parking prices. The capabilities that are developed in this study are marked with deep-red colour (author created).

4.4 Optimised parking prices: Example pricelist

The number of cars entering the ÜC through the Suur-Sõjamäe-Lõõtsa intersection between the daytime hours from 6.00–19.00 is about 900-1000. It is assumed that around 30% of them are parking on the "K" parking lot. The present parking prices at ÜC parking lots have the rates as follows:

- 2 h with a parking clock are for free;
- 24 h parking: 6 eur;
- every starting 30 min cost: 1 eur.

The dynamic and optimised parking prices are plan to follow the parameters, such as:

1. Season: winter, spring, summer, autumn;
2. Weekdays: Mondays and Fridays should have lower prices, since on these days people often prefer home-office;
3. Daytime hours.

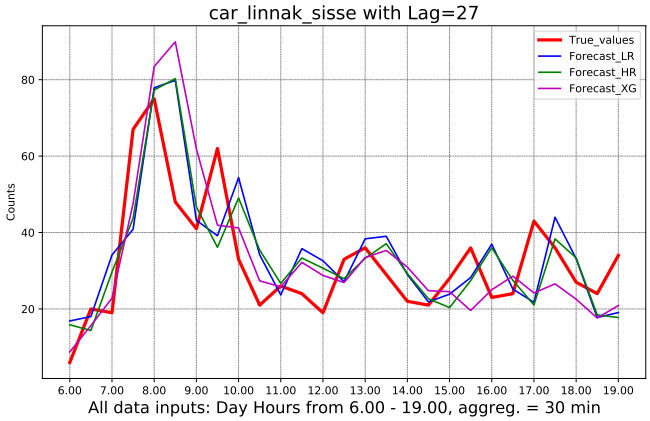
The dynamic and optimised parking arrangement at ÜC parking lots could be:

- 15 min with a parking clock are for free;
- 24 h parking: 6.50 eur;
- daytime hours between 7.00–10.00 and 12.00–15.00 every starting 30 min Price-1 (higher than 1 eur);
- daytime hours between 10.00–12.00 every starting 30 min Price-2 (1 eur).
- daytime hours between 15.00–7.00 every starting 30 min Price-3 (lower than 1 eur);

The example of optimised dynamic pricelist linked to expected incoming cars amount at Suur-Sõjamäe-Lõõtsa road intersection is given in Figure 13. The price list represented in Figure 13a considers the working days, such as Tuesday, Wednesday and Thursday, and excludes seasonality, meaning that the amount of cars predicted are produced by generalised ML model, which considered all calendar seasons at once.

ML_model	Address	Forecast Season	Weekday	Timestamp	Forecasted number incoming cars	Total number cars	Price type
ML0: XG	Suur-S.-Lõõtsa	spring, summer, autumn, winter	Tuesday, Wednesday, Thursday	6.00	9	9	Price_3
				6.30	16	25	
				7.00	23	48	
				7.30	**48**	96	Price_1
				8.00	**83**	179	
				8.30	**90**	269	
				9.00	**62**	331	
				9.30	**42**	373	
				10.00	**41**	414	
				10.30	27	441	Price_2
				11.00	26	467	
				11.30	32	499	
				12.00	29	528	
				12.30	27	555	Price_1
				13.00	33	588	
				13.30	35	623	
				14.00	31	654	
				14.30	25	679	
				15.00	25	704	
				15.30	20	724	Price_3
				16.00	25	749	
				16.30	29	778	
				17.00	24	802	
				17.30	27	829	
				18.00	23	852	
				18.30	18	870	
				19.00	21	**891**	

(a) Example price list (author created).



(b) Incoming cars (author created).

Figure 13. The example of optimised dynamic pricelist linked to expected incoming cars amount at Suur-Sõjamäe-Lõõtsa intersection. The amount of forecasted cars entering the ÜC through the Suur-Sõjamäe-Lõõtsa intersection, using the ML model developed in this study.

4.5 Business Information Model and Business Rules

The target group of the information system under development is one-time/random visitors of ÜC. The system planned promotes the data analytics capability. It does not change the established management of parking in parking lots, which is beneficial as it does not require large investments. The parking pricing optimised by traffic data analysis will not affect the cus-

tomers having monthly contract of parking, thus allowing to have control and management of regular campus customers. The same as in the existing information system, the start and end of parking are implemented by a mobile operator: message/call or using a mobile application connected to mobile operator. Traffic data analysis will include a forecast for the upcoming day. The forecast can take into account the season (summer, autumn, winter, spring), day of the week, working hours. The forecast will include the distribution of the number of expected incoming vehicles during the working day at a particular road-intersection of ÜC. This forecast is going to be delivered to the finance department, where the parking pricing for a particular parking lot will be calculated and optimised. Further, the optimised parking prices are delivered to ÜC database, from where they are displayed on the screen, located in front of the entrance to the parking lot. The expected results of promoting of traffic data analytics capability by parking pricing optimisation for random visitors of ÜC are: an increase of parking service profit of ÜC, an increase of satisfaction and reputation of campus community members, as well as a decrease in the level of motorisation and CO2 emissions from the transportation.

Business Glossary

AGREEMENT - An official document that is signed either by PARKING LOT OWNER and CUSTOMER or by PARKING LOT OWNER and PARKING OPERATOR.

CUSTOMER - A person (individual or legal) who is ÜC visitor or has parking agreement with ÜC.

CUSTOMER TYPE - A set of rights depending on existence of contract.

CONTROLLER - a person controlling the PARKING, assigns PENALTY in case vehicle is not parked.

CAR - vehicle that is parked, linked to CUSTOMER.

INCOMING CARS FORECAST - An expected number of cars within certain time-interval. Depends on intersection address, season, weekday.

MOBILE OPERATOR payment - A system that allows to pay for parking in the parking lot with a mobile invoice.

PARKING LOT OWNER - A legal entity owning the parking territory.

PARKING LOT - A parking lot whose entry and exit are not restricted by an entry/exit blocking system.

PARKING REGULATIONS - The rules set for CUSTOMERS in PARKING LOT and displayed on SCREEN before entering the PARKING LOT.

PARKING - Leaving of vehicle in the PARKING LOT for longer than 15 minutes.

PARKING INTERVAL- chronological parking event.
PARKING OPERATOR - A company that arranges parking at the parking lot.
PRICE - An optimised price based on INCOMING CARS FORECAST. It depends on the location of the PARKING LOT in the ÜC area. The price of parking may be specified by season of the year, weekday, daytime hours.
START/END OF PARKING - sending an SMS/calling MOBILE OPERATOR or using Mobile App linked to MOBILE OPERATOR.

Business Rules
R1. A PARKING LOT OWNER has one or more PARKING LOTS .
R2. Each PARKING LOT has one PARKING LOT TYPE .
R3. Each PARKING LOT has one or more PRICES .
R4. A PARKING LOT OWNER signs one or more AGREEMENTS with PARKING OPERATOR .
R5. Each PRICE is linked to one PRICE TYPE .
R6. Each PRICE may or may not be oprimised by PARKING PRICE OPTIMISATION .
R7. Each PARKING PRICE OPTIMISATION may or may not have INCOMING CARS FORECAST .
R8. A CUSTOMER is associated with one CUSTOMER TYPE .
R9. A CUSTOMER may or may not have an AGREEMENT .
R10. A CUSTOMER AGREEMENT have one or more AGREEMENT CARS.
R11. Every PARKING is always associated with one CUSTOMER .
R12. Each PARKING is always associated with one specific CAR .
R13. Each MOBILE DEVICE is associated with one MOBILE OPERATOR .
R14. The PARKING is started/terminated through the one MOBILE DEVICE .
R15. A CUSTOMER may have one or more MOBILE DEVICES .
R16. Each PARKING consists of one or more PARKING INTERVALS .
R17. PARKING INTERVAL is counted by one MOBILE OPERATOR .
R18. PARKING LOT OWNER issues one or more PARKING OPERATOR INVOICES .
R19. PARKING LOT OWNER issues one or more CUSTOMER INVOICES .
R20. MOBILE OPERATOR issues one or more MOBILE INVOICES .
R21. PARKING OPERATOR makes payment to PARKING LOT OWNER .
R22. A CUSTOMER (non-agreement) pays one or more MOBILE INVOICES .
R23. A CONTROLLER is associated with one PARKING OPERATOR.
R24. A CONTROLLER issues one or more PENALTIES.

R25. PARKING LOT OWNER may sign one or more AGREEMENTS with CUSTOMER .

R26. A CUSTOMER makes payment to PARKING LOT OWNER .

R27. Each PARKING is associated with one PARKING LOT .

R28. A MOBILE DEVICE is associated with one or more MOBILE INVOICES .

R29. A CAR may be parked via one or more MOBILE DEVICES .

R30. A CAR may be assigned one or more PENALTY .

R31. PARKING OPERATOR signs MOBILE PARKING AGREEMENT with one or more MOBILE OPERATORS .

Business Information Model (BIM) demonstrating the relations of business entities, parties and events of parking service with dynamic and optimised parking prices is given in Figure 14.

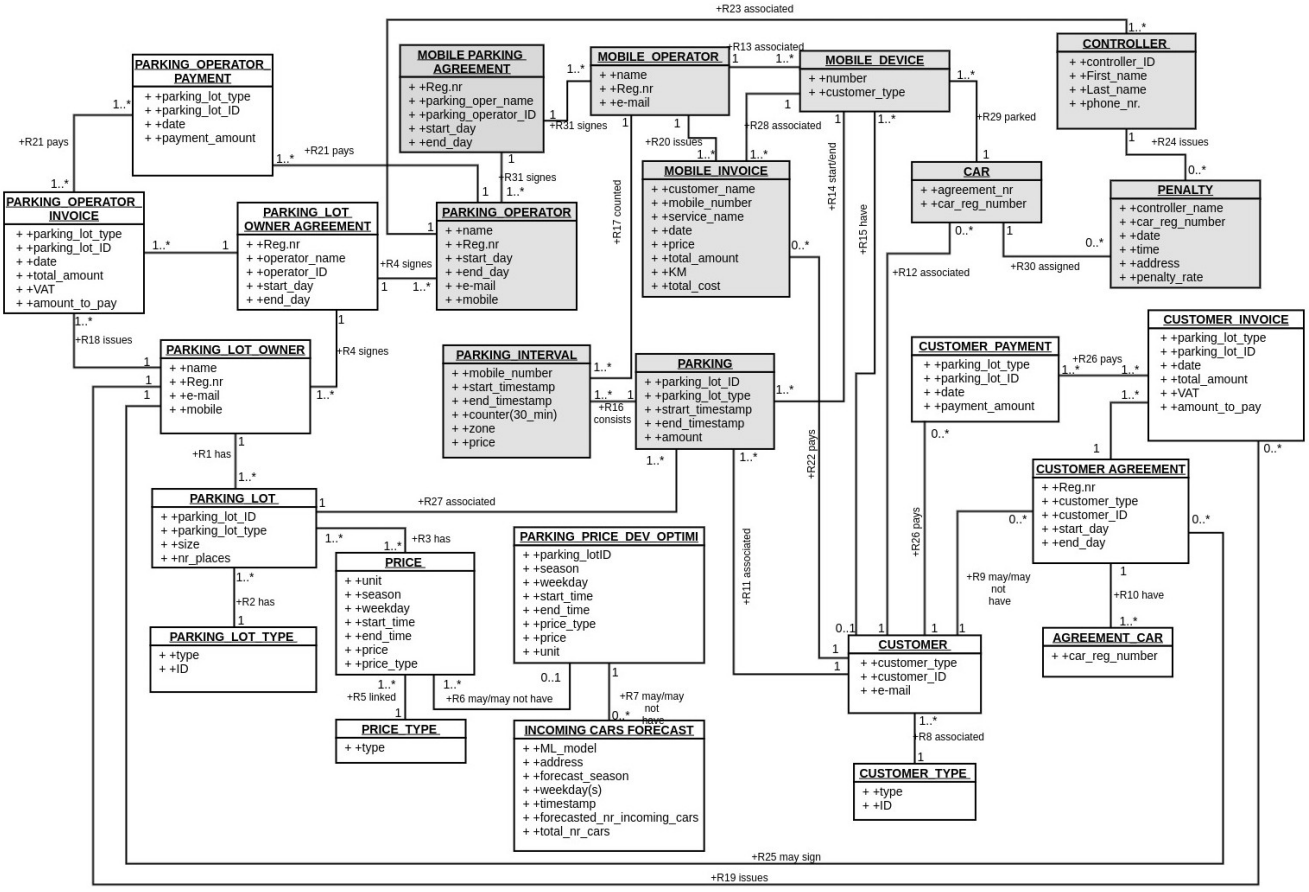


Figure 14. Business Information Model (BIM) reflecting the relations of business entities, parties and events of parking service with dynamic parking pricing. Grey coloured entities correspond to external system (author created).

4.6 Use Cases

The use cases (UC) that form the parking information system using the development of dynamic and optimised pricing are represented in Figure 15.

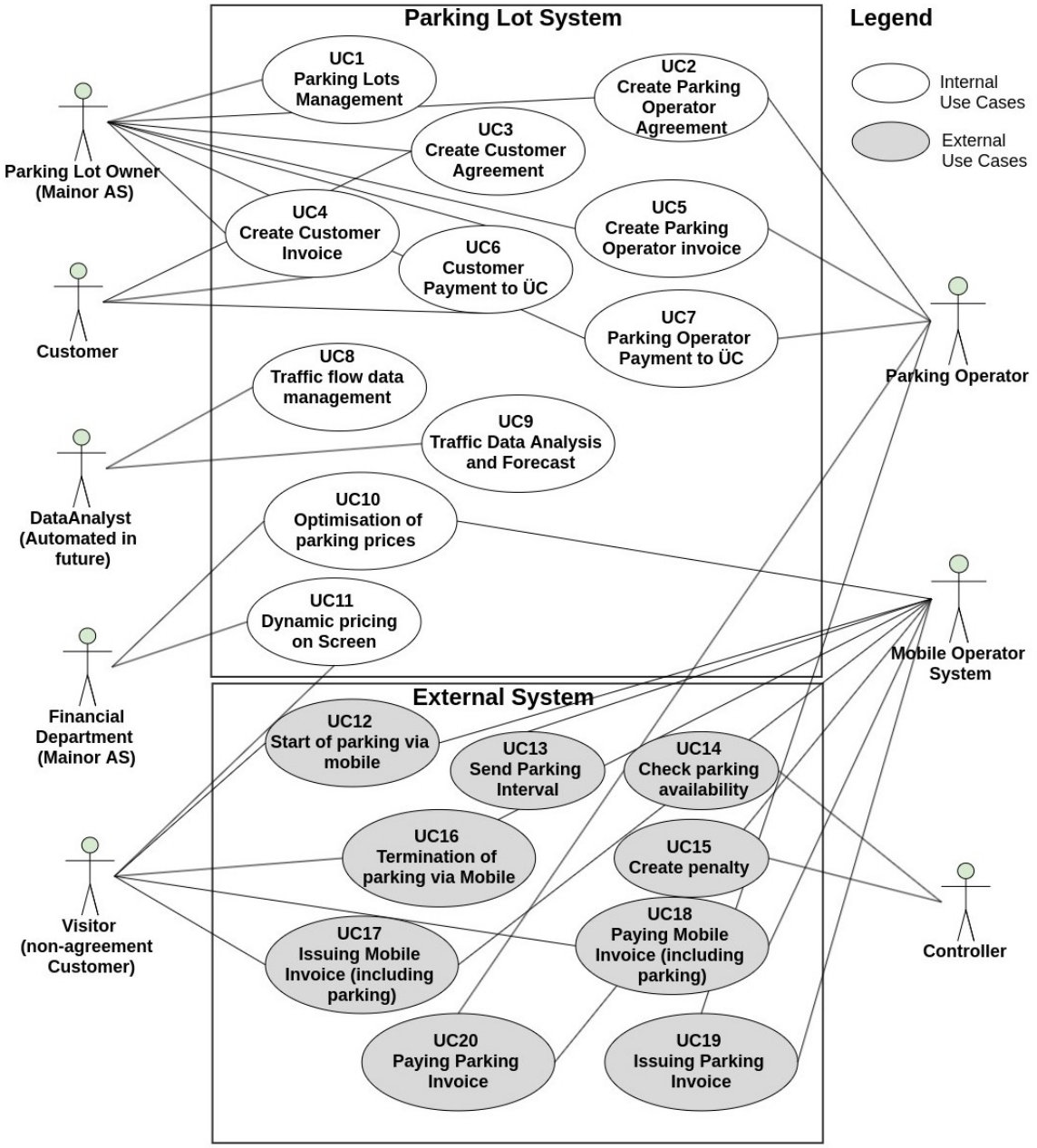


Figure 15. The use cases (UC) that form the parking information system using the development of dynamic pricing. UCs marked by gray colour correspond to external system (author created).

4.7 Layered model of realisation of dynamic and optimised prices of parking (business service and process)

The layered model of parking with dynamic and optimised prices including the business roles, events, services, processes, application services and components is given in Figure 16. The layered view in Figure 16 demonstrates the integration of the capabilities necessary for the development of dynamic pricing of parking. The capabilities marked by deep-red colour are the ones that are developed in the present work.

