

# **TRAFFIC PREDICTION BASED ON PROBABILISTIC GRAPHICAL METHOD**

MOHAN RAJU JAYASHANKARAMMA SHANKAR RAJ  
August, 2020

SUPERVISORS:  
Dr. Frank Badu Osei  
Dr. Michael Yang





# TRAFFIC PREDICTION BASED ON PROBABILISTIC GRAPHICAL METHOD

MOHAN RAJU JAYASHANKARAMMA SHANKAR RAJ  
Enschede, The Netherlands [July - 2020]

Thesis submitted to the Faculty of Geo-Information Science and Earth Observation of the University of Twente in partial fulfilment of the requirements for the degree of Master of Science in Geo-information Science and Earth Observation.  
Specialization: Geoinformatics

SUPERVISORS:  
Dr. Frank Badu Osei  
Dr. Michael Yang

THESIS ASSESSMENT BOARD:  
Prof. Dr. IR. Alfred Stein (Chair)  
Dr. IR. M. Van Keulen (External Examiner, faculty EEMCS, UT)  
Dr. Frank Badu Osei  
Dr. Michael Yang

PROCEEDURAL ADVISOR:  
J.P.G. Bakx MSc

#### DISCLAIMER

This document describes work undertaken as part of a programme of study at the Faculty of Geo-Information Science and Earth Observation of the University of Twente. All views and opinions expressed therein remain the sole responsibility of the author, and do not necessarily represent those of the Faculty.

# ABSTRACT

Road traffic congestion is considered as a ubiquitous, a phenomenon which occurs during peak hours in densely covered areas like in Amsterdam and Rotterdam. This includes the complexity of planning infrastructure for transportation. One of the critical issues involved is to identify design elements (Traffic lanes, road type and driving lane). On the other hand, the road network in an urban area is a distinctive network. The study of such systems includes the evaluation and optimization of the traffic state. Needless to say, more traffic on the outer roads has a more significant impact on the traffic condition within the city: A considerable deal of start and end of traffic journey on the urban road network. Here the question arises is how traffic was developed over time within the given network? This research proposes a novel framework which is a threefold approach to address the traffic congestion on the Dutch highways. The entire framework is developed based on a Bayesian network (BN) approach. The network structure was modelled based on the probabilistic dependency, which helps in identifying causes for congestion. Firstly, the BN was established over the set of random variables which is called as ‘contributing variables’ such as Time of the day, Speed and vehicle characteristics. The discrete variables are further evaluated individually to identify the major causes of traffic congestion.

Furthermore, the association between contributing variables and congestion occurrence was quantified by measuring the odds ratio. Secondly, using the same network model, the traffic flow predictions are performed by using continuous variables. For the continuous variables, the data was downloaded for 30 & 60-minute intervals, respectively. The accuracy of the result of the mentioned 2-time intervals was measured based on RMSE, MAE and MAPE. Thirdly, the characteristics of Amsterdam and Rotterdam road network was studied based on the qualitative and quantitative aspect. A qualitative aspect includes the visual inspection that describes the texture and gradient of the road network. Quantitative aspect involves the descriptive statistics that describe node density, street length, number of nodes and average street length.

The entire study made use of two different data sources.

- For prediction and congestion diagnosis- historical dataset extracted from NDW, which is an open-access database which attempts to collect, process, store and distribute all the traffic-related data.
- For analysing road network characteristics – OSM database was considered.

The results from the BN model states that Firstly - the contributing variables such as Travel Time, Time of the day and speed as a major impact on traffic congestion occurrence. Secondly – 30 minutes predictions are better, as the RMSE value was recorded lower compare to 60 minutes interval. Thirdly, based on the road network statistics - Amsterdam as many nodes, