REPUBLIC OF TURKEY
YILDIZ TECHNICAL UNIVERSITY
GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

MATHEMATICAL MODELLING METHODOLOGY AND A TRAFFIC MODEL BY DATA ANALYSIS AND MACHINE LEARNING

Serdar DEMİR

MASTER OF SCIENCE THESIS
Department of Mathematics
Mathematics Program

Supervisor
Assoc. Prof. Dr. Özgür YILDIRIM

January, 2023

REPUBLIC OF TURKEY
YILDIZ TECHNICAL UNIVERSITY
GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

MATHEMATICAL MODELLING METHODOLOGY AND A TRAFFIC MODEL BY DATA ANALYSIS AND MACHINE LEARNING

A thesis submitted by Serdar DEMİR in partial fulfillment of the requirements for the degree of **MASTER OF SCIENCE** is approved by the committee on 23.01.2023 in Department of Mathematics, Mathematics Program.

Assoc. Prof. Dr. Özgür YILDIRIM
Yıldız Technical University
Supervisor

Approved By the Examining Committee

Assoc. Prof. Dr. Özgür YILDIRIM, Supervisor
Yıldız Technical University _____

Assoc. Prof. Dr. Mutlu AKAR, Member
Yıldız Technical University _____

Assoc. Prof. Dr. Şebnem ÖZDEMİR, Member
İstinye University _____

I hereby declare that I have obtained the required legal permissions during data collection and exploitation procedures, that I have made the in-text citations and cited the references properly, that I haven't falsified and/or fabricated research data and results of the study and that I have abided by the principles of the scientific research and ethics during my Thesis Study under the title of Mathematical Modelling Methodology and a Traffic Model by Data Analysis and Machine Learning supervised by my supervisor, Assoc. Prof. Dr. Özgür YILDIRIM. In the case of a discovery of false statement, I am to acknowledge any legal consequence.

Serdar DEMİR

Signature

This research was supported by Research Fund of Yıldız Technical University by grant No. FYL-2021-4399.

To my current and prospective family

ACKNOWLEDGEMENTS

I would like to thank my esteemed advisor, Assoc. Prof. Dr. Özgür Yıldırım, who has supported me throughout my master's degree; Assoc. Prof. Dr. Şebnem Özdemir, my esteemed head of the department, who make me feel her support despite all her workload; and my collaborate Res. Asst. Merve Gizem Çömlekçi, from whom I have been learning a lot during my research assistantship.

I would like to thank my father Dr. Burhan Demir, who was also in the academy at the past and went through similar paths hence shares his experiences; and my dear mother Yasemin and my beloved brother Sedat for all their supports.

Lastly and most importantly, I would like to thank Lecturer Busenaz Bahadır, who is by my side under all circumstances and makes me feel her support, whom I believe to be my most valuable asset in this life and whom I love more than myself.

Serdar DEMİR

TABLE OF CONTENTS

LIST OF ABBREVIATIONS viii

LIST OF FIGURES ix

ABSTRACT xi

ÖZET xii

1 INTRODUCTION 1

1.1 Literature Review 1

1.2 Objective of the Thesis 2

1.3 Hypothesis 2

2 MATHEMATICAL MODELLING 3

2.1 Definiton of Mathematical Modelling 3

2.2 Mathematical Modelling Procedure 4

2.2.1 Identifying the problem 6

2.2.2 Listing the factors, variables and assumptions 6

2.2.3 Formulating the problem 7

2.2.4 Solving the problem 8

2.2.5 Interpreting the mathematical solution 8

2.2.6 Comparing the results with the real world 8

2.2.7 Presenting the results 9

2.2.8 Completion of the mathematical modelling process 9

2.3 Categorization of Mathematical Models 9

2.3.1 Deterministic and Stochastic Models 10

2.3.2 Empirical and Mechanistic Models 11

2.3.3 Discrete and Continuous Models 11

2.3.4 Dynamic and Static Models 12

2.3.5 Linear and Non-linear Models 13

3 TRAFFIC FLOW THEORY 14

3.1 Traffic Model Types and Key Concepts 14

3.1.1 Fundamental Diagram 16
3.1.2 Microscopic Models 18
3.1.3 Macroscopic Models 22
3.1.4 Mesoscopic Models 23

4 THE TRAFFIC MODEL WITH DATA ANALYSIS AND MACHINE LEARNING 25
4.1 Data Analysis and Machine Learning 25
4.1.1 Data Analysis 25
4.1.2 Machine Learning 27
4.2 Regression Analysis 28
4.2.1 Linear Regression 29
4.2.2 Polynomial Regression 31
4.2.3 Other Types of Regression Analysis 33
4.2.4 Error Calculation for Regression Models 34
4.3 The Model 35
4.3.1 The Data Sets and Techniques Used in the Model 36
4.3.2 Preprocessing of the Data and Statistical Features of the Data . 38
4.3.3 Construction of the Model 42

5 CONCLUSIONS AND FUTURE SCOPE OF WORK 50

REFERENCES 51

PUBLICATIONS FROM THE THESIS 56

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
CA	Cellular Automata
csv	Comma seperated value
CTM	Cell Transmission Model
dsv	Delimiter seperated value
FD	Fundamental Diagram
GM	General Motors
HO	Higher-Order
IMM	İstanbul Metropolitan Municipality
LWR	Lighthill-Whitham-Richards
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
ML	Machine Learning
MSE	Mean Squared Error
RMSE	Root Mean Squared Error
SQL	Structured Query Language
tsv	Tab seperated value

LIST OF FIGURES

Figure 2.1 Modelling cycle for accessible models for computers [4] 5
Figure 2.2 Differences between deterministic and stochastic models for pure birth models [33] 11
Figure 3.1 Time chart of Traffic Flow Theory [38] 15
Figure 3.2 Fundamental diagrams shown in different graphs [5, 6] 16
Figure 3.3 Comparison between fundamental diagram in different planes developed by traffic theory developers [5, 6, 38–41] 17
Figure 3.4 A flow density curve and a bilinear fundamental relation example by Castillo [42] 17
Figure 3.5 Categorization of lane-changing models [58] 21
Figure 4.1 Flow Chart for Data Analysis Process [67] 26
Figure 4.2 Our selected location [70] 35
Figure 4.3 Excel table form of "Hourly Traffic Density Data Set" for January 2020 [72] 36
Figure 4.4 Possible parameters for hourly weather data [73] 37
Figure 4.5 Covid-19 Shutdowns and Related Data 37
Figure 4.6 Statistical Features of the Data Set 39
Figure 4.7 Pearson Correlation Table 40
Figure 4.8 Kendall's Correlation Table 40
Figure 4.9 Spearman's Correlation Table 40
Figure 4.10 Data set after modifications 41
Figure 4.11 Traffic flow in our selected location before and during Covid-19 . . 41
Figure 4.12 Statistical features of our data set after scaling 44
Figure 4.13 Spearman's rank correlation coefficient table after scaling 44
Figure 4.14 Number of vehicles estimation: scaled data set 45
Figure 4.15 Number of vehicles estimation: non-scaled data set 45
Figure 4.16 Average speed estimation: scaled data set 45
Figure 4.17 Average speed estimation: non-scaled data set 46
Figure 4.18 Traffic flow estimation: scaled data set 46
Figure 4.19 Traffic flow training set predictions 47
Figure 4.20 Traffic flow test set predictions 47

Figure 4.21 Number of vehicles training set predictions 48
Figure 4.22 Number of vehicles test set predictions 48
Figure 4.23 Average speed training set predictions 49
Figure 4.24 Average speed test set predictions 49

ABSTRACT

Mathematical Modelling Methodology and a Traffic Model by Data Analysis and Machine Learning

Serdar DEMİR

Department of Mathematics
Master of Science Thesis

Supervisor: Assoc. Prof. Dr. Özgür YILDIRIM

In this thesis, it is aimed to construct a traffic model predicting the average speed of the vehicles, the number of vehicles and the traffic flow for a specific coordinate with the help of machine learning algorithms. For this purpose, firstly, general information about mathematical modelling is shared and then traffic flow theory is discussed. After that, two fundamental components of our model which are data analysis and machine learning are mentioned and the model is constructed. There are various sub-branches of machine learning algorithms including supervised learning, which is suitable for our data set. Regression analysis, which is a subclass of supervised learning, was used in the model. Since regression analysis is a common study area for mathematics, statistics and data science, it is examined from these three aspects. After having applied the fundamental principles of data analysis, data set is used in order to construct our machine learning model. In the model, it is aimed to estimate the average speed, the number of vehicles and the traffic flow of vehicles for different day and hours for the chosen coordinate. The results of the model are compared with the real world data and the accuracy of the model is calculated with various error measurement methods.

Keywords: Mathematical modelling, traffic flow theory, data analysis, machine learning, regression analysis.

YILDIZ TECHNICAL UNIVERSITY
GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

ÖZET

Matematiksel Modelleme Metodolojisi ve Veri Analizi ve Makine Öğrenmesi Yardımıyla Bir Trafik Modeli

Serdar DEMİR

Matematik Anabilim Dalı
Yüksek Lisans Tezi

Danışman: Doç. Dr. Özgür YILDIRIM

Bu tezde, makine öğrenmesi algoritmaları yardımıyla belirli bir koordinattaki araçların ortalama hızını, araç sayısını ve trafik akışını tahmin eden bir trafik modeli oluşturulması amaçlanmıştır. Bu amaç doğrultusunda öncelikle matematiksel modellemeyle alakalı genel bilgiler verilmiş ve ardından trafik akış teorisi anlatılmıştır. Bu aşamanın ardından tezdeki modelin iki temel bileşenini oluşturan veri analizi ve makine öğrenmesinden bahsedilmiş ve modelin oluşturulma aşamasına geçilmiştir. Makine öğrenmesinin çeşitli alt dalları olup bunlardan biri elimizdeki veri seti için uygun olan denetimli öğrenmedir. Modelimizde denetimli öğrenmenin bir alt dalı olan regresyon analizi kullanılmıştır. Regresyon analizi matematiğin, istatistiğin ve veri biliminin çalışma alanına dahil olduğu için bu üç açıdan da incelenmiştir. Veri analizinin temel ilkelerinin uygulanmasının ardından veri seti makine öğrenmesi modelini oluşturmak için kullanılmıştır. Oluşturulan modelde İstanbul'un belirli bir koordinatı için farklı gün ve saatlere göre ortalama hız, araç sayısı ve trafik akışı tahminlenmiştir. Elde edilen sonuçlar gerçek verilerle karşılaştırılmış ve modelin doğruluğu çeşitli hata ölçüm yöntemleriyle hesaplanmıştır.

Anahtar Kelimeler: Matematiksel modelleme, trafik akış teorisi, veri analizi, makine öğrenmesi, regresyon analizi.

YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

1
INTRODUCTION

1.1 Literature Review

Mathematical modelling has a long history and the beginning of it can even be considered as the invention of numbers when is approximately 3000 BC [1]. Evolution of mathematics also leads improvement of mathematical modelling and more complex models take place of primitive ones as time goes by. The reader should see references [1–4] in order to have more knowledge about history of mathematical modelling.

Traffic flow theory and machine learning, on the other hand, pretty shorter history in comparison to mathematical modelling. Greenshield [5, 6] is generally considered as the founder of the traffic flow theory with his development of Fundamental Diagram (FD) (also called as *Fundamental Relation*) in 1934 which he further develops in 1935. These works are the first studies about traffic flow theory and fundamental diagram still constitues a core part of the traffic flow theory.

Lighthill, Whitham and Richards are also important for the traffic flow theory with his pioneer constructions for their traffic models [7–9]. Their model is generally called as LWR (Lighthill-Whitham-Richards) model and it is pretty important in terms of opening a new age in traffic flow theory, by developing a brand new model type which is a macroscopic traffic model.

Beginning of the machine learning history is sometimes accepted as 1943 with the first work on this field by the work of McCulloch and Pitts [10]. This article is considered as the very beginning of machine learning history which examines neural networks. After this article, there are also important milestones such that the work of Alan Turing [11] who is the one of the most important contributors to the Artificial Intelligence (AI) and develops *Turing Test*. In addition to them, Arhtur Samuel [12] is the first person to use "machine learning" term and he also contributes to the machine learning by his studies on creation of checkers playing machine which leaded to a development of machine learning.

There are some studies that models the traffic by regression analysis [13–17] and there are some studies in the last years that measures the effects of weather conditions [18–22] and Covid-19 pandemic shutdowns [23, 24]. Some of the studies are concentrated on İstanbul while the others are models that are developed for other cities. Additionally, they generally investigates these effects separately. However, in this thesis, we will investigate both weather and Covid-19 pandemic effects together on the traffic by using data analysis and regression analysis.

1.2 Objective of the Thesis

In this thesis, it is aimed to create a traffic model for a specific location in Istanbul by using regression analysis with the help of data analysis and machine learning algorithms. The model predicts the average speed of the vehicles, number of vehicles and traffic flow at the given location based on previous traffic data, previous weather data and pandemic shutdown data that are previously acquired. The purpose is to use minimum number of variable to predict these three dependent variables with the maximum possible accuracy.

1.3 Hypothesis

Weather conditions and pandemic shutdowns have an impact on traffic flow and for a specific location, average speed of the vehicles, number of vehicles and traffic flow can be determined by previous traffic, weather and some additional data with the help of a regression model.

2
MATHEMATICAL MODELLING

In this chapter, required knowledge for the rest of the thesis about mathematical modelling and its techniques are presented.

2.1 Definiton of Mathematical Modelling

All mathematical models have a common purpose which is to solve some occurring problems on the present conditions on that date. However, complexity of these models differs significantly, and hence, their eligibility to solve the problems efficiently is different. Therefore, it is useful to give a definition of mathematical modelling which contains this common purpose, then it is logical to clarify points elaborately.

Definition 2.1. (Mathematical Modelling): Mathematical modelling is a process that is a transition of a real world problem into mathematics. [2, 25]

Generally, every real world problem experience a simplification during this transition. The main reason is that some problems have too many variables to handle or they are too complicated to create a meaningful mathematical representation without simplifying.

This transition generally appears as a construction of a relationship between variables. However, there is no restriction about what kind of relationship should be constructed. While sometimes quite simple linear functions demonstrate this relation, it is also possible to have enormously difficult partial differential equations to indicate it. Surprisingly, it is not necessary for a model to have a rigorous equation to represent this relationship. Some models are just constructed with a mathematical mindset by building logical connections between variables which puts forth a relationship and this situation is generally enough for them to qualify as a mathematical model.

While the strictness of mathematics arise in a lot of areas, it may be seen that this definition is not as firm as a definition of *limit* or a *field*, for example. Since this

definition is not a rigid definition, we can consider "mathematical model" as a quite general term. Therefore, the definition allows us to consider a lot of systems or formulas as a mathematical model.

Since this definition does only define the aim of mathematical modelling, steps or methods should be clarified in order to have a better understanding about how mathematical modelling process is conducted. For that reason, the procedure for mathematical modelling will be investigated in the next section. Because, methods, steps and the process will be a decider about the validity and the quality of the model.

2.2 Mathematical Modelling Procedure

In the previous section, we define mathematical modelling as a "transition of a real world problem into mathematics" in definition (2.1). Now, we will dive deeper to discover how this transition is carried out.

Mathematical modelling has some basic principles that every model should have. Steps of constructing the model can be slightly different from model to model or source to source, however, these fundamental principles are always taken into account.

According to Witelski and Bowen [26], mathematical modelling has two main steps consisting of the *formulation phase* and the *solution phase*. In the formulation phase, the problem is defined with the help of principles or laws in branches of science and the necessary elements for the model such as factors, parameters, initial conditions, boundary conditions are determined. Then, in the solution phase, the problem is solved analytically and/or numerically. If solution cannot be acquired by investigating the problem completely, problem is reduced into the solvable parts then they are reassembled. After these two steps, mathematical model is considered as completed. Apart from these two steps, results obtained in the solution phase is also compared with the real world in order to verify the model.

Certainly, it is possible to divide this process into more parts. For instance, Marion [3] offers a 4-stages allocation about mathematical modelling process which includes building, studying, testing and using. It is generally optimistic to expect the modeling process to go smoothly. Therefore, it is suitable to correct the errors in the studying and testing phase by returning to the building phase. In this case, if any change is experienced in the model, the studying and testing phases must be repeated.

Schichl [4] analyzes this process a little differently and divide this process into 5 parts which are real-world problem, construction of model, collecting data, computing the

solution and interpreting the results. However, when it comes to the models accessible for computers, his process proposal is not limited to this and he shared a much more detailed guideline as follows:

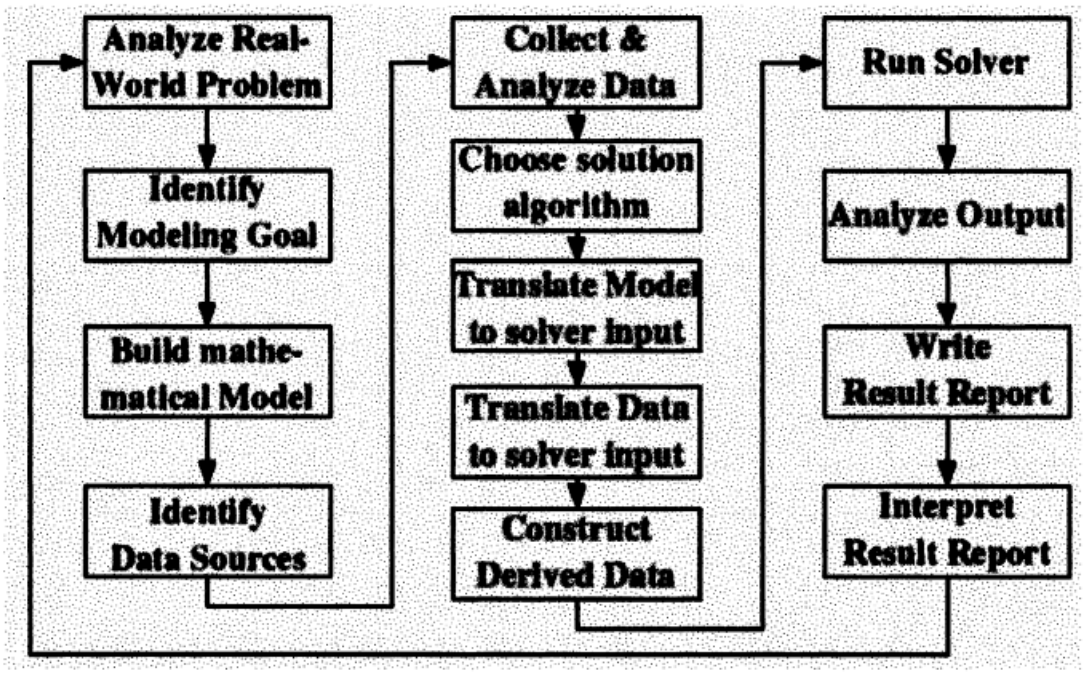


Figure 2.1 Modelling cycle for accessible models for computers [4]

All mentioned fragmentations of the process are reasonable ones and they all can be used in order to discuss mathematical modelling process. The most detailed one is the cycle (2.1), however it is a specific situation and the cycle stands for the models with computer involvement. We want to use a neither too simple nor too complicated and a widely-admitted modelling process allocation and there is a process which perfectly satisfy our need. Below, there is a mathematical modelling process divided into 7 stages which are often considered as steps of mathematical modelling process and also used by some authorities [2, 25, 27, 28].

