Sofia University "St. Kliment Ohridski"

Faculty of Mathematics and Informatics

Department of "Software Technologies"

ABSTRACT

for the acquisition of an educational and scientific degree "Doctor",

professional direction: 4.6 Informatics and computer sciences,

doctoral program: Software Technologies - Software Engineering

on topic

"Aided decision making for public transportation optimizations using Big Data"

Ph.D. Candidate: **Georgi Kalinov Yosifov**

Supervisor:

Assoc. Prof. Milen Yordanov Petrov, Ph.D.

Sofia, 2022

I. GENERAL CHARACTERISTICS OF THE DISSERTATION

Motivation and Relevance of the Topic

Sofia is a European city and the largest populated place on the territory of the Republic of Bulgaria. According to data from the Bulgaria National Statistical Institute, despite the country's declining population (НСИ, 2021), in the capital, forecasts are that by 2030 it will grow by another 25,000 people (НСИ, 2018). An increase is also observed in the number of cars on the city's roads. According to data from the Sofia Municipality, from 2011 to 2020, their number has almost doubled - from 462,043 to 833,280 (Сантова, 2021). This is the situation in many other big cities as well. The growing population, growing traffic and the growing need for quality and unobstructed movement in the urban environment, in Sofia and cities like it, create great challenges for the administration and business. There are studies that show the potential benefits of implementing transport optimization policies by reducing urban traffic levels, which in financial terms reach billions of BGN for larger countries (INRIX, 2020). To solve this problem, mechanisms, and tools for monitoring the road situation must be developed. With their help, critical road sections could be identified for which specific measures should be taken.

At the same time, with the increased attention to the sensitivity of the collection and management of personal data and regulations such as the European General Data Protection Regulation (GDPR) or the Californian California Consumer Privacy Act (CCPA), special attention needs to be paid to the ways in which these mechanisms and tools operate.

The dissertation seeks a solution to a problem concerning both the accurate determination of traffic levels in an urban environment and the use of data that are not defined as sensitive for the citizens.

Objective of the dissertation work, main tasks, and research methods

The aim of the dissertation work is to support decision-making for the optimization of urban transport by determining, researching, and forecasting traffic load levels, using data collected from positional data of periodic public transportation vehicles used as probes in the traffic.

To achieve the set goal, the following tasks are set for implementation:

- To create a methodology for classification and analysis of the current state of data collection methods for determining traffic levels.
- To compile an overview of the different methods for analyzing and predicting traffic levels at a future point in time.
- To prepare a novel algorithm based on the compiled methodology, with the help of which to determine the traffic levels in an urban environment.
- To create experimental scenarios to study the qualities and limitations of the proposed algorithm.
- To develop tools required to support the algorithm data processing and management in the experimental scenarios.
- To explore different ways of forecasting traffic levels in an urban environment at a future point in time.
- To make a comparative analysis of the performance of the different forecasting methods and to determine the best performing one.

Practical applicability and benefits

Current or predicted traffic load data in an urban environment, can be used to extract the necessary information for the optimization of public transport. They can be published as open data, for use by the public or be useful in the creation of state or municipal policies related to the construction of infrastructure and distribution of public funds. By correlating it with data from CO_2 emission sensors, air pollution levels could be monitored, and future peaks predicted. This information can also be taken into account by routing software used by private individuals or companies in the shipping

industry, so that drivers can choose the most optimal routes. All these applications are part of the practical benefits of the current work.

Publications

The achievements of this dissertation have been published in a total of three publications. In all three publications, the PhD student is the first author. All of them have been published in international scientific journals with publishers ACM and Springer, with "impact rank", and are indexed in the scientific database SCOPUS.
All publications have been presented at international scientific conferences, with the last two presented in London, UK at Computing Conference 2022 and International Congress on Information and Communication Technology (ICICT) 2022, and the first at CompSysTech 2020, organized in Ruse, Bulgaria.

As of 09.2022, the publications have been cited a total of 4 times (verified using Google Scholar data).

List of publications on the topic of the dissertation

1. Georgi Yosifov, Milen Petrov, Predicting Traffic Indexes on Urban Roads based on Public Transportation Vehicle Data in Experimental Environment, Lecture Notes in Networks and Systems, editor/s:Janusz Kacprzyk , Publisher:Springer, 2022, pages:159-168, ISSN (print):2367-3370, ISSN (online):2367-3389, ISBN:978-3-031-10466-4, doi:10.1007/978-3-031-10467-1_8, Ref, IR , SCOPUS, SJR (0.17 - 2020), SCOPUS Quartile: Q4 (2022), (INSPEC, WTI Frankfurt eG, zbMATH, SCImago), PhD

2. Georgi Yosifov, Milen Petrov, Review of urban traffic detection approaches with accent of transportation in Sofia, Bulgaria, Proceedings of Seventh International Congress on Information and Communication Technology. Lecture Notes in Networks and Systems, vol 465., editor/s:Yang, XS., Sherratt, S., Dey, N., Joshi, A., Publisher:Springer, 2022, pages:509-517, ISSN (print):978-981-19-2396-8, ISSN (online):978-981-19-2397-5, doi:https://doi.org/10.1007/978-981-19-

2397-5_47, Ref, IR , SCOPUS, SJR (0.17 - 2020), SCOPUS Quartile: Q4 (2022), др.(INSPEC, WTI Frankfurt eG, zbMATH, SCImago), PhD

3. Georgi Yosifov, Milen Petrov, Traffic flow city index based on public transportation vehicles data, International Conference on Computer Systems and Technologies - CompSysTech'20 (CompSysTech'2020), editor/s:Vassilev T.,Trifonov R., Publisher:Association for Computing Machinery, 2020, pages:201-207, ISBN:978-145037768-3, doi:10.1145/3407982.3408007, Ref, IR , SCOPUS, SJR (0.182 - 2020), ACM Digital Library, PhD

Conference reports

1. Sectional report, Georgi Yosifov, Predicting Traffic Indexes on Urban Roads based on Public Transportation Vehicle Data in Experimental Environment

2. Sectional report, Georgi Yosifov, Review of urban traffic detection approaches with emphasis of transportation in Sofia, Bulgaria

3. Sectional report, Georgi Yosifov, Investigation of the correlation between the number of vehicles and the urban traffic load index based on positional data from periodic public transportation vehicles

4. Sectional report, Georgi Yosifov, Traffic flow city index based on public transportation vehicles data, International Conference on Computer Systems and Technologies

II. STRUCTURE AND SCOPE OF THE DISSERTATION

The dissertation consists of six chapters. The text is written in 180 pages and contains 76 figures and 17 tables. 108 literary sources and internet pages have been cited. The work is supplemented with five appendices. Each chapter is divided into thematic sections, helping to describe the problems solved by the work. At the end of the

dissertation, the scientific publications on the subject are added, the contributions are listed, and a declaration of originality is attached.

In the following sections, we present the specific content of each chapter. The sixth and final chapter is concluding, summarizing the work, and proposing directions for future research. The caption numbers of the figures and tables in the abstract match those of the dissertation work.

1. Introduction

The chapter consists of three thematic parts. The first part examines the development of the population and transport in the city of Sofia, Bulgaria. Then it shows the economic effect that reduced traffic can have by presenting data from the 2020 COVID-19 pandemic as analyzed by INRIX (INRIX, 2020). The importance and different ways in which the level of traffic can be determined and how this information could be used to improve the quality of life are also discussed. Particular attention is paid to data collection methods and their sensitivity and regulatory requirements (GDPR European Union, 2020).

The second part defines the set goal and tasks of the dissertation, and the third part describes the structure of the document.

2. **Comparative analysis of the current state of research in methods of data collection and determination of public transport traffic levels**

The second chapter consists of five parts. The first part examines the trends of the publications indexed in the scientific database Scopus by certain keywords related to the topic of the dissertation.

In the second part, "Methodology for classification and analysis of the current state of data collection methods and determination of traffic levels." various methods of gathering information that can be used for traffic level analysis are considered and a categorization of each against selected set of criteria is presented.

The third part reviews methods for predicting traffic load levels with various examples and cited studies on the topic.

Part four examines some of the types of neural networks that are used in this dissertation. The last part presents the contributions of the chapter and where they are presented.

3. Index of urban traffic load, based on positional data from periodic public transportation vehicles

The introductory part sets out the task of the chapter - to verify the hypothesis whether data from positional sensors of public urban transport vehicles can be used to calculate the real-time traffic level in an urban environment.

The second part focuses on analyzing time and positional data from the public transport in two different cities - Edinburgh, Scotland and Sofia, Bulgaria. In the following part, the conclusions from the research are highlighted.

In part four, referring to the results of the previous one, a novel algorithm is proposed for calculating traffic load indices based on positional data from public transport vehicles in 30-minute time intervals.

In parts five and six, possible ways of collecting periodic public transport data to be used as input to the algorithm and the visualization of its results after execution are proposed.

In the seventh part, experimental scenarios are described that validate the use of the algorithm and determine a high positive correlation between the algorithm indexes and the observed real traffic level. The chapter ends with a conclusion and presentation of the contributions.

4. Description of the created software tools for conducting the experiments

The chapter contains five parts, the last two being the conclusion and contributions. The first part presents the functional and non-functional

requirements for the software modules for calculating the traffic load indexes according to the algorithm presented in the previous chapter.

The second part describes the software modules, describing their interrelationships, input and output file formats, class diagrams, and preview screens.

In the third part, the created software for measuring transit times of public transport from the experiment carried out in the city of Sofia is described and shown.

5. Traffic Load Index Prediction

The fifth chapter contains a total of five parts, the last two being a description of the contributions and a conclusion. The first part describes the purpose of the created experiment and its stages. It will be used to determine the index and compare different traffic level prediction methods calculated by the algorithm presented in this work.

In the second part, the experimental scenario is described and the various data preparation operations to be used by the machine learning mechanisms chosen to predict the results are also given.

The third part describes the results of the experiment, showing examples and giving comparisons of the performance of the different selected algorithms for single-step models (predicting for one time interval ahead in the future) or multi-step models (predicting for several time intervals ahead).

III. CONTENTS OF THE DISSERTATION

1. CHAPTER 1. Introduction

Sofia is a European city and the largest populated place on the territory of the Republic of Bulgaria. According to data from the Bulgaria National Statistical Institute, despite

the country's declining population (НСИ, 2021), in the capital, forecasts are that by 2030 it will grow by another 25,000 people (НСИ, 2018). An increase is also observed in the number of cars on the city's roads. According to data from the Sofia Municipality, from 2011 to 2020, their number has almost doubled - from 462,043 to 833,280 (Сантова, 2021). This is the situation in many other big cities as well. The growing population, growing traffic and the growing need for quality and unobstructed movement in the urban environment, in Sofia and cities like it, create great challenges for the administration and business. There are studies that show the potential benefits of implementing transport optimization policies by reducing urban traffic levels, which in financial terms reach billions of BGN for larger countries (INRIX, 2020). To solve this problem, mechanisms, and tools for monitoring the road situation must be developed. With their help, critical road sections could be identified for which specific measures should be taken.

At the same time, with the increased attention to the sensitivity of the collection and management of personal data and regulations such as the European General Data Protection Regulation (GDPR) or the Californian California Consumer Privacy Act (CCPA), special attention needs to be paid to the ways in which these mechanisms and tools operate.

2. CHAPTER 2. Comparative analysis of the current state of research in methods of data collection and determination of public transport traffic levels

This chapter examines the trends on the subject as seen in the scientific literature. A methodology is then created to categorize the studies into two main directions of the current dissertation work:

1. Determining the level of traffic in an urban environment
2. Forecasting the level of traffic at a future moment.

When examining the number of articles published in scientific publications in the Scopus database by keywords related to the current work - "traffic congestion ", "urban traffic", "public transport", "public transport big data", "traffic prediction", "congestion

prediction", for the years between 2000 and 2020, it can be seen, that they all follow an increasing trend, which shows the interest that researchers have in the subject.

2.1. Methodology for classification and analysis of the current state of data collection methods and determining traffic levels.

Due to the importance of the topic, many scientific studies have been done to detect and measure traffic congestions. There are multiple ways to collect data and visualize the current dynamics of traffic in a city. The main ones found these days and covered here are:

- Smart devices / phones
- GPS in a car
- Video surveillance
- Positional data from public transportation vehicles

2.1.1. Smart devices / phones

One option to detect traffic is to use location data from mobile devices of passengers and drivers (Martín et al., 2019), (Idachaba & Ibhaze, 2016), (Tu et al., 2021). Another one is the use of data from the telephone antennas of mobile devices, giving not so precise information, but still showing promising results (S. Li et al., 2020) .

The problem with this data is that it is mostly controlled by private companies or individuals and not made available to researchers or governments for analysis. This data could also be considered sensitive information for the people providing it and therefore it needs to go through pre-anonymization, which would add an additional layer of complexity that could potentially lead to data compromise and security risks.

2.1.2. GPS in a car

Another option would be to use the GPS and internet connectivity of personal vehicles to achieve a similar result(D'Este et al., 1999), and experiments were also done in that direction with cars of taxi companies (Kan et al., 2019). This approach shares many of the pros and cons of mobile devices and could be a companion source of information.

2.1.3. Video surveillance

Another possible approach is the use of data from cameras, through the analysis of picture and video stream (Buch et al., 2011; Nemade, 2016), (Nemade, 2016). By using a convolutional neural network (CNN) and image analysis, researchers were able to achieve 89.5% traffic level classification accuracy (Kurniawan et al., 2018).

However, this comes at the cost of using complex algorithms, expensive computing power (Esteve et al., 2007) and the need of considering various factors such as lighting, weather conditions, etc. (Stetsenko & Stelmakh, 2020), (J. Li et al., 2021). Economically, in the absence of an existing infrastructure, a large investment would be required to install those monitoring devices with sufficient presence to cover major road arteries.

2.1.4. Positional data from public transportation vehicles

There are studies that use the bus transport as samples in the traffic to determine the traffic situation. Carli et. al. defines metrics for measuring congestion in an urban environment using GPS technology in a given area. The study does not discretize the classification of traffic congestion but compares the values with the best obtained so far in the time series (Carli et al., 2015). Another study that uses buses as samples is done by Kumar & Sivanandan, 2019. In it, buses are equipped with GPS devices and compare the unique characteristics of buses against other vehicles on the road. The study also defines a congestion index (CI) (Kumar & Sivanandan, 2019).

Table 1 below shows a summary of the above approaches. They are classified against the following criteria we define:

- **Availability** – corresponds to the volume of data being recorded, not to the ease of access of that data.
    - High - coverage of the entire city
    - Medium – coverage on main roads

    - Low – limited coverage
- **Confidentiality** – data may be available in abundance, but due to local regulations (such as the European GDPR), may not be accessible or, if accessible, processing may not be permitted.
    - High - no confidential personal data of people is used
    - Medium - not all cases use people's personal confidential data
    - Low - confidential personal data of people is used
- **Price** – costs of implementing both data collection and data processing.
    - High – Price of a single reporting module and its installation is over BGN 1,000.
    - Average – Price of a single reporting module and its installation is between BGN 100 and BGN 1,000.
    - Low – Price of a single reporting module and its installation is below BGN 100.

Table 1 Categorization of possible data collection approaches against defined criteria

№	Approach	Availability	Privacy concerns	Cost
1.	Car GPS	Moderate	High	Moderate
2.	Smart devices/ phones	High	High	Low
3.	CCTV	Moderate	High	High
4.	Public transport positioning data	Moderate	Low	Low

From the table, it is clear that the approach of the current study (Public transport positioning data) would provide us with a low-cost method with a high level of privacy, while having enough data available to perform practical traffic measurements on the main roads.

2.2. Overview of Methods for Forecasting Traffic Load Levels

There are various ways of predicting the traffic level that have been investigated in the literature. They could be divided into two main groups - those that use deep machine learning methods, such as neural networks (in their various varieties) and those that use various computational and statistical models such as Kalman filters (Kalman, 1960), ARIMA (Lan & Miaou, 1999), Exponential filters (Ross, 1982), Box–Jenkins method (Hamed et al., 1995) and others. There are also studies that use a combination of the above models (Wang et al., 2022). A disadvantage of purely statistical models is that assumptions are made that do not always match the dynamics of data collected from the real world. For this reason, in the last decade, many studies chosen for the purpose of traffic prediction have been done with deep machine learning through neural networks (de Medrano & Aznarte, 2020).

3. CHAPTER 3. Index Of Urban Traffic Load, Based on Positional Data from Periodic Public Transportation Vehicles

3.1. Introduction

In a modern city, most major thoroughfares have one type of intermittent public transportation or another. The Introductory chapter discussed the benefits of using public transport data to measure the state of the road situation - it has an almost constant presence on the road and each bus can be considered as a measuring agent - a sample in the traffic. They are also by definition anonymous, as no personal data of individuals is being used, only the positioning data of a particular bus.

The aim of the current chapter is to investigate the hypothesis that the data from the positioning sensors of these transport vehicles can be used to calculate the real-time traffic level in an urban environment. The level of traffic will be defined by a traffic index to be calculated at half-hourly intervals. The index should determine the traffic level in a distinctive road segment.

3.2. Analysis of time and positional data from public transport

Based on the tasks set above, data on the distribution of vehicle's transit times through different road segments in two European capitals - Edinburgh, Scotland and Sofia, Bulgaria - have been analyzed. The times will be examined in their averaged value at half-hourly intervals and the histograms, slope and distributional properties of the data will be determined.

3.2.1. A study of data from Edinburgh, Scotland

3.2.1.4. Results

In 10 out of 10 or 100% of the examined segments, a right-skewed probability distribution was observed. A commonly used technique for approximating the probability distribution of a data set to a symmetric form is to apply a data transformation (Vadali, 2017).

After applying a logarithmic transformation to the data, we can inspect the Q-Q plot of the distribution (Figure 18), where, except for single values at its two ends (the extreme values), most of the points lie on the Q-Q line. This signals that the distribution of the sample data after transformation closely approximates that of a theoretical normal distribution.

This is also supported by the histogram (Figure 19), which is shown below. A symmetrical graph is observed here. By using the skewness function from the "moments" library of the R programming language, we can compare the skewness of the "before" and "after" histograms. Calculating it gives the values of 0.905724 for the data before the logarithmic transformation and -0.007795457 for the one after. Values close to zero indicate that the graph is symmetric, while values larger or smaller, depending on the sign, indicate that the graph is skewed to the left or right. In this case, the logarithmic data value is almost symmetrical.

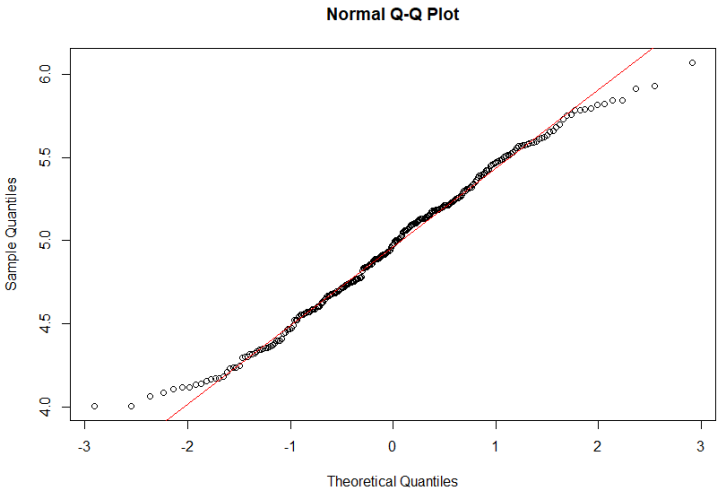

Figure 18 Q-Q plot of the natural logarithm of the average times spent in the segment

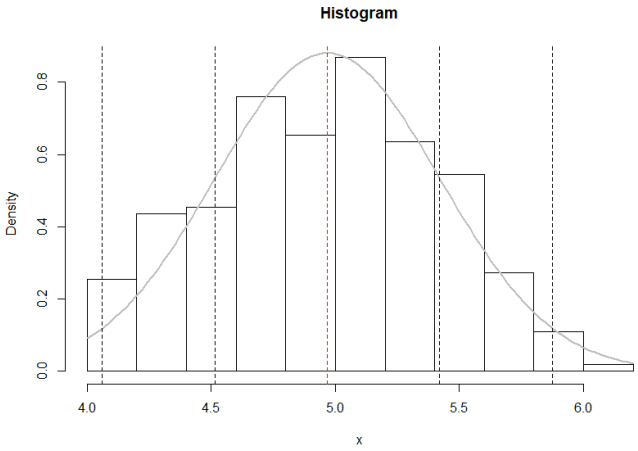

Figure 19 Histogram of the natural logarithm of the average times spent in the segment

Conducting the same experiment for all studied segments of the real data set, a similar trend is observed. On average, the percent improvement in the slope of the histograms for the 10 segments is **74.02%**. The mean value of the kurtosis function on the data before the transformation for all segments is 7.4406588 and after it is 3.0526822, given a value of "3" for a theoretical normal distribution (Pearson, 1905).

3.2.1. Study of data from the city of Sofia, Bulgaria

The previous section showed the statistical analysis in processing the data from Edinburgh, Scotland. Although random road segments were chosen to be studied, there is a possibility that the analysis is only valid for this particular city and may not be applicable in other cases. To make sure that this is not the case, it is necessary to find an example from another city or country that supports the results found.

In the introductory chapter, it was described that a city like Sofia would be suitable for a study like the current one, as it has a significant number of public transport vehicles covering its main roads. Unfortunately, however, the positional data of buses and trolleybuses in the capital of Bulgaria is not publicly available, even retrospectively. For this reason, for the purposes of the research, alternative approaches are needed to extract data for the city of Sofia.

One of these approaches is the use of cameras to monitor the meteorological situation. The current study does not need specific coordinates to calculate the time it takes for the vehicle to cross a road segment, and the time itself can be detected directly from a video recording of a selected road segment. For the purposes of the experiment in the current study, video materials were taken from the website https://weather-webcam.eu (Weather Webcam EU, 2022), as the screen was recorded in real time with the launched web page through the application Zoom (Zoom Video Communications, 2022).

The aim of the research is to be able to assemble traffic footage for at least two days, which can then be processed by detecting the times it takes for public transport vehicles to pass through the selected road segment and examine their characteristics. The total number of recorded times after conducting the experiment is 320. The first point on

each of the two days is recorded in the range 05:30 AM - 06:00 AM and the last in the range 11:30 PM – 00:00 AM. A transformation is then applied to the data based on the hypothesis made - taking their logarithmic values, averaged every half hour and studying their distribution.

On Figure 26 the Q-Q plot comparing the distribution of values with the theoretical normal distribution is plotted. Despite the small number of records, it can be seen that the two distributions are approaching each other.

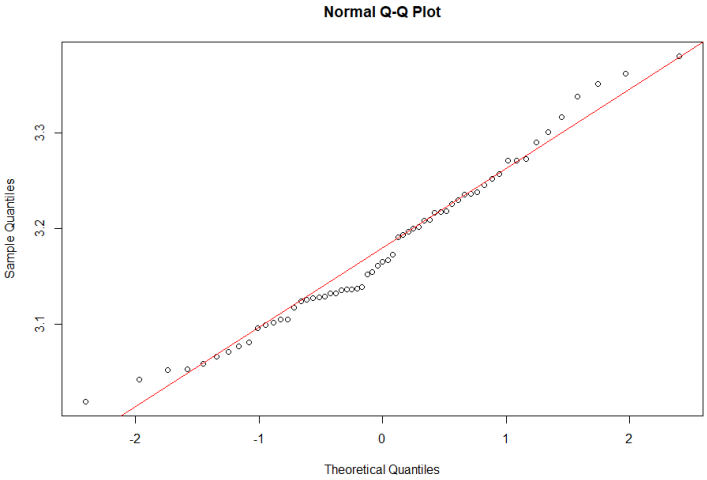

Figure 26 The Q-Q plot of the non-extreme logarithm of the half-hourly mean times measured from Druzhba district

3.3. Conclusion from the study of data from Edinburgh and Sofia

In the previous two sections it was shown, by examining real transit times through different and heterogeneous segments, that the logarithm of their mean value taken at half-hour intervals approximates a normal distribution. In the following sections, an algorithm is formalized that takes this observation into account and, by applying statistical methods, calculates 6 traffic load index values. The correlation between the

thus calculated index and the actual traffic level is then investigated through a set of experiments.

3.4. Proposed algorithm

Based on the studied data from the cities of Edinburgh and Sofia, the following algorithm for calculating the traffic index is proposed. The definitions used in this section are shown in Table 4.

Table 4 Definitions

Symbol	Definition
$\mathbf{T_j}$	30 min time interval
$\mathbf{CP_i}$	Checkpoint – the road is divided by checkpoints
$\mathbf{BS_i}$	A bus stop $\{ BS_0, \dots BS_n \} \subset \{ CP_0, ..., CP_m \}$
$\mathbf{S_i}$	A segment is defined by two consecutive checkpoints
$\mathbf{t_kS_i}$	The time for which a bus has travelled thru segment S_i
$\mathbf{T_jS_i = \{ t_0S_i, .. t_kS_i \}}$	Set of all times recorded by buses for segment S_i at time interval T_j
$\mathbf{MT_jS_i}$	The mean of the times in the set T_jS_i
HDS	Historic dataset
HDS[i]	Historic dataset for segment i
$\mathbf{IS_i}$	Traffic index of segment i
CI	City traffic index

The defined traffic index contains 6 discrete values from 0 to 5 described in Table 5.

Table 5 Traffic index

Index value	Description
0	No traffic
1	Light traffic
2	Normal traffic - low
3	Normal traffic - high
4	High traffic
5	Very high traffic

"Bus route" is defined to be the pre-selected route for that bus line. The bus routes are naturally separated by their bus stops. An additional segmentation of the bus route

using control points (CP) is also presented. The path between two consecutive control points is called a "segment" (S).

The 24 hours of the day are divided into 30-minute intervals to be used in subsequent calculations. These intervals are denoted as T_i. When a bus passes through a segment S_i, the travel time is measured and added to the set of values for that segment for that time interval. Each time the vehicle passes, the arithmetic mean time for the segment is recalculated. When the time interval is over, an arithmetic mean time is taken for it and stored in the historical data set (HDS). This data is used to calculate the traffic index for that segment. HDS consists of groups of values for each segment. In each group of values, the natural logarithm of the arithmetic mean times for each elapsed time interval is stored.

Note that each segment is compared to itself. This is done due to the specific nature of it - how many crosswalks there are, how busy they are, whether there are traffic lights, roundabouts, speed limits, etc.

When a new time interval begins - denoted by Tj, a new set is created for each segment S_i, namely T_jS_i. The arithmetic mean time for the previous interval in this segment is automatically added to this set, so initially the set will look like $T_jS_i = \{MT_{j-1}\ S_i\}$. This is done because no time slot is isolated and no sudden change in traffic is expected between two consecutive time slots.

At any moment, the traffic index for a specific segment can be calculated. Assuming that the segment is Si and the time slot we are currently in is T_j. The arithmetic mean of the times stored in the set T_jS_i can be calculated, namely $t = MT_jS_i$. The standard deviation σ can then be calculated as well as the arithmetic mean μ for the historical data set for the segment (HDS[i]). Then the natural logarithm of the currently measured time t is compared with the standard deviation of the arithmetic mean, and it is found in which range of the data it lies. If t is between the minimum time and μ-2σ - then the index for S_i is 0. Similarly, if it is between μ-2σ and μ-1σ then it is 1; if it is between μ-1σ and μ, then the value will be 2; if it is between μ and μ+1σ it will be 3, then if it is between μ+1σ and μ+2σ it will be 4 and finally if it is greater than μ+2σ it will be 5.

The calculation of the City Index (CI) for the time interval T_j is performed by averaging the indices for all segments in the time interval T_j.

$$CI = \sum_{i=0}^{n} \frac{\text{calc}Index(\log(\text{MTjSi}))}{n}$$

The described approach has some limitations. Only roads that have certain criteria can be used to calculate the congestion index:

- There must be routes for public transportation vehicles present on this road.
- There should not be a dedicated fast (bus) lane for public transport on this road.

3.7. Experimental part

3.7.1. Study of correlation between the number of vehicles and traffic index calculated by the algorithm

The purpose of the below described experiment is to investigate the correlation between the number of vehicles that have passed during a specific time interval and the results calculated by the algorithm described in this chapter. If a strong positive correlation is demonstrated, this would mean that the algorithm is working correctly and correctly reflects the traffic situation.

3.7.1.2. Experiment scenario

The specialized software "Simulation of Urban MObility" (SUMO) is used to compile the following scenario of the experiment. A section with a length of 1000 m was selected from the city of Sofia, constituting a busy part of Blvd “Doctor G.M. Dimitrov" through which 3 lines of public transport pass at the time of writing the dissertation, it also does not have a separate bus lane. The chosen road part is further divided into 10 segments.

3.7.1.5. Results of the experiment

For each segment of the simulation, two graphs describing the input and output characteristics of the program are calculated. For illustration purposes, segment 2 is presented in this section. On Figure 37 the number of vehicles passing through the given segment in a particular half hour is illustrated. Along the abscissa is the sequential number of the half-hour of the day - for example, 0 would correspond to 00:00 AM to 00:30 AM, 18 would correspond to 09:00 AM to 09:30 AM, and so on. The ordinate indicates the number of unique vehicles that passed through the segment. The characteristics of the input data are reflected by three peaks – morning, early afternoon, and evening.

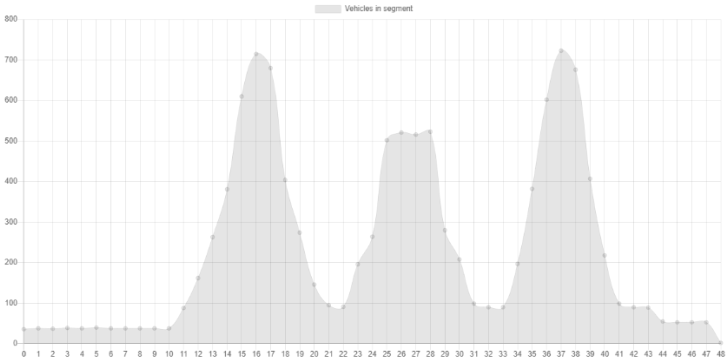

Figure 37 Number of vehicles that passed through the segment in a specific half hour

On Figure 38 the calculated values from our traffic amount algorithm are shown. On the abscissa the half-hour segments are again listed, and on the ordinate the value of the index in this period. The reason that the first value shown is for 05:30 AM is due to the working hours of the public transport. This graph again shows three traffic peaks in the morning, afternoon, and evening hours.

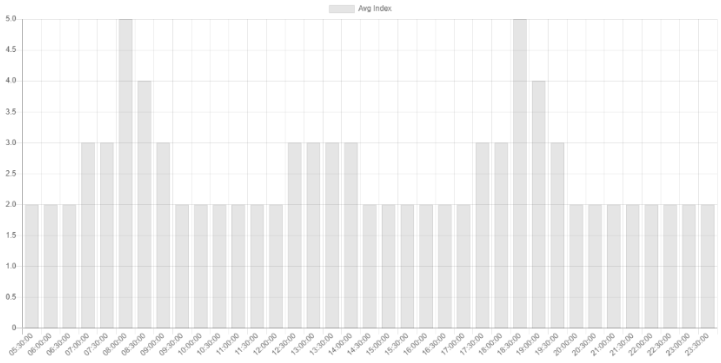

Figure 38 Calculated traffic index for a segment in a given half hour

From the calculated comparison of the results of the experiment, the average Pearson correlation coefficient between the data for all segments of the simulation is 0.8202 or according to the classification - there is a high positive correlation. Looking at the data in more detail, we see that approximately 10% of the segments have a very high level of correlation, and 80% a high level of correlation between the number of vehicles in a given road segment and the level of traffic, which is calculated by the algorithm defined in this chapter.

3.7.2. Investigating the results of the algorithm over time

An increase in road traffic during peak hours would also mean a greater flow of passengers in need of public transportation. The increase in passengers creates a need that urban transport must fill, namely, to provide enough vehicles, in a short enough period of time, to satisfy the demand. To solve this issue, the transport company could be flexible in choosing how many buses cover a given route at a given time.

An increase in the number of busses also means increase in the number of measured values for the segments. This would mean that more data would be collected in a half-hour period on a busy part of the day than on an off-peak part. To solve this case study, the algorithm presented in this paper groups the results collected in half-hour time

intervals and saves the arithmetic average time for each, which it then uses to calculate the traffic Index.

The purpose of the experiment presented below is to investigate the positive impact of this grouping on the measured results.

3.7.2.1. Experiment scenario

Two variants of the algorithm were created:

- Standard - which groups the measured values into half-hourly ranges and uses them to calculate the Index
- Modified – which uses all measured values for the calculations without grouping

For each of the high-traffic time ranges, several experiments were done with a different factor of increasing the number of buses, respectively 0, 5, 10, 100 times, and for each factor, each variant of the algorithm - Standard and Modified, has been run with the same initial data.

3.7.2.2. Results

Unlike the "Standard" algorithm, the "Modified" one shows a degradation of the results with the increase in the coefficient. With the Standard algorithm, the three traffic peaks are observed - two larger ones in the morning and evening and then a smaller one in the midday hours for all factors of increment. However, looking at the results of the scenarios with increased factors under the Modified algorithm, we notice that the midday peak has disappeared from it, and the evening peak has drastically decreased. In the results for factor of 10, from the maximum Index of 5 as previously recorded, the index was reduced to 4, while in the extreme case with a factor of 100, we see that the evening peak has reached an Index level of only 3.

The above experiment demonstrates the robustness of the algorithm to changes in the number of buses serving a route and its ability to provide consistent results that represent accurately the traffic conditions.

3.7.3. Applying the Traffic Index Calculation Algorithm to Real Data

3.7.3.1. Applying the algorithm to segments of the city of Edinburgh, Scotland

"Princess Street" is located in the central part of the city of Edinburgh. It is a tourist street in close proximity to many attractions and also the city's central railway station. During the period of recording, 569 unique vehicles passed through it. The total number of records that have been processed is 8495. Figure 41 shows arithmetically averaged Indexes over several days, calculated by the traffic algorithm ran on that segment.

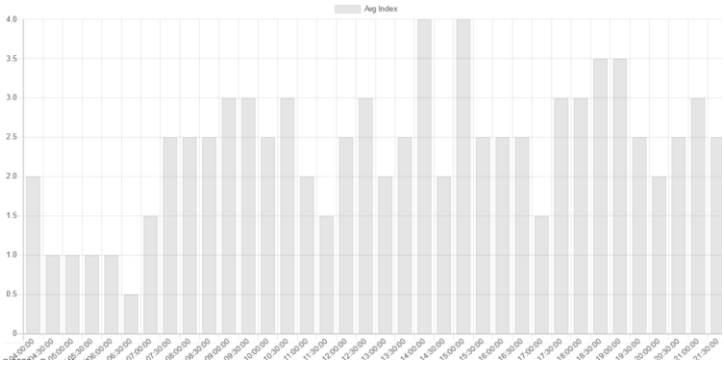

Figure 41 Traffic distribution for the second selected segment of the City of Edinburgh – "Princess Street"

Three traffic peaks are observed in the figure. The first is in the morning hours in the period 09:00 AM - 10:30 AM, the second is and the biggest peak is in the afternoon hours 02:00 PM - 03:00 PM, while the third is in the evening hours of 06:30 PM – 07:00 PM.

3.7.3.2. Applying the algorithm to the data from the city of Sofia, Druzhba district

This section will showcase the results returned by the algorithm for a selected segment with records of public transport times from Druzhba district, Sofia. On Figure 42 and Figure 43 the results of the algorithm from the averaged indexes of the two days considered are shown. The highest amount of traffic is observed in the period 04:30 PM. to 07:30 PM., with peaks at 05:30 PM and 06:00 PM. Higher-than-normal traffic is also reported at 09:30 AM - 10:30 AM, and the rest of the time traffic is normal for the section.

Figure 42 Map visualization of the calculated index of the segment in Druzhba district, Sofia

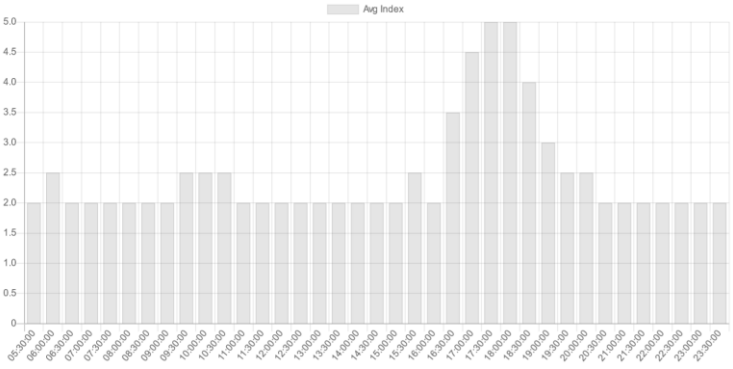

Figure 43 Traffic distribution for the second selected segment of the city of Sofia, Druzhba district

3.8. Conclusion

In this chapter, two datasets related to the position or time of passage of vehicles from the periodic public transport in the cities of Edinburgh and Sofia were analyzed and a statistical analysis was applied to them.

Based on this analysis, an assumption has been made, which has been subsequently validated through experiments, that the measured time of passage of the public transport vehicle has a relationship with the number of vehicles on the road. Then an algorithm has been presented to calculate traffic indexes for both the entire city and a specific road segment. Definitions, applications, and limitations of the algorithm have been given. A map visualization mechanism for the state of traffic in the city, both historically and in real time has been presented.

To validate the hypothesis and the created algorithm, a set of experiments were prepared. The correlation between the number of vehicles and the results of the algorithm was investigated and a high positive correlation was found between them. The robustness of the algorithm in the presence of a heterogeneous number of vehicles

in different time ranges was also investigated. This showed that the method fulfills its stated objectives.

Thus, the presented algorithm can be used by the public administration to make decisions about infrastructure changes and make future investments. Alternatively, it could be used to find the fastest routes for the city in real time.