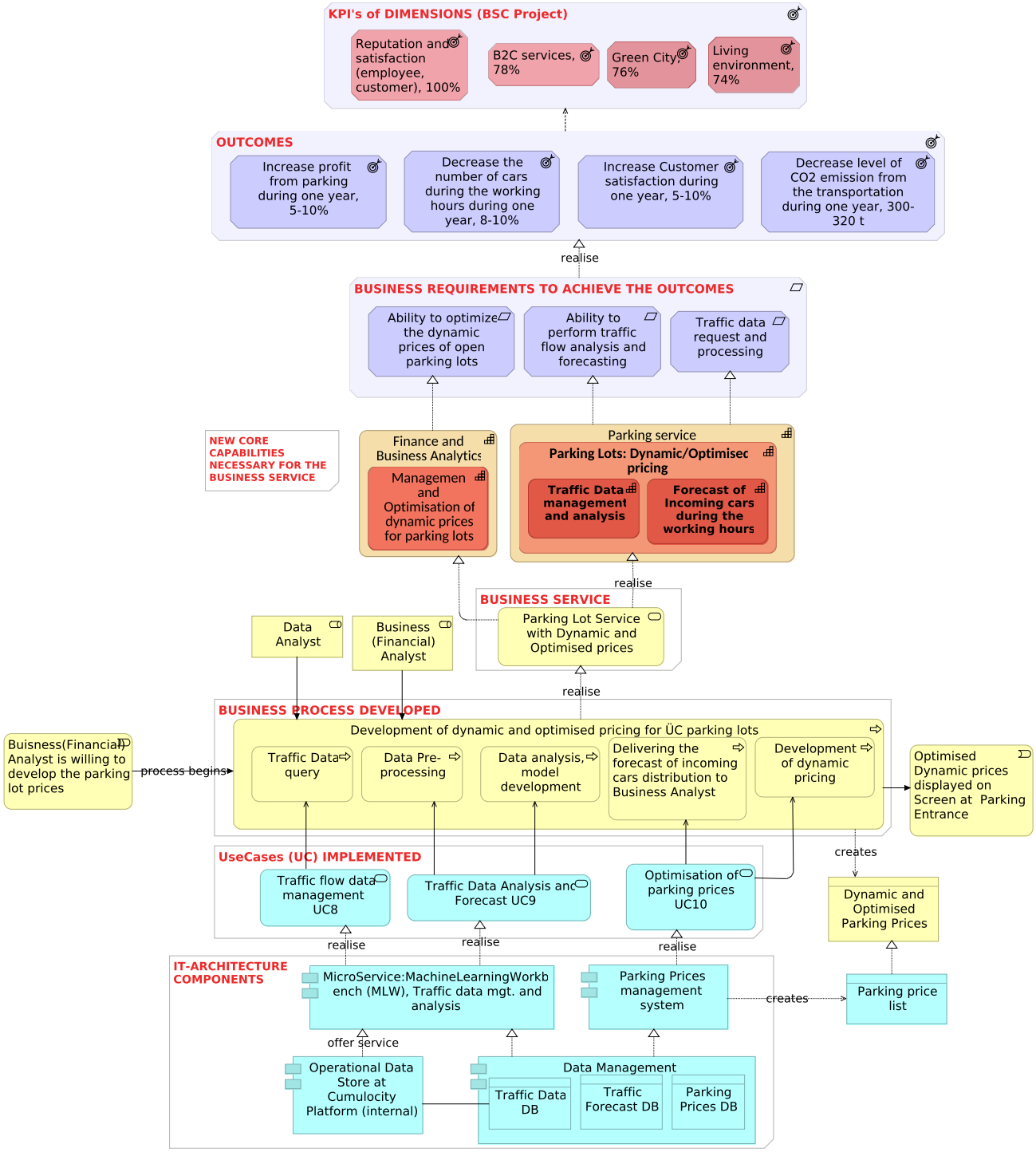


Figure 16. Aligned IT solution of the payment process at parking lot based on the dynamic pricing through the business architecture to the business outcomes. The model also includes business roles, events, services, processes, application services and components (author created).

4.8 Business processes: TO-BE

The configuration of the process flow given in Figure 17 demonstrates the TO-BE process flow.

The advantages of TO-BE solution is that it promotes the data analytics capability, which supports and implements the business requirement of the design of data-driven service(s). On top of that, the TO-BE process does not require considerable investments related to technological solutions, as it does not change the established management of parking in parking lots.

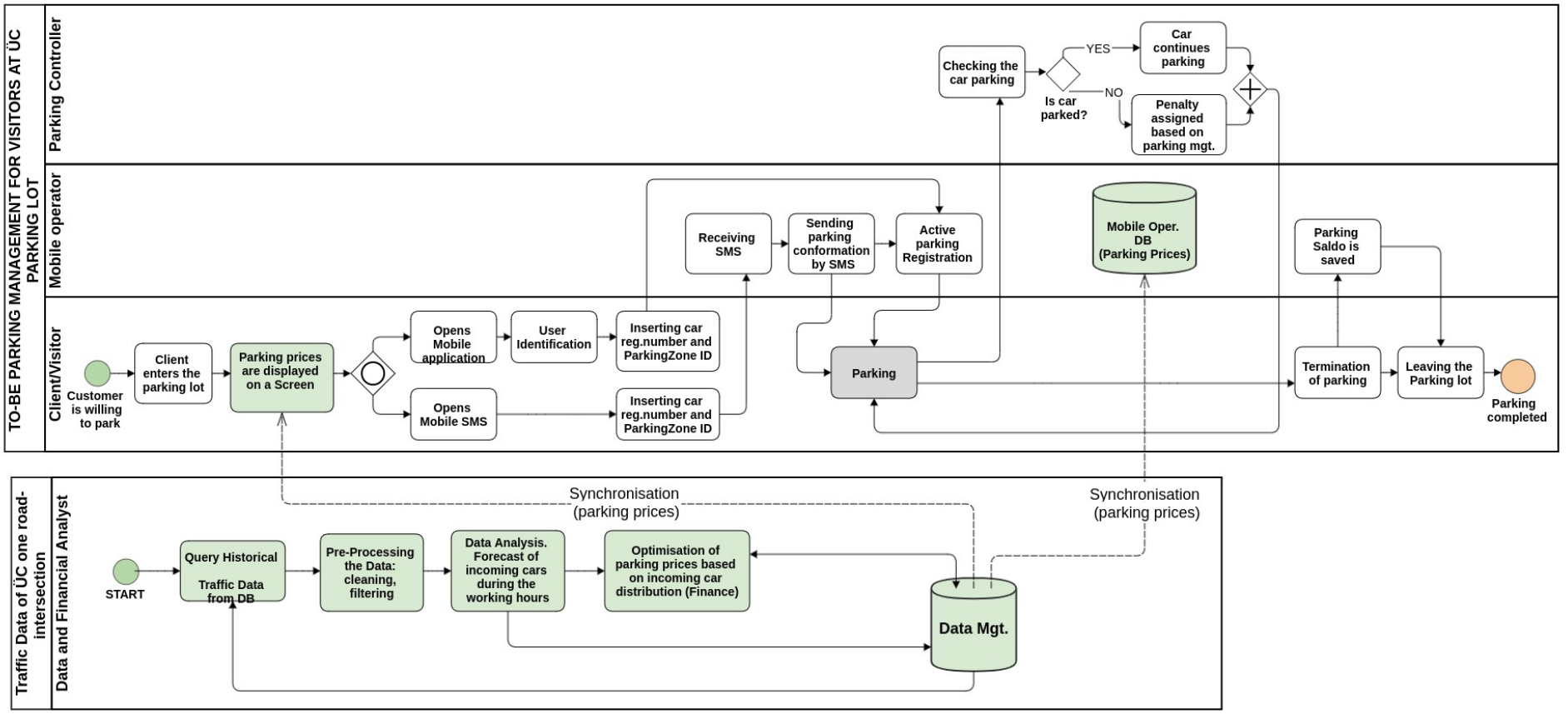


Figure 17. TO-BE process flow of parking at parking lot with dynamic and optimised prices.
Activities that are developed in this study are marked with green colour (author created).

4.9 Architectural vision of the Information System

The component diagram demonstrating the architectural vision of the information system enabling the implementation of the parking service on ÜC parking lots employing dynamic and optimised parking prices is represented in Figure 18. The IT-architecture given in Figure 18 includes also the components showing the access to time series database collecting the traffic flow data of Ülemiste City. The technology generating the traffic flow data allows the integration with Cumulocity platform.

In the present work the traffic data were pulled from the Fyma time series database, and the training and testing of ML models developed were implemented in Jupyter Notebook on local PC. In Figure 18 the components coloured by pink are those which are external. The components coloured by green are those that represent the present and developed connections and integrations.

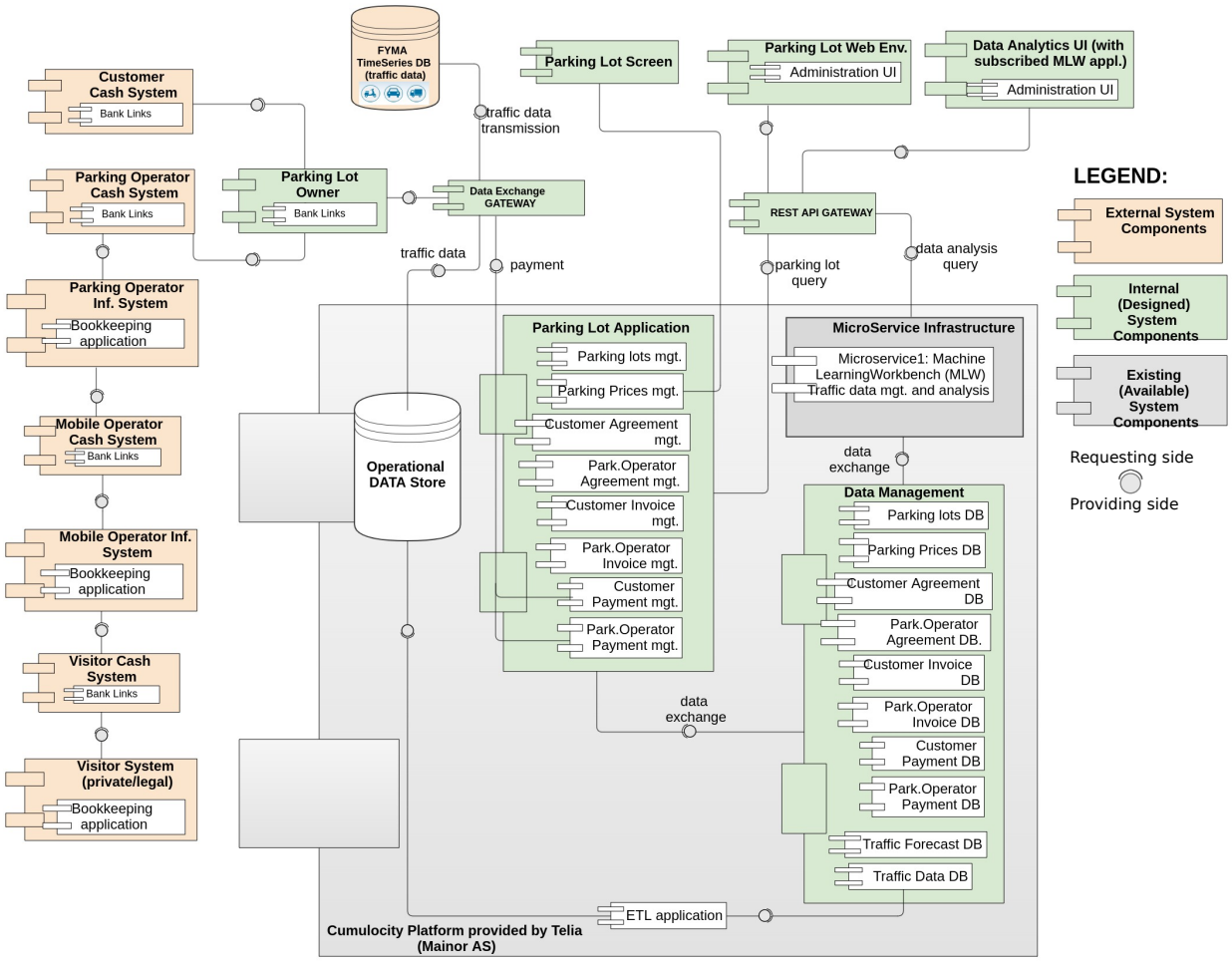


Figure 18. Architectural vision of the information system enabling traffic data analysis, calculation and optimisation of dynamic parking prices and digital display of these prices on the screen (author created).

5. ÜC traffic analysis as a method for UC8, UC9 implementation

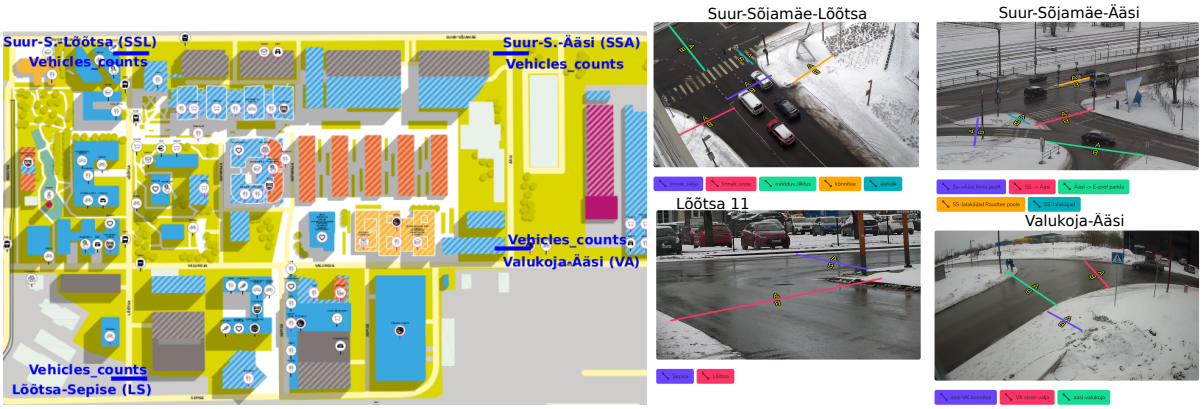
This chapter reflects the traffic data management and analysis, as well as the building of machine learning models forecasting the amount of cars entering the Ülemiste City through the Suur-Sõjamäe-Lõõtsa road-intersection during the daytime hours. This Chapter also addresses the study of the effect of weather properties on the models accuracy.

5.1 Traffic data of Ülemiste City

The data used in this work were collected from two main sources: Fyma and EMHI Tallinn Airport measuring station databases. Fyma database includes the time series data of the counts of objects, such as, cars, vans, trucks, etc. The data needed manual pre-processing, for example, due to the selected working days and hours, as well as the selected vehicle types.

Four main cameras at the Ülemiste City (UC) are capturing the vehicles that are coming in and out of the city, Figure 19a. The planning of the Ülemiste City infrastructure can be supported by employing the data generated by ML object detection algorithm applied on live video streams from the cameras located at the road intersections. The ML model can distinguish between the objects such as: cars, vans, buses, trucks, motorbikes, bicycles, tractors, persons. by detecting and counting the objects on the counter lines. The representation of the road intersections, camera views and the counter lines are given in Figure 19b. The available traffic data can be examined from the following perspectives (categorical variables): temporal, spatial, compositional, directional. These factors may be considered, when modelling the infrastructure of the whole Ülemiste City. These categorical variables can be encoded using, for example, one-hot, dummy, binary encodings [17, 18].

The present study focuses on the modelling of traffic flow of one road intersection: Suur-Sõjamäe-Lõõtsa (SSL). The traffic data were received by queering the Fyma time series database. The time series were given in UTC (Coordinated Universal Time) and were stored in .csv files.



(a) The location of the cameras [15]. (b) Camera views and the counter lines [16].

Figure 19. The location of the cameras in the Ülemiste City. Camera views and counter lines for object detection.

The time stamp format used was: yyyy-mm-dd hh:mm:ss. Further on, a date time in UTC was changed to Estonian local time Europe/Tallinn, meaning the UTC Offset +3:00 hours during Eastern European Summer Time and +2:00 hours during Eastern European Time. The missing values in data were replaced with zeros, meaning no vehicles within the considered time periods.

5.2 Weather Data

Most studies considering weather properties in the modelling prefer to include temperature, such as, in the study by [19]. Usually, temperature has high correlation with such target variable values as, for example, electricity consumption, traffic amount. However, the number of cars and/or bicycles can also be affected by weather parameters, such as, wind speed, relative humidity. On top of that, the addition of different weather parameters and/or combination of weather parameters may improve the accuracy of ML model. In the study by [19], it was found that in addition to temperature, the inclusion of wind speed and power effective cloud cover to ML models improved electricity consumption forecasts, supporting the assumption that temperature was not the only variable affecting the electricity loads.

In the present work the weather data were received from EMHI Tallinn Airport measuring station. The files with weather properties by months with the aggregation of 30 minutes were stored in .csv format, the timestamp was changed to yyyy-mm-dd hh:mm:ss and converted to Estonian local time. The data included the following weather properties:

- Time in UTC format;
- Wind direction in degrees, i.e. average wind direction during 10 minutes. "VRB" means wind with variable direction;
- Wind speed in knots: average wind speed during 10 minutes;
- Visibility. If visibility > 10000 m, the data is in kilometres; if visibility < 10000 m then the data is in meters. For example: "10" = 10 km and more, "3000" = 3000 m, "400" = 400 m;
- Weather phenomena: SN-moderate snow, SHSN-blowing heavy snow, RA-rain, SHRA-showers, FZRA-Freezing Rain, DZ-drizzle, FZDZ-Freezing drizzle, PL-ice rain, SG-snow blades, SHGS-snow gravel/ice gravel, SHGR-hail, TS-thunderstorms no precipitation, TSRA-Thunderstorms with showers, TSGR-thunderstorms with hail, VCTS-thunderstorms around the airport.
- The lower boundary of the lower cloud layer in feet, and the amount of cloudiness in octants.
- Air temperature: in degrees Celsius;
- Relative humidity: in percent;
- Precipitation in millimetres per day. Precipitation is measured in the period from 01.05 until 31.10.

5.3 Approahes for Time series

Uni-variate time series can be modelled by typical classical methods of analysis. The analysis and modelling of the multivariate data is more complex and classical methods usually are ineffective in this case. The application of machine learning for uni-variate and multivariate time series is where the classical methods may fail. Compared to static data, when the values in the future are "predicted", the time series values are "forecasted".