Here are the steps:

1. Identifying the real problem.
2. Listing the factors, variables and assumptions.
3. Formulating the problem.
4. Solving the problem.
5. Interpreting the mathematical solution.

6. Comparing the results with the real world.

7. Presenting the results.

In some sources, it is usual to encounter that 3rd and 4th stages which are formulating and solving are combined as a single stage which is also a case in [25].

Before investigating these 7 stages closely, there is a point that should be indicated. In the modelling process, generally not everything goes perfectly as it is also indicated by Marion [3], and if there is a mistake in the model which is observed in a stage, it is necessary to re-build the model by turning back to the construction. In these circumstances, it is vital to apply all stages again, because, theoretically the model is a new one and it should be evaluated by a brand new process.

2.2.1 Identifying the problem

This is the very first step of the mathematical modelling process and here, the modeller should clarify the objective of the model precisely and this objective should be indicated at the beginning of this modelling cycle [29].

Decisions about the model is made upon questions. Questions about the modelling phase can broadly be defined under three categories which are *evaluation questions*, *detection questions* and *design questions* [26]. Evaluation questions concern about predicting other features of the system in a quantiative way when the required knowledge about the system available. Detection questions, on the other hand, appears when some information is not available and they are used to find them in a reverse manner. Finally, design questions are interested in optimization which are used to find the best solution for demanded situation.

Apart from these questions, the modeller should also decide the type of the model in terms of classifications [29], which will be mentioned in the next sections. After these questions and determining the class of the model, the problem is able to be described more clearly and purpose of the mathematical model will be determined in a sharper way.

2.2.2 Listing the factors, variables and assumptions

This stage is related with what will be included to the model. Real world problems has an advanced level of complexity which makes formulation or solution phase harder. At this point, the modeller should determine how much the real world problem will be simplified. This simplification usually loosen the connection between the model and

the reality, however, it is not possible to adapt the model to the real world completely. Most of the time, mathematical form or representation of every effective variable on the problem cannot be determined easily and this will make the construction phase challenging. On the other hard, sometimes problems occur in the solution phase rather than the formulation phase. The desire of reflecting the real world completely to the model generally results with a hard-to-solve or an unsolvable model.

Collecting the data and analysing the theoretical laws, rules, equations or the data is generally the core part of this stage [29]. Sometimes, more data is required to meaningful relationship between variables, in this case, data collection must be restarted. Data should be collected and displayed objectively, data sources must be shared and the data should be correctly analysed in order to construct an accurate relationship between variables. [25].

Bender [30] propose that variables can be separated into three categories which are neglected variables, exogenous variables and endogenous variables. Here, neglected ones are straightforward and they are simply disregarded. Exogenous variables are independent variables whose behaviour is not designed in the model, however, endogenous variables are designed and studied in the model.

After these variables are determined, the modeller should make some assumptions about how these variables or factors related with each other and how they affect the solution of the model. Based on these assumptions, the model is ready to be constructed mathematically.

2.2.3 Formulating the problem

This stage is where the problem is translated into mathematics. After variables are specified, it is now time to build the relationship between them mathematically. This relationship might be linear or non-linear, variables may be identified as discrete or continuous or the equations for the model can be formed using physical equations or probabilistic functions. They will all shape the character of the mathematical model.

This part is often considered as the hardest part of the mathematical modelling process [2]. It is generally due to the fact that not everything in the universe can be easily expressed in mathematics in a correct way. After this formulation, mathematical modelling construction will actually finish and it will be time to solve the problem.

2.2.4 Solving the problem

Techniques used in the solution phase usually depends on what kind of model is constructed. Sometimes, especially for basic models, solution phase is easily conducted. However, complexity of the model generally make solution more sophisticated to handle. Solution of the equations or systems in the mathematical model is performed analytically and/or numerically and there are a lot of techniques that can be used in this stage with respect to the model type.

Since computers are generally better at computation relatively to the human brain, computers are also preferred to be used when intensive computation is needed in the solution phase of the model. Softwares, programming languages or even a single library in a programming language can dramatically increase the speed of this calculation.

It is also conventional to use some graphical or tabular form of the solution [29]. This situation also helps the interpretation stage by making the results clearer.

2.2.5 Interpreting the mathematical solution

In this step, solution obtained in the previous part is interpreted. This interpretation is usually about deciding whether the behaviour of the model is expected or not. [27]. It should also be checked that if the solution is valid mathematically, additionally, this part is closely related with the comparison of results with the real world. The modeller should check if the results of the solution is consistent and it satisfy the assumptions that is made in the previous parts. If there exist any unforeseen interpretation, it is needed to go back to the second phase, listing the factors, variables and assumptions, then construct the model by trying to remove inconsistencies from the model.

2.2.6 Comparing the results with the real world

In this stage, validation and evaluation of the model are performed. [2, 27] If there exist a correlation between solutions and observed results in the real world, and if the solution is valid mathematically, it can be said that mathematical model describe the real world situation properly. However, if there is no correlation of the correlation is a loose one, the modeller should back to construction phase of the model for both making new assumptions and creating a new mathematical formula.

Depending on the type of the model, there are several types of comparison tool to compare the results with the real world. Hypothesis testing, error calculations, cross validation, existence and uniqueness of the solution, etc. are all methods to

verification or validation tools for the models, however, methods for verification is different with respect to the type of the models.

2.2.7 Presenting the results

After the previous steps completed and the model is verified by comparing the results with real world, modelling process is generally finalized by presenting the results. This presentation might be in quite different shapes. An article, a conference paper, a report or even a presentation is a used way to present the results.

2.2.8 Completion of the mathematical modelling process

After all these steps, mathematical modelling process is accepted to be completed. As it is indicated earlier, the one should back to early stages when s/he encounters unexpected results in some stage or results are not coherent with the real world results.

This process is more or less similar for all mathematical models. However, how these principles are applied to the model might be quite different from model to model. For that reason, it is also needed to share some widely-used classifications of mathematical modelling in order to have better view how application of this process differs with respect to the type of models.

2.3 Categorization of Mathematical Models

Mathematical models has several classifications with respect to different authorities in terms of distinct properties of the models. These sub-classes can be composed in diverse ways. For instance, whereas some classifications take mathematical tools and methods into account, some others regard mathematical features of parameters as a separator.

Naturally, this classification process results with an outcome. Since there are different classifications for different aspects, a model will belong to a class in each of these classifications. As an example, creating a model by utilizing physical equations or a model by using data will result two different models which are mechanistic and empirical, respectively. However, these models are also discrete or continuous models and they are also linear or nonlinear models in other type of classifications.

Neural networks, statistical models, computational models, model theory or operations research can all be taken into account as a sub-section of our broad mathematical modelling. These models and all model types belong to some classes

that will be mentioned later, however, apart from types of the models that will be mentioned later, some models has some specific names for them regarding the methods used in it. For example, *compartmental modelling* stands for the models that have an input and an output such that population models which are concentrated on population growth or decline. As another example, agent-based models can be given which are actually computational and stochastic models. *Analytical hierarchical process* is a process that can additionally be accepted as a mathematical model which is also related with operations research.

Some models barely have mathematical tools or methods in it or they just have an analytical thinking process which is related with mathematics itself. Nonetheless, these models can be considered as a mathematical model owing to the fact that mathematical modelling is a quite general term as it is indicated earlier. Also, taking into consideration the things above, listing the most commonly used classifications will be helpful for emphasizing the substantial points in mathematical modelling.

2.3.1 Deterministic and Stochastic Models

Mathematical models are classified into two categories with respect to the type of data as deterministic and stochastic. In deterministic models, all data in the model are known correctly, whereas in stochastic models, data are partially or completely have some probabilistic behaviour [31]. The main difference between these two types of models is about the outcome they predict. Deterministic models predict always the same outcome while stochastic models might predict possible outcomes depending on the properties of the data with the effects of statistical aspect of the model [3].

Construction techniques for these two models are different due to the nature of these model types, additionally, stability of results are different as well. Even if deterministic models have unique solutions, minor perturbations of initial conditions may cause devastating changes in the solution of the model, which yields an instability; on the contrary, in stochastic models, state variables are defined by probabilistic distributions and hence less uncertainty is experienced in stochastic models [32].

As the general nature of these two types of models are taken into consideration, it is expected and actually observed that they are used in diverse situations. Since stochastic models are concentrated on describe patterns and making forecasts, physical concepts are generally not involved in these type of models; therefore, deterministic models are generally used where physical coherence is needed and hence, it is often preferred by scientist and engineers [32].

Pure birth model is an example of models in which both deterministic and stochastic methods can be used in it. The table about differences between these two types in this kind of models shared in the table below:

Characteristic	Deterministic model	Stochastic model
independent variables	time	time and number of individuals
number value x(t)	number	random value
state function x = x(t)	function	random process
general law kxdt	population growth over time dt at number x	probability of the birth of an individual in time dt for the number x
initial number x(0)	number	number
state equation	differential equation	differential-difference system
result mexp (-kt)	law of population change	expected value

Figure 2.2 Differences between deterministic and stochastic models for pure birth models [33]

2.3.2 Empirical and Mechanistic Models

Another classification for mathematical models are about the knowledge used in it. In mechanistic models, laws or rules in science are used and the model is constructed theoretically which generally considers a single stage in the mathematical modelling process [3, 34]. On the other hand, empirical models are not interested in the mechanism of the model and they are constructed with the help of the data available by selecting variables and trying to fit the data to the real world situation [29].

There are, actually, no model which is entirely empirical or mechanistic. Because, it is not possible to define a real world phenomenon only empirically or mechanically which also yields that every empirical model will include some mechanical parts and every mechanical model will have some empirical parts [35].

2.3.3 Discrete and Continuous Models

The classification regarding discrete and continuous models is determined by the features of the variables whether they are discrete or continuous. First of all, discrete variables take values that are integers or more specifically natural numbers, however, continuous variables may take all values such as real numbers within an interval [25]. This situation leads different model types with respect to variables in the model.

The fact above uncovers a significant point that discrete variables have immediate

changes while continuous variables do not have. For example, number of students in a classroom can be measured by natural numbers and this number will increase by 1 if another student comes to the classroom and will decrease by 1 if some student leaves the classroom. However, continuous variables change over time. As an example, an accelerating vehicle or a vehicle slowing down will experience a velocity change, however, even if this velocity change appears incredibly fast, it does not take place immediately. As a result, this change is observed over time, although, this time interval might be quite short.

In a probabilistic and statistical view, discrete and continuous random variables have different distributions that have different characteristics. While discrete random variables have *probability mass function*, continuous random variables have *probability distribution function* for their distributions [36] Additionally, their *cumulative distribution functions*. have different characteristics and these differences affect how model is constructed.

Although the things indicated above put forth the differences between these two class of models, it should be noted that discrete and continuous models are not as diverse as black and white. For example, being continuous of a variable depends on our measurement, such that time is generally a continuous variable and it can be partitioned into discrete parts by creating time intervals such as days, hours or minutes. In this setup, time might be considered as a discrete variable because it will have jumps on every day or every hour. In some models, deciding whether a variable is a discrete or continuous is a straightforward choice, whereas it sometimes depends on our choice [25].

There are also some methods which can be used in order to approximate discrete models to continuous models and vice versa. For example, *normal distribution* is continuous, however, with the help of *central limit theorem*, it became applicable also on discrete distributions and as a result *binomial distribution* which is discrete is approximated with the *normal distribution* [37]. Moreover, *Finite Difference Method* is an example of how a continuous model can be approximated as a discrete model [2]. These approximations create flexibility to handle the models more efficiently and provide a better fit to the real world data.

2.3.4 Dynamic and Static Models

Another classification for mathematical models is about whether they are dynamic or static and this distinction is decided by the state of the system. In dynamic models, the system changes over time and as a result the knowledge that we get from the

system changes; however, in static models, the system does not change and/or model is constructed as if the system stays in the same conditions [31].

2.3.5 Linear and Non-linear Models

Classification of linear and non-linear models are determined by the type of the data and the linear character of the mathematical functions in the model [31]. As can be expected, a linear function will result a linear mathematical model, whereas a non-linear one creates a non-linear model. This linearity may arise in different ways, for example, the linear model can consist of a linear regression or a linear differential equation. This distinction is important also for that these two class of mathematical models use different programming tools when it comes to the mathematical programming models [31].

As in the case in 2.3.3 for discrete and continuous models, there are also some techniques to define a linear function in a non-linear way and vice versa. For example, it is possible to approximate non-linear models by linear models with the help of *linearization*, which is a process of approximation of a function linearly. There will generally be no need to approximate a linear function to a non-linear one, since the linear functions are usually easier to handle and already give appropriate results.

Next section we will discuss traffic flow theory in which we will see some applications of mathematical modelling techniques. Additionally, this will be a key part in our model.

3
TRAFFIC FLOW THEORY

İstanbul is one of the most populated cities in the world and it has one of the highest density among all cities. In addition to this situation, when the fact that most of the population live in specific parts of the İstanbul is taken into account, traffic issue in İstanbul is inevitable. For that reason, it is aimed to create a traffic model for İstanbul. Therefore, some general principles and methods will be mentioned about traffic flow theory before constructing our own model.

3.1 Traffic Model Types and Key Concepts

Traffic flow theory is usually investigated under 4 main categories which are fundamental diagram, microscopic modelling, mezoscopic modelling and macroscopic modelling. They all have some different approaches to the traffic flow theory, for example, microscopic models concentrates on individual cars on the traffic whereas macroscopic models analyses the traffic as a whole. Although they have their own techniques to modelize the traffic flow, they also have some common principles.

Conventionally, there are three main variables drawing the attention which are generally encountered in majority of the models in each type above. These variables are *flow(q), speed(v)* and *density(ρ)* and they have the following relationship:

$$q(x,t) = v(x,t).\rho(x,t) \tag{3.1}$$

This equation (3.1) is the main form of the equation for flow-speed-density relationship and it is most widely used in the construction of the fundamental diagram and it lightens the way through the development of traffic flow theory. It is quite usual to encounter with different versions of this equation, however, it should be noted that there are numerous equations in the traffic flow theory which are derived from the equation above.