4. CHAPTER 4. Description of created software tools for conducting the experiments

For conducting the experiments described in this thesis, specialized software has been developed, the purpose of which is to perform the activities defined by the algorithm for calculating the traffic indexes, as well as presenting the results. .NET Framework and the C# programming language have been used for the purpose of creating the software.

4.1. Requirements for the software for conducting the experiments

The software for conducting the experiments should meet the following functional requirements:

- To implement the traffic index calculation algorithm. The main and primary purpose of the software is to calculate the index as described in this document.
- The software should generate all necessary output files for conducting the experiments. For this purpose, standard formats and types and structure of the generated files must be defined.
- The software should provide a graphical visualization module of the algorithm results that uses as input the output files generated by the algorithm execution module.
- The visualization module should provide a map representation of the tested segments. Adding the color coding above the map visualization and the ability to check the segment index in a specific interval is also desirable.

- City index calculation is another thing the software should be able to do. The user of the software should be given the option to select the time interval in which to display the index.

As well as the following non-functional requirements:

- The software should be divided into separate modules that can be developed, updated, and maintained independently of each other.
- The module for the algorithm should be abstracted of the type of input data. The necessary abstractions must be created to ensure reusability of the code in different scenarios – generated data, real data, data in different coordinate systems, etc.

4.2. Software modules

On Figure 45 software modules and their interrelationships are shown. The software has two foundational modules:

• TrafficCongestionAlgorithm – a module that contains the main components for calculating the traffic index. Here, the main models and processing components have been defined, as well as the interrelationships between them. The module has been made in a way that allows it to be agnostic about the format of the input data.
• DataVisualisation – this is the module that is responsible for data visualization. It provides a map representation as well as auxiliary graphs to evaluate the result.

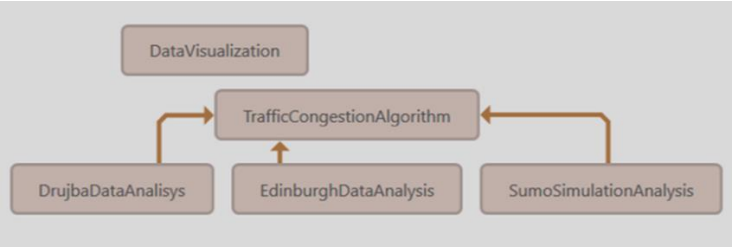

Figure 45 System software modules

Three program entry point modules are defined depending on what inputs the experiment will be run with. They are the following:

- EdinburghDataAnalysis – This is the module that uses as input data real values recorded from the capital of Scotland - Edinburgh and feeds them to the algorithm. Its results are described in section "3.7.3 Applying the Traffic Index Calculation Algorithm to Real Data" of this document.
- SumoSimulationAnalysis – This is the module that uses as input the generated file from the SUMO simulation discussed in section "3.7.1. Study of correlation between the number of vehicles and traffic index calculated by the algorithm " of this document.
- DrujbaDataAnalysis – This is the module that processes the recorded times from Drujba district, Sofia and applies the algorithm described in the previous chapter to them.

4.3. Software for measuring transit times of public transport through Druzhba district, Sofia.

For the purposes of facilitating the measurement of the times of vehicles in the city of Sofia, Druzhba district, an application was also developed, which has the following functional characteristics:

- To contain the functionalities of a stopwatch, with the possibility of starting, stopping, restarting the time, as well as stopping with recording the measured time.
- Ability to display the previous recorded times in reverse chronological order - the most recent at the top.
- Ability to save all recorded times to a CSV file.
- On startup to automatically read the recorded times from the last created CSV file.

The software is written in C# with Windows Forms - interface creation technology.

5. CHAPTER 5. Traffic Load Index Prediction

An experiment has been conducted, considering the traffic in 10 road segments. The experiment is with data prepared by the specialized traffic simulation software SUMO in an urban environment for simulated duration of 365 days, and various time series forecasting techniques are applied and compared, and a proposal is made to use a specific methodology based on the results of their performance.

5.1. Stages of the experiment

The process of experiment preparation and execution consists of five main stages, with the output data of each stage becoming input for the next one. The stages are as follows:

- The first stage is the preparation of the experiment - generation of a schedule of the vehicles participating in the traffic based on rules with added randomness factors.
- The second stage represents traffic simulation based on the schedule generated by the first.
- The third stage consists of the execution of the algorithm for calculating the traffic index in the time intervals generated by the simulation.
- The fourth stage represents the transformation of the output of the algorithm into a suitable format for consumption by the machine learning platform.
- The fifth and final stage uses the traffic index time series data to make predictions about the state of traffic at a future point in time.

5.2. Experiment scenario

The experiment consists of examining and comparing the results of different machine learning models and applying them to the available data generated in the first four stages of the previous section of this chapter.

The focus is on solving two main tasks:

- Forecasting the traffic load index for the next time slot in a certain road segment

- Forecasting of traffic indexes in several consecutive time intervals based on a set of measurements in the past period for a certain road segment

For each of the tasks, different machine learning models are trained, and their results are compared. Finally, based on these results, a proposal is made to use the one that gave the best results.

5.3. Results of the experiment

In this section, the results of the various experiments are discussed in detail and a comparative analysis of their performance is made. In all models, mean squared error is used to calculate the loss function, and mean absolute error is used for metrics.

5.3.1. One-step models

The first type of models that have been considered provides a prediction of a single value of the traffic load index in the future.

5.3.1.1. Baseline model

In order to understand and compare the results obtained from the performance of the machine learning models, a simple baseline model has to be used. A baseline model has been chosen that has the property of taking the value in the current interval and returning it as a prediction for the next one, assuming that there will be no sharp drops and rises in the traffic level in the two adjacent time intervals.

5.3.1.2. Multistep dense model

The first model to be described is a neural network with two consecutive dense layers. To make a prediction, the neural network uses the 16 previous states of the index as input.

5.3.1.3. A model using temporal convolution

Traditionally, convolutional neural networks (CNNs) (Tealab, 2018) have been primarily used for object recognition from an image, but recently studies are emerging that show they can also be successfully used for timeseries predictions. The current experiment scenario uses similar data as the multistep model using highly connected layers – the input is the 16 previous indices, and the output is the prediction for the next half hour.

5.3.1.4. A recurrent neural network with long-short term memory (LSTM)

Recurrent Neural Networks (RNNs) are neural networks well suited to working with timeseries because they keep internal state from one time step to the next. For the next model for predicting the traffic index in the next interval, a type of RNN, namely Long Short-Term Memory (LSTM) is chosen.

5.3.1.5. Analysis of the results of the execution of the one-step models

The results obtained from the implementation of the models described in the above section can be viewed in a systematic way. What is observed is that there is nearly a 50% improvement of all three models compared to the baseline, with the recurrent neural network yielding the best results, but it needs to be considered that this model is also more demanding in terms of time and resources to train (Figure 69).

In Table 16 the results of the different single-step models and their mean absolute error values are shown.

Table 16 Results of the one-step models

Model	Mean absolute error
Baseline	0.4052
Multi step dense	0.2146
Conv	0.2013
LSTM	0.1881

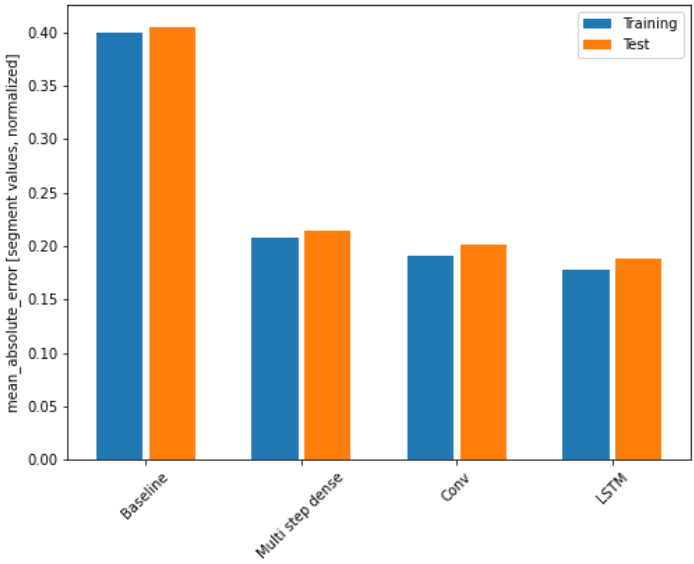

Figure 69 Comparison of mean absolute error of one-step models

5.3.2. Multistep models

Unlike single-step models, multi-step models have the property of predicting several steps ahead in the timeseries. In this section of the chapter, the results of various multistep models will be examined, finally analyzing, and comparing their performances.