Simpler ML methods of time series forecasting perform more effective than the advanced ones, including neural network models [20]. The authors in [21] considered RNN and LSTM methods among the less accurate ones, meaning that the research progress did not necessarily guarantee improvements in forecasting performance.

The real world time series usually demonstrate both trends and seasonality. The trends in the time series can be noticed by increasing or decreasing the values over time, demonstrating the upward or downward trend, respectively. Regularity of variations in the observations over the

same/similar time interval is a property pointing to seasonality. For example, the variations in the observations are repeating every week, month, etc. The regularity of the period of change defines the difference between seasonal and cyclical behaviour. In case of cyclical behaviour the time between the periods is not precise, and this is in contrary to seasonality, when the period is strictly regular, i.e. exact amount of time between the peaks and troughs in the data. Often, time series demonstrate both cyclical and seasonal behaviour. The difference between seasonal and cyclic behaviour can be measured with reasonable accuracy. First, the regularity of peaks in the data can be measured, and then the deviation of time peaks from the average distance between them can be calculated. Time series with highly pronounced seasonality typically demonstrate clear peaks in both partial auto-correlation and auto-correlation (self-correlation) functions. Compared to auto-correlation, i.e. correlation between the time series values that are $\mathrm{lag_n}$ intervals apart, the partial auto-correlation is the correlation between the time series values that are $\mathrm{lag_n}$ intervals apart, as well as accounting for the values of the intervals between. A good practical way to distinguish between the cyclical or seasonal behaviour is to consider the physical nature of the data, meaning if the periodicity is caused directly on time then the data are most likely seasonal. For example, temperature has seasonality as it directly depends on the time of the year, i.e. yearly seasonal pattern. If the reason for the change is mainly due to the previous values of the time series, and not directly to time, then most likely the behaviour is cyclical. For example, when the value of a stock goes up, more people invest, which causes the price to go up, and vice versa. This kind of behaviour has cyclical pattern.

To make time series forecasting, the given time series data should be transformed to stationary, i.e. having no seasonality and trend. Stationary time series data have constant mean, variance values, as well as constant auto-correlation.

5.4 Modelling Methodology

ML methods can be split into three main branches: supervised, unsupervised and reinforcement learning. This study focuses on unsupervised and supervised, i.e. clustering and ML methods, respectively. The data used in unsupervised learning has no labels and is unclassified, meaning that the hidden patterns, distributions should be found while training of the model.

5.4.1 ML methods used in this study

Hard Clustering: k-means

There are two main flavors of clustering methods [7]: hard clustering methods assign each data point to exactly one cluster; soft clustering methods assign each data point to several different clusters with varying degrees of belonging. Hard clustering can be interpreted as a special case of soft-clustering, meaning that if a data point belongs to a cluster then degree of belonging is 1 and in case of no belonging it is 0. The clustering methods learn a reasonable cluster assignments for data points based on the intrinsic geometry of the entire data-set. An index of the cluster to which the i-th data point $\mathbf{x}^{(i)}$ belongs to can be denoted as $y^{(i)} \in \{1, \ldots, k\}$ and can be considered as a label [7].

k-means is a hard clustering method, which uses Euclidean geometry as a notion of similarity. A data-set is partitioned into k non-overlapping clusters $\mathcal{C}_1, \ldots, \mathcal{C}_k$, characterised by the cluster mean:

$$\mathbf{m}^{(c)} = (1/|\mathcal{C}_c|) \sum_{\mathbf{x}^{(i)} \in \mathcal{C}_c} x^{(i)}, \tag{5.1}$$

where $|\mathcal{C}_c|$ denotes the number of data points in the cluster C_c. A data point $\mathbf{x}^{(i)}$ can be assigned to that cluster $y^{(i)}$, whose mean is closest to $\mathbf{x}^{(i)}$. However, in order to determine the cluster means $\mathbf{m}^{(c)}$, one should know the cluster assignments $y^{(i)}$ already in the beginning. This instance is solved by the k-means algorithm as follows:

- Input: data points $\mathbf{x}^{(i)} \in \mathbb{R}^n$, for $i = 1, \ldots, m$ and number k of clusters is pre-defined;
- Initialisation: choose initial cluster means $\mathbf{m}^{(1)}, \ldots, \mathbf{m}^{(k)} \in \mathbb{R}^n$;
- Repeat until stopping condition is met by updating cluster assignments and cluster means, respectively.

Linear Regression

Consider predicting the numeric label y based on n features of a data point:
$\mathbf{x} = (x_1, \ldots, x_n)^T \in \mathbb{R}^n$. This can be implemented by learning a predictor function $h(\mathbf{x})$ such that $y \approx h(\mathbf{x})$ [7]. By using the linear regression, the predictor function is restricted to be a linear function, meaning that the hypothesis space includes linear predictor functions:

$$\mathcal{H} = \{h^{(\mathbf{w})}(\mathbf{x}) = \mathbf{w}^T\mathbf{x} \text{ for some } \mathbf{w} \in \mathbb{R}^n\}. \tag{5.2}$$

For two vectors $\mathbf{w} = \big(w_1, \ldots, w_n\big)^T \in \mathbb{R}^n$ and $\mathbf{x} = \big(x_1, \ldots, x_n\big)^T \in \mathbb{R}^n$, an inner (dot) product between those vectors is denoted as $\mathbf{w}^T\mathbf{x} = \sum_{r=1}^{n} w_r x_r$.

To measure the quality of a particular predictor $h^{(\mathbf{w})}(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$ obtained for some particular choice of the weight vector $\mathbf{w} \in \mathbb{R}^n$, the labelled data points are used. Labelled data points with features $\mathbf{x}^{(i)}$ and labels $y^{(i)}$ can be compared with the predicted labels, which typically incur a non-zero prediction error: $y^{(i)} - h^{(\mathbf{w})}(\mathbf{x}^{(i)})$. To measure the error or loss incurred by the prediction, a loss function should be defined $\mathcal{L}(y, \hat{y})$, which can be chosen freely depending on the application at hand. For numeric labels y, a popular choice for the loss is the squared error loss: $\mathcal{L}(y, \hat{y}) = (y - \hat{y})^2$. Given a set of labelled data points $\big(\mathbf{x}^{(i)}, y^{(i)}\big)$, the average squared loss can be calculated

$$\mathcal{E}(\mathbf{w}) = (1/m) \sum_{i=1}^{m} (y^{(i)} - \mathbf{w}^T\mathbf{x}^{(i)})^2. \tag{5.3}$$

The optimal weight vector $\mathbf{w}_{\rm opt}$ is any weight vector which achieves the minimum value of $\mathcal{E}(\mathbf{w})$, i.e.,

$$\mathcal{E}(\mathbf{w}_{\rm opt}) = \min_{\mathbf{w} \in \mathbb{R}^n} \mathcal{E}(\mathbf{w}).$$

An optimal predictor is then obtained as $h(\mathbf{x}) = \mathbf{w}_{\rm opt}^T\mathbf{x}$. The average squared loss:

$$\mathcal{E}(\mathbf{w}_{\rm opt}) = (1/m) \sum_{i=1}^{m} (y^{(i)} - \mathbf{w}_{\rm opt}^T\mathbf{x}^{(i)})^2$$

incurred by the optimal predictor $h(\mathbf{x}) = \mathbf{w}_{\rm opt}^T\mathbf{x}$ is also known as the training error.

Linear Regression with Huber losss

An important property of ML methods is their robustness to (small) perturbations in the data . Linear predictor is heavily affected by corrupting only one single data point [7]. The reason for this sensitivity is rooted in the properties of the squared error loss function used by Linear Regression. By using a different loss function it is possible to learn a linear predictor that is robust against few outliers. One such robust loss function is known as Huber loss $\mathcal{L}(\hat{y}, y)$.

Given a data point with label y and a predicted label $\hat{y} = h(\mathbf{x})$ the Huber loss is defined as

$$\mathcal{L}(y, \hat{y}) = \begin{cases} (1/2)(y - \hat{y})^2 & \text{for } |y - \hat{y}| \leq c \\ c(|y - \hat{y}| - c/2) & \text{else.} \end{cases}$$

A parameter c in Huber loss represents two important special cases and should be adapted based on the application. The first special case when c is chosen very large, such that the condition $|y - \hat{y}| \leq c$ is always satisfied. In this case, the Huber loss becomes the squared error loss $(y - \hat{y})^2$ (with an additional factor 1/2). The second special case when c is very small (close to 0), such that the condition $|y - \hat{y}| \leq c$ is never satisfied. In this case, the Huber loss becomes the absolute loss $|y - \hat{y}|$ scaled by a factor c.

Gradient Boosting (GB) and Extreme Gradient Boosting (XGB)

The main difference between the GB and XGB is that in GB an ensemble of DTs is built sequentially, while in XGB the DTs are built in parallel. The boosting stands for making a weak learning algorithm (e.g. one DT) to a strong one by using the ensemble of DTs. Gradient is referring to gradient descent that is utilised to minimise the loss function.

A specific choice of weight vector (or parameter) results in a specific predictor map $h^{(\mathbf{w})}(\mathbf{x})$, such that finding a good predictor becomes equivalent to finding a weight vector [7]. By solving the: $\min_{\mathbf{w} \in \mathbb{R}^d} f(\mathbf{w})$, where $f(\mathbf{w})$ is referred also as the loss function, the weights resulting in minimum training error cab be obtained. By selecting the linear predictor functions and squared error (SE) loss one can get convex [22] and differentiable loss function $f(\mathbf{w})$, which can be minimised by gradient descent (GD) algorithm. GD is an iterative algorithm that gradually improves the current approximation of $\mathbf{w}^{(k)}$ for the $\mathbf{w}_{\text{opt}}$. A single gradient step can be summarised as follows:

- compute predictions for the data points in a batch, given the current weights;
- compute mean SE (MSE) loss;
- compute gradient of the loss function;
- update the weights - change the weights values to the opposite direction from gradient;
- weight = weight - $\mathrm{l}_{\text{rate}} \cdot$ gradient.

Gradient boosting algorithms are a group of ML methods that combine many weak learning

models together to create a strong predictive model. Gradient boosting systems have two necessary parts: a weak predictor and an additive component. Gradient boosting systems use DTs as their weak predictors. The additive component comes from the fact that trees are added to the model over time, which means that existing trees are not modified, their values remain fixed. The gradient boost algorithm depends on the loss function, which must be differentiable. Classification algorithms often use log loss, while regression may use squared errors. Tuning the hyper-parameters of the model requires attention. For instance, test the performance of the model on the training set at different learning rates and then use the best learning rate for prediction. Taking random sub-samples of the training data-set, a technique called stochastic GB (SGD), can help prevent over-fitting. This method significantly reduces the strength of correlation between the trees.

Extreme Gradient Boosting (XGBoost) is an enhanced and tuned version of the gradient boosting DTs system built with performance and speed in mind. XGBoost is an efficient implementation of the SGD algorithm. It is an ensemble of DTs, where new trees fix errors of those trees that are already part of the model. Each time a new weak predictor is added to the model, the weights of previous predictors are frozen. Trees are added until no further improvements can be made to the model. The most important factor behind the success of XGBoost is its scalability in all scenarios. XGBoost is designed for classification and regression on tabular data-sets.

XGBoost can also be used for time series forecasting, which first requires to transform the time series data to a supervised learning problem. In case of time series, the model evaluation is implemented by so called "walk-forward validation", since the evaluation using the k-fold cross validation would result in optimistically biased results. Time series data can be formulated as supervised learning using previous time steps as input variables and using the next time step as output variable. In Figure 20 the time step column is dropped and the measures/observation column is shifted . This representation is called a "sliding window", since the window of input and expected output is shifted forward in time to create new samples for the supervised learning model. A sliding window method uses the values of the prior time steps to predict the values in the next time steps. This method is also called as "lag" method. A window width or size of the lag corresponds to the the number of previous time steps. A sliding window approach is the basis how the time series data is transformed to supervised learning. A walk-forward validation method employs true observation values as input and makes the forecast.

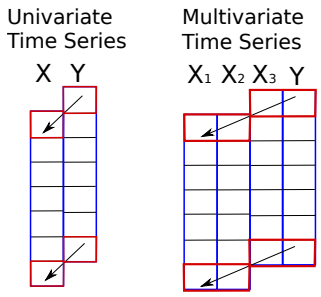


Figure 20. Dropping of the time step column and shifting of the measures/observation column (author created).

5.4.2 Tools Overview

This section gives a brief overview of tools used in this study, like programming language, libraries and source code. High level flow chart of traffic flow forecasting system is represented in Figure 21.

Programming Language

Data pre-processing was made by R-programming language [23] and Python [24]. Filtering of the needed working days and hours, i.e. Tuesday, Wednesday, Thursday from 6.00-19.00 with the aggregation of 30 minutes, was implemented by `R` employing the following libraries: `ggplot2`, `dplyr`, `lubridate`, `readr`, `tidyr` [25, 26, 27, 28, 29]. Python libraries, such as, `Scikit-learn`, `Pandas`, `Matplotlib`, `NumPy`, `XGBoost`, `statsmodels`, `tslearn` [17, 18, 30, 31, 32, 33, 34] were used in Data transformation and ML modelling.

Source Code

Folder ylemiste holds four main folders:

- `/api-fyma`. This folder holds Python scripts to make API queries to Fyma time series databases, and to make a conversion to local time;
- `/data-traffic`. This folder holds three sub-folders, such as:
  – `/data-raw`, holds raw traffic data received from Fyma time series databases in .csv format;
  – `/data-local`, holds traffic data in local time by months in .csv format;
  – `/data-TKN`, holds "filtering-TKN-1year.Rnw" file with R-script to filter the neces-

sary weekdays and working hours. The file is applied on monthly traffic data in local time and the filtered traffic data is produced in .csv files by months.

- `/data-emhi`. This folder holds three sub-folders, such as:
  – `/data-raw`, holds raw weather data received from EMHI in .xls format;
  – `/data-local`, holds weather data in local time by months in .csv format;
  – `/data-TKN`, holds "filtering-TKN.Rnw" file with R-script to filter the necessary weekdays and working hours. The file is applied on monthly weather data in local time and the filtered weather data is produced in .csv files by months.
- `/analysis-traffic`, holds two sub-folders:
  – `preprocessing-modelling`, holds Jupyter Notebooks [35], such as:
  "SSL − data − transformation − traffic + weather.ipynb";
  "SSL − clustering + BM + ML0.ipynb";
  "SSL − ML1 − ML2 − ML3 − ML4.ipynb".
  – `/summaries`, holds .odt file with calculated MAEs.

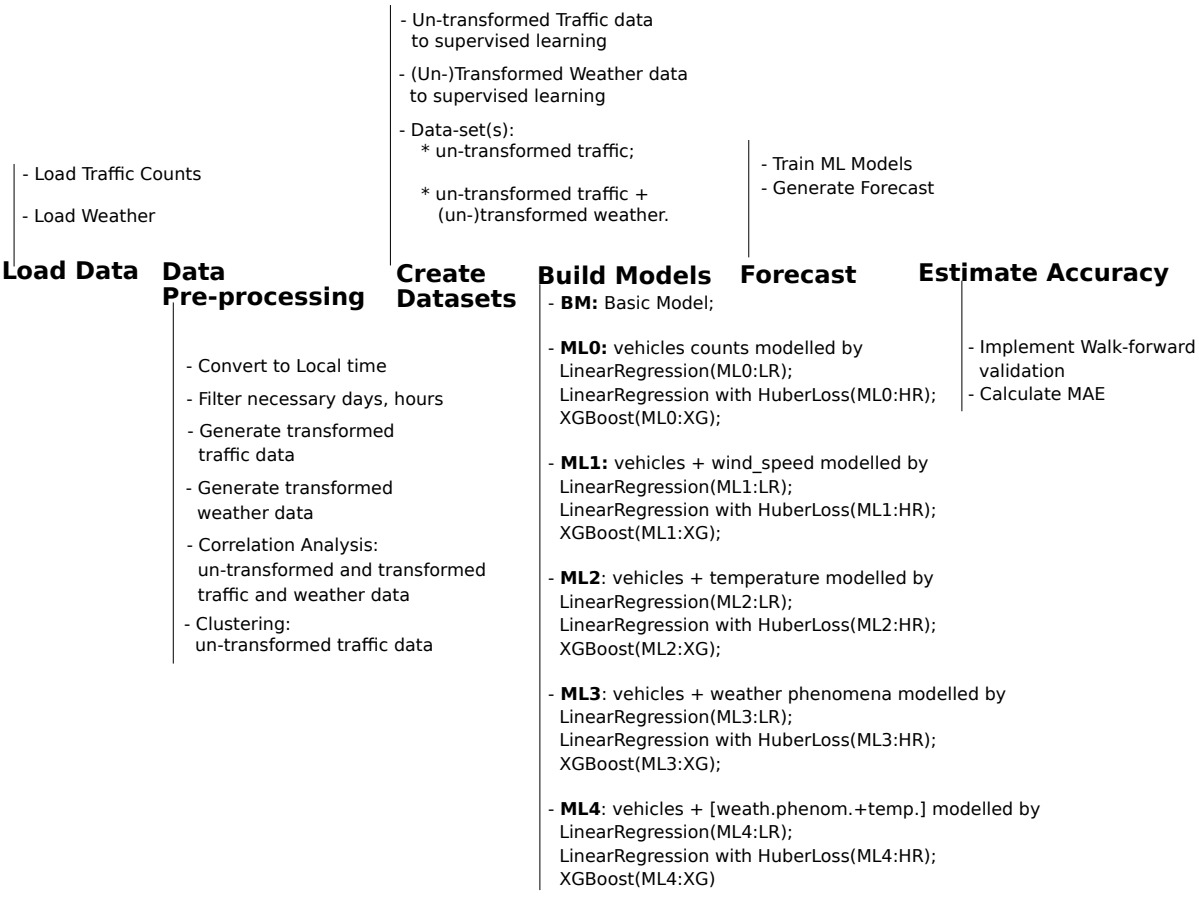


Figure 21. High level flow chart of traffic flow forecasting system (author created).