Here, it is useful to have a time chart of models. Therefore, a time chart is shared below:

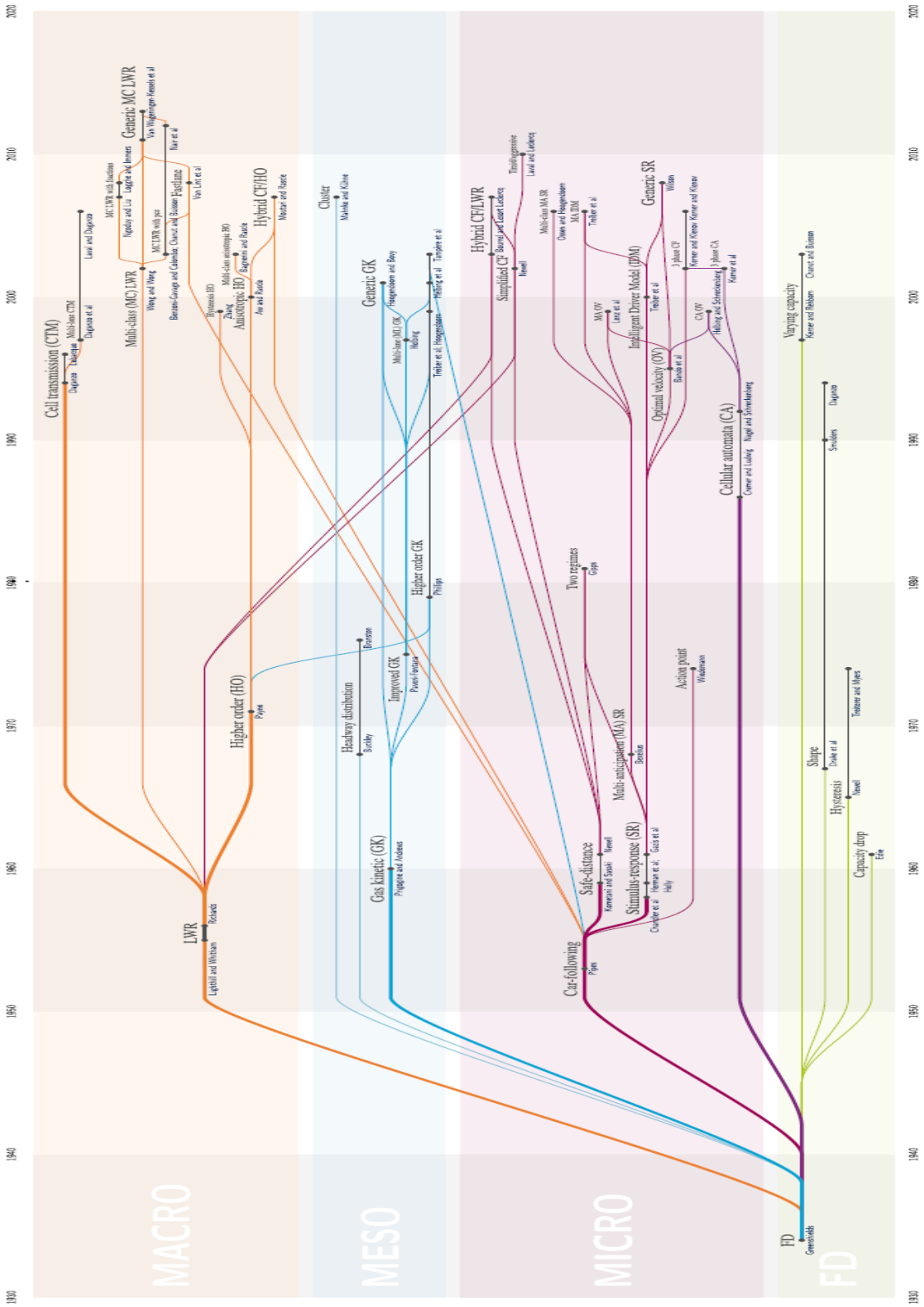


Figure 3.1 Time chart of Traffic Flow Theory [38]

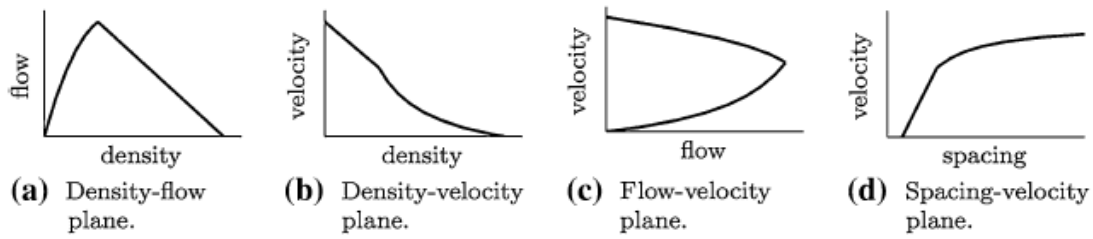
In order to be familiarized deeper to the traffic flow models, aforementioned 4

categories above will investigated in the following sections.

3.1.1 Fundamental Diagram

Greenshield [5, 6] is generally considered as the founder of traffic flow theory by his construction of *fundamental diagram*. In his studies, he constructs a relation between two variables which are average distance between vehicles (which can be named as *spacing* and denoted as *s* which can also be the fourth variable in addition to three previously mentioned ones) and the speed. Additionally, it is also possible to derive other variables from this relation such as density and flow which demonstrate average number of vehicles per unit length and average number of vehicles per unit time, respectively [38].

Fundamental diagram for different planes can be seen in the following graphs:



(a) Density-flow plane. **(b)** Density-velocity plane. **(c)** Flow-velocity plane. **(d)** Spacing-velocity plane.

Figure 3.2 Fundamental diagrams shown in different graphs [5, 6]

Afterwards, studies on the fundamental diagram expanded and this situation helped improvement of the fundamental diagram. They all represent fundamental diagrams, however, they are slightly different. This difference is mainly due to development of the traffic flow theory and different perspectives about it.

In addition to Greenshields [5, 6] there are well-known contributors of traffic flow theory who are Daganzo [39], Drake [40], Smulders [41] and Castillo [42] and they played role in improvement of both fundamental diagram and flow-speed-density relationship. Their diagrams slightly differs, however, they all represent the fundamental diagram and obey the main principles of flow-speed-density relationship.

In the following graphs, different fundamental diagrams constructed by different contributors are shared in order to make the comparison more observable:

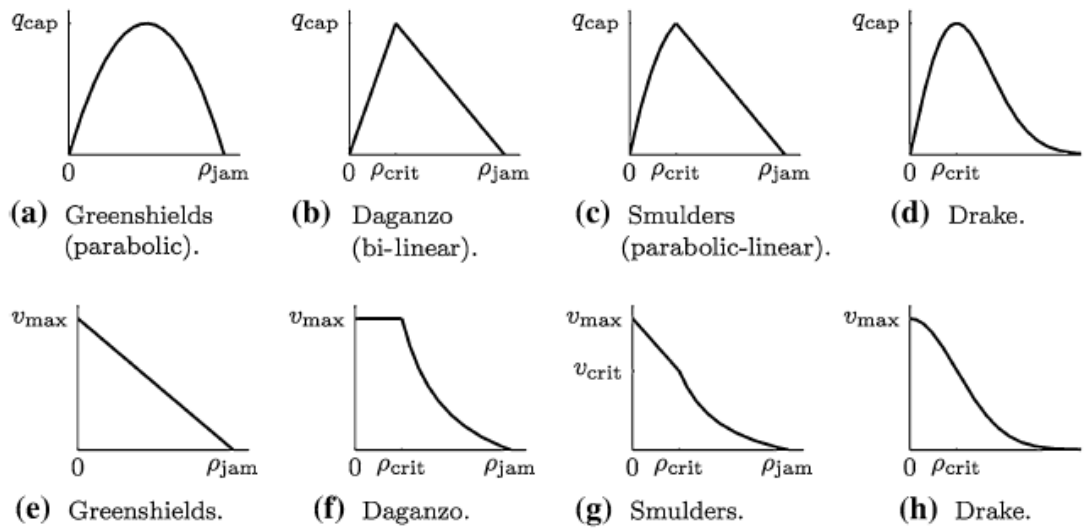


Figure 3.3 Comparison between fundamental diagram in different planes developed by traffic theory developers [5, 6, 38–41]

In addition to previously mentioned ones, Castillo [42] contributes the traffic flow theory by developing fundamental diagram and he is one of the latest contributors to the theory. He also constructs a fundamental diagram as following:

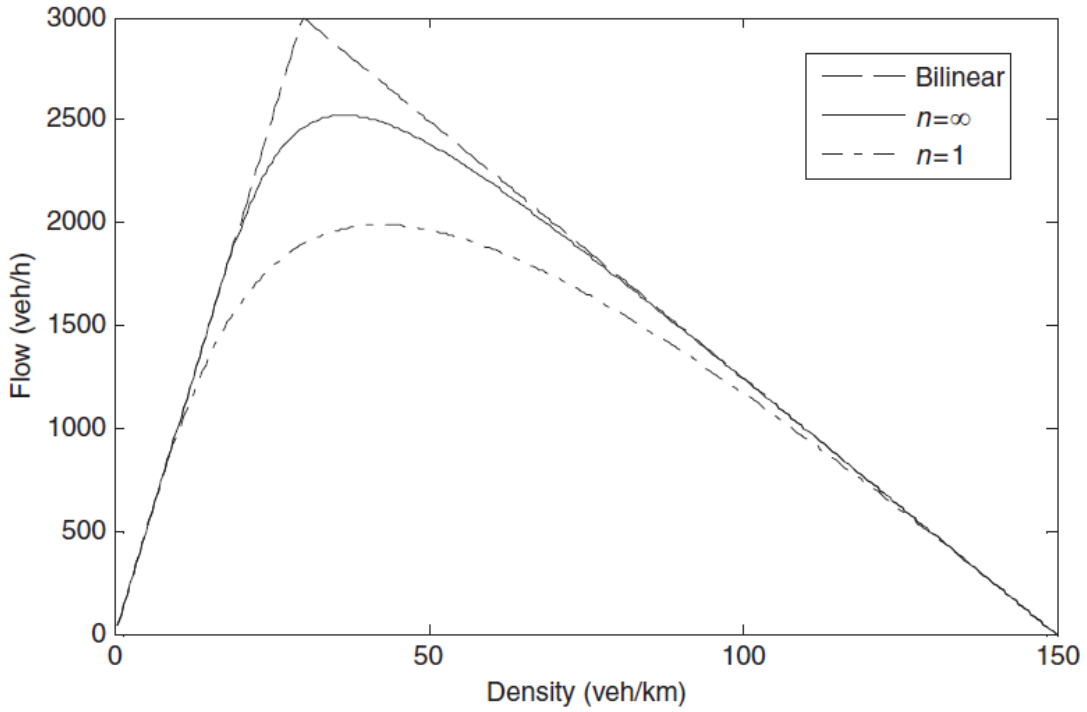


Figure 3.4 A flow density curve and a bilinear fundamental relation example by Castillo [42]

Castillo [42] also states that even if formulations can differ, there are some features which are frequently supposed for the flow-speed-density relationship:

- speed is between 0 and V_{max}
- Density is between 0 and P_{max}
- speed is 0 at maximum density and speed is maximum at zero density
- Flow is 0 at zero density and maximum density
- The fundamental diagram is strictly concave

The fundamental diagram and flow-speed-density relationship yields an important result. Because of the flow-speed-density relationship, maximum flow will appear at average speed and density values [43]. This is generally considered as one of the biggest challenges in the traffic flow theory because when density is low, an increase in density will result an increase in flow; however, when density is high, an increase in density will result a decrease in flow.

Taking everything into consideration, fundamental diagrams are tremendously significant for all subclasses of traffic flow theory [38]. For example, macroscopic and mezoscopic models pronouncedly contain the fundamental diagram and microscopic models uses principal assumptions on the fundamental diagram. Therefore, it should be kept in mind that main principles in this section will somehow apply all the models in the following sections.

3.1.2 Microscopic Models

Microscopic models examines the traffic by concentrating the vehicles individually [43]. These models assert that drivers adjust their behavior with respect to the leading vehicle. Microscopic models are generally classified under two categories depending on the type of this behavior: car-following models and lane-changing models.

Car-following models concentrates on the vehicles continue its way by not changing lane, while lane-changing models examines the vehicles changing lane as its name also indicates [38]. Both two classes have some sub-classes which are constructed by derived models and obeying the original method of his class, car-following or lane changing. It should be noted that majority of microscopic models compose of car-following models. Additionally, lane changes are also investigated in macroscopic models, however, we will investigate it under the microscopic models because microscopic lane-changing models are more common.

3.1.2.1 Car following models

Car following models can be considered as the most commonly used microscopic models and they have various extensions. These models concentrated on the distance between cars and they have three main subbranches which are safe-distance models, stimulus-response models and action point models [38].

In order to indicate their order, vehicles are enumerated in macroscopic models [38]. For example, the vehicle which is considered is numbered as *n*, its leader is *n-1* and its follower is *n+1*. While modelling the movement of the vehicle, *distance(x)*, *speed(v)* and *acceleration* are widely used and their equations are shared below:

$$v = \frac{dx}{dt} \tag{3.2}$$

$$a = \frac{dx}{dt} = \frac{d^2x}{dt^2} \tag{3.3}$$

where x is the position of the vehicle in front.

The first car-following model is constructed by Pipes [44] in 1953 and it is developed as a safe-distance model [38]. Safe-distance models investigates the traffic depending on the measurement of safe following distance between vehicles. Pipes investigates the safe distance between cars with the following equation that measures the distance between nth vehicle and (n+1)th vehicle.

$$x_{n-1} = x_n + d + Tv_n + l_{n-1}^{veh} \tag{3.4}$$

In this equation, d is the distance between two vehicles, Tv_n is the legal distance between vehicles and l_{n-1}^{veh} is the length of the vehicle in front, x_{n-1} and x_n are the positions of the $(n-1)th$ and nth vehicles respectively.

Stimulus-response models are another type of microscopic models which is widely used and their invention arise mainly because of the lack of ability of safe-distance models about reflecting the reality completely in some cases. Chandler et al. [45], Herman et al. [46], Gazis et al. [47] are all important contributors to the stimulus-response models and this model type sometimes referred as GHR (Gazis, Herman, Rothery) or GM (General Motors) model.

Stimulus-response models describe driver reactions more realistically which depends on distance and speed changes with respect to the vehicle in front [48]. This model

type also includes finite reaction time into the model and they are also applicable to the cases in which traffic flow is relatively busy. According to this model type, drivers tend not to leave so much space in front, so other vehicles cannot enter the gap, however, they also pay attention to keep the safe distance between vehicles.

Stimulus response models can be formulated as following [49]:

$$a(n+1)_{t+\Delta t} = \lambda_{m,l}[v(n+1)_t]^m \frac{v(n)_t - v(n+1)_t}{[x(n)_t - x(n+1)_t]^l} \tag{3.5}$$

In this equation (3.5), Δt represent reaction time, $\lambda_{m,l}$ represents a sensitivity parameter, m and l are calibration exponents [49]. This equation can also be simplified for $m = 0$ and $l = 0$ as follows:

$$a(n+1)_{t+\Delta t} = \lambda_{0,0}[v(n)_t - v(n+1)_t] \tag{3.6}$$

Another type of microscopic models are the action point models which are sometimes mentioned as psychophysical models. Action point models are produced because the only human element was finite reaction time in safe-distance and stimulus-response models, however, there are other effects on driver behaviour in reality because humans are not eligible to satisfy some requirements in the other two models [48]. For instance, drivers may not be able to realize stimulus below a perception threshold or they may not use acceleration and brake pedals in a precise way. In action point models, these factors are also included into the model.

Michaels [50] played an important role in construction of these models and action point models developed by studies of Brackstone and McDonalds [51, 52]. Brackstone and McDonalds reveal in their studies that it is difficult to come to a conclusion in validation of these models which is fundamentally due to the fact that calibration of the elements of the model are not successful [48].

Another microscopic model that is needed to mention is Cellular Automata (CA) which is younger than the other microscopic models. Main idea in this model is to partition the road into cells and determining whether there is a vehicle in this cell or not [38]. These model type allows computers to do simulation in an easier way [53]. Cremer and Ludwig [54] were the creators of this model type and another CA model developed by Nagel and Schreckenberg [55] is sometimes considered as prototype of CA models [53].

3.1.2.2 Lane-changing models

Driving behaviour is one of the key elements of microscopic models and lane changes plays a significant role in this behaviour. These lane changes can be considered as a fundamental driving task and it affects both traffic flow characteristics and traffic safety [56].

Lane changing is a pretty sophisticated behavior in the real world because of different reasons such as driver distraction, environmental effects or geometric properties of the road [57]. This situation yields a significant result that creating a model reflecting the real world is a tough problem and one of the most helpful thing to solve this issue is to have a large data set that includes improved measurements.

Below, a general categorization of lane-changing models are shared:

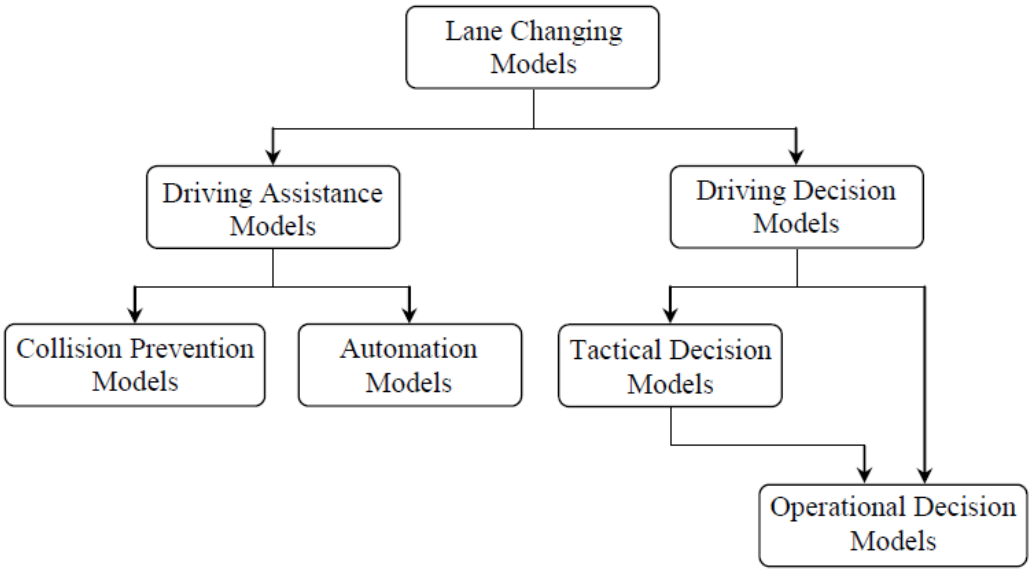


Figure 3.5 Categorization of lane-changing models [58]

Driving assistance models evaluates lane changing performance of vehicles based on the angle of steering wheel and lateral motions [58]. Collision prevent models aim road safety and they are constructed in order to help drivers about a safe lane changing during their maneuvers. Automation models are developed to carry out driving tasks partially or completely, these models also include side crash avoidance systems.

Driver decision models investigates lane changing behaviour under different conditions. It is possible to add another subclass to the figure (**??**) by dividing these driving decision behaviours into three parts which are strategical, operational and strategical [59, 60]. Strategical behaviours include planning the route and determining the goals, tactical behaviours is related with the calculations about

executing of the decision, finally, operational behaviours is where maneuvers are performed. All these steps occur just in seconds and all these behaviours are parts of a process.

3.1.3 Macroscopic Models

Macroscopic models are the models that investigate the traffic as a whole by concentrating on large scale conditions. There are three main variables in this model type, which are also mentioned in the beginning of this chapter and they are flow, density and speed.

Lighthill and Whitham [7, 8] and Richards [9] constructed a traffic model which is called as LWR (Lighthill-Whitham-Richards) model. This model type is sometimes referred as kinematic wave model and it can be considered as a prototype for macroscopic traffic models.

LWR model can be derived by obeying three rules which are the following [61]:

1. Constitutive Law: $q = kv$
2. Conservation Law: $\frac{\partial k}{\partial t} + \frac{\partial q}{\partial x} = 0$
3. Fundamental Diagram: $v = V(k)$

The second rule and the flow-density relationship results the LWR model such that:

$$k_t + Q(k)_x = 0 \tag{3.7}$$

which is a partial differential equation.

LWR model is a simple and useful model type and this situation caused this model type to lead the construction of other models, however, it has some drawbacks. For example, LWR model does not express the nonequilibrium situations such that existence of traffic jams, because mean flow speed is supposed to be equal to the value at equilibrium at each time instant [62]. In order to overcome this situation, Higher-Order (HO) models developed.

Payne [63] is known as the first developer of the Higher-Order models. He derived this model by using stimulus-response model which is a car-following model. The model takes its name from usage of a fundamental diagram and two coupled partial

differential equations. Payne identifies three different terms in his model in order to remove the lack of ability that LWR experience. These terms are convection, relaxation and anticipation which can be defined as following [64]:

- Convection: Express changes in the mean speed which arises because of inflowing and outflowing vehicles.
- Relaxation: Express the tendency of traffic flow in order to relax to an equilibrium speed.
- Anticipation: Express anticipation of drivers on downstream traffic conditions changing spatially.

Payne's equation can be seen as follows [64]:

$$\partial_t V + V\partial_x V = \frac{V^e(r) - V}{T} - \frac{c_0^2}{r}\partial_x r \tag{3.8}$$

Aw and Rascle [65] developed further Payne's model by adding a pressure term $p(\rho)$ and the equation can be seen as follows [43]:

$$\frac{\partial}{\partial t}(v + p(\rho)) + v\frac{\partial}{\partial x}(v + p(\rho)) = 0 \tag{3.9}$$

There is also another type of macroscopic models which is called as the Cell Transmission Model (CTM). In Cell Transmission Models, two new variables are introduced which are supply and demand. This model is also a kinematic wave model and it can be described as discretization of LWR model [43].

Overall, all macroscopic models concentrates on large scale conditions rather than concentrating on the vehicles individually on the contrary to microscopic models. Now, we will investigate another type of traffic model which lies between microscopic models and macroscopic models: mesoscopic models.

3.1.4 Mesoscopic Models

Mesoscopic models are introduced in order to fill the gap between microscopic models and macroscopic models and number of mesoscopic models are relatively smaller in comparison to microscopic and macroscopic models. There are two main types of mesoscopic models which are gas-kinetic models and hybrid models.

In gas-kinetic models, traffic is modelled by imitating the movement of gases. This model type is constructed by Prigogine and Andrews[66] in 1960 and it has been developed by improved gas-kinetic models, higher order gas kinetic models and generik gas-kinetic models throughout the history of traffic flow theory.

Prigogine and Andrews [66] utilizes Maxwell-Boltzmann collision equation, which is an equation in kinetic theory of the gases, in order to model the traffic flow by observing the behaviour patterns of small particles and gases. In their work, relaxation term and the collision term are derived and a time-independent equation equation is obtained. With this process, it is aimed to obtain information about changes that causes a relatively free traffic flow to turn a traffic jam.

The other type mesoscopic models is hybrid models are generally considered as simulation models and they combine different approaches that microscopic models and macroscopic models have. Therefore, features of hybrid models reflects different characteristics. This situation can be seen as a huge advantage, however, it brings a disadvantage that hybrid models are flexible models but they are highly complex.

In next section, machine learning algorithms and statistical methods will be shared in order to investigate how mathematical modelling using machine learning and statistics constructed.

4
THE TRAFFIC MODEL WITH DATA ANALYSIS AND MACHINE LEARNING

In this chapter, we will construct a multivariate polynomial regression model for estimating average speed of the vehicles, number of vehicles within a time interval, and traffic flow on a specific location. In order to clarify some points, a brief information about machine learning and regression analysis will be given which will be beneficial in the construction process of our model.

4.1 Data Analysis and Machine Learning

Data analysis can be seen as a prerequisite of a construction of a machine learning model. Therefore, general information about data analysis will be shared in the next section and then machine learning will be investigated.

There are some important tools that can be used for data analysis and machine learning. Data analysis can be done even through basic spreadsheet applications such as Microsoft Excel, Google Spreadsheet, Apple Numbers and many others. There are also some dashboard applications that are used for revealing analytics from the data and share the results with a better visualization. There are a lot of examples of these kind of applications such that Microsoft Power BI, IBM Cognos, Tableau, etc. Additionally, there are two main programming languages for machine learning which are Python and R. In our traffic model, Excel and Python will be used for the data analysis and Python will be used for model construction by machine learning but it should be noted that other mentioned applications are also eligible to do these operations.

4.1.1 Data Analysis

Data analysis can be most basically defined as extracting information from the data in order to understand the relationship between variables. This extraction process

contains different techniques and can be partitioned into different stages. There is not a strict classification of these stages but an example of a data analysis flow chart can be seen below:

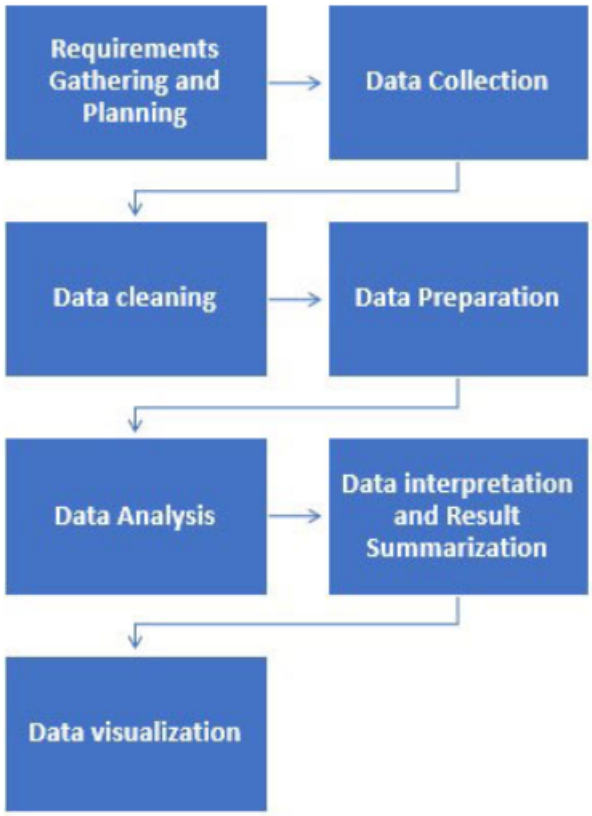


Figure 4.1 Flow Chart for Data Analysis Process [67]

Data analysis process starts even before the data collection. First of all, problem should be defined as in the case of mathematical modelling process (2.2). Then, expected qualitative and quantitative characteristics of the data should be clarified such that what kind of attributes are needed or how much data will be sufficient. After that, data is collected from different sources. After these stages above, if there are enough and useful data, it is time to clean and prepare the data for the analysis. However, during preprocessing or in any other stages it may be encountered that some additional data is needed in order to make a better analysis.

If it is assumed that all the data collection process is done, then it is time to *clean* the data. Presumably no data is perfect and it requires to experience a cleaning phase before the analysis. Data cleaning generally refers to excluding irrelevant, corrupted or dirty data [67]. There might be inconsistencies, inaccuracies, empty records or simply unnecessary information in the data. In the cleaning phase, these kind of data is removed from the data set.

In some of the cases, removing a complete row is not a good option even if it has blank cells in it. This is because, there might be quite precious information in the other columns that would likely to be beneficial for the analysis. In these cases, some techniques are used to fill the blank cells such that filling with mean, median or most frequent value. Although this has some advantages about not to lose data in other columns, it changes the general characteristics of the data and it may lead wrong inferences.

After this cleaning phase, data needs to be prepared for the analysis. In this step, some operations such that seperating some columns, partitioning or combining the records, transforming type of the data into another form might be needed. Then, data analysis takes place which can be done through different applications such that Excel or Python as in our case.

After revealing insights from the data, now it is time to interpret this result and it might be a good idea to create a report about it. In this reporting phase, there is another important step which is data visualization. This step makes it easier to show hidden aspects of the analysis with simple and effective graphs and generally dashboards.

Once these steps are concluded, data analysis process is considered to be done. Now, we will discuss machine learning which is quite related with data analysis.

4.1.2 Machine Learning

History of Artificial Intelligence (AI) is a less than a century as it is discussed in 1.1. This situation causes that notions and definitions rapidly change in AI because it is still evolving pretty rapidly. Therefore, machine learning had some different meanings throughout the history but it now refers to an implementation method of AI. However, we have a pretty general definition about machine learning which is still valid for today and worthy to share without change.

Samuel [12] defines machine learning in 1959 as "The field of study that gives computers the ability to learn without being explicitly programmed". This definition is probably the most well-known definition about machine learning and still applies to the today's world. Now, another point is yet to clarify after this definition which is to investigate how this learning process is conducted.

Machine learning usually classified into subcategories and each of these categories have different learning processes. It is even possible to see some little differences in the same subcategory. The main classification of machine learning algorithms can be seen as follows:

- Supervised Learning
- Semi-supervised Learning
- Unsupervised Learning
- Reinforcement Learning

Each of these learning techniques have different purposes and different usage areas. Since our model will be about regression analysis and regression analysis is a supervised learning algorithm, we will share a very brief information about the other ones and we will concentrate on regression analysis which is a subfield of supervised learning.

Supervised learning is a machine learning method that uses labeled data [68] and it has two subclasses which are classification and regression. Unsupervised learning on the other hand uses unlabeled data and mainly used for clustering. Semi-supervised learning uses both labeled and unlabeled data and reinforcement learning is the last type of machine learning algorithms in which agents learn on his own with a reward or punish system.

There some libraries in Python that are vitally important and useful for machine learning such as: Pandas, NumPy, SciPy, Scikit-Learn. In addition to these libraries, Seaborn and MatPlotlib libraries are extremely useful for creating visual tools. Keras, TensorFlow libraries are useful for deep learning processes.

There are some additional techniques used in the machine learning modelling process and these will be shared during the construction phase of the model. Now, we will examine regression analysis before building the model.

4.2 Regression Analysis

Regression analysis has different types which are generally used for different purposes. In the model, multivariate polynomial regression will be used. Linear regression and polynomial regression are related subjects, therefore, detailed information about these two regression types will be shared. Additionally, other regression types will be briefly mentioned.

Error calculation is another important part for regression analysis, because it is not possible to model a phenomenon with a 100-percent accuracy. Therefore, it is significant to measure how model reflects the reality which can be measured by various

error calculation techniques. Hence, different error measurement techniques will be shared in the following sections.

4.2.1 Linear Regression

Linear regression is the most fundamental version of the regression analysis and it is aimed to find linear relationship between independent variable(s) and a dependent variable in this type of regression. It is beneficial for demonstrating linear relationship, however, high level of correlation between independent variable(b) and the dependent variable is needed in order to have an accurate result in linear regression.

4.2.1.1 Simple Linear Regression

Simple linear regression is a regression analysis method which analyses the relationship between one independent and one dependent variable. For x as a independent variable and y as a dependent variable, equation that shows the relationship between variables can be denoted as follows:

$$y = \beta_0 + \beta_1 x \tag{4.1}$$

Here, β_0 is the *y-intercept* and β_1 is the coefficient. In other words, when we take $x = 0$, the equation $y = \beta_0$ will be satisfied and β_0 will be the point on the *y-axis*. The term β_1, on the other hand, denotes the change in y when x is incremented by 1 units. Since this equation represents a line equation, β_1 is also the slope of this line.

In real life, simple linear regression generally does not appear as in equation (4.1), because it is almost impossible to have a perfect equation. Therefore, the margin between real and calculated results should be indicated somehow and this is done through adding an error term, which is generally shown with ϵ. Simple linear regression equation with an error term can be seen as follows:

$$y = \beta_0 + \beta_1 x + \epsilon \tag{4.2}$$

There are a lot of different techniques to calculate this error term ϵ and these techniques will be investigated in section (4.2.4).

It should also be noted that simple linear regression does not have to be in a one-dimensional form and their components may be vectors or more generally

matrices. Therefore, equations (4.1) and (4.2) are also valid for multi-dimensional cases.

Apart from this mathematical and statistical aspect, simple linear regression can additionally be constructed using machine learning algorithms. For example, in Python, it can be built with the help of NumPy and Scikit-Learn libraries. "LinearRegression()" method is the core part for both linear regression and polynomial regression which can also be seen in the code of our model.

4.2.1.2 Multiple Linear Regression

Multiple linear regression is another regression method which investigates the relationship between more than one independent variables and a dependent variable. An equation representing multiple linear regression can be seen as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n \tag{4.3}$$

In equation (4.3), coefficients have a similar meaning as it is mentioned in (4.2.1.1). Here, y is the dependent variable, β_0 is the *y-intercept*, $\beta_1, \beta_2, ..., \beta_n$ are coefficients and $x_1, x_2, ..., x_n$ are independent variables for $1, 2, ..., n \in \mathbb{N}$.

An error term is also generally added to the multiple linear regression equation for the same reasons in equation (4.2), then equation becomes as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \epsilon \tag{4.4}$$

Simple linear regression and multiple linear regression has a common purpose that they put forth relationship between variables. On the other hand, there are some differences between them. Seemingly, the only difference between these two equations (4.2) and (4.4) is the number of independent variables (and number of coefficients as a result), however, this increase in the number of parameters (independent variables) has some favorable and unfavorable results.

Identifying a relationship mathematically between more than one independent variables and single dependent variable is a quite different process different than the situation contains single independent and single dependent variable. In linear regression models, it is expected to see some correlation between the dependent variable and independent variable(s), however, multiple linear regression is more tolerant than simple linear regression about decline of this correlation. This situation

implies that multiple linear regression is more likely to produce better results when there is a weaker correlation between dependent variable and independent variable(s) in comparison to simple linear regression.

Increment of the number of variables is not always a good thing, although the situation above seems to give us more space when creating a regression model. Generally more data and processing power is needed to constructed multiple linear regression models but more importantly multicollinearity is the main problem in this kind of models. Multicollinearity stands for the situation when independent variables in a model are correlated to each other and it results a decline in accuracy and reliability of the model. Multiple linear regression models frequently suffer from *multicollinearity*, which is also a problem for multivariate polynomial regression that will also be discussed in 4.2.2.

4.2.2 Polynomial Regression

In the previous section, we examined the most fundamental type of regression analysis: linear regression. Most of the problems in the world cannot be modelled with a linear regression regardless of whether it is simple or multiple linear regression. This is mainly because of that real world problems generally occur in a more complex way and parameters are more likely to have non-linear relationships to dependent variables, therefore they cannot be simplified that much. At this point, polynomial regression models starts to be more beneficial.

In polynomial regression, fluctuations in the correlation of dependent variable and independent variable(s) are more tolerated than linear regression. This is mainly because of that polynomial regression capture these changes in an easier way such that independent variable(s) can be positively related with the dependent variable in some interval while it is negatively correlated for another interval. Additionally, strength of the correlation may change. Polynomial regression is more adaptable in these kind of fluctuations.

It is correct that polynomial regression has some advantages that is mentioned above, however, it also have some deficits. *Overfitting* and *underfitting* are two significant problems in polynomial regression and they occur in some conditions such that complexity of the model, existence of outliers, multicollinearity (for multiple polynomial regression), insufficient data, etc.

Overfitting refers to the situation that model has a high complexity and it models most of the details of the data set rather than concentrating on the underlying pattern of the data set. This situation causes model to capture some *noises* that generally does

not reflect the original distribution of the data set. Overfitting generally appear when number of variables or degree of the model increases. When this increment occurs beyond the optimum level, the model starts to memorize the data set rather than learning the main features of the data set. As a result, accuracy of the model decreases.

Underfitting, on the other hand, occurs when the complexity of the model is too low and the model cannot capture the underlying pattern of data set. This might be because of one of the reasons that is mentioned above, however, the most important factor is to have a lower degree polynomial than it should be. Below, there is a graph that compares overfitting, perfect fit and underfitting:

Overall, in order to avoid overfitting and underfitting, degree of the polynomial should be exactly as it should be, neither too high nor too low. Additionally, for multiple polynomial cases, variables should also be selected elaborately. After having identified this important problem, now it is time to give specific information about two types of polynomial regression which are simple linear regression and multiple polynomial regression.

4.2.2.1 Single Polynomial Regression

Simple polynomial regression that when there is a single independent and single dependent variable. This type of regression is useful when the correlation between The equation for simple polynomial regression can be defined as follows:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + ... + \beta_n x^n \tag{4.5}$$

Here, y is the dependent variable, β_0 is the *y-intercept*, $\beta_1, \beta_2, ..., \beta_n$ are coefficients and $x_1, x_2, ..., x_n$ are independent variables for $1, 2, ..., n \in \mathbb{N}$.

As in the linear regression cases, when we add an error term ϵ, we get the following equation:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + ... + \beta_n x^n + \epsilon \tag{4.6}$$

It should also be noted that, it is again possible to have a multi-dimensional form of these equations (4.5)and (4.6) when the components of these equations are matrices.

4.2.2.2 Multiple Polynomial Regression

Multiple or multivariate polynomial regression is a regression analysis method that constructs an equation between more than one independent variable and a dependent variable. This type of regression is extremely complex, because increasing the degree and the number of independent variables will dramatically increase the number of coefficients. More coefficients does not only imply necessity of more processing power but also may affect the model in a bad way by redundantly increasing the complexity.

It may be a little bit awkward to give n-degree equation of multiple polynomial regression, because when the degree of a multivariate polynomial regression increases, the number of terms increases dramatically. However, quadratic form is shared below:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2 + \epsilon \tag{4.7}$$

Multiple polynomial regression models suffer from different problems. Overfitting and underfitting is the main problem in this type of regression which is caused by the degree of the polynomial and number of variables in the equation. Moreover, multicollinearity is more severely appears in multiple polynomial regression.

4.2.3 Other Types of Regression Analysis

There are also other types of regression analysis. The other main types of the regression include logistic regression, LASSO (Least Absolute Shrinkage and Selection Operator) regression, ridge regression, ElasticNet regression, support vector regression, bayesian regression and and a lot of additional subbranches. Nevertheless, we will emphasize two main important types of regression analysis which are Lasso regression and Ridge regression, because we also utilize these two types in the construction of our model.

Lasso regression and ridge regression are also called as L1-Regularization and L2-Regularization, respectively. They aim to minimize the error by adding a penalty term. These regularization methods increases the accuracy of the model by adding different terms to the regression equation.

These regularization methods are generally a good choices when multicollinearity appears and they are used complementarily [69]. The difference between these two types of regression, while ridge regression adds the sum of the squares of the coefficients to the regression equation, LASSO regression adds the sum of absolute values of the coefficients. In addition to these two methods, ElasticNet regression is

also a regression type that is a mixture of these aforementioned two types of regression and it is used in order to increase the accuracy of the regression models.

Since our model will be a multiple linear regression model, multicollinearity is a quite expected problem. Hence, these regularization methods will also be used as a last step in our model.

4.2.4 Error Calculation for Regression Models

In statistics and machine learning, there are a lot of measurement techniques to determine the accuracy of the model and these techniques have different usage areas. There are a lot of useful statistical measures that are widely used such as t-test, f1-test, chi-squared test, hypothesis testing, confidence interval etc., however, they are not completely suitable for evaluation models. For that reason, our model will include some measurements that are generally used for measuring the accuracy of regression models and they are shared in 4.2.4.

All the error measurements below aims to express the level of accuracy of the model with different techniques. The main idea in these calculations is to measure the difference between actual value and the predicted value and then determine whether the model has a good fit to the real data or not. It is intended to use 6 different error measurement techniques in our model which are:

- Mean Absolute Error (MAPE)
- Mean Squared Error (MSE)
- Mean Absolute Percentage Error (MAPE)
- R^2 score
- Root Mean Squared Error (RMSE)
- Median Absolute Error (MedAE)

In sklearn.metrics library, there are functions to calculate them directly. Each of these error measurement methods are useful ones and they are widely used. It should be noted that percentage errors are better to indicate the accuracy of the model because it does not depend on the numerical value of the variables and they measure the error as a percentage. In other words, having 10 MSE means almost nothing without knowing what is the set of interval of the target variable. For a target variable that takes values

between 0 and 15, 10 MSE is awful, however, for a variable that is in the interval 0 and 1000000, 10 MSE is unbelieveably good.

Now it is time to start constructing our model.

4.3 The Model

In this section, we will construct our own traffic model which aims to predict average speed, number of vehicles and traffic flow in a specific location. While constructing this model, fundamental principles of mathematical modelling and traffic flow theory is used and machine learning algorithms is utilized, especially regression analysis.

This model aims to construct a relationship between traffic data and weather conditions at the given point and pandemic effects on the traffic. There are various works about affecting factors of the traffic. In the Literature Review (1.1) section we also emphasised that there are some studies in the recent years that discuss the effects of weather and Covid-19 effects on traffic, separately. Now, we are going to try to investigate their combined effect on the traffic for a specific location in İstanbul.

For the location selection, the location below is selected because of various reasons. First of all, bridges play a significant role in İstanbul traffic and the number of vehicle in this point is not the highest one but highly above average when all locations are considered. Secondly, standard deviation of average speed and number of vehicles is close to the average of all coordinates which makes this location suitable for modelling for further generalizations. Location is shared below:

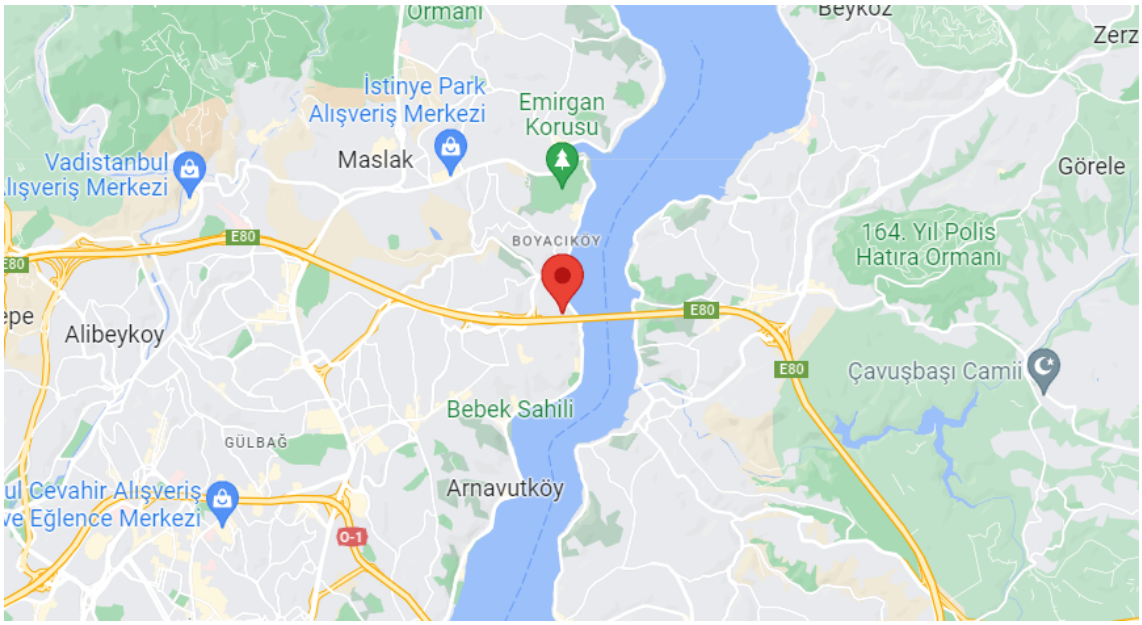


Figure 4.2 Our selected location [70]

Before starting modelling immediately, we will first analyze the data especially about Covid-19 pandemics effects on the traffic. Firstly, we will compare the traffic situation between existence and non-existence of Covid-19 pandemics and pandemic shutdowns by sharing some insights. After that we will share the correlation tables and maps to reveal relationships between variables. Then, we will create our model by determining the parameters.

4.3.1 The Data Sets and Techniques Used in the Model

The data on the web services are generally stored as demiliter seperated values (dsv). Data in different columns are seperated with a delimiter such as commas, tabs, semicolons or colons. There are different types of dsv files, however, comma seperated values(csv) are the most used one and tab seperated values (tsv) are their follower. It is generally preferred to stored the data as csv or tsv because of its a lot of convenience. The main advantage of demiliter seperated values is that they are supported by many of the spreadsheet applications or databases.

Although data is stored as csv, it is not so suitable for humans to work with because of its visual aspects. We generally prefer to work with tabular data, data shown in the table, and it is possible to do so by importing these csv files into Python, SQL or spreadsheet applications like Excel to obtain a data frame in order to have a better visualization of the table.

Our data set consist of main three different components. First one is "Hourly Traffic Density Data Set" that is obtained through İstanbul Metropolitan Municipality Open Data Portal [71]. There are 17 months of data here, however, May 2021 is incomplete and only consist of 3 days of data. Therefore, we acquired 16 months of data here, for the time interval between January 2020 - April 2021. The traffic data stored as csv and we transformed it to an Excel table which is shared below:

	DATE_TIME	LONGITUDE	LATITUDE	GEOHASH	MINIMUM _SPEED	MAXIMUM _SPEED	AVERAGE _SPEED	NUMBER_OF_ VEHICLES
2	3.01.2020 13:00	28.8446044921875	41.1135864257812	sxk9bk	9	78	53	105
3	16.01.2020 07:00	29.1522216796875	41.0092163085938	sxk9r3	6	64	27	68
4	17.01.2020 20:00	28.4710693359375	41.0421752929688	sxk1rz	53	96	72	11
5	14.01.2020 22:00	29.3609619140625	40.8169555664062	sxkbhn	17	128	73	216
6	17.01.2020 20:00	28.8446044921875	40.9707641601562	sxk906	16	82	55	123
7	20.01.2020 10:00	29.2950439453125	41.1080932617188	sxkcfe	68	152	93	88
8	24.01.2020 16:00	28.7127685546875	40.9927368164062	sxk3jq	6	91	49	299
9	27.01.2020 13:00	28.8995361328125	41.0696411132812	sxk99s	6	46	15	167
10	6.01.2020 20:00	28.8226318359375	40.9982299804688	sxk3pz	6	114	59	287

Figure 4.3 Excel table form of "Hourly Traffic Density Data Set" for January 2020 [72]

Additionally, all data sets for these 16 months can be found on [71].

The other component of our data set includes weather data, from the open-meteo.com "Historical Weather API" servise at [73]. Here is a view of selections for this website:

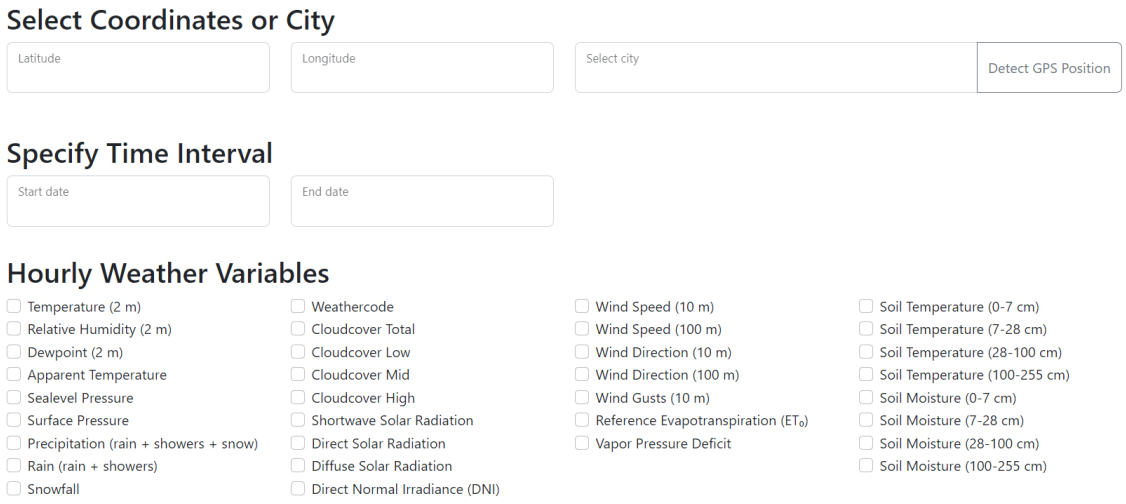


Figure 4.4 Possible parameters for hourly weather data [73]

There are a lot of options for hourly weather variables and it is possible get the data by using the Application Programming Interface (API) of the website. It is possible to include all the data above, however, this would be quite unnecessary. Nonetheless, we first download all the available data for the relevant time interval to see that how the weather and traffic is correlated. After this step, we checked correlation between variables and most of the variables that have less than 0.1 Pearson correlation coefficient, Kendall's rank correlation coefficient or Spearman's rank correlation coefficient are eliminated but some of them is left in order to emphasize some important points.

Our third and last component of data set consist of Covid-19 pandemic period data and it includes information about pandemic shutdowns, workdays, publick holidays and the days that face to face education is held:

	time	public_holiday	face to face education	c19_shutdown	workday
2	1.01.2020 00:00	1	0	0	0
3	1.01.2020 01:00	1	0	0	0
4	1.01.2020 02:00	1	0	0	0
5	1.01.2020 03:00	1	0	0	0
6	1.01.2020 04:00	1	0	0	0
7	1.01.2020 05:00	1	0	0	0
8	1.01.2020 06:00	1	0	0	0
9	1.01.2020 07:00	1	0	0	0
10	1.01.2020 08:00	1	0	0	0

Figure 4.5 Covid-19 Shutdowns and Related Data

After obtaining all these data and inserting our data about pandemics, data is

combined through by Excel table merge features. It should be noted that there are a lot of alternatives ways to do this. SQL can do this by left, right, inner or outer "JOIN" and Python is also eligible to do that by Pandas library which includes merge and join methods. Even in the Excel, there are a lot of different features such that PowerPivot connections or vlookup, index and match functions.

4.3.2 Preprocessing of the Data and Statistical Features of the Data

We merged these three data sets and now we have the following data set that includes:

- date-time: Date and time of the record
- min-speed: Minimum speed of the vehicles at given location
- max-speed: Maximum speed of the vehicles at given location
- avg-speed: Average speed of the vehicles at given location
- no-of-veh: Number of total vehicles that passes through the location within a 1 hour time interval
- flow: Measurement of traffic flow that is calculated with the flow-speed-density relationship formula and equals to avg-speed*no-of-veh
- day: Day of the week, enumerated from 1 to 7 by starting from Monday and ending on Sunday. Obtained by applying "=HOUR()" Excel formula on date-time column. Numerical value is obtained instead of text because it is aimed to be used in the regression.
- hour: Hour of the day, enumerated from 0 to 23 such that 0 indicates 00:00-01:00, 1 indicates 01:00-02:00 and so on. Obtained by applying "=WEEKDAY()" Excel formula on date-time column. Numerical value is obtained instead of text because it is aimed to be used in the regression.
- public-holiday: One-hot encoded variable that shows either there is public holiday on that day or not, takes values either 0 or 1
- c19-shutdown: It shows whether Covid-19 shutdown exist on that hour interval or not in İstanbul, takes values either 0 or 1, one-hot encoded
- f2f-education: The days face to face education is held
- workday: One-hot encoded variable that shows either the day is a work day or not, 1 for Monday-Tuesday-Wednesday-Thursday-Friday and 0 for Saturday-Sunday

- temp: Temperature at the given location at that time interval
- rel-hum: Relative humidity at the given location at that time interval
- precipitation: Precipitation measurement in milimeters at the given location at that time interval
- evapo: Evapotranspiration level at the given location at that time interval
- vpd: Vapour Pressure Deficit level at the given location at that time interval

In order to see statistical properties of the data, "DataFrame.describe()" method can be used from Pandas library. Result is is shared below:

	min-speed	max-speed	avg-speed	no-of-veh	day	hour	public-holiday	f2f-education	c19-shutdown	work day	temp	rel-hum	precipitation	evapo	vpd
count	11368	11368	11368	11368	11368	11368	11367	11368	11368	11368	11368	11368	11368	11368	11368
mean	13,82	138,1	78,27	227,7	4	11,51	0,04	0,09	0,21	0,69	14,49	74,43	0,08	0,12	0,53
std	11,97	18,18	15,68	133,9	1,99	6,9	0,18	0,28	0,41	0,46	7,42	13,99	0,28	0,14	0,51
min	1	75	27	11	1	0	0	0	0	0	-4,2	32	0	0	0
25%	6	124	68	97	2	6	0	0	0	0	8,7	65	0	0,02	0,18
50%	9	138	82	237	4	12	0	0	0	1	13,3	76	0	0,05	0,35
75%	17	151	91	341,3	6	17	0	0	0	1	21,1	86	0	0,16	0,67
max	87	227	114	645	7	23	1	1	1	1	32,8	100	5,7	0,7	3,33

Figure 4.6 Statistical Features of the Data Set

Now, we will check the correlation between these variables. Below, we will share all three main correlation measurements which are Pearson correlation coefficient, Kendall's rank correlation coefficient or Spearman's rank correlation coefficient. Each of these correlation coefficients takes values in [-1,1] such that 1 implies a perfect positive correlation while -1 implies a perfect negative correlation. 0 implies that there is no correlation between variables. When the absolute value of the correlation coefficient approaches zero, correlation between variables decreases.