5.3.2.1. Baseline of multistep models

A model that returns the last result as the next predictions is chosen as the baseline. It is expected here to get a worse result than with the single-step models, since the probability of changing the traffic in a period of 4 hours is greater.

5.3.2.2. A model using linear projection

The linear model is the simplest and will be examined first. It is a neural network that takes as input the latest traffic index report and the time dimensions and based on these it predicts what the next 8 will be by linear projection.

5.3.2.3. A neural network with dense layers

The structure of this neural network differs from the linear one in that there is one layer with 512 neurons between the input and output layers. However, this configuration is similar to linear in that, like it, it only takes the last traffic index and time dimensions and uses them to determine the next 8 indices.

5.3.2.4. Temporal Convolutional Neural Network

Having looked at the results of two models that use only the latest result to make their prediction, we will now look at those that work with a fixed number of historical steps back and use this information to calculate the traffic index. Such a model is the temporal convolutional neural network. For the current experiment, it is configured to

run on the last 16 records, and again predict the next 8. The convolutional layer of the network is composed of 256 neurons, and the activation used is ReLU.

5.3.2.5. Recurrent Neural Network

For the next model, the results of a recurrent neural network and specifically one with long-short-term memory (LSTM) has been used. The network is configured to accumulate input data for a period of 48-time intervals backward and generate information for the next 4 hours.

5.3.2.6. Autoregressive recurrent neural network

The final model that will be used to predict the time series of traffic indices is an autoregressive recurrent neural network. The difference between this model and the previous ones is that in it the result is generated in steps and each generated step is added as an input to generate the next one.

5.3.2.7. Performance analysis of multistep models

From the results of the multistep models presented in Figure 76 and Table 17, it can be seen that all models except the linear one, give a fivefold improvement in the error measured at the baseline. The model with the best performance is the autoregressive recurrent neural network, which is also the only model with an error below 0.2. However, this model is the most computationally intensive, which should be taken into account when calculations are run on more segments of a real environment.

Table 17 Comparison of the mean absolute error of the multistep models

Model	Mean absolute error
Last	1.0802
Linear	0.6338
Dense	0.2429
Conv	0.2371
LSTM	0.2093
AR LSTM	0.1985

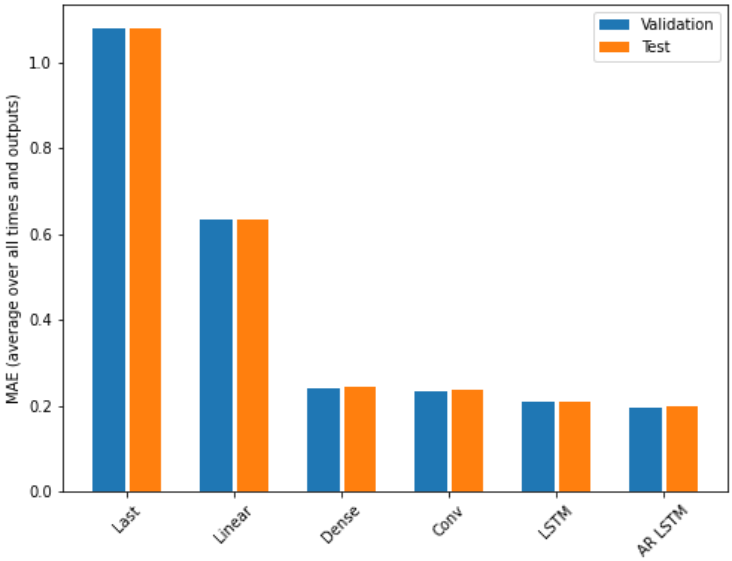

Figure 76 Comparison of the mean absolute error of the multistep models

6. CHAPTER 6. Conclusion and future work

6.1. Summary of the implementation of the initial objectives

The thesis examines various mechanisms for calculating and predicting urban traffic, with a special emphasis on the study of traffic using positional data from public transportation vehicles, used as probes in traffic. A methodology for classification and analysis of the current state of methods for collecting data and determining traffic levels has been compiled, as well as a methodology for classification and analysis of the current state of methods for predicting traffic levels in the future. Heterogeneous sources of transport information were studied, and conclusions were drawn about the data and their distributions and properties through applied statistical analysis. Based on this information, an algorithm was created that categorizes traffic in an urban environment, determining the traffic index. Using a number of experimental scenarios, the qualities of the developed algorithm were investigated and validated, showing that there is a strong positive correlation between the values of the calculated index and the number of vehicles on the roadway. The robustness of the algorithm to a variable number of traffic samples is also investigated and applied to demonstrate the results. Different methods of data collection that can be used for the input stream of the algorithm are discussed, and a visualization of the results is proposed. The specially developed software solutions with the help of which the experiments were made are briefly presented. After the successful calculation of the traffic index, different ways of forecasting traffic levels in an urban environment at a future point in time are investigated, comparing the performance of the different prediction ways, and determining the most accurate one.

The aim of the dissertation work to support decision-making for the optimization of public transport by determining, researching, and predicting traffic load levels, using data collected from positional coordinates of periodic public transport used as a traffic sample is fulfilled, as the proposed approach meets all the requirements set at the beginning of the study. These results would help to optimize transport in the public environment, which would lead to cost reduction and proper future planning.

6.2. Directions for future research

The studied algorithm presented in this way does not take into account the influence of neighboring road segments and their traffic on each other, but only considers them in isolation. The study of such an influence could not only improve the evaluation of the traffic levels but could also be used to determine the influence of the road used by intermittent urban transport vehicles on such neighboring roads that are not being used, thus mitigating the constraints of the algorithm and expanding its scope of action.

Before starting to apply the algorithm presented in this dissertation on real city data, there is the question of optimal road segmentation. Not only is this a prerequisite of the eventual launch of the large-scale system, but mechanisms for network modification should also be supported, in the event of changed/added transport lines. Such a task would be impractical to solve manually, since for the scale of a city, even the size of Sofia, it would mean the segmentation and maintenance of thousands of kilometers of transport routes.

Also, having hundreds or thousands of vehicles sending their positional coordinates every second requires a special system architecture capable of receiving, recording, and processing them, while being robust and always available. The system should also support machine learning modules to perform the functions of predicting traffic at a future point in time. Creating such an architecture is a task that would be a good addition to the current work.

In chapter 5 of the dissertation, single-step and multi-step models for forecasting traffic at a future time were considered. The maximum forecasting period described in the chapter consists of 8 consecutive time intervals or a duration of 4 hours. However, there are various applications that require longer computations forward in time and for which different types of statistical analysis and mechanisms are required, which are not presented in this work, but would be useful in the future.

IV. DISSERTATION CONTRIBUTIONS

A. Scientific and applied contributions

1. Through a review of scientific literature, various types of methods for determining the traffic load in an urban environment have been examined and categorized. An analysis of their characteristics by selected categories was made.
2. Created statistical analysis of data on the transit times of vehicles from periodic urban transportation vehicles, calculated from two heterogeneous datasets from Edinburgh, Scotland and Sofia, Bulgaria, through selected road segments.
3. An algorithm was developed based on the statistical analysis, with the help of which the level of traffic on a road segment can be indirectly determined. Through a number of experiments, the qualities of the presented algorithm have been determined and verified.
4. Made a comparative analysis of the results of single-step and multi-step machine learning models to determine traffic load levels at a future time.

B. Applied Contributions

1. Developed set of software solutions for collecting, processing, calculating, and visualizing the level of traffic load, by means of the presented algorithm, offering the possibility of modular integration to support different types of input data.

BIBLIOGRAPHY (SAMPLE)

Buch, N., Velastin, S. A., & Orwell, J. (2011). A Review of Computer Vision Techniques for the Analysis of Urban Traffic. *IEEE Transactions on Intelligent Transportation Systems*, *12*(3), 920–939. https://doi.org/10.1109/TITS.2011.2119372

Carli, R., Dotoli, M., Epicoco, N., Angelico, B., & Vinciullo, A. (2015). Automated evaluation of urban traffic congestion using bus as a probe. *IEEE International Conference on Automation Science and Engineering*, *2015-Octob*, 967–972. https://doi.org/10.1109/CoASE.2015.7294224

de Medrano, R., & Aznarte, J. L. (2020). A spatio-temporal attention-based spot-forecasting framework for urban traffic prediction. *Applied Soft Computing*, *96*, 106615. https://doi.org/10.1016/j.asoc.2020.106615

D'Este, G. M., Zito, R., & Taylor, M. A. P. (1999). Using GPS to measure traffic system performance. *Computer-Aided Civil and Infrastructure Engineering*, *14*(4), 255–265. https://doi.org/10.1111/0885-9507.00146

Esteve, M., Palau, C. E., Martínez-Nohales, J., & Molina, B. (2007). A video streaming application for urban traffic management. *Journal of Network and Computer Applications*, *30*(2), 479–498. https://doi.org/10.1016/j.jnca.2006.06.001

GDPR European Union. (2020). *GDPR checklist for data controllers*. Proton Technologies AG. https://gdpr.eu/checklist/

Hamed, M. M., Al-Masaeid, H. R., & Said, Z. M. B. (1995). Short-Term Prediction of Traffic Volume in Urban Arterials. *Journal of Transportation Engineering*, *121*(3), 249–254. https://doi.org/10.1061/(ASCE)0733-947X(1995)121:3(249)

Idachaba, F., & Ibhaze, A. (2016). GSM/GPS Assisted Road and Traffic Congestion Detection System. *International Journal of Applied Engineering Research*, *11*(24), 11610–11613. https://www.researchgate.net/publication/315795939_GSMGPS_Assisted_Road_and_Traffic_Congestion_Detection_System/references

INRIX. (2020). *INRIX 2020. Global Traffic Scoreboard*. https://inrix.com/scorecard/

Kalman, R. E. (1960). A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, *82*(1), 35–45. https://doi.org/10.1115/1.3662552

Kan, Z., Tang, L., Kwan, M.-P., Ren, C., Liu, D., & Li, Q. (2019). Traffic congestion analysis at the turn level using Taxis' GPS trajectory data. *Computers, Environment and Urban Systems*, *74*, 229–243. https://doi.org/10.1016/j.compenvurbsys.2018.11.007

Kumar, S. V., & Sivanandan, R. (2019). Traffic congestion quantification for urban heterogeneous traffic using public transit buses as probes. In *Periodica Polytechnica Transportation Engineering* (Vol. 47, Issue 4, pp. 257–267). https://doi.org/10.3311/PPtr.9218

Kurniawan, J., Syahra, S. G. S., Dewa, C. K., & Afiahayati. (2018). Traffic Congestion Detection: Learning from CCTV Monitoring Images using Convolutional Neural Network. *Procedia Computer Science*, *144*, 291–297. https://doi.org/10.1016/j.procs.2018.10.530

Lan, C.-J., & Miaou, S.-P. (1999). Real-Time Prediction of Traffic Flows Using Dynamic Generalized Linear Models. *Transportation Research Record: Journal of the Transportation Research Board*, *1678*(1), 168–178. https://doi.org/10.3141/1678-21

Li, J., Xu, Z., Fu, L., Zhou, X., & Yu, H. (2021). Domain adaptation from daytime to nighttime: A situation-sensitive vehicle detection and traffic flow parameter estimation framework. *Transportation Research Part C: Emerging Technologies*, *124*, 102946. https://doi.org/10.1016/j.trc.2020.102946

Li, S., Li, G., Cheng, Y., & Ran, B. (2020). Urban arterial traffic status detection using cellular data without cellphone GPS information. *Transportation Research Part C: Emerging Technologies*, *114*, 446–462. https://doi.org/10.1016/j.trc.2020.02.006

Martín, J., Khatib, E. J., Lázaro, P., & Barco, R. (2019). Traffic monitoring via mobile device location. *Sensors (Switzerland)*, *19*(20), 4505. https://doi.org/10.3390/s19204505

Nemade, B. (2016). Automatic Traffic Surveillance Using Video Tracking. *Procedia Computer Science*, *79*, 402–409. https://doi.org/10.1016/j.procs.2016.03.052

Pearson, K. (1905). "DAS FEHLERGESETZ UND SEINE VERALLGEMEINER-UNGEN DURCH FECHNER UND PEARSON." A REJOINDER. *Biometrika*, *4*(1–2), 169–212.

Ross, P. van. (1982). EXPONENTIAL FILTERING OF TRAFFIC DATA. *Transportation Research Record*.

Stetsenko, I. v., & Stelmakh, O. (2020). Traffic lane congestion ratio evaluation by video data. *Advances in Intelligent Systems and Computing*, *1019*, 172–181. https://doi.org/10.1007/978-3-030-25741-5_18

Tealab, A. (2018). Time series forecasting using artificial neural networks methodologies: A systematic review. *Future Computing and Informatics Journal*, *3*(2), 334–340. https://doi.org/https://doi.org/10.1016/j.fcij.2018.10.003

Tu, W., Xiao, F., Li, L., & Fu, L. (2021). Estimating traffic flow states with smart phone sensor data. *Transportation Research Part C: Emerging Technologies*, *126*, 103062. https://doi.org/10.1016/j.trc.2021.103062

Vadali, S. (2017). *Day 8: Data transformation — Skewness, normalization and much more*. https://medium.com/@TheDataGyan/day-8-data-transformation-skewness-normalization-and-much-more-4c144d370e55

Wang, B., Wang, J., Zhang, Z., & Zhao, D. (2022). *Traffic Flow Prediction Model Based on Deep Learning* (pp. 739–745). https://doi.org/10.1007/978-981-16-5963-8_100

Weather Webcam EU. (2022). *Уеб камера от София от ж.к. Дружба 1 с панорама към Балкана*. https://weather-webcam.eu/sofia-drujba-letishte-stara-planina-live-kamera/

Zoom Video Communications, I. (2022). *Zoom*. https://zoom.us/

НСИ. (2018). *Прогноза за населението по области и пол*. https://www.nsi.bg/bg/content/2996/прогноза-за-населението-по-области-и-пол

НСИ. (2021). *Естествен прираст на 1 000 души от населението по статистически райони, области и местоживеене*. https://www.nsi.bg/bg/content/2989/естествен-прираст-на-1-000-души-от-населението-по-статистически-райони-области-и

Сантова, А. (2021). *Двойно повече са станали колите в София за 10 години*. Капитал. https://www.capital.bg/politika_i_ikonomika/bulgaria/2021/02/08/4171361_dvoino_poveche_sa_stanali_kolite_v_sofiia_za_10_godini/

Declaration of originality of results

I hereby declare that this dissertation contains original results obtained by me with the support and assistance of my supervisors. The results obtained by other scientists are described in detail and cited in the bibliography.

This dissertation has not been applied for the acquisition of an educational and scientific PhD in another school, university, or scientific institute.

Signature: _____