5.5 Pre-processing

This section demonstrates exploratory analysis by investigating the properties of the data via scatter and/or histogram plots, data transformation methods and their application on available data. Clustering analysis also performs as an exploratory analysis that is meant to identify structures within the data.

5.5.1 Transformation Methods

A large portion of the performance of ML algorithms is due to the right choice and pre-processing of the data. It is relevant to avoid non-important variables (features), since it may cause noise in the data, involve highly correlated features and/or features having low correlation with the target value. Choosing the most important features may reduce training and evaluation time, reduces complexity of the model, improves prediction and reduces over-fitting [36]. It is recommended that all features should follow or be close to Normal distribution, and data should have no significant outliers. Although, it is recommended not to throw away outliers, unless one has evidence that they are errors. If outliers are still present, the algorithms that are robust to outliers can be used. For instance, covariance or mean are sensitive to outliers. In this case mean can be replaced with median. Simple normalisation called "scaling" is typically applied in a way, such as, for each data point x_i from a set $\mathbf{X}$, a scaled value is computed as:

$$x_i^{'} = \frac{x_i - \min(\mathbf{X})}{\max(\mathbf{X}) - \min(\mathbf{X})}. \tag{5.4}$$

It is also common to centre the values around e.g. 0 or their arithmetic mean, median, centre of mass etc. "Standardisation" to zero mean and unit variance can be implemented, such as, given the mean μ and standard deviation σ for a variable, then for each data point x_i, the standardised variable value is:

$$x_i^{'} = \frac{x_i - \mu}{\sigma}. \tag{5.5}$$

Using the variance σ^2 instead of σ is called variance scaling.

The necessary number of features n and data points m minimising the over-fitting depends on the hypothesis space one is using [7]. d is an effective dimension of a hypothesis space and the ratio d/m leads or not to over-fitting. For linear hypothesis space, the effective dimension of d is equal to the number of features. The ratio of d/m can be brought below the critical value of 1 by increasing the m data points and using smaller hypothesis space d, meaning the

decreasing of the number of features n. As a rule of thumb, in case of linear hypothesis space it is recommended to have $m \geq n \cdot 10$.

Trends and seasonality should be removed from the time series to use the data in prediction models. There are various ways to de-trend a time series, for example:

- Power transformation;
- Local smoothing: applying moving window functions;
- Linear regression;
- Differencing a time series.

There are various ways to remove seasonality, such as, for example:

- Average de-trended values;
- Differencing a time series;
- Use the loess method.

Data Transformations used in this study

- Power Transformations: square, cube root, which are usually used to fix the right-skewed data: $x_i^{'} = \sqrt{x_i}$, $x_i^{'} = \sqrt[3]{x_i}$
- Moving Window Function. Rolling mean over a selected period is calculated and then subtracted from the original time series to get de-trended ones.

```
rolling-mean = data.rolling(window-size).mean()
data-detrended = data-rolling-mean
```

- Linear Regression. Linear regression model can be applied to remove trend following the further steps:
  – Fit a linear regression model to time series data;
  – Use a fit model to predict time series values from beginning to end;
  – Subtract predicted values from original time series to remove the trend.

```
least_squares = OLS(data.values, list(range(data.shape[0])))
result = least-squares.fit()
fit = pd.Series(result.predict(list(range(data.shape[0]))),
index = data.index)
data_ols-detrended = data - fit
```

- Differencing Over Power Transformed time series. Differencing is applied to power transformed time series by shifting its value by 1 period and subtracting it from original power transformed time series. It is common to try shifting time series by different time periods to remove seasonality and get stationary time series.

```
data-pow = data.apply(lambda x : x ** 0.5)
data-pow-shift = data-pow-data-.pow.shift(periods=periods)
```

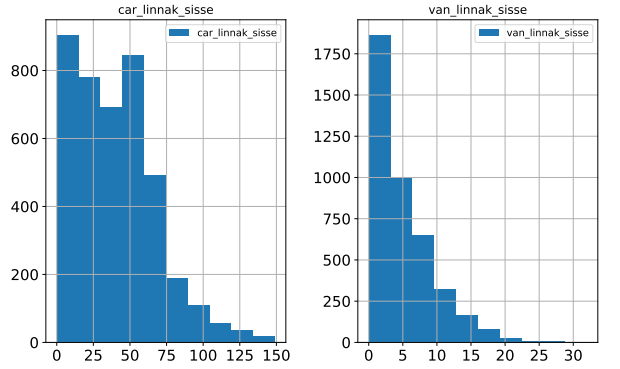
5.5.2 Traffic and Weather Data statistics

In this section, the histograms of both transformed and un-transformed traffic and weather data are presented. The transformation methods, such as: power, moving window, linear regression, differencing over power transformed time series were applied on traffic and weather data.

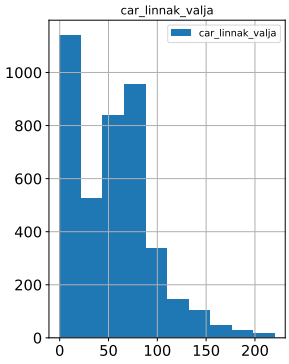
Traffic Data

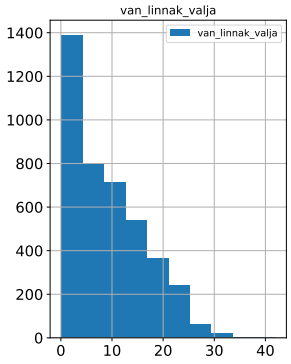
The traffic lanes were distinguished between the incoming and outgoing. The light traffic was considered for both directional lanes due to very small amount of data. There were some periods, when the vehicle counts were equal to zero. This behaviour is related to testing, adjustment of ML object detection model applied on video streams from camera. Besides, it can be connected to interruptions in the Internet connection. Considering this matter, it was decided to extract only valuable traffic data, meaning to exclude the periods with zero values. The histograms of the valuable data of cars, vans and light traffic are represented in Figure 22. Due to insufficient and heterogeneous distribution of the light traffic data, it was decided to exclude it from the analysis and modelling.

The histogram plots represented in Figure 22 demonstrate right-skewness (right-tail, positive skewness). Typical transformations could be, such as, power: square, cube root, moving window, i.e. rolling mean over a selected period, linear regression.

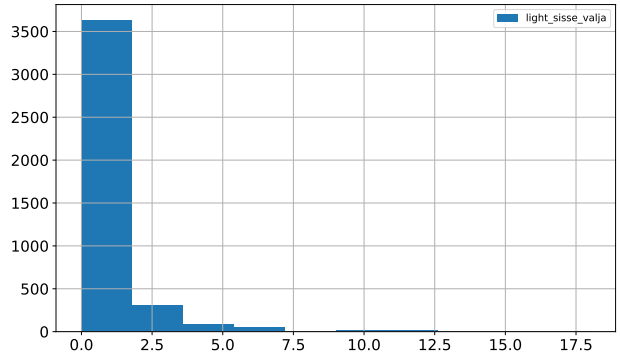


(a) Incoming cars (left) and vans (right).





(b) Outgoing cars (left) and vans (right).



(c) Incoming/Outgoing light traffic.

Figure 22. Histograms of valuable data of cars, vans and light traffic within the period from 03.2021–04.2022 on Tuesdays, Wednesdays and Thursdays between 6.00–19.00 (author created).

Weather Data

Weather phenomena data were given as categorical variables, for example, moderate snow (SN), rain (RN), Freezing Rain (FZRA), etc. It was decided to reduce all categories of weather phenomena to four main categories, such as: no phenomena (NP: 1), snow (SN: 2) , rain (RN: 3) and drizzle (DZ: 4). The histograms of weather properties are represented in Figure 23.

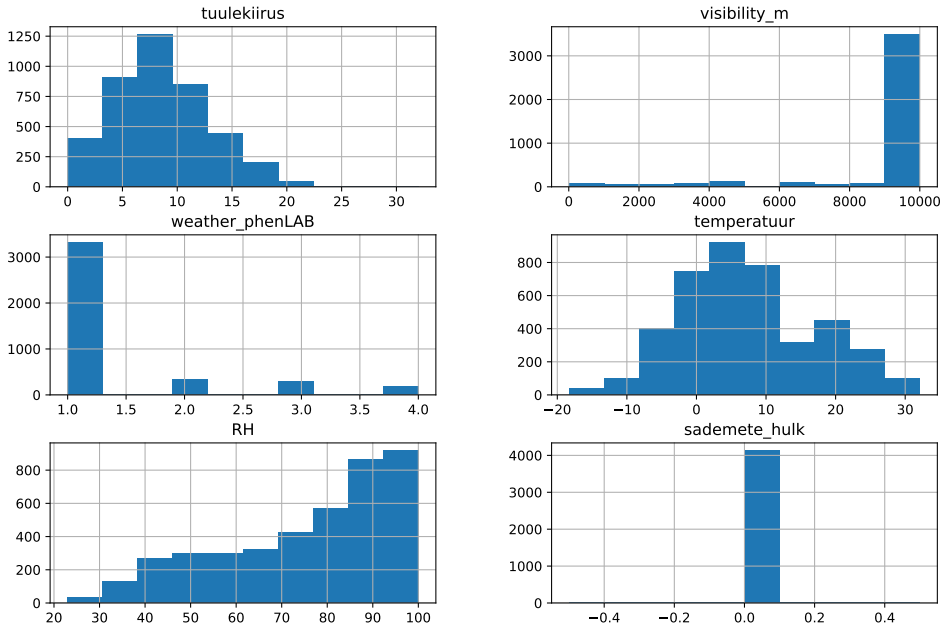


Figure 23. Histogram plots of weather properties (author created).

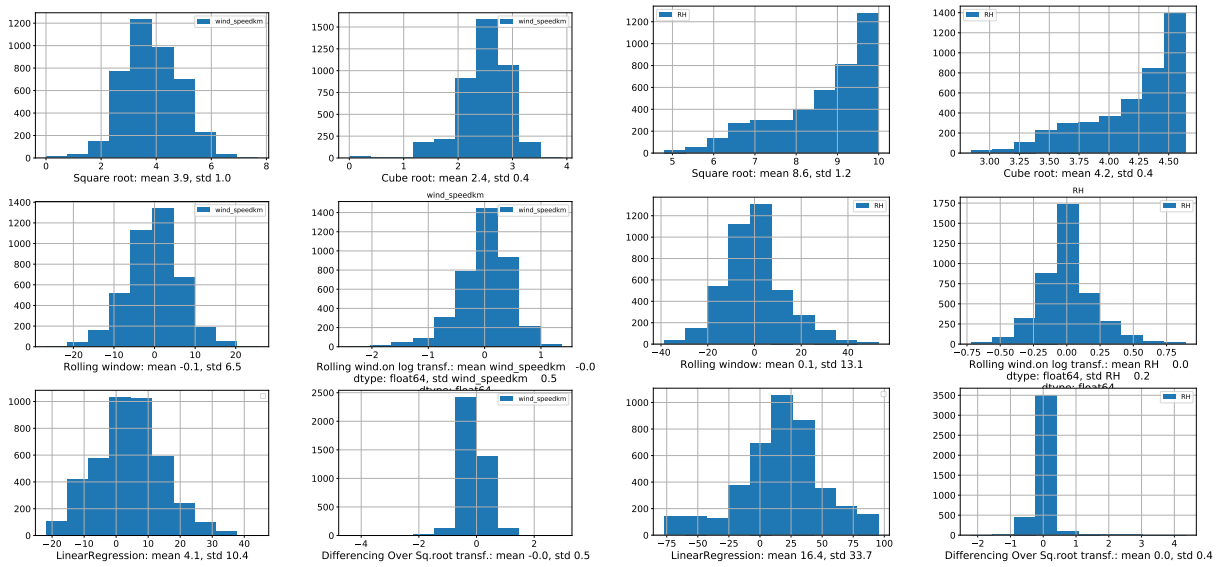
Generally, there are two popular ways of encoding the categorical variables, i.e. dummy or one-hot encoding. In case one categorical variable has n values then one-hot encoding converts it into n variables, while dummy encoding converts it into $n-1$ variables. In case of several categorical variables k each of which has n values, one-hot encoding will produce kn variables, while dummy encoding give $kn-k$ variables. A shortcoming of dummy and/or one-hot encodings is that they are expanding the feature/hypothesis space without adding much information, meaning that in case of many different categories a lot of additional columns/variables will be generated. There will be several columns with zero values and few of them with ones, and this introduces sparsity in the data-set. Another option for feature encoding is binary encoding. This encoding type works well, when there are many different categories. The feature is first converted into numerical value using an ordinal encoder and then transformed to binary number, which is split into different columns. A possible shortcoming of binary encoding is that it uses explicit parameter sharing between categories, which may complicate model training and performance [37]. The weather phenomena categories were binary encoded, such as:

- NP labelled as "1" -> 001;
- SN labelled as "2" -> 010;
- RN labelled as "3" -> 011;

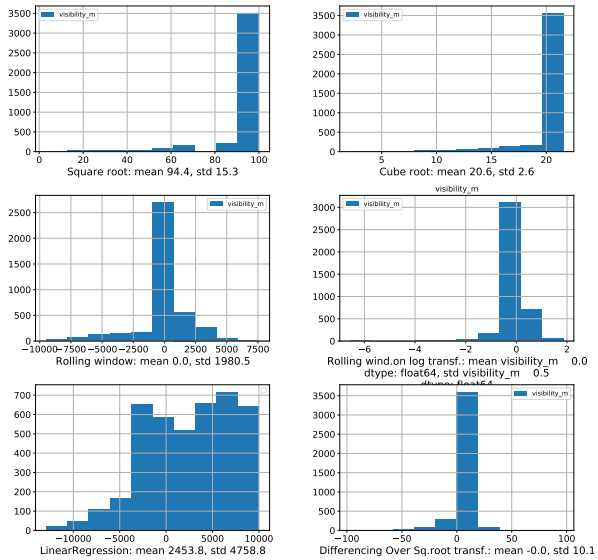
- DZ labelled as "4" -> 100.

Binary encoded weather phenomena was not transformed, since it was received from categorical data. The air temperature was used in its original form. Bi-modality in the air temperature distribution is its inherent property. The application of the transformation may distort this phenomenon.

Transformed wind speed data, RH and visibility by Power Transformations (PowSq, PowCu), Moving Window Function (RolWin), Linear Regression (LR) and Differencing Over Power Transformed (Pow-Shift) time series are represented in Figure 24.



(a) Wind speed. (b) RH.



(c) Visibility.

Figure 24. Transformed wind speed data, RH and visibility by Power Transformations (PowSq, PowCu), Moving Window Function (RolWin), Linear Regression (LR) and Differencing Over Power Transformed (Pow-Shift) time series (author created).

Correlation Analysis

Correlation coefficient indicates the strength of linear relationship. The coefficients are varying between -1 and 1, meaning that in case of 0 there is no relationship, in case of 1 or -1 there is a direct/strongest relationship. Strong positive correlation means that for every positive in-

crease in one variable, there is a positive increase of a fixed proportion in the other. Strong negative correlation means that for every positive increase in one variable, there is a negative decrease of a fixed proportion in the other. Thus, if the correlation coefficient between the target value and the variable is 0, then this most likely can mean that this variable will not provide any additional information to the model. The correlation heat matrix with the target values (lag=27[one day], lag=54[two days], lag=81[three days equiv. to one week]), lagged traffic data, original and transformed by linear regression weather data is given in Figure 25.

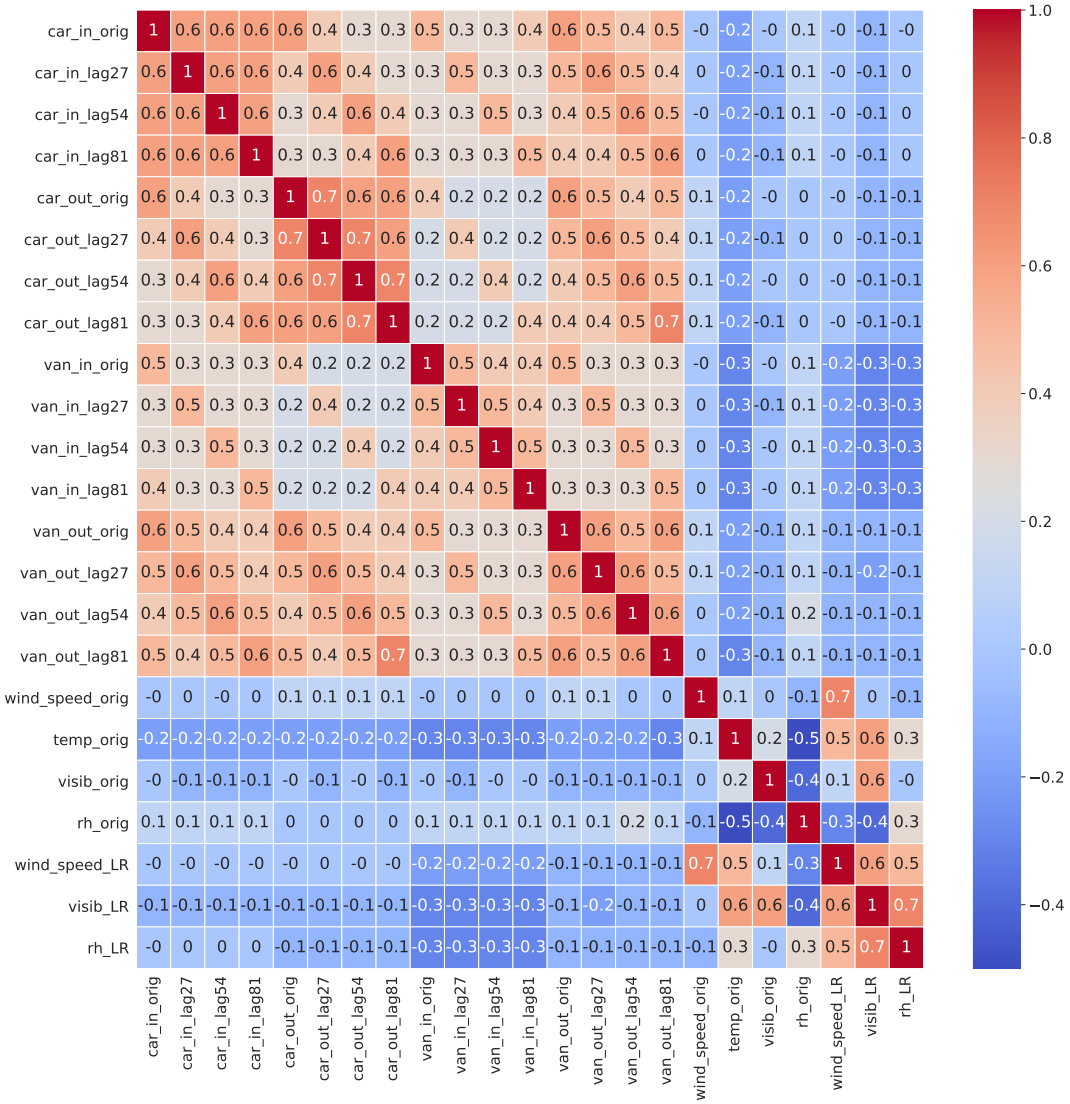


Figure 25. The correlation heat matrix with the target values, lagged traffic data, original and transformed by linear regression weather data (author created).

The correlations of the original and lagged traffic data were as follows:

- car-in-lag27 (one day lag), car-in-lag-54 (two days lag), car-in-lag81 (three days lag or one week lag): 0.6, 0.6, 0.6, respectively;
- van-in-lag27 (one day lag), van-in-lag-54 (two days lag), van-in-lag81 (three days lag or one week lag): 0.5, 0.4, 0.4, respectively;
- car-out-lag27 (one day lag), car-out-lag-54 (two days lag), car-out-lag81 (three days lag or one week lag): 0.7, 0.6, 0.6, respectively;
- van-out-lag27 (one day lag), van-out-lag-54 (two days lag), van-out-lag81 (three days lag or one week lag): 0.6, 0.6, 0.6, respectively.

From Figure 25 can be seen that some transformed variables have higher correlation coefficient compared to their original values. Air temperature was correlated with both original and transformed traffic data. The highest correlation was with the incoming vans, i.e. $\mathrm{cor} = -0.3$. The wind speed transformed by linear regression had higher correlation with the original incoming vans, i.e. $\mathrm{cor} = -0.2$. The same tendency was in the case of visibility and relative humidity transformed by linear regression. For visibilityLR the correlation with the original incoming vans was $\mathrm{cor} = -0.3$. For relative humidityLR the correlation with the original incoming vans was i.e. $\mathrm{cor} = -0.3$ The correlation between air temperature and original and transformed relative humidity was quite high, i.e. $\mathrm{cor} = -0.5$, meaning that these two variables are linearly dependent–collinear, and it is not reasonable to combine them together in the model. The correlation between air temperature and visibility was around $\mathrm{cor} = 0.2$, meaning that there may be some linear dependence between them. The lowest relation was between the air temperature and wind speed, i.e. $\mathrm{cor} = 0.1$.

It was decided to select one day lag, i.e. lag=27, since there was the highest correlation with outgoing cars, i.e. $\mathrm{cor} = 0.7$. Another reason of selecting the one day lag was related to EMHI weather forecast details, i.e. the model issuing the forecasts with a step of 1 hour only for upcoming 54 hours. It was decided to use weather phenomena instead of visibility and relative humidity, since it may give better estimation and/or the feeling of weather conditions outside, which is relevant when considering a person behavioural pattern, while choosing a mean of transportation.

The decision to include a variable to the model should not be reasoned solely on the values of the correlation coefficients. The correlation matrix presented gives linear correlations, although the relationships between the variables can also be non-linear. Thereby, a variable to be included to the model, on the top of the correlation coefficient, should be based on the nature of a physical phenomenon and its possible effect on the target variable. In the present work it

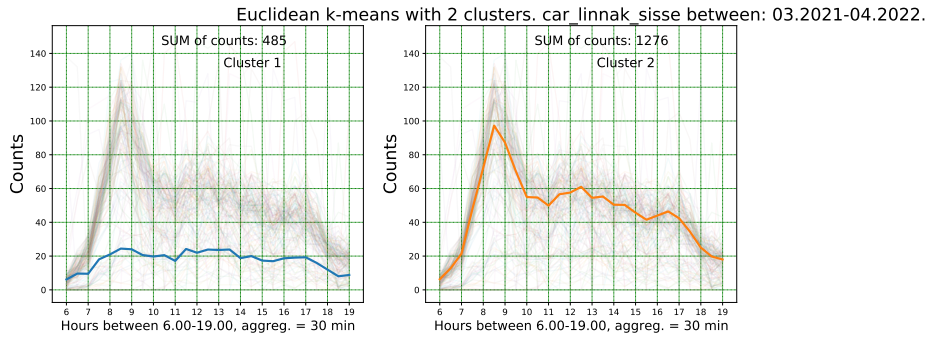
was assumed that weather phenomena, temperature and wind speed may most probably have an effect on the incoming/outgoing cars and vans.

Clustering analysis with k-means

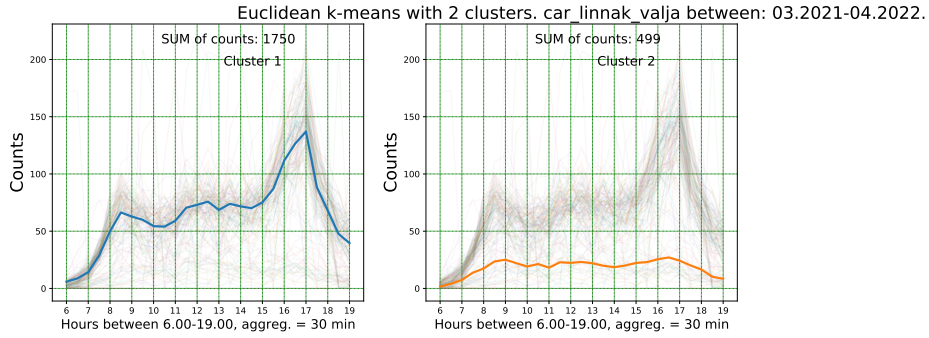
Clustering enables to estimate the average distribution of traffic flow for more distant time periods in future considering seasonal patterns, e.g grouping of traffic counts of certain days and weeks. The clustering analysis was applied on cars and vans data-sets. The clustering of incoming/outgoing cars and vans was made daily and weekly. The results of the clustering are given in Figures 26, 27. While implementing the clustering daily and weekly, it was possible to extract the certain days and weeks of the respective clusters, such as, for example, for incoming cars clustered weekly with $k = 2$ clusters the representation was:

Cluster 1: 11-Apr-2021, 18-Apr-2021, 25-Apr-2021, 02-May-2021, 09-May-2021, 16-May-2021, 23-May-2021, 30-May-2021, 06-Jun-2021, 13-Jun-2021, 04-Jul-2021, 11-Jul-2021, 26-Sep-2021, 03-Oct-2021, 10-Oct-2021, 17-Oct-2021, 24-Oct-2021, 31-Oct-2021, 07-Nov-2021, 14-Nov-2021, 21-Nov-2021, 28-Nov-2021, 05-Dec-2021, 12-Dec-2021, 19-Dec-2021, 26-Dec-2021, 02-Jan-2022, 09-Jan-2022, 16-Jan-2022, 23-Jan-2022, 30-Jan-2022, 13-Feb-2022, 20-Feb-2022, 20-Mar-2022, 27-Mar-2022, 03-Apr-2022, 10-Apr-2022, 17-Apr-2022

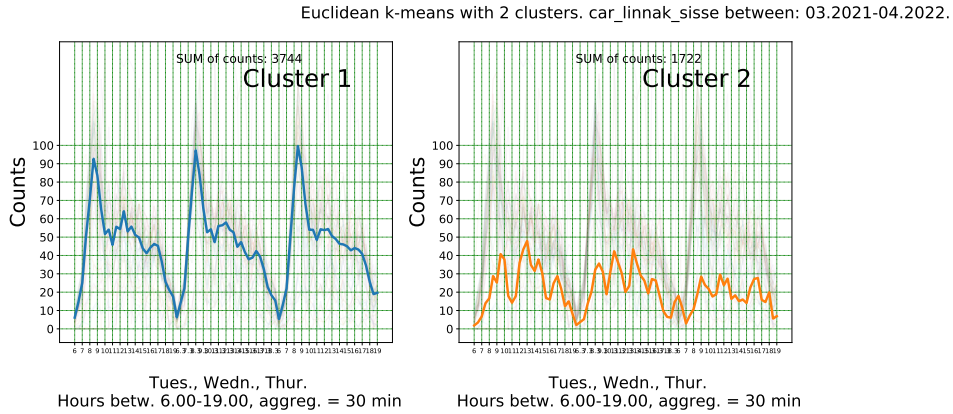
Cluster 2: 20-Jun-2021, 27-Jun-2021, 18-Jul-2021, 25-Jul-2021, 01-Aug-2021, 08-Aug-2021, 15-Aug-2021, 12-Sep-2021, 19-Sep-2021, 06-Feb-2022, 27-Feb-2022, 06-Mar-2022, 13-Mar-2022.



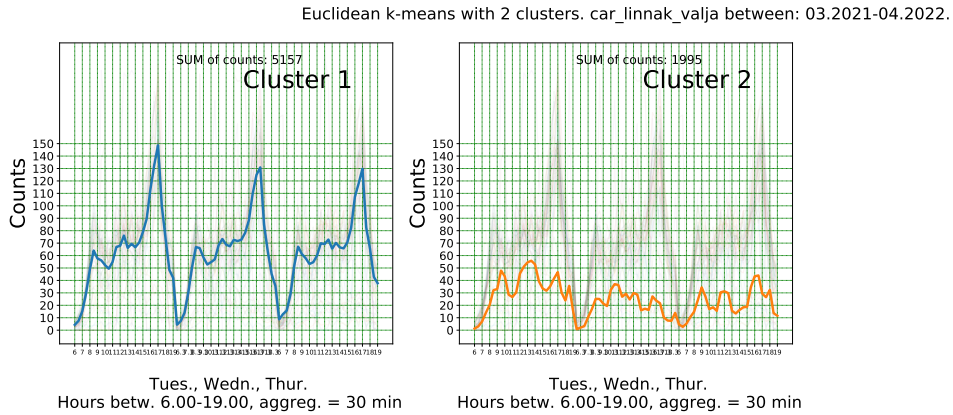
(a) Incoming cars daily.



(b) Outgoing cars daily.

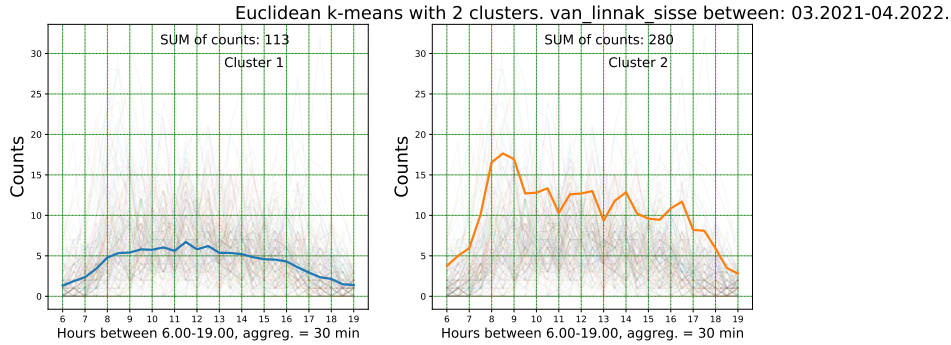


(c) Incoming cars weekly.

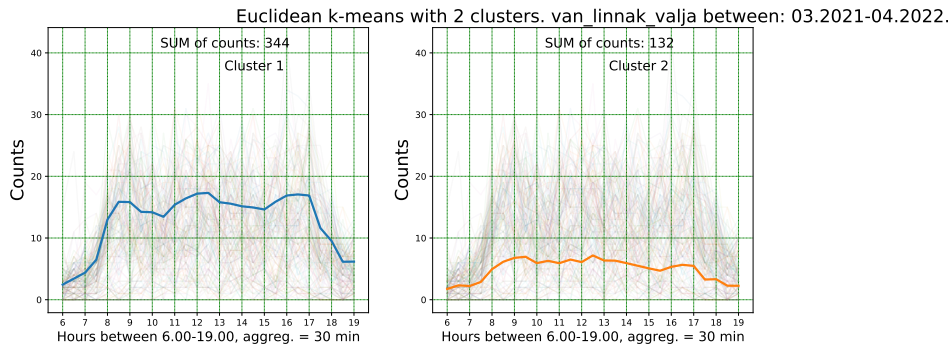


(d) Outgoing cars weekly.

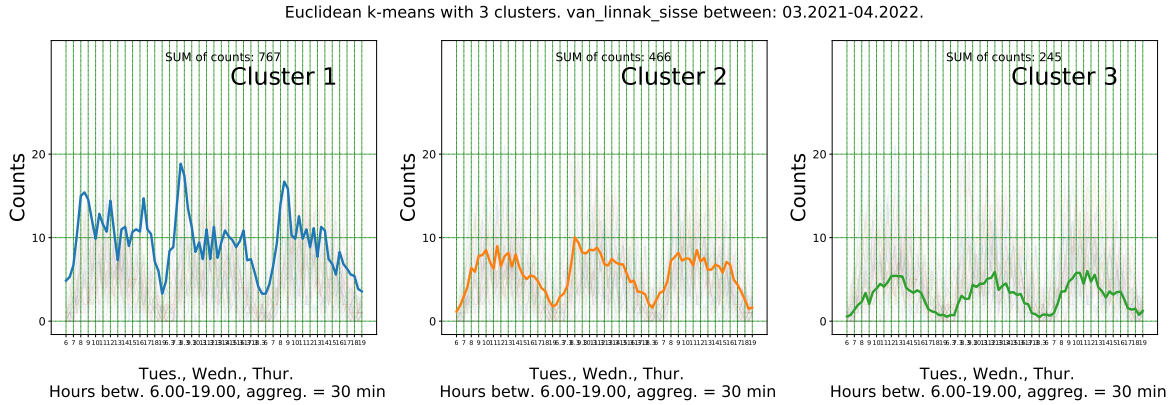
Figure 26. Clustering of incoming/outgoing cars daily and weekly (author created).



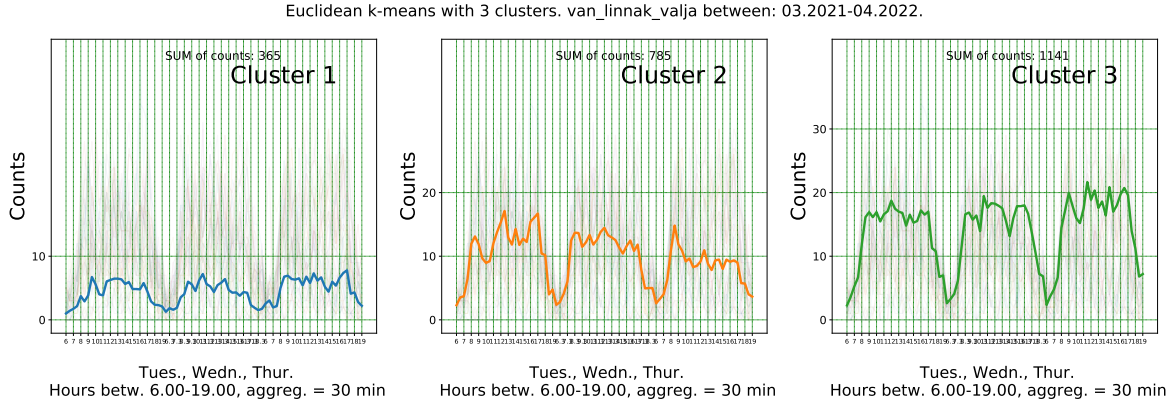
(a) Incoming vans daily.



(b) Outgoing vans daily.



(c) Incoming vans weekly.



(d) Outgoing vans weekly.

Figure 27. Clustering of incoming/outgoing vans daily and weekly (author created).

5.6 Models and Results

The feature vector of a data point included the histories of vehicles counts and weather observations. It was created by sliding window method with the lag=27, which was equivalent to one day. The lag of one week, i.e. lag=81, was omitted due to the lower correlation. The weather phenomena was not "dropped/lagged", since it was binary encoded. Thereby, it was added to the "dropped/lagged" traffic data by shifting it for one day, i.e. shift=27. The impact of each weather parameter on the performance of ML models developed was evaluated by adding it separately to the traffic data-sets. The addition of several weather parameters was based on their correlation, meaning that two or more parameters could be used if their correlation were low. The models developed in the present study were as following:

Basic (naive) model (BM) The forecast is based on the approach, such as, the present day counts of the cars that are available are considered and the same number of cars for the next day is forecasted;

ML0:LR Feature vector consisted of vehicles counts. The learning algorithm applied was LinearRegression;

ML0:HR Feature vector consisted of vehicles counts. The learning algorithm applied was LinearRegression with Huber Loss;

ML0:XG Feature vector consisted of vehicles counts. The learning algorithm applied was XGBoost;

ML1:LR Feature vector consisted of vehicles counts and transformed wind speed by linear regression. The learning algorithm applied was LinearRegression;

ML1:HR Feature vector consisted of vehicles counts and transformed wind speed by linear regression. The learning algorithm applied was LinearRegression with Huber Loss;

ML1:XG Feature vector consisted of vehicles counts and transformed wind speed by linear regression. The learning algorithm applied was XGBoost;

ML2:LR Feature vector consisted of vehicles counts and original air temperature. The learning algorithm applied was LinearRegression;

ML2:HR Feature vector consisted of vehicles counts and original air temperature. The learning algorithm applied was LinearRegression with Huber Loss;

ML2:XG Feature vector consisted of vehicles counts and original air temperature. The learning algorithm applied was XGBoost;

ML3:LR Feature vector consisted of vehicles counts and binary encoded weather phenomena. The learning algorithm applied was LinearRegression;

ML3:HR Feature vector consisted of vehicles counts and binary encoded weather phenomena. The learning algorithm applied was LinearRegression with Huber Loss;

ML3:XG Feature vector consisted of vehicles counts and binary encoded weather phenomena. The learning algorithm applied was XGBoost;

ML4:LR Feature vector consisted of vehicles counts and original air temperature + binary encoded weather phenomena. The learning algorithm applied was LinearRegression;

ML4:HR Feature vector consisted of vehicles counts and original air temperature + binary encoded weather phenomena. The learning algorithm applied was LinearRegression with Huber Loss;

ML4:XG Feature vector consisted of vehicles counts and original air temperature + binary encoded weather phenomena. The learning algorithm applied was XGBoost;

Each ML model was applied on traffic data-sets consisting of:

- Incoming cars;
- Outgoing cars;
- Incoming vans;
- Outgoing vans.

and the weather property(ies) selected. The feature vectors of a data point depending on the model are represented in Figure 28.

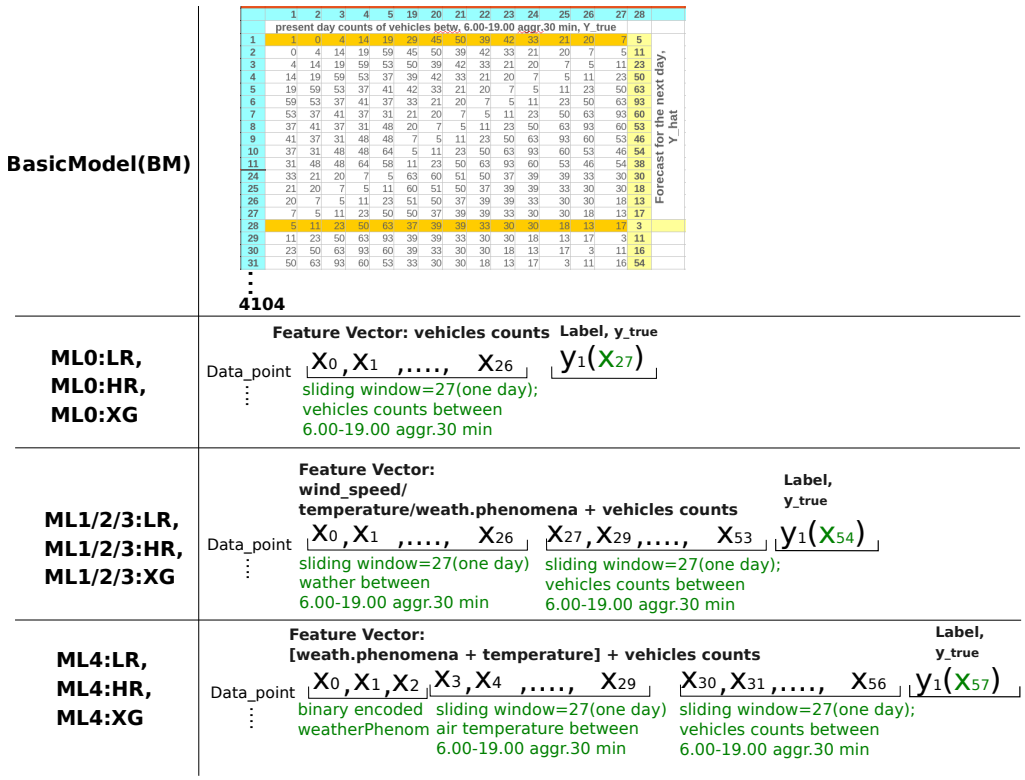


Figure 28. Feature vector of a data point depending on ML model developed (author created).

The performance of each model developed was evaluated by mean absolute error (MAE), which is an average absolute error defined as:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|. \tag{5.6}$$

MAE is the mean absolute difference between the actual and predicted value. The mean absolute error uses the same scale as the measured data. Mean absolute error is a common measure of forecast error in time series analysis. The diagrams representing the performance of BM, ML0/1/2/3/4 are given in Figure 29. The results summarising the MAEs of the described ML models are given in Table 9, Table 10, Table 11.

From the Tables 9, 10, 11 it can be noticed that ML0:XGB improved the accuracy of BM of incoming/outgoing cars by 31.34 and 42.14%, respectively, and outgoing vans by 36.18%. The exception was with incoming vans, where the improvement was 4.38%. ML0:LR and ML0:HR improved the accuracies of BM in a similar manner by varying between 24.73–

35.07 % in the case of incoming/outgoing cars. The accuracy improvement in the case of incoming vans varied between 7.5–8.44% and in the case of outgoing vans it varied between 36.62–38.82%.

From the viewpoint of weather properties, the LR model demonstrated a slight positive effect on all weather properties considered, and it varied around 1%. The exception was in the case of outgoing vans, where the positive effect varied between 3.11–3.81%. The HR model demonstrated the positive effect of wind speed, air temperature and weather phenomena on incoming/outgoing cars and it varied between 1.07–3.1%. The most pronounced positive effect was in the case of outgoing vans, where the improvement varied between 3.58–5.73%. XGBoost regressor demonstrated the positive effect of weather phenomena on incoming cars, i.e. 5.28%, and air temperature together with weather phenomena on outgoing vans, i.e. 3.44%.

Table 9. Linear regression (LR) MAEs of BM and ML0/1/2/3/4:LR and the effects of of ML0:LR compared to BM, as well as ML1/2/3/4:LR relative to ML0:LR. The positive effect demonstrates the improvement and the negative one degradation of MAE. Red colored values correspond to the positive effects higher than 3%.

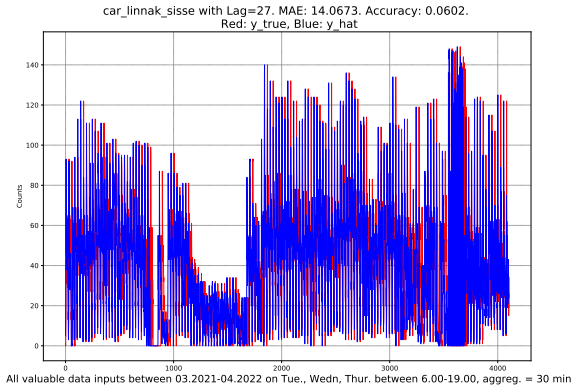
	BM	**ML0:LR**	eff.,%	ML1:LR	eff.,%	ML2:LR	eff.,%	ML3:LR	eff.,%	ML4:LR	eff.,%
car-in	14.07	10.59	24.73	10.62	-0.28	10.78	-1.79	10.47	1.13	10.85	-2.46
car-out	18.39	11.95	35.02	11.79	1.34	11.77	1.51	11.97	-0.17	11.88	0.59
van-in	3.20	2.96	7.5	3.01	-1.69	2.98	-0.68	2.95	0.34	2.98	-0.68
vanout	4.56	2.89	36.62	2.80	**3.11**	2.79	**3.46**	2.79	**3.46**	2.78	**3.81**

Table 10. Linear regression with Huber loss (HR) MAEs of BM and ML0/1/2/3/4:HR and the effects of ML0:HR compared to BM, as well as ML1/2/3/4:HR relative to ML0:HR. The positive effect demonstrates the improvement and the negative one degradation of MAE. Red colored values correspond to the positive effects higher than 3%.

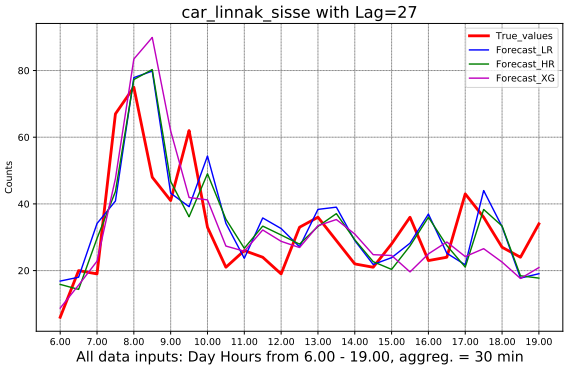
	BM	**ML0:HR**	eff.,%	ML1:HR	eff.,%	ML2:HR	eff.,%	ML3:HR	eff.,%	ML4:HR	eff.,%
car-in	14.07	10.29	26.87	10.18	1.07	10.31	-0.19	10.03	2.53	10.36	-0.68
car-out	18.39	11.94	35.07	11.69	2.09	11.57	**3.1**	11.79	1.26	11.62	2.68
van-in	3.20	2.93	8.44	2.94	-0.34	2.92	0.34	2.93	0.0	2.94	-0.34
vanout	4.56	2.79	38.82	2.63	**5.73**	2.67	**4.3**	2.69	3.58	2.66	**4.66**

Table 11. XGBoost (XG) MAEs of BM and ML0/1/2/3/4:XG and the effects of ML0:XG compared to BM, as well as ML1/2/3/4:XG relative to ML0:XG. The positive effect demonstrates the improvement and the negative one degradation of MAE. Red colored values correspond to the positive effects higher than 3%.

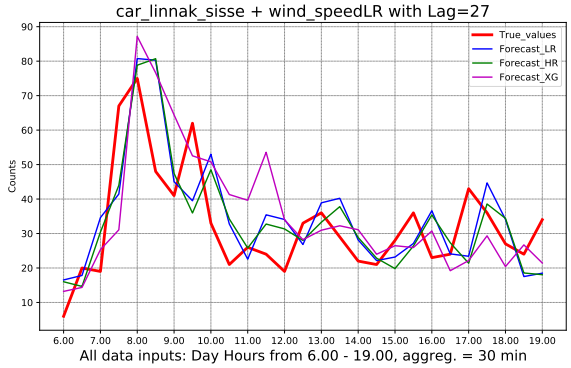
	BM	**ML0:XG**	eff.,%	ML1:XG	eff.,%	ML2:XG	eff.,%	ML3:XG	eff.,%	ML4:XG	eff.,%
car-in	14.07	9.66	31.34	10.97	-13.56	10.07	-4.24	9.15	**5.28**	9.67	-0.1
car-out	18.39	10.64	**42.14**	10.64	0.0	11.98	-12.59	10.58	0.56	10.91	-2.54
van-in	3.20	3.06	4.38	3.17	-3.59	3.21	-4.9	3.24	-5.88	3.05	0.33
vanout	4.56	2.91	36.18	2.98	-2.41	2.98	-2.41	3.24	-11.34	2.81	**3.44**



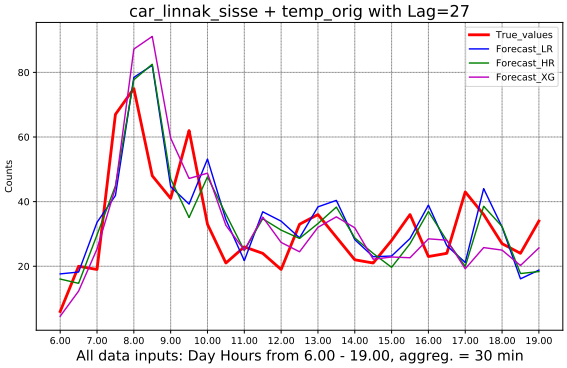
(a) BM: Incoming cars.



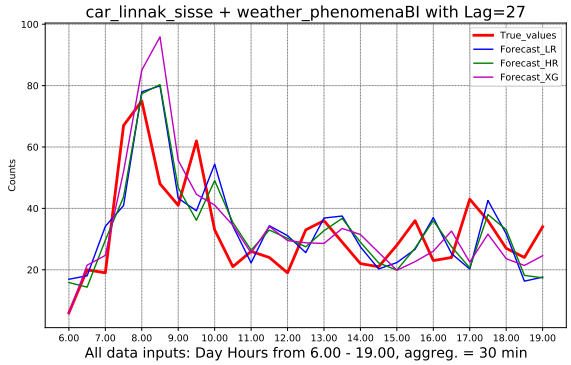
(b) ML0: Incoming cars.



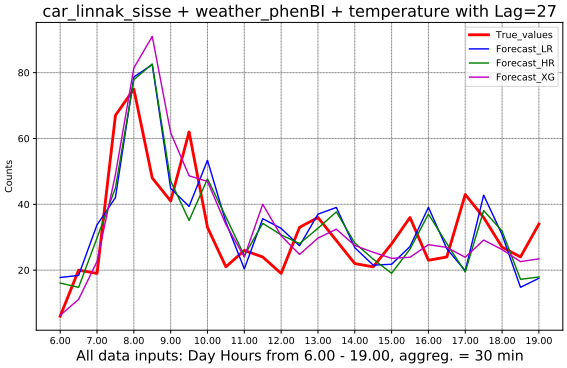
(c) ML1: Incoming cars.



(d) ML2: Incoming cars.



(e) ML3: Incoming cars.



(f) ML4: Incoming cars.

Figure 29. Representation of the performance of the models for incoming cars: BM, ML0, ML1, ML2, ML3 and ML4, Figure 28 (author created).

5.7 Importance of traffic counts and weather features.

The feature importance was evaluated in XGBoost model employing:

- 'Gain' importance type;
- SHAP based importance.

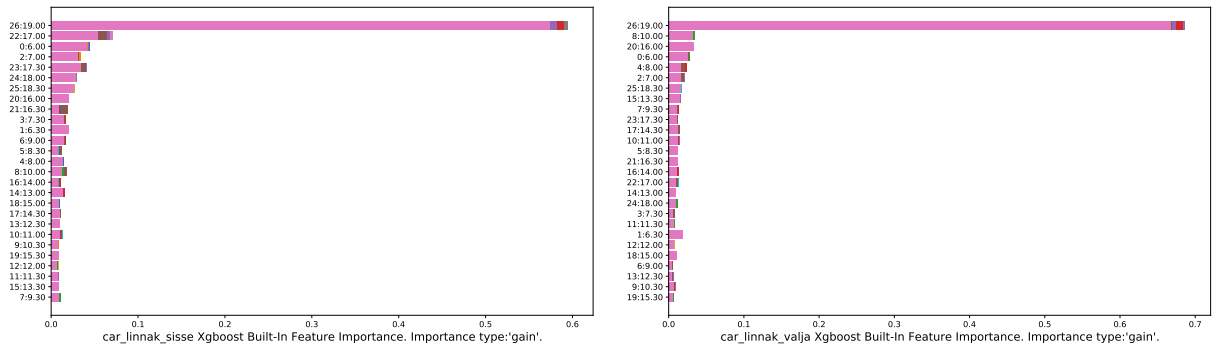
The 'Gain'importance type uses the average gain across all splits the feature is used in. Typically, as a rule of thumb 'Gain' importance type is applied, as it is more representative, meaning that usually there is no interest in the mere appearance of feature-specific splits, but how much those splits helped. 'Gain' refers to the relative contribution of the corresponding feature to the model, calculated by taking into account the contribution of each feature for each tree in the model. "Gain" means the improvement in accuracy that a feature brings to the branches it resides on. A higher value of 'Gain' compared to another feature means that it was more important for generating a prediction. The 'Gain' is the most relevant attribute to interpret the relative importance of each feature.

SHAP (SHapley Additive exPlanations) method splits the forecast into the parts to reveal the meaning of each feature. It is based on the Shapley vector [38], which originates from game theory. Shapley vector defines the principle of optimal distribution of gain between the players in the theory of cooperative games. It represents a distribution, where the gain of each player is equal to his average contribution to the welfare of the total coalition. Thereby, using the Shapley value, all possible combinations and options are revealed and the features, which are relevant, can be identified. SHAP can be utilised, where it is necessary to cluster big data and find relationships between groups of features.

The evaluation of feature importance of incoming/outgoing cars and vans in XGBoost model employing 'Gain' importance type are given in Figure 30, where the time periods between 6.00–19.00 with the aggregation of 30 min are indicated as (feature)0: 6.00, (feature)1: 6.30, (feature)3: 7.00 and so on. The evaluation of feature importance by SHAP for incoming cars and vans are given in Figure 31.

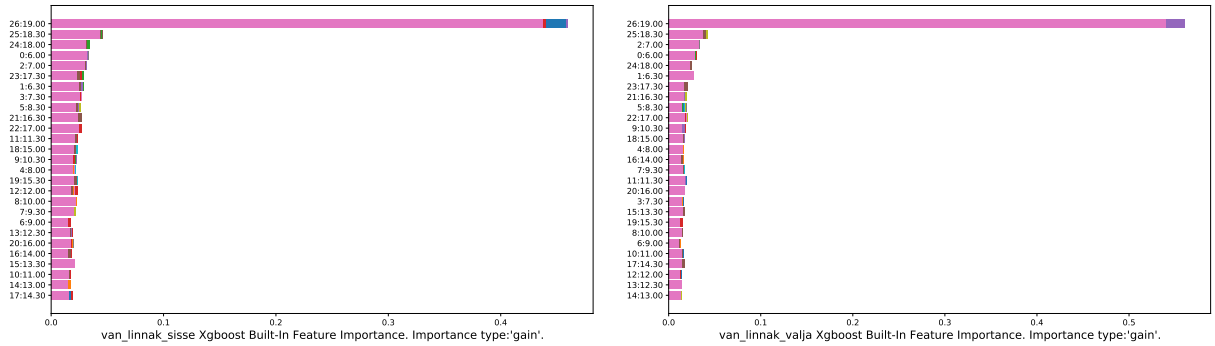
The importance of incoming cars and vans

In Figures 30a,b first four features correspond to daytime hours at 19.00, 17.00, 6.00, 7.00 for incoming cars, and the daytime hours at 19.00, 10.00, 16.00, 6.00 for outgoing cars. In Figures 30c,d first four features correspond to the daytime hours at 19.00, 18.30, 18.00, 6.00 for incoming vans, and the daytimes at 19.00, 18.30, 7.00, 6.00 for outgoing vans.



(a) Incoming cars.

(b) Outgoing cars.

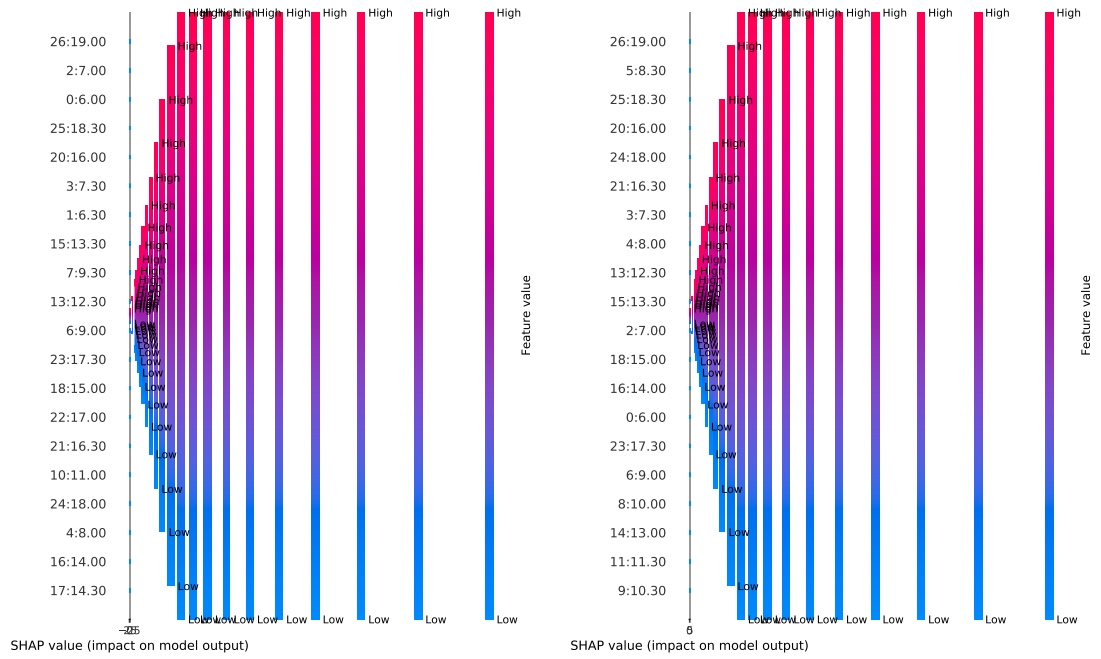


(c) Incoming vans.

(d) Outgoing vans.

Figure 30. Feature importance of incoming/outgoing cars and vans in XGBoost model employing 'Gain' importance type (author created).

As it can be seen from the Figure 31, the distributions of the feature importances estimated by SHAP are slightly different compared to the ones estimated by 'Gain' importance type. When considering the features with high importance, it can be noticed that in view of incoming cars the daytime hours from 6.00–7.30, at 9.30, 12.30, 13.30, 16.00 were relevant. In the case of incoming vans the relevant daytime hours appeared from 7.30–8.30, 12.30, 13.30, 16.00, 16.30 are relevant.



(a) Incoming cars.

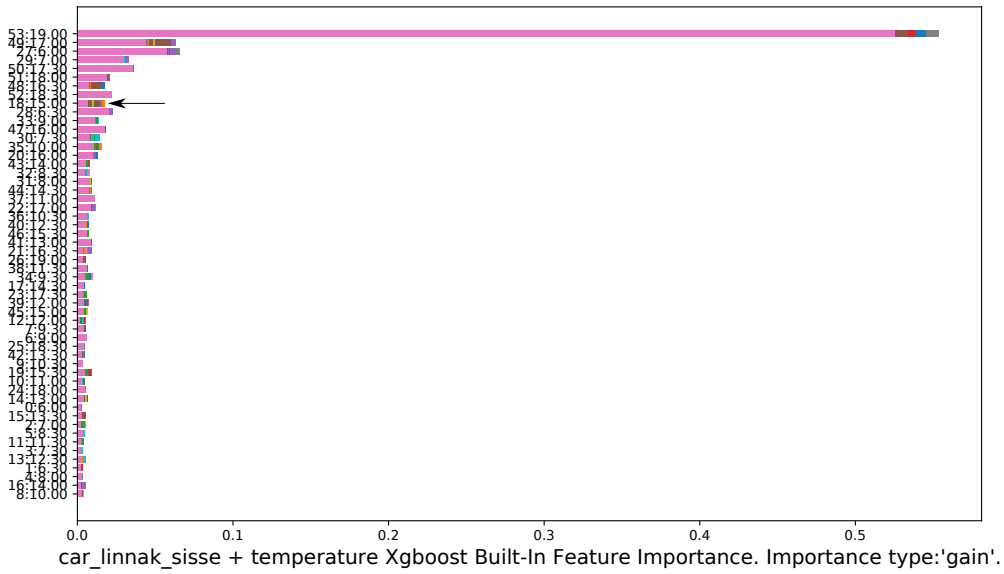
(b) Incoming vans.

Figure 31. Feature importance by SHAP in ML0/1/2/3:XG for incoming cars and vans, Figure 28 (author created).

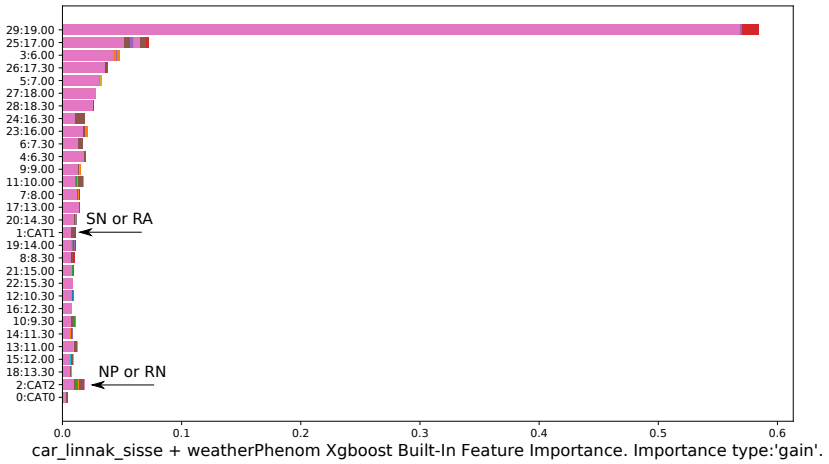
The effect of importance of weather properties

The effect of selected weather parameters on model performance was also evaluated in XGBoost model by 'Gain' importance type and SHAP.

The features in Figures 32 can be read so that from the feature 0 to feature 26 there are weather properties and from feature 27 to feature 53 there is a traffic counts history, see also Figure 28. Both, the weather parameters and the history of traffic counts used correspond to time periods between 6.00–19.00 with the aggregation of 30 minutes. In case of weather phenomena the features 0, 1, 2 correspond to binary encoded categories and starting from feature 3 to feature 29 there is a traffic counts history. The evaluation of traffic counts history combined with wind speed, air temperature and weather phenomena in XGBoost model by 'Gain' importance type and SHAP are represented in Figures 32, 33 and Figures 34, 35, respectively.



(a) Incoming cars + temperature.



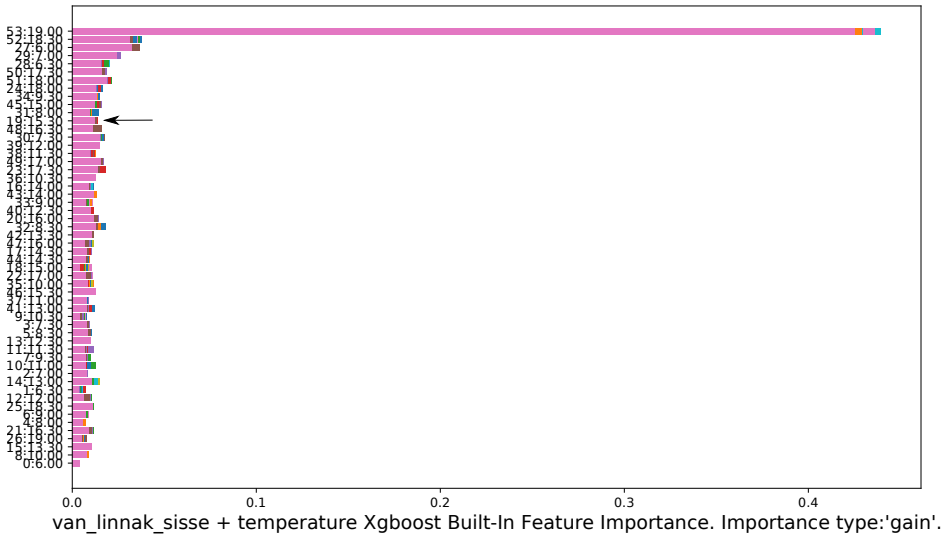
(b) Incoming cars + weather phenomena.

Figure 32. Incoming cars. The evaluation of importance of traffic counts history combined with air temperature and weather phenomena in XGBoost model by 'Gain' importance type (author created).

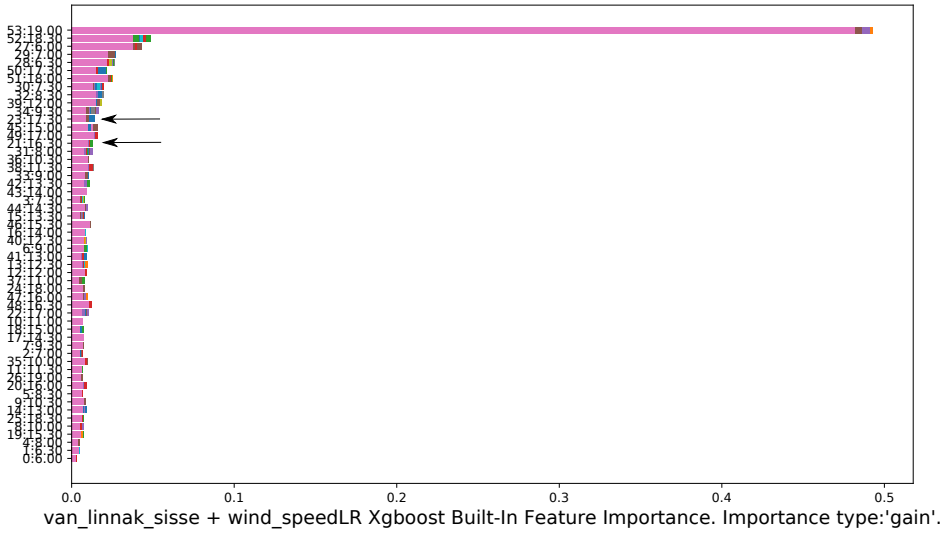
'Gain' importance type

In Figure 32a the first four features correspond to car counts at 19.00, 17.00, 6.00, 7.00 for incoming cars. The air temperature at 15.00 was within the first 9 features for incoming cars. In Figure 32b the first four features correspond to incoming car counts at 19.00, 17.00, 6.00, 17.30. CAT1 (SN or RN) was within the first 17 features. The wind speed was not within the

first 15 features meaning that it was not relevant in the case of incoming cars.



(a) Incoming vans + temperature.



(b) Incoming vans + wind.

Figure 33. Incoming vans. The evaluation of importance of traffic counts history combined with air temperature and wind speed in XGBoost model by 'Gain' importance type (author created).

In Figure 33a first four features correspond to incoming van counts at 19.00, 18.30, 6.00, 7.00. In Figure 33b the first four features correspond to van counts at 19.00, 18.30, 6.00, 7.00. The air temperature at 15.30 for incoming and at 16.00, 14.30, 16.30 for outgoing vans was within

the first 14 features. Wind speed at 17.30 and 16.30 was within the first 15 features in the case of incoming vans.

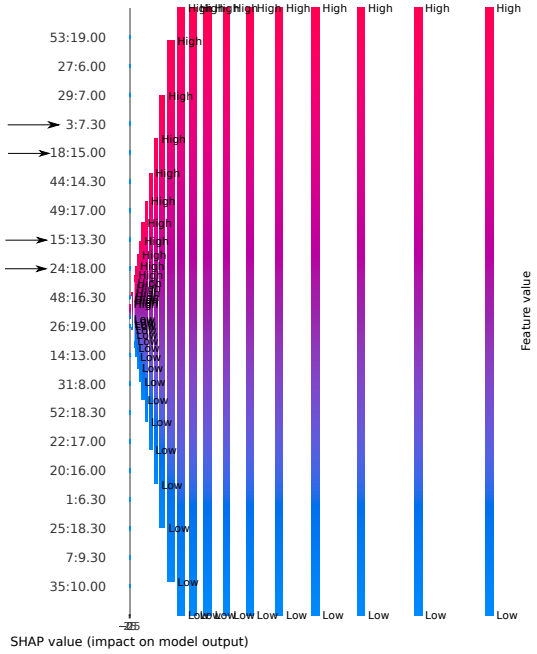
SHAP importance type
Interestingly, the SHAP values on the high-importance side in Figure 34a calculated using the traffic and wind speed data-set included several wind speed features. The air temperature at 9.00 and 8.00 was relevant for incoming cars, while weather phenomena features were not present on the high-importance side of the diagram, Figures 34c. The weather phenomena CAT1 (RA or SN) was close to relevant feature values for incoming cars Figure 34c.

In Figure 35a, the wind speed was within the high importance features, Figure 35b, while weather phenomena was not present within the relevant ones, Figure 35c. The air temperature at 12.30 and 17.00 was relevant for incoming vans, Figures 34b.

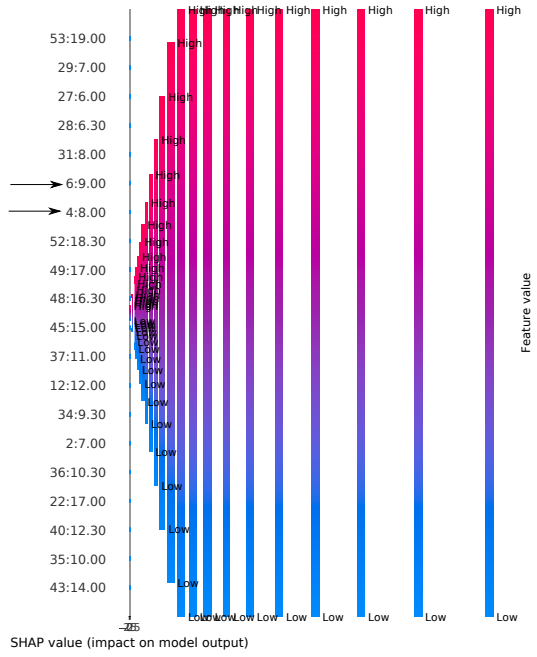
Summarising feature importances
SHAP evaluation method distinguished inherent daytime hours more explicitly compared to importance 'Gain'. Most probably, the daytime hours at 6.00 and 19.00 were highlighted as relevant, since they specified the beginning and end of the next day.

SHAP evaluation demonstrated the relevant daytime hours for incoming cars from 6.00–7.30, at 9.30, 12.30, 13.30 and at 16.00. In the case of incoming vans the relevant daytime hours appeared from 7.30–8.30, at 12.30 and after 16.00. Based on the 'Gain' importance type, the most relevant daytime hours in the view of incoming cars were at 7.00, 7.30, 9.00, 10.00, 13.00 and after 16.00 in the evening. For the outgoing cars the most significant daytime hours were at 7.00, 8.00, 10.00, 13.30 and after 16.00. In the case of incoming/outgoing vans the relevant daytime hours appeared before 8.30 and after 15.00.

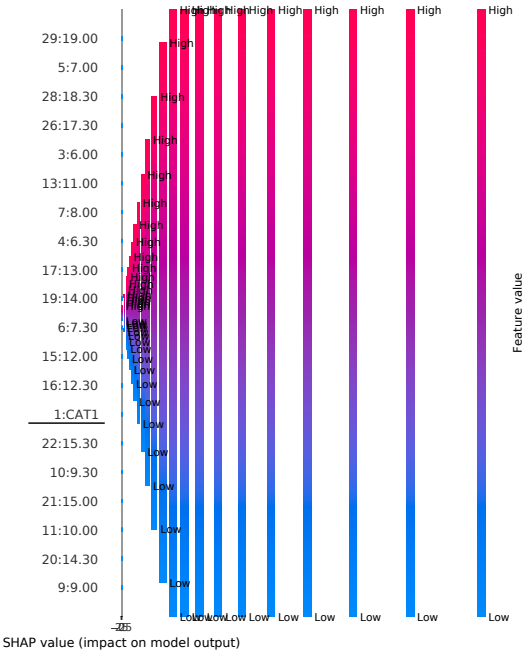
To sum up common results received by both feature importance evaluations considering weather properties, the air temperature and weather phenomena had a moderate effect on incoming/outgoing car forecasts, while in the case of incoming/outgoing vans the effect of air temperature and weather phenomena was more pronounced. For incoming/outgoing cars the CAT1 (SN or RN) had a higher effect based on both 'Gain' and SHAP evaluations. For incoming/outgoing vans the CAT2 (NP or RN) had an effect based on 'Gain' importance type.



(a) Incoming cars + wind speed.