Although they all measure the correlation level between variables, Spearman's correlation coefficient is more suitable for our multivariate polynomial regression model. Nevertheless, we will share all of the correlation tables below:

	min-speed	max-speed	avg-speed	no-of-veh	day	hour	public-holiday	f2f-education	c19-shutdown	workday	temp	rel-hum	precipitation	evapo	vpd
min-speed	1	0,32	0,57	-0,54	0,05	-0,41	0,05	-0,05	0,39	-0,09	-0,23	0,3	0,02	-0,3	-0,27
max-speed	0,32	1	0,67	-0,53	0,13	-0,21	0,09	-0,12	0,35	-0,19	-0,06	0,26	-0,12	-0,31	-0,23
avg-speed	0,57	0,67	1	-0,63	0,1	-0,46	0,09	-0,11	0,39	-0,17	-0,17	0,39	-0,04	-0,34	-0,34
no-of-veh	-0,54	-0,53	-0,63	1	-0,25	0,36	-0,13	0,11	-0,5	0,35	0,17	-0,29	-0,02	0,32	0,27
day	0,05	0,13	0,1	-0,25	1	0,01	0,02	-0,15	0,27	-0,75	-0,01	-0,01	0	0	0
hour	-0,41	-0,21	-0,46	0,36	0,01	1	0,02	0	-0,04	-0,01	0,14	-0,35	-0,01	0,23	0,27
public-holiday	0,05	0,09	0,09	-0,13	0,02	0,02	1	-0,06	0,12	-0,28	0,11	-0,04	-0,02	0,07	0,07
f2f-education	-0,05	-0,12	-0,11	0,11	-0,15	0	-0,06	1	-0,16	0,21	-0,23	0,07	0,05	-0,1	-0,15
c19-shutdown	0,39	0,35	0,39	-0,5	0,27	-0,04	0,12	-0,16	1	-0,36	-0,27	0,1	0,04	-0,14	-0,18
workday	-0,09	-0,19	-0,17	0,35	-0,75	-0,01	-0,28	0,21	-0,36	1	-0,02	0,01	0,01	-0,02	-0,02
temp	-0,23	-0,06	-0,17	0,17	-0,01	0,14	0,11	-0,23	-0,27	-0,02	1	-0,43	-0,18	0,59	0,77
rel-hum	0,3	0,26	0,39	-0,29	-0,01	-0,35	-0,04	0,07	0,1	0,01	-0,43	1	0,24	-0,76	-0,82
precipitation	0,02	-0,12	-0,04	-0,02	0	-0,01	-0,02	0,05	0,04	0,01	-0,18	0,24	1	-0,14	-0,19
evapo	-0,3	-0,31	-0,34	0,32	0	0,23	0,07	-0,1	-0,14	-0,02	0,59	-0,76	-0,14	1	0,84
vpd	-0,27	-0,23	-0,34	0,27	0	0,27	0,07	-0,15	-0,18	-0,02	0,77	-0,82	-0,19	0,84	1

Figure 4.7 Pearson Correlation Table

	min-speed	max-speed	avg-speed	no-of-veh	day	hour	public-holiday	f2f-education	c19-shutdown	workday	temp	rel-hum	precipitation	evapo	vpd
min-speed	1	0,33	0,58	-0,43	0,04	-0,36	0,04	-0,01	0,31	-0,08	-0,17	0,27	0,03	-0,33	-0,27
max-speed	0,33	1	0,53	-0,35	0,1	-0,15	0,07	-0,1	0,29	-0,16	-0,04	0,19	-0,11	-0,28	-0,15
avg-speed	0,58	0,53	1	-0,45	0,08	-0,33	0,07	-0,1	0,36	-0,15	-0,12	0,27	-0,05	-0,32	-0,24
no-of-veh	-0,43	-0,35	-0,45	1	-0,18	0,22	-0,1	0,09	-0,41	0,28	0,1	-0,2	-0,01	0,3	0,19
day	0,04	0,1	0,08	-0,18	1	0	0,02	-0,13	0,24	-0,66	-0,01	-0,01	-0,02	0	0,01
hour	-0,36	-0,15	-0,33	0,22	0	1	0,02	0	-0,04	-0,01	0,09	-0,24	-0,02	0,22	0,22
public-holiday	0,04	0,07	0,07	-0,1	0,02	0,02	1	-0,06	0,12	-0,28	0,09	-0,03	-0,03	0,05	0,06
f2f-education	-0,01	-0,1	-0,1	0,09	-0,13	0	-0,06	1	-0,16	0,21	-0,19	0,06	0,06	-0,07	-0,13
c19-shutdown	0,31	0,29	0,36	-0,41	0,24	-0,04	0,12	-0,16	1	-0,36	-0,2	0,08	0,04	-0,16	-0,16
workday	-0,08	-0,16	-0,15	0,28	-0,66	-0,01	-0,28	0,21	-0,36	1	-0,02	0	0,01	-0,01	-0,02
temp	-0,17	-0,04	-0,12	0,1	-0,01	0,09	0,09	-0,19	-0,2	-0,02	1	-0,27	-0,2	0,35	0,57
rel-hum	0,27	0,19	0,27	-0,2	-0,01	-0,24	-0,03	0,06	0,08	0	-0,27	1	0,24	-0,64	-0,71
precipitation	0,03	-0,11	-0,05	-0,01	-0,02	-0,02	-0,03	0,06	0,04	0,01	-0,2	0,24	1	-0,11	-0,27
evapo	-0,33	-0,28	-0,32	0,3	0	0,22	0,05	-0,07	-0,16	-0,01	0,35	-0,64	-0,11	1	0,63
vpd	-0,27	-0,15	-0,24	0,19	0,01	0,22	0,06	-0,13	-0,16	-0,02	0,57	-0,71	-0,27	0,63	1

Figure 4.8 Kendall's Correlation Table

	min-speed	max-speed	avg-speed	no-of-veh	day	hour	public-holiday	f2f-education	c19-shutdown	workday	temp	rel-hum	precipitation	evapo	vpd
min-speed	1	0,47	0,75	-0,58	0,05	-0,48	0,05	-0,01	0,37	-0,1	-0,24	0,38	0,04	-0,45	-0,38
max-speed	0,47	1	0,72	-0,53	0,14	-0,22	0,09	-0,12	0,35	-0,19	-0,06	0,28	-0,14	-0,4	-0,22
avg-speed	0,75	0,72	1	-0,64	0,11	-0,45	0,09	-0,12	0,43	-0,18	-0,17	0,39	-0,06	-0,46	-0,35
no-of-veh	-0,58	-0,53	-0,64	1	-0,25	0,36	-0,13	0,11	-0,5	0,35	0,15	-0,29	-0,01	0,43	0,28
day	0,05	0,14	0,11	-0,25	1	0,01	0,02	-0,15	0,27	-0,75	-0,02	-0,01	-0,02	0	0,01
hour	-0,48	-0,22	-0,45	0,36	0,01	1	0,02	0	-0,04	-0,01	0,14	-0,36	-0,03	0,33	0,33
public-holiday	0,05	0,09	0,09	-0,13	0,02	0,02	1	-0,06	0,12	-0,28	0,11	-0,04	-0,03	0,06	0,07
f2f-education	-0,01	-0,12	-0,12	0,11	-0,15	0	-0,06	1	-0,16	0,21	-0,24	0,07	0,06	-0,08	-0,16
c19-shutdown	0,37	0,35	0,43	-0,5	0,27	-0,04	0,12	-0,16	1	-0,36	-0,25	0,1	0,04	-0,19	-0,2
workday	-0,1	-0,19	-0,18	0,35	-0,75	-0,01	-0,28	0,21	-0,36	1	-0,03	0	0,01	-0,01	-0,02
temp	-0,24	-0,06	-0,17	0,15	-0,02	0,14	0,11	-0,24	-0,25	-0,03	1	-0,39	-0,26	0,49	0,75
rel-hum	0,38	0,28	0,39	-0,29	-0,01	-0,36	-0,04	0,07	0,1	0	-0,39	1	0,29	-0,81	-0,88
precipitation	0,04	-0,14	-0,06	-0,01	-0,02	-0,03	-0,03	0,06	0,04	0,01	-0,26	0,29	1	-0,14	-0,33
evapo	-0,45	-0,4	-0,46	0,43	0	0,33	0,06	-0,08	-0,19	-0,01	0,49	-0,81	-0,14	1	0,8
vpd	-0,38	-0,22	-0,35	0,28	0,01	0,33	0,07	-0,16	-0,2	-0,02	0,75	-0,88	-0,33	0,8	1

Figure 4.9 Spearman's Correlation Table

Here, we are going to do some modifications in our data set. Firstly, we will exclude some of the columns which are not planned to be used in our machine learning model because they seem to have low correlation with average speed or number of vehicles. Some variables will stay even if they are less likely to be used. Also, we will insert the variable *flow* in this stage which will be calculated as $flow = "avg - speed" x "no - of - veh"$.

Minimum speed and maximum speed will also be excluded even if they have strong correlation with average speed and number of vehicles, because we would like to predict the traffic without knowing any existence information on that hour. Hence,

none of the independent variable will be an element traffic data set, which also implies that average speed and number of vehicles will not be used to predict each other.

After these operations, we have the following data set:

date_time	flow	avg-speed	no-of-veh	day	hour	c19-shutdown	workday	temp	rel-hum	evapo	vpd
01.01.20 00:00	10721	71	151	3	0	0	0	5,6	88	0,01	0,11
01.01.20 01:00	10384	88	118	3	1	0	0	5,7	86	0,01	0,13
01.01.20 02:00	10586	79	134	3	2	0	0	5,8	84	0,02	0,15
01.01.20 03:00	6966	86	81	3	3	0	0	5,7	82	0,02	0,17
01.01.20 04:00	5810	83	70	3	4	0	0	5,6	78	0,03	0,2
01.01.20 05:00	7128	88	81	3	5	0	0	5,6	74	0,03	0,24
01.01.20 06:00	6664	98	68	3	6	0	0	5,7	71	0,03	0,26
01.01.20 07:00	7980	84	95	3	7	0	0	5,5	75	0,03	0,22
01.01.20 08:00	11224	92	122	3	8	0	0	5,5	80	0,01	0,18
01.01.20 09:00	10920	91	120	3	9	0	0	5,7	83	0,03	0,16
01.01.20 10:00	12960	90	144	3	10	0	0	6,4	86	0,04	0,13
01.01.20 11:00	14790	85	174	3	11	0	0	6,8	86	0,05	0,14
01.01.20 12:00	18480	84	220	3	12	0	0	7,1	84	0,07	0,16
01.01.20 13:00	18924	83	228	3	13	0	0	7,2	77	0,08	0,24
01.01.20 14:00	20405	77	265	3	14	0	0	7,6	72	0,1	0,3
01.01.20 15:00	20440	70	292	3	15	0	0	7,7	69	0,11	0,32
01.01.20 16:00	18538	62	299	3	16	0	0	7,5	68	0,11	0,33
01.01.20 17:00	16038	54	297	3	17	0	0	7,2	68	0,08	0,32
01.01.20 18:00	17538	74	237	3	18	0	0	6,8	69	0,06	0,3

Figure 4.10 Data set after modifications

It is also important to see pandemic effects on the traffic flow by comparing two dates when one of them is before Covid-19 pandemic and one of them is during Covid-19 pandemic. Therefore, we compared January and February 2020 versus January and February 2021, since this is the best possible option for our data set.

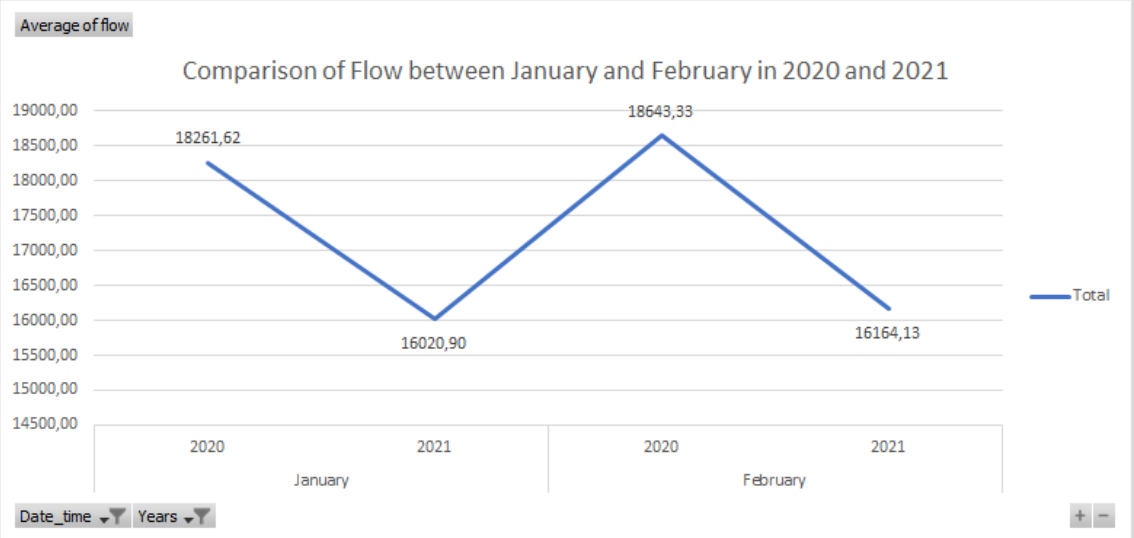


Figure 4.11 Traffic flow in our selected location before and during Covid-19

Now, we will create our multiple polynomial regression model based on this data set. The data set consists of 11368 records for now but the number of records will decrease after outlier elimination, which will be conducted in the next section.

4.3.3 Construction of the Model

Some fundamental data preprocessing and data analysis job is conducted in previous sections, now, it is time to use machine learning algorithms and make a model that predicts average speed, number of vehicles and traffic flow.

In the last section, we determined our possible parameters based on the correlations between variables. We believe they are so much affecting factors on the traffic and correlation coefficients also support our claim. Therefore, our intention is to use several independent variable to predict one dependent variable at a time (average speed, number of vehicles, traffic flow). Hence, we will actually construct three different models to predict it.

Traffic flow will be calculated with the most fundamental equation which is $q = v\rho$ as it is indicated in the last section. We are going to make some simplifications in this point that number of vehicles will be counted as the density, however, this will not change anything for the validity of the model. Because, we would divide number of vehicles by the area, which is actually same as scaling the number of vehicles. Therefore, only numerical result would change in the model however the main principles of traffic flow theory will stay same.

We also needed another simplification that the data set we have does not include the information length of the vehicles. Additionally, public transportation usage information is only shared with monthly statistics so they cannot be implemented to model in a proper way. It would be possible to add this by partitioning this into hourly time intervals, however, this would harm our results by altering the characteristics of our data set. Therefore, we determine to ignore this effects.

So far, all the irrelevant data is cleaned and correlation between variables are observed. As a result, we have a clue that we have the following variables that affect the traffic and might be useful in the prediction phase which are: day, hour, Covid 19 pandemic shutdowns, evapotranspiration, vapour pressure deficit, relative humidity, work days, before and after work hours are our main variables. It may also be said that public holidays and face to face education has an effect on the traffic, it is correct but there is so little public holidays and these days can also be included to the model since they are not "work days". Additionally, adding "face to face education" variable is also questionable because there were too few days where it was held because of the Covid-19 pandemic period.

A regression analysis does not have to possess all the steps above, on the other hand, these steps may be extended further, it generally depends on the data set and

complexity of the model.

Eliminating outliers is generally an easy task but determining how this elimination will be conducted is a challenging issue, because not every method is applicable to all cases. The most widely used method for outlier eliminations are Z-score method and Inter Quartile Range (IQR) method.

In Z-Score method, a value that is greater than 3 or less than -3 is generally considered as an outlier. IQR method handles outliers a little differently, and it eliminates values that lies 1,5*IQR below the first quartile and 1,5*IQR above the third quartile. Both methods are usable but specific features of the data set is decisive which one to use. It should also be noted that eliminating outliers does not always provide a good result because some valuable information might be lost during this operation.

We eliminated outliers by using IQR method, however, elimination is not applied to all variables. Because, for weather variables there are a lot of value that are zero or close to zero in precipitation, evapotranspiration and vapour pressure deficit columns and this situation affects the statistical distribution of the data and hence quartiles. Therefore, precipitation, evapotranspiration and vapour pressure deficit outliers intentionally left in the data set. However, minimum speed, maximum speed, average speed, number of vehicle and flow outliers are excluded.

Another significant point is feature scaling. It is generally recommended to do this after the training and test split, however, it does not make a significant difference for our data set since we tried both of the methods before creating the model. We will use "shuffle" and "stratify" hyperparameters in the train/test split phase in order to make sure that our training and test set have similar distributions therefore chronological place of feature scaling does not matter.

There are two main methods for feature scaling which are standardization and normalization. Standardization method scales the values by adjusting their mean as 0 and standard deviation as 1 whereas normalization scales the values between 0 and 1 which is also called as min-max scaling.

After using IQR outlier elimination method and normalization, we have the following statistical features of the data:

	s-avg-speed	s-no-vehicles	s-flow	new-flow	s-day	s-hour	s-temp	s-rel-hum	s-evapo	s-vpd	c19_shutdown	workday
count	10601	10601	10601	10601	10601	10601	10601	10601	10601	10601	10601	10601
mean	0,59	0,36	0,36	0,19	0,5	0,52	0,51	0,62	0,17	0,16	0,18	0,7
std	0,21	0,2	0,19	0,11	0,33	0,29	0,2	0,21	0,21	0,16	0,39	0,46
min	0	0	0	0	0	0	0	0	0	0	0	0
25%	0,44	0,17	0,21	0,1	0,17	0,26	0,35	0,47	0,03	0,06	0	0
50%	0,64	0,38	0,36	0,17	0,5	0,52	0,48	0,63	0,09	0,11	0	1
75%	0,77	0,53	0,49	0,25	0,83	0,78	0,69	0,78	0,24	0,21	0	1
max	1	1	1	0,61	1	1	1	1	1	1	1	1

Figure 4.12 Statistical features of our data set after scaling

	s-avg-speed	s-no-vehicles	s-flow	new-flow	s-day	s-hour	s-temp	s-rel-hum	s-evapo	s-vpd	c19_shutdown	workday
s-avg-speed	1	-0,61	-0,31	-0,03	0,11	-0,42	-0,15	0,37	-0,42	-0,33	0,39	-0,18
s-no-vehicles	-0,61	1	0,91	0,71	-0,27	0,29	0,12	-0,27	0,39	0,24	-0,46	0,37
s-flow	-0,31	0,91	1	0,93	-0,27	0,18	0,1	-0,19	0,34	0,18	-0,42	0,35
new-flow	-0,03	0,71	0,93	1	-0,22	0,1	0,07	-0,1	0,24	0,11	-0,33	0,27
s-day	0,11	-0,27	-0,27	-0,22	1	0	-0,01	-0,01	0	0,01	0,28	-0,75
s-hour	-0,42	0,29	0,18	0,1	0	1	0,11	-0,34	0,28	0,3	0,01	0
s-temp	-0,15	0,12	0,1	0,07	-0,01	0,11	1	-0,39	0,49	0,76	-0,23	-0,03
s-rel-hum	0,37	-0,27	-0,19	-0,1	-0,01	-0,34	-0,39	1	-0,81	-0,87	0,08	0,01
s-evapo	-0,42	0,39	0,34	0,24	0	0,28	0,49	-0,81	1	0,8	-0,14	-0,02
s-vpd	-0,33	0,24	0,18	0,11	0,01	0,3	0,76	-0,87	0,8	1	-0,17	-0,02
c19_shutdown2	0,39	-0,46	-0,42	-0,33	0,28	0,01	-0,23	0,08	-0,14	-0,17	1	-0,36
workday3	-0,18	0,37	0,35	0,27	-0,75	0	-0,03	0,01	-0,02	-0,02	-0,36	1

Figure 4.13 Spearman's rank correlation coefficient table after scaling

Now, we will create our model for three different dependent variables which are average speed, number of vehicles and traffic flow. We tried excessive amount of scenarios and recorded best results in terms of MAPE and R^2-Score. We constructed the model by using different independent variables, different degrees of polynomials, different train/test allocation ratio. Below there are six different tables in which they include best results for the relevant case. They investigates five different situations that scaled flow estimation, number of vehicle estimation, scaled number of vehicle estimation, average speed estimation, scaled average speed estimation.

After several trials, it is observed that best result is obtained when hour, day, Covid-19 shutdowns and evapotranspiration is used as independent variables with a 80-percent training set and a 20-percent test set. The other variables are also good candidates to predict our target variables, however, they are correlated with our main 4 independent variables which leads multicollinearity. In a same manner, it is also observed that increasing the degree of the polynomial beyond 10 degree results with an overfitting model. Therefore, 0.8-0.2 train/test ratio and polynomial degree of 5-6-7-8-9 will be investigated in proceeding tables and then coefficient and intercept of the regression equation will be shared of the best models for each of average speed, number of vehicles and flow.

	NUMBER OF VEHICLES									
	1st group		2nd group		3rd group		4th group		5th group	
	Training Set	Test set	Training Set	Test set	Training Set	Test set	Training Set	Test set	Training Set	Test set
Normalization	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Weight	0.8	0.2	0.8	0.2	0.8	0.2	0.8	0.2	0.8	0.2
Degree	5	5	6	6	7	7	8	8	9	9
Variables										
Hour	+	+	+	+	+	+	+	+	+	+
Day	+	+	+	+	+	+	+	+	+	+
C19-shutdowns	+	+	+	+	+	+	+	+	+	+
Evapo	+	+	+	+	+	+	+	+	+	+
Rel-hum	-	-	-	-	-	-	-	-	-	-
Workday	-	-	-	-	-	-	-	-	-	-
Vpd	-	-	-	-	-	-	-	-	-	-
Error										
MAE	0.0787	0.0791	0.0758	0.0762	0.0707	0.0715	0.0683	0.0699	0.0672	0.0692
MSE	0.0109	0.0110	0.0101	0.0101	0.0093	0.0094	0.0087	0.0090	0.0085	0.0088
MAPE	42,8711%	48,1012%	40,3598%	45,7232%	35,8734%	40,6263%	34,2343%	40,0326%	32,9858%	38,7421%
MedAE	0.0595	0.0611	0.0585	0.0599	0.0519	0.0534	0.0488	0.0514	0.0484	0.0518
RMSE	0.1046	0.1047	0.1007	0.1006	0.0963	0.0968	0.0931	0.0947	0.0919	0.0938
Model Fit										
R2-Score	0.7370	0.7419	0.7563	0.7616	0.7773	0.7765	0.7923	0.7861	0.7975	0.7900

Figure 4.14 Number of vehicles estimation: scaled data set

	NUMBER OF VEHICLES									
	1st group		2nd group		3rd group		4th group		5th group	
	Training Set	Test set	Training Set	Test set	Training Set	Test set	Training Set	Test set	Training Set	Test set
Normalization	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO
Weight	0.8	0.2	0.8	0.2	0.8	0.2	0.8	0.2	0.8	0.2
Degree	5	5	6	6	7	7	8	8	9	9
Variables										
Hour	+	+	+	+	+	+	+	+	+	+
Day	+	+	+	+	+	+	+	+	+	+
C19-shutdowns	+	+	+	+	+	+	+	+	+	+
Evapo	+	+	+	+	+	+	+	+	+	+
Rel-hum	-	-	-	-	-	-	-	-	-	-
Workday	-	-	-	-	-	-	-	-	-	-
Vpd	-	-	-	-	-	-	-	-	-	-
Error										
MAE	49.0392	48.825	47.1151	47.4205	44.4251	43.9112	44.0561	44.1893	42.2655	42.9015
MSE	4341.4179	4316.265	4008.7892	4056.546	3697.5812	3649.412	3591.0451	3646.922	3394.076	3446.722
MAPE	37.1093%	38.8022%	34.9378%	37.6085%	31.8853%	33.4370%	31.8242%	34.1155%	29.6047%	29.7186%
MedAE	36.5871	36.2466	35.7457	36.7136	32.4796	32.7837	32.2992	32.8083	30.1393	30.9448
RMSE	65.8894	65.6983	63.315	63.691	60.8077	60.4104	59.9253	60.3897	58.2587	58.7088
Model Fit										
R2-Score	0.7586	0.7567	0.7771	0.7713	0.7944	0.7943	0.8003	0.7944	0.8113	0.806

Figure 4.15 Number of vehicles estimation: non-scaled data set

	AVERAGE SPEED									
	1st group		2nd group		3rd group		4th group		5th group	
	Training Set	Test set	Training Set	Test set	Training Set	Test set	Training Set	Test set	Training Set	Test set
Normalization	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Weight	0.8	0.2	0.8	0.2	0.8	0.2	0.8	0.2	0.8	0.2
Degree	5	5	6	6	7	7	8	8	9	9
Variables										
Hour	+	+	+	+	+	+	+	+	+	+
Day	+	+	+	+	+	+	+	+	+	+
C19-shutdowns	+	+	+	+	+	+	+	+	+	+
Evapo	+	+	+	+	+	+	+	+	+	+
Rel-hum	-	-	-	-	-	-	-	-	-	-
Workday	-	-	-	-	-	-	-	-	-	-
Vpd	-	-	-	-	-	-	-	-	-	-
Error										
MAE	0.0918	0.0950	0.0895	0.0934	0.0880	0.0922	0.0870	0.0920	0.0862	0.0929
MSE	0.0160	0.0170	0.0155	0.0166	0.0153	0.0166	0.0151	0.0166	0.0148	0.0171
MAPE	25.3168%	29.2422%	24.7278%	28.8013%	24.3421%	28.5266%	24.1559%	28.5357%	23.9319%	28.5698%
MedAE	0.0679	0.0684	0.0660	0.0680	0.0634	0.0659	0.0616	0.0645	0.0613	0.0664
RMSE	0.1265	0.1306	0.1246	0.1290	0.1236	0.1287	0.1228	0.1290	0.1217	0.1307
Model Fit										
R2-Score	0.6385	0.6075	0.6489	0.6168	0.6549	0.6187	0.6593	0.6170	0.6652	0.6066

Figure 4.16 Average speed estimation: scaled data set

	AVERAGE SPEED									
	1st group		2nd group		3rd group		4th group		5th group	
	Training Set	Test set	Training Set	Test set	Training Set	Test set	Training Set	Test set	Training Set	Test set
Normalization	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO
Weight	0.8	0.2	0.8	0.2	0.8	0.2	0.8	0.2	0.8	0.2
Degree	5	5	6	6	7	7	8	8	9	9
Variables										
Hour	+	+	+	+	+	+	+	+	+	+
Day	+	+	+	+	+	+	+	+	+	+
C19-shutdowns	+	+	+	+	+	+	+	+	+	+
Evapo	+	+	+	+	+	+	+	+	+	+
Rel-hum	-	-	-	-	-	-	-	-	-	-
Workday	-	-	-	-	-	-	-	-	-	-
Vpd	-	-	-	-	-	-	-	-	-	-
Error										
MAE	6.6584	6.9949	6.5011	6.8433	6.3993	6.7591	6.3419	6.7679	6.3414	6.7338
MSE	85.5623	89.8387	83.1439	87.5976	816827.0000	86.6119	80.8427	88.8796	80.4548	86.3812
MAPE	9.5420%	9.8224%	9.3226%	9.6098%	9.1801%	9.5107%	9.1030%	9.5142%	9.0960%	9.4775%
MedAE	4.8609	5.3060	4.6983	5.0400	4.5927	4.8386	4.5064	4.7940	4.5757	4.8016
RMSE	9.2500	9.4783	9.1183	9.3594	9.0378	9.3066	8.9913	9.4276	8.9697	9.2941
Model Fit										
R2-Score	0.6526	0.6322	0.6624	0.6413	0.6684	0.6454	0.6718	0.6361	0.6733	0.6463

Figure 4.17 Average speed estimation: non-scaled data set

	FLOW									
	1st group		2nd group		3rd group		4th group		5th group	
	Training Set	Test set	Training Set	Test set	Training Set	Test set	Training Set	Test set	Training Set	Test set
Normalization	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Weight	0.8	0.2	0.8	0.2	0.8	0.2	0.8	0.2	0.8	0.2
Degree	5	5	6	6	7	7	8	8	9	9
Variables										
Hour	+	+	+	+	+	+	+	+	+	+
Day	+	+	+	+	+	+	+	+	+	+
C19-shutdowns	+	+	+	+	+	+	+	+	+	+
Evapo	+	+	+	+	+	+	+	+	+	+
Rel-hum	-	-	-	-	-	-	-	-	-	-
Workday	-	-	-	-	-	-	-	-	-	-
Vpd	-	-	-	-	-	-	-	-	-	-
Error										
MAE	0.0816	0.0820	0.0790	0.0778	0.0747	0.0744	0.0713	0.0717	0.0696	0.0713
MSE	0.0117	0.0122	0.0110	0.0107	0.0101	0.0099	0.0093	0.0092	0.0089	0.0091
MAPE	40.3227%	41.1918%	37.7685%	37.8364%	34.2479%	34.8593%	32.5596%	34.7609%	31.7482%	34.2676%
MedAE	0.0625	0.0623	0.0614	0.0589	0.0562	0.0555	0.0528	0.0539	0.0515	0.0541
RMSE	0.1081	0.1103	0.1047	0.1033	0.1006	0.0996	0.0962	0.0958	0.0945	0.0951
Model Fit										
R2-Score	0.6846	0.6699	0.7041	0.7107	0.7272	0.7309	0.7501	0.7512	0.7590	0.7547

Figure 4.18 Traffic flow estimation: scaled data set

Depending on the information above, we select three models for each dependent variables which are:

1. Traffic flow estimation with 9-degree multivariate polynomial with scaled data set. Independent variables: day, hour, evapotranspiration, Covid-19 shutdowns:



Figure 4.19 Traffic flow training set predictions

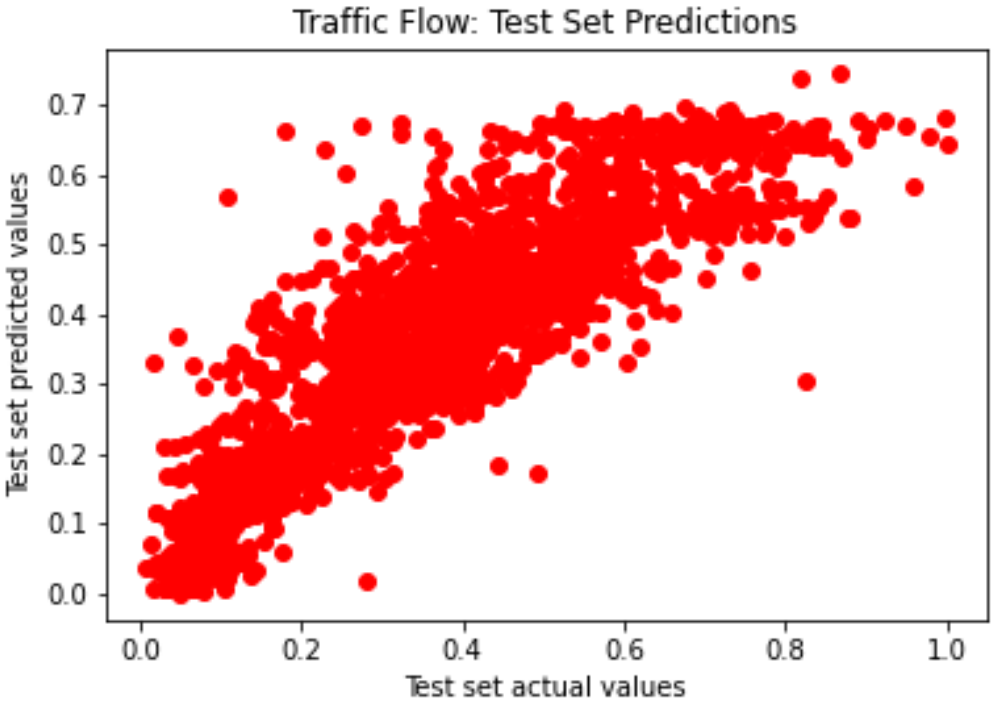


Figure 4.20 Traffic flow test set predictions

2. Number of vehicle estimation with 9-degree multivariate polynomial with nonscaled data set. Independent variables: day, hour, evapotranspiration,

Covid-19 shutdowns:

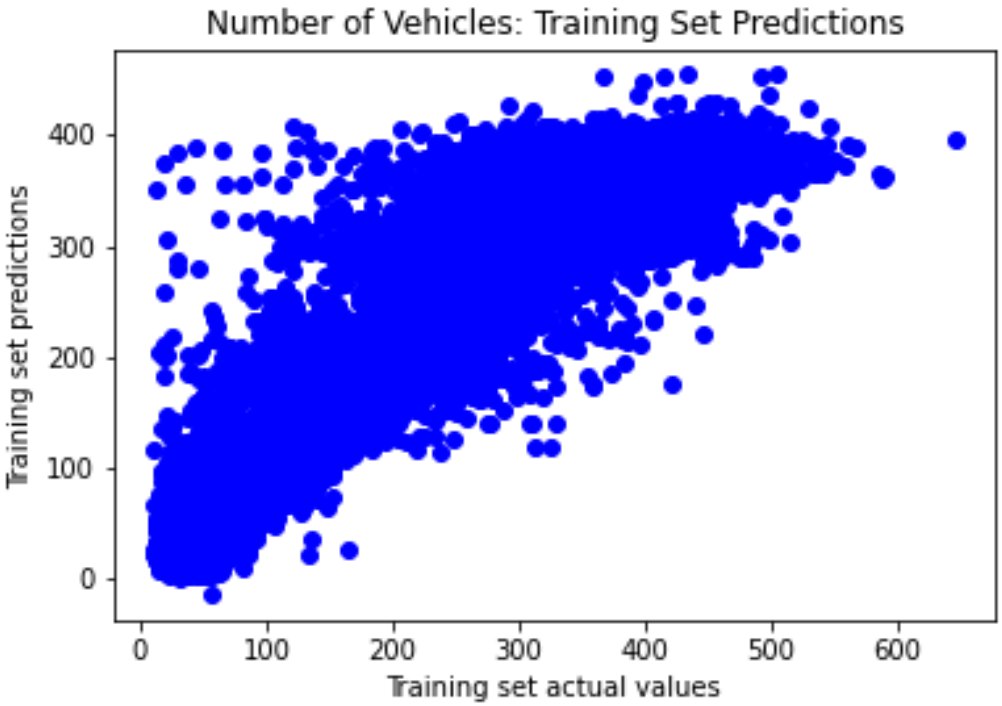


Figure 4.21 Number of vehicles training set predictions

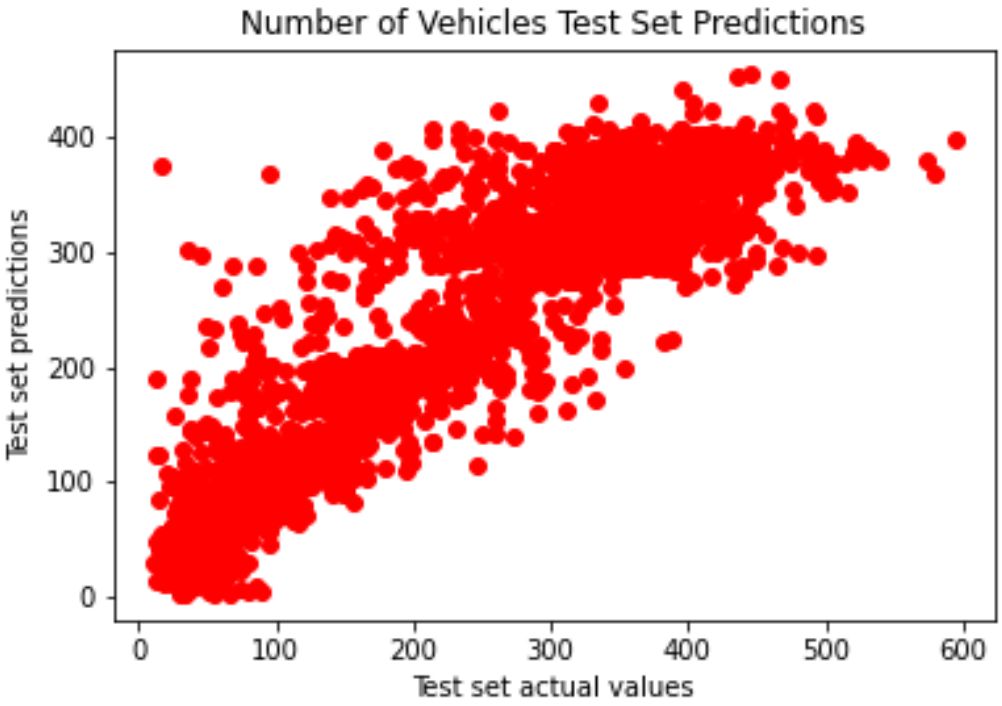


Figure 4.22 Number of vehicles test set predictions

3. Average speed estimation with 9-degree multivariate polynomial with nonscaled

data set. Independent variables: day, hour, evapotranspiration, Covid-19 shutdowns:

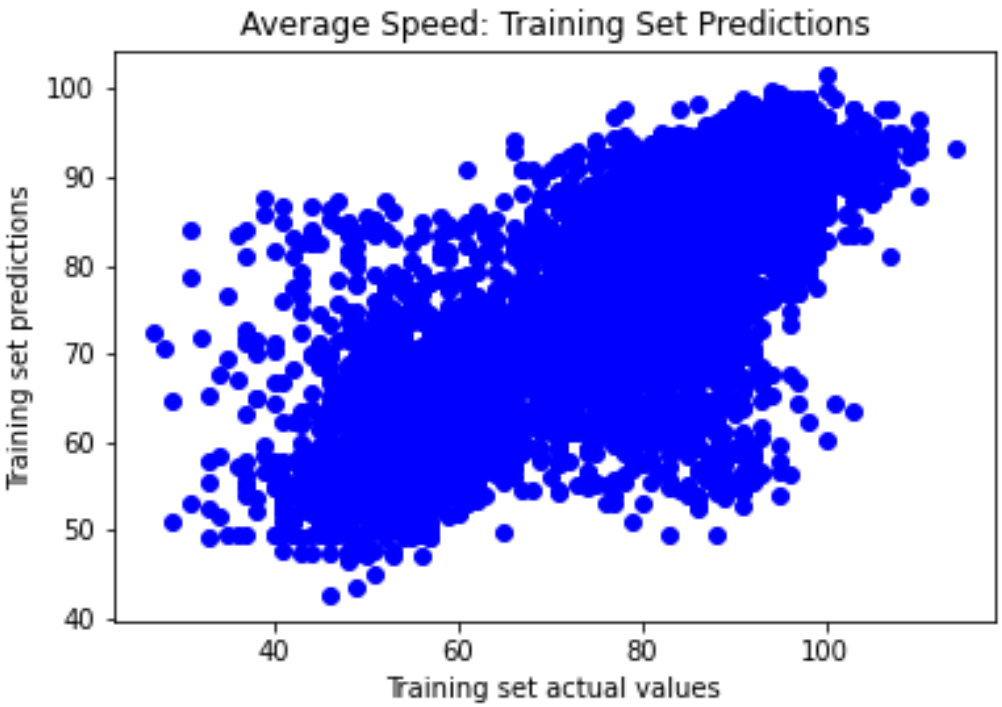


Figure 4.23 Average speed training set predictions

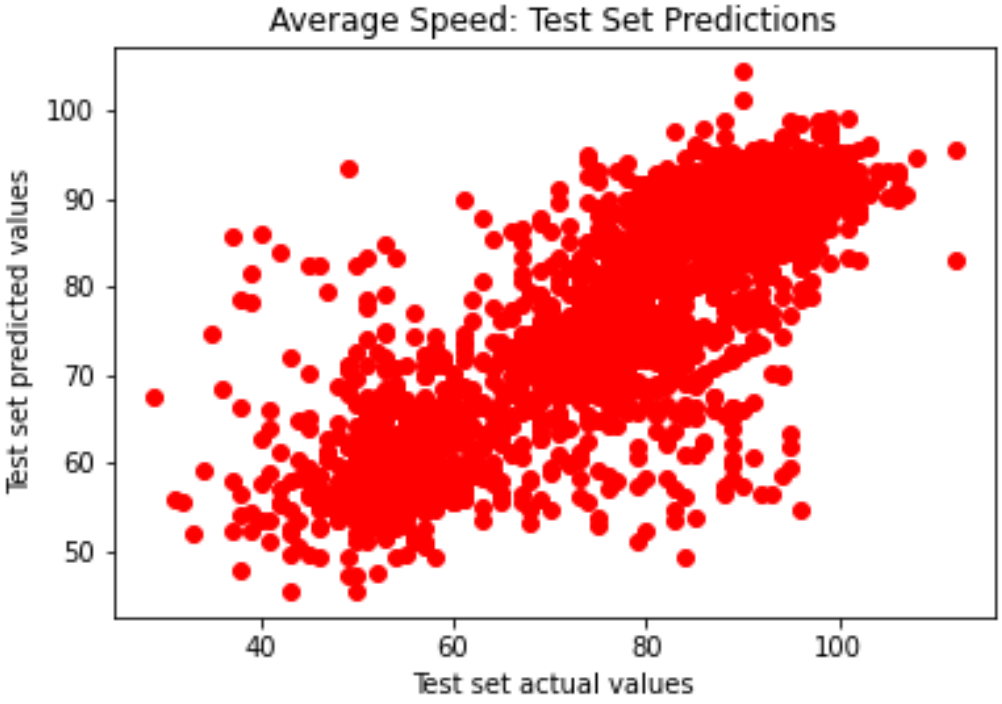


Figure 4.24 Average speed test set predictions

5
CONCLUSIONS AND FUTURE SCOPE OF WORK

In this thesis, we created a traffic model for a specific location in İstanbul is created. This model predicts the average speed, number of vehicles and traffic flow on this specific location based on day, hour, weather and Covid-19 pandemic shutdowns. Comparisons show that Covid-19 pandemics dramatically affect the traffic intensity, also, there are some key variables in weather forecasts such as evapotranspiration, vapour pressure deficit, relative humidity that affect the traffic intensity.

Our results suggests that they are relatively high correlation between some weather conditions mentioned above and traffic variables, additionally there are also relatively high correlation between Covid-19 pandemic shutdowns and traffic variables. Error measurements also claims that our model predicts these three values (speed, number of vehicles and flow) with a reasonable margin for a specific location.

We worked on a different locations and we have similar error results which is likely to demonstrate that our model is generalizable with small changes to adapt large scale. The most important thing is to apply some sort of encoding for different coordinates when making a generalization of this model. However, it is better to use this model type for every specific location separately, because regression analysis is not the best way for making calculations based on different coordinates.

When the volume of our data set and our intention to keep number of variables as minimum level are considered, the results that is obtained by the model can be considered as satisfactory. Furthermore, it should be noted that performance of the model is likely to improve with a larger data set. For a future work, it is aimed to study a similar traffic model with a larger data set and more coordinates, presumably with a different machine learning algorithm.

REFERENCES

[1] J. Kallrath, *Modeling languages in mathematical optimization*. Springer Science & Business Media, 2004, vol. 88.

[2] J. Caldwell, Y. Ram, "Mathematical modelling: Concepts and case studies," 1999.

[3] G. Marion, D. Lawson, "An introduction to mathematical modelling," *Edinburgh: Bioinformatics and Statistics Scotland, University of Bristol*, 2008.

[4] H. Schichl, "Models and the history of modeling," in *Modeling languages in mathematical optimization*, Springer, 2004, pp. 25–36.

[5] B. D. Greenshields, J. Thompson, H. Dickinson, R. Swinton, "The photographic method of studying traffic behavior," in *Highway Research Board Proceedings*, vol. 13, 1934.

[6] B. Greenshields, J. Bibbins, W. Channing, H. Miller, "A study of traffic capacity," in *Highway research board proceedings*, National Research Council (USA), Highway Research Board, vol. 1935, 1935.

[7] M. J. Lighthill, G. B. Whitham, "On kinematic waves i. flood movement in long rivers," *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, vol. 229, no. 1178, pp. 281–316, 1955.

[8] M. J. Lighthill, G. B. Whitham, "On kinematic waves ii. a theory of traffic flow on long crowded roads," *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, vol. 229, no. 1178, pp. 317–345, 1955.

[9] P. I. Richards, "Shock waves on the highway," *Operations research*, vol. 4, no. 1, pp. 42–51, 1956.

[10] W. S. McCulloch, W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The bulletin of mathematical biophysics*, vol. 5, no. 4, pp. 115–133, 1943.

[11] A. Turing, "Computing machinery and intelligence," 1950.

[12] A. L. Samuel, "Machine learning," *The Technology Review*, vol. 62, no. 1, pp. 42–45, 1959.

[13] I. Alam, D. M. Farid, R. J. Rossetti, "The prediction of traffic flow with regression analysis," in *Emerging Technologies in Data Mining and Information Security*, Springer, 2019, pp. 661–671.

[14] L. Pun, P. Zhao, X. Liu, "A multiple regression approach for traffic flow estimation," *IEEE access*, vol. 7, pp. 35 998–36 009, 2019.

[15] D. Branston, H. Van Zuylen, "The estimation of saturation flow, effective green time and passenger car equivalents at traffic signals by multiple linear regression," *Transportation Research*, vol. 12, no. 1, pp. 47–53, 1978.

[16] F. Yasli, H. I. Turkmen, M. A. Guvensan, "Long-term traffic speed estimation via regression using weekly day patterns," in *2019 Innovations in Intelligent Systems and Applications Conference (ASYU)*, IEEE, pp. 1–6.

[17] Z. Zhang, X. Yang, "Freeway traffic speed estimation by regression machine-learning techniques using probe vehicle and sensor detector data," *Journal of transportation engineering, Part A: Systems*, vol. 146, no. 12, e04020138–e04020138, 2020.

[18] M. Cools, E. Moons, G. Wets, "Assessing the impact of weather on traffic intensity," *Weather, Climate, and Society*, vol. 2, no. 1, pp. 60–68, 2010.

[19] B. L. Smith, K. G. Byrne, R. B. Copperman, S. M. Hennessy, N. J. Goodall, "An investigation into the impact of rainfall on freeway traffic flow," in *83rd annual meeting of the Transportation Research Board, Washington DC*, Citeseer, 2004.

[20] Z. H. Khan, S. A. A. Shah, T. A. Gulliver, "A macroscopic traffic model based on weather conditions," *Chinese Physics B*, vol. 27, no. 7, p. 070 202, 2018.

[21] C.-H. Chou, Y. Huang, C.-Y. Huang, V. S. Tseng, "Long-term traffic time prediction using deep learning with integration of weather effect," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Springer, 2019, pp. 123–135.

[22] D. Akin, V. P. Sisiopiku, A. Skabardonis, "Impacts of weather on traffic flow characteristics of urban freeways in istanbul," *Procedia-Social and Behavioral Sciences*, vol. 16, pp. 89–99, 2011.

[23] S. Toksoy, A. Atalay, "Assessing the impact of covid-19 pandemic on urban transportation and air quality in istanbul,"

[24] U. Alcan, F. Kaçar, "Impact analysis of covid-19 pandemic on istanbul traffic with big data tools," *Electrica*, vol. 22, no. 2, pp. 226–236, 2022.

[25] D. Edwards, M. Hamson, *Mathematical modelling skills*. Springer, 1996.

[26] T. Witelski, M. Bowen, *Methods of mathematical modelling*. Springer, 2015.

[27] D. N. Burghes, "Modelling with differential equations," Tech. Rep., 1981.

[28] W. Blum, "Quality teaching of mathematical modelling: What do we know, what can we do?" In *The proceedings of the 12th international congress on mathematical education*, Springer, Cham, 2015, pp. 73–96.

[29] D. A. Towers, D. Edwards, M. Hamson, *Guide to mathematical modelling*. Bloomsbury Publishing, 2020.

[30] E. A. Bender, *An introduction to mathematical modeling*. Courier Corporation, 2000.

[31] J. M. G. Sánchez *et al.*, "Modelling in mathematical programming," *International Series in Operations Research and Management Science, Springer*, no. 978-3, pp. 030–57 250, 2021.

[32] P. Renard, A. Alcolea, D. Ginsbourger, "Stochastic versus deterministic approaches," *Environmental modelling: Finding simplicity in complexity*, pp. 133–149, 2013.

[33] S. Serovajsky, *Mathematical modelling*. Chapman and Hall/CRC, 2021.

[34] A. K. Thakur, "Model: Mechanistic vs empirical," in *New trends in pharmacokinetics*, Springer, 1991, pp. 41–51.

[35] I. Nestorov, M. Rowland, S. Hadjitodorov, I. Petrov, "Empirical versus mechanistic modelling: Comparison of an artificial neural network to a mechanistically based model for quantitative structure pharmacokinetic relationships of a homologous series of barbiturates," *AAPS PharmSci*, vol. 1, no. 4, pp. 5–13, 1999.

[36] S. M. Moghadas, M. Jaberi-Douraki, *Mathematical Modelling: A Graduate Textbook*. John Wiley & Sons, 2018.

[37] R. W. Hall, "Discrete models/continuous models," *Omega*, vol. 14, no. 3, pp. 213–220, 1986.

[38] F. van Wageningen-Kessels, H. Van Lint, K. Vuik, S. Hoogendoorn, "Genealogy of traffic flow models," *EURO Journal on Transportation and Logistics*, vol. 4, no. 4, pp. 445–473, 2015.

[39] C. F. Daganzo, "The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory," *Transportation Research Part B: Methodological*, vol. 28, no. 4, pp. 269–287, 1994.

[40] J. Drake, J. Schofer, A. May, "A statistical analysis of speed-density hypotheses," *Traffic Flow and Transportation*, 1965.

[41] S. Smulders, "Control of freeway traffic flow by variable speed signs," *Transportation Research Part B: Methodological*, vol. 24, no. 2, pp. 111–132, 1990.

[42] N. Caceres, L. M. Romero, F. G. Benitez, J. M. del Castillo, "Traffic flow estimation models using cellular phone data," *IEEE Transactions on Intelligent Transportation Systems*, vol. 13, no. 3, pp. 1430–1441, 2012.

[43] F. Kessels, Kessels, Rauscher, *Traffic flow modelling*. Springer, 2019.

[44] L. A. Pipes, "Car following models and the fundamental diagram of road traffic," *Transportation Research/UK/*, 1966.

[45] R. E. Chandler, R. Herman, E. W. Montroll, "Traffic dynamics: Studies in car following," *Operations research*, vol. 6, no. 2, pp. 165–184, 1958.

[46] R. Herman, E. W. Montroll, R. B. Potts, R. W. Rothery, "Traffic dynamics: Analysis of stability in car following," *Operations research*, vol. 7, no. 1, pp. 86–106, 1959.

[47] D. C. Gazis, R. Herman, R. W. Rothery, "Nonlinear follow-the-leader models of traffic flow," *Operations research*, vol. 9, no. 4, pp. 545–567, 1961.

[48] S. Hoogendoorn, V. Knoop, "Traffic flow theory and modelling," *The transport system and transport policy: an introduction*, pp. 125–159, 2013.

[49] L. Elefteriadou *et al.*, *An introduction to traffic flow theory*. Springer, 2014.

[50] A. S. Michaels, W. R. Vieth, J. A. Barrie, "Solution of gases in polyethylene terephthalate," *Journal of Applied Physics*, vol. 34, no. 1, pp. 1–12, 1963.

[51] M. Brackstone, M. McDonald, "Development and deployment of inter-urban att test scenarios," in *Intelligent Transportation: Realizing the Future. Abstracts of the Third World Congress on Intelligent Transport SystemsITS America*, 1996.

[52] M. Brackstone, M. McDonald, "Car-following: A historical review," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 2, no. 4, pp. 181–196, 1999.

[53] W. Knospe, L. Santen, A. Schadschneider, M. Schreckenberg, "Empirical test for cellular automaton models of traffic flow," *Physical Review E*, vol. 70, no. 1, p. 016 115, 2004.

[54] M. Cremer, J. Ludwig, "A fast simulation model for traffic flow on the basis of boolean operations," *Mathematics and computers in simulation*, vol. 28, no. 4, pp. 297–303, 1986.

[55] K. Nagel, M. Schreckenberg, "A cellular automaton model for freeway traffic," *Journal de physique I*, vol. 2, no. 12, pp. 2221–2229, 1992.

[56] H. Zhou, Y. Sun, X. Qin, X. Xu, R. Yao, "Modeling discretionary lane-changing behavior on urban streets considering drivers' heterogeneity," *Transportation letters*, vol. 12, no. 3, pp. 213–222, 2020.

[57] M. Rahman, M. Chowdhury, Y. Xie, Y. He, "Review of microscopic lane-changing models and future research opportunities," *IEEE transactions on intelligent transportation systems*, vol. 14, no. 4, pp. 1942–1956, 2013.

[58] S. Moridpour, M. Sarvi, G. Rose, "Lane changing models: A critical review," *Transportation letters*, vol. 2, no. 3, pp. 157–173, 2010.

[59] J. A. Michon, "Psychonomie onderweg," *Inaugural speech at University of Groningen*, vol. 2, 1971.

[60] R. Sukthankar, S. Baluja, J. Hancock, "Evolving an intelligent vehicle for tactical reasoning in traffic," in *Proceedings of International Conference on Robotics and Automation*, IEEE, vol. 1, 1997, pp. 519–524.

[61] W.-L. Jin, *Introduction to Network Traffic Flow Theory: Principles, Concepts, Models, and Methods*. Elsevier, 2021.

[62] J. Popping, "An overview of microscopic and macroscopic traffic models," 2013.

[63] H. Payne, "Models of freeway traffic and control. mathematical models of public systems. simulation councils," *Inc., Vista, CA, USA*, 1971.

[64] S. P. Hoogendoorn, P. H. Bovy, "State-of-the-art of vehicular traffic flow modelling," *Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering*, vol. 215, no. 4, pp. 283–303, 2001.

[65] A. Aw, M. Rascle, "Resurrection of" second order" models of traffic flow," *SIAM journal on applied mathematics*, vol. 60, no. 3, pp. 916–938, 2000.

[66] I. Prigogine, F. C. Andrews, "A boltzmann-like approach for traffic flow," *Operations Research*, vol. 8, no. 6, pp. 789–797, 1960.

[67] R. Dixit, *Data Analysis with Python: Introducing NumPy, Pandas, Matplotlib, and Essential Elements of Python Programming*. BPB Online, 2022.

[68] A. Jung, *Machine Learning: The Basics*. Springer Nature, 2022.

[69] M. Learning, *Machine Learning*. Machine Learning, 2020.

[70] *Google maps 41°05'29.8"n 29°03'12.0"e*, (Date last accessed 15/01/2023). [Online]. Available: `https://www.google.com/maps/place/41%C2%B005'29.8%22N+29%C2%B003'12.0%22E`.

[71] *Hourly traffic density data set*, (Date last accessed 15/01/2023). [Online]. Available: https://data.ibb.gov.tr/en/dataset/hourly-traffic-density-data-set.

[72] *Traffic density data for january 2020*, (Date last accessed 15/01/2023). [Online]. Available: https://data.ibb.gov.tr/en/dataset/hourly-traffic-density-data-set/resource/db9c7fb3-e7f9-435a-92f4-1b917e357821.

[73] *Historical weather api*, (Date last accessed 15/01/2023). [Online]. Available: https://open-meteo.com/en/docs/historical-weather-api.

PUBLICATIONS FROM THE THESIS

Conference Papers

1. Ö. Yıldırım, S. Demir, "Mathematical modelling of traffic with regression analysis and machine learning." *In Sixth International Conference on Analysis and Applied Mathematics* , p.24, 2022.

Projects

1. Ö. Yıldırım, S. Demir, Mathematical Modelling and Solution Methods for the Problems Generated, Scientific Research Project: No. FYL-2021-4399, Researcher.