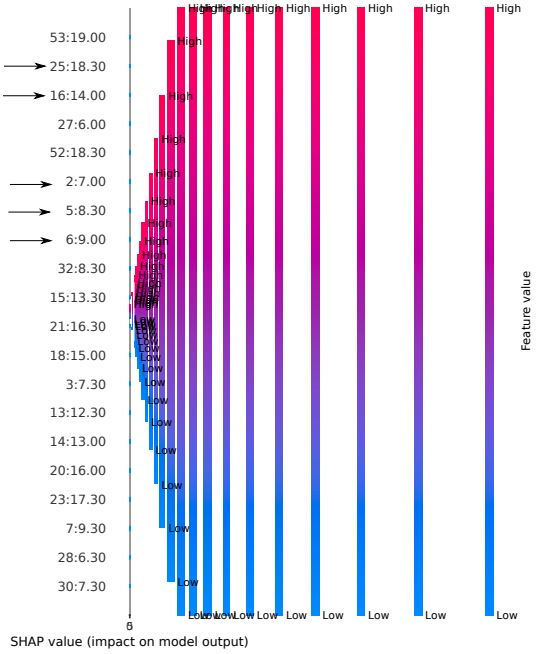


(b) Incoming cars + air temperature.

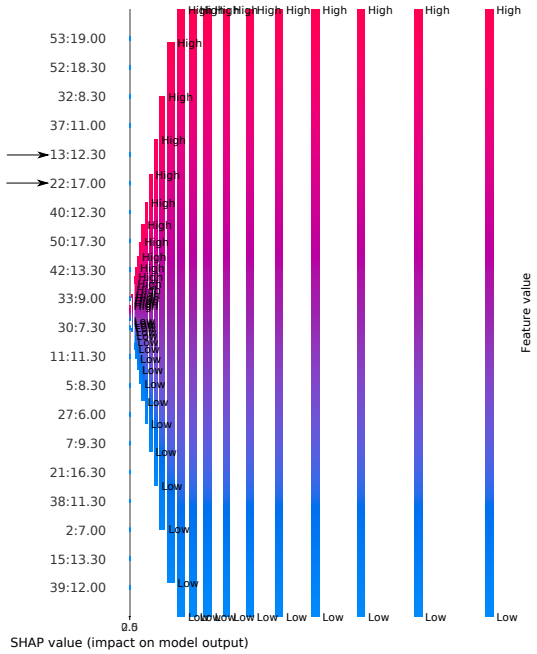


(c) Incoming cars + weather phenomena.

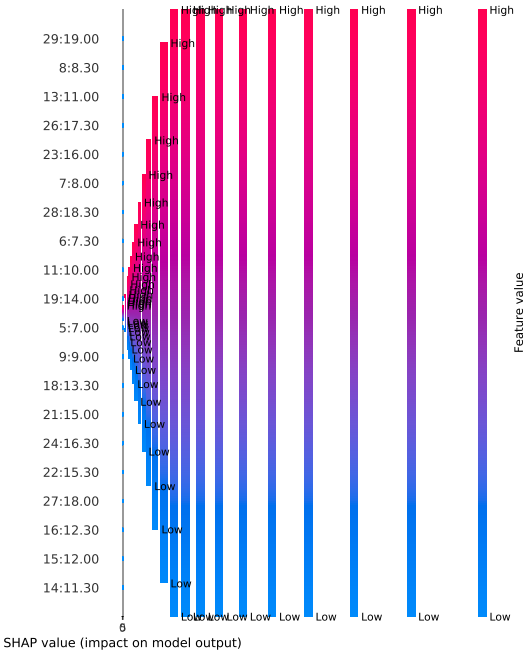
Figure 34. Feature importance by SHAP in ML0/1/2/3:XG for incoming cars, Figure 28 (author created).



(a) Incoming vans + wind speed.



(b) Incoming vans + air temperature.



(c) Incoming vans + weather phenomena.

Figure 35. Feature importance by SHAP in ML0/1/2/3:XG for incoming vans, Figure 28 (author created).

5.8 Evaluation and Limitations of results

In the future, it is planned to automate the Data Analyst tasks, meaning that the query, pre-processing, analysis and building of machine learning models will be implemented as an additional module of the system developed. Further development and addition of non-functional requirements are planned as a future step and will be implemented through the user experience research.

The counts of vehicles calculated by object detection ML model were the input for the ML models developed in this study. The accuracy of generated data-sets depends on the technical parameters, such as, height, viewing angle, stability of a video-camera, as well as on the constant presence of an Internet connection, since the ML object detection model is applied on live video-streams. The factors described can be the reasons for the appearance of zero values in the data-sets. In this study, such constantly appearing zero values, most likely associated with the lack of an Internet connection, were excluded. However, possible single zero values could still remain in the data. This could lead to the fact that during the training, the models could learn the misleading patterns in the data.

Another constraint can be related to the limited sample size. Valuable data used in the models corresponded to 153 days, which still is not enough to exclude possible over-fitting given the length of the feature vectors.

Summary

Ülemiste City (ÜC) strives to develop an environment that allows both rapid economic growth and development of companies and start-ups, as well as the sustainable and green environment within the Ülemiste City area. The present study focused on the development of information system providing the data-driven parking service at ÜC parking lots. The information system created enables the optimisation of dynamic parking prices based on the management and analysis of traffic flow data in ÜC. The target group of the service developed is one-time/random visitors of ÜC. The parking pricing optimised by traffic data analysis will not affect the customers having monthly contract of parking, thus allowing to have control and management of regular campus customers. The aim of developing the parking service with optimised prices was to increase the profit from parking service by 5-10%, as well as to decrease the motorisation level by 8-10% in the ÜC area through the management and control of the amounts of random (one-time) visitors of ÜC during the daytime hours. The current daytime parking at ÜC parking lots is for free or has a negligible cost, which significantly increases the number of actively moving vehicles within the ÜC area. The infosystem developed needs minimal investments, which can be considered as an advantage. The system employs ÜC traffic flow data management and analysis capabilities, which are the inputs for financial and business analytics, supporting the development and optimisation of dynamic parking prices. Traffic flow data management and analysis are used as the methods implementing the Use Case (UC) UC8 and UC9.

The study also focuses on the development of machine learning (ML) models forecasting the traffic flow–amount of cars and vans incoming or outgoing the Ülemiste City (ÜC) through the Suur-Sõjamäe-Lõõtsa intersection on Tuesday, Wednesday and Thursdays between the working hours from 6.00–19.00 with the 30 minutes granularity. In the future, it is planned to automate the Data Analyst tasks, meaning that the query, pre-processing, analysis and building of machine learning models will be implemented as an additional module of the system developed.

During the work the following outcomes were received:

- Ülemiste City generalised strategic goals were highlighted. Objectives of mobility program of ÜC were defined. ÜC Balanced Scorecard (BSC) Project representing the strategic development areas with sub-domains and priority programs were developed. ÜC Strategic View and Business Architecture were created.
- Detailed system analysis of the parking service with dynamic and optimised prices was implemented, i.e. evaluation and shortcomings of AS-IS process, strategic model of the service, value stream and capability analysis were implemented, business information model and use case diagram were created. The architectural vision of the information system providing the parking service with dynamic and optimised prices was developed;
- Expected effect of the implementation of dynamic and optimised parking at ÜC parking lots during one year will be:
  1. Increase in profit of 5-10% from parking service;
  2. Decrease in the number of cars during the working hours by 8-10%;
  3. Decrease in the level of CO2 emission from the transportation by 300-320 t.
- Cluster analysis implemented gave an estimate of the average distribution of the traffic flow, and also revealed possible seasonal patterns;
- The Basic (naive) model (BM) developed can forecast the incoming/outgoing car and van counts with an absolute error not exceeding 10%; Machine learning developed improved the accuracy of BM by 26–29% in average. XGBoost improved the accuracy of BM for incoming/outgoing cars the most;
- The inclusion of weather parameters resulted in the accuracy improvement of the ML models within the range from 1 to 6% compared to the ML models including the traffic counts only. The air temperature and weather phenomena demonstrated higher relevance compared to wind speed; XGBoost demonstrated the effect of weather phenomena (snow or rain) for incoming cars. LinearRegression and LinearRegression with Huber loss demonstrated the effect of wind speed, air temperature, weather phenomena (no phen. or rain) for outgoing vans.
- Feature importance evaluation by 'Gain' importance type and SHAPE highlighted the relevant daytime hours, which contributed to the optimisation of dynamic parking prices.

The traffic data were generated by external ML object detection model applied on videos streamed through the cameras adjusted on the road intersections of the ÜC. The data on weather properties were received from EMHI Tallinn Airport measuring station. Weather variables, which assumed to correlate with the traffic flows were: wind speed, air temperature and weather phenomena (snow, rain, drizzle). The data used for building the models included historical traffic and the respective weather data of 153 days. The models tested in this work

included the Basic (naive) model (BM) and three ML models, such as: Linear Regression (LR), Linear Regression with Huber Loss (HR) and Extreme Gradient Boosting (XGB).

Based on the study findings the following useful aspects can also be highlighted:

- Feature importance evaluation From the viewpoint of data-driven parking service at ÜC parking lots:
  1. The relevant daytime hours based on the feature importance evaluation:
     – Incoming cars: until 7.30, at 9.30, 12.30, 13.30 and at 16.00;
     – Incoming vans: 7.30–8.30, at 12.30 and after 16.00;
     – Outgoing cars: 7.00, 8.00, 10.00, 13.30 and after 16.00;
     – Outgoing vans: 6.00–7.00, after 18.00, 19.00.
  2. The traffic flow amounts forecasted will be employed to evaluate the CO2 emission from the transportation within the ÜC area necessary for GHG reporting [5];
  3. ML models built are ready to be run using an integrated Jupyter Notebook on Machine Learning Workbench (MLW), which is a micro-service application on Cumulocity IoT platform used by ÜC.
- From the viewpoint of traffic flow data analysis:
  1. Cluster analysis implemented identified uncorrelated groups in traffic counts, enabling to estimate the average distribution of traffic flow considering seasonal patterns;
  2. The Basic (naive) model (BM) developed can forecast the incoming/outgoing car and van counts with an absolute error not exceeding 10%;
  3. ML models developed improved the accuracy of BM by 26–29% in average. XGBoost improved the accuracy of BM for incoming/outgoing cars the most;
  4. The inclusion of weather parameters resulted in the accuracy improvement of the ML models within the range from 1 to 6% compared to the ML models including the traffic counts only. The air temperature and weather phenomena demonstrated higher relevance compared to wind speed;
  5. XGBoost demonstrated the effect of weather phenomena (snow or rain) for incoming cars. LR and HR demonstrated the effect of wind speed, air temperature, weather phenomena (no phen. or rain) for outgoing vans.

Based on the listed main results the problems of the Thesis were solved and the goals were achieved.

The Author of this work (M.E.) has made the queries to time series database to receive the traffic counts in the ÜC. The Author of this work made pre-processing of the traffic data, tested the models, and composed all the necessary scripts in `R` and `Python` programming languages.

Bibliography

[1] S. C. K. Tekouabou, E. B. Diop, R. Azmi, R. Jaligot, and J. Chenal. Reviewing the application of machine learning methods to model urban form indicators in planning decision support systems: Potential, issues and challenges. *Journal of King Saud University - Computer and Information Sciences*, 2021. URL https://www.sciencedirect.com/science/article/pii/S131915782100210X.

[2] M. Laos. *Mainori lugu*. Mainor AS, 2014.

[3] Ülemiste City, Mainor AS. Ülemiste city radar, 2023.

[4] European Comission. EU taxonomy for sustainable activities, 2020. URL https://ec.europa.eu/info/business-economy-euro/banking-and-finance/sustainable-finance/eu-taxonomy-sustainable-activities_en. [Online; accessed 29-July-2022].

[5] GHG Protocol. A corporate accounting and reporting standard, 2004. URL https://ghgprotocol.org/sites/default/files/standards/ghg-protocol-revised.pdf. [Online; accessed 29-July-2022].

[6] F. G. Mohammadi, F. Shenavarmasouleh, M. H. Amini, and H. R. Arabnia. *Data Analytics for Smart Cities: Challenges and Promises*, chapter 2, pages 13–27. John Wiley & Sons, Ltd, 2022. URL https://doi.org/10.1002/9781119748342.ch2.

[7] A. Jung. *Machine Learning: The Basics*. Machine Learning: Foundations, Methodologies, and Applications. Springer Singapore, 2022. URL https://books.google.ee/books?id=1IBaEAAAQBAJ.

[8] Z. Liu, Y. Liu, Q. Meng, and Q. Cheng. A tailored machine learning approach for urban transport network flow estimation. *Transportation Research Part C: Emerging Technologies*, 108:130–150, 2019. URL https://doi.org/10.1016/j.trc.2019.09.006.

[9] M. I. Jordan and T. M. Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015. URL https://www.science.org/doi/abs/10.1126/science.aaa8415.

[10] P.-L. Hsu. Machine learning-based data-driven traffic flow estimation from mobile data.

Master's thesis, 2021. URL https://www.diva-portal.org/smash/get/diva2:1590038/FULLTEXT01.pdf.

[11] The Open Group. The open group architecture framework, 2009.

[12] Business Architecture Guild. A guide to the business architecture body of knowledge (bizbok guide), 2016.

[13] The Open Group. Archimate® 3.2 specification, 2022.

[14] Wikipedia contributors. Representational state transfer — Wikipedia, the free encyclopedia, 2022. URL https://en.wikipedia.org/w/index.php?title=Representational_state_transfer&oldid=1124435026. [Online; accessed 2-December-2022].

[15] Ülemiste City. Ülemiste city parking, 2023.

[16] Fyma. Fyma ai video object recognition, 2023.

[17] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. URL https://scikit-learn.org/stable/auto_examples/applications/plot_cyclical_feature_engineering.html.

[18] W. McKinney. Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 56 – 61, 2010. URL https://pandas.pydata.org/pandas-docs/stable/user_guide/timeseries.html#time-date-components.

[19] M. Sinimaa. Modeling and short-term electricity demand forecasting:estonia case study. Master's thesis, 2020.

[20] J. Brownlee. *Deep Learning for Time Series Forecasting: Predict the Future with MLPs, CNNs and LSTMs in Python*. Machine Learning Mastery, 2018. URL https://books.google.ee/books?id=o5qnDwAAQBAJ.

[21] S. Makridakis, E. Spiliotis, and V. Assimakopoulos. Statistical and machine learning forecasting methods: Concerns and ways forward. *PLOS ONE*, 13(3):1–26, 03 2018. URL https://doi.org/10.1371/journal.pone.0194889.

[22] Wikipedia contributors. Convex function — Wikipedia, the free encyclopedia, 2022. URL https://en.wikipedia.org/w/index.php?title=Convex_function&oldid=1094479150. [Online; accessed 29-July-2022].

[23] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. URL http://www.R-project.org.

[24] G. Van Rossum and F. L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009.

[25] H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. URL https://ggplot2.tidyverse.org.

[26] H. Wickham, R. François, L. Henry, and K. Müller. *dplyr: A Grammar of Data Manipulation*, 2022. https://dplyr.tidyverse.org, https://github.com/tidyverse/dplyr.

[27] G. Grolemund and H. Wickham. Dates and times made easy with lubridate. *Journal of Statistical Software*, 40(3):1–25, 2011. URL https://www.jstatsoft.org/v40/i03/.

[28] H. Wickham, J. Hester, and J. Bryan. *readr: Read Rectangular Text Data*, 2022. https://readr.tidyverse.org, https://github.com/tidyverse/readr.

[29] H. Wickham and M. Girlich. *tidyr: Tidy Messy Data*, 2022. https://tidyr.tidyverse.org, https://github.com/tidyverse/tidyr.

[30] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007. doi: 10.1109/MCSE.2007.55.

[31] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. Fernández del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. URL https://doi.org/10.1038/s41586-020-2649-2.

[32] T. Chen and C. Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4232-2. URL http://doi.acm.org/10.1145/2939672.2939785.

[33] S. Seabold and J. Perktold. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010.

[34] R. Tavenard, J. Faouzi, G. Vandewiele, F. Divo, G. Androz, C. Holtz, M. Payne, R. Yurchak, M. Rußwurm, K. Kolar, and E. Woods. Tslearn, a machine learning toolkit for time series data. *Journal of Machine Learning Research*, 21(118):1–6, 2020. URL http://jmlr.org/papers/v21/20-091.html.

[35] T. Kluyver, B. Ragan-Kelley, F. Pérez, B. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. Hamrick, J. Grout, S. Corlay, P. Ivanov, D. Avila, S. Abdalla, and C. Willing. Jupyter notebooks – a publishing format for reproducible computational workflows. In F. Loizides and B. Schmidt, editors, *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, pages 87 – 90. IOS Press, 2016.

[36] P. Duboue. *The Art of Feature Engineering: Essentials for Machine Learning*. Cambridge University Press, 2020. URL https://books.google.ee/books?id=_BzhDwAAQBAJ.

[37] C. Seger. An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing, 2018. URL https://www.diva-portal.org/smash/get/diva2:1259073/FULLTEXT01.pdf.

[38] C. Molnar. *Interpretable Machine Learning*. 2 edition, 2022. URL https://christophm.github.io/interpretable-ml-book.

Appendix 1 – Non-exclusive licence for reproduction and publication of a graduation thesis[1]

I Marika Eik

1. Grant Tallinn University of Technology free licence (non-exclusive licence) for my thesis "Design of Data-Driven Parking Service: Case Study of Ülemiste City", supervised by Juri Belikov and Alari Krist

1.1. to be reproduced for the purposes of preservation and electronic publication of the graduation thesis, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright;

1.2. to be published via the web of Tallinn University of Technology, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright.

2. I am aware that the author also retains the rights specified in clause 1 of the non-exclusive licence.

3. I confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights, the rights arising from the Personal Data Protection Act or rights arising from other legislation.

18.05.2023

1. The non-exclusive licence is not valid during the validity of access restriction indicated in the student's application for restriction on access to the graduation thesis that has been signed by the school's dean, except in case of the university's right to reproduce the thesis for preservation purposes only. If a graduation thesis is based on the joint creative activity of two or more persons and the co-author(s) has/have not granted, by the set deadline, the student defending his/her graduation thesis consent to reproduce and publish the graduation thesis in compliance with clauses 1.1 and 1.2 of the non-exclusive licence, the non-exclusive license shall not be valid for the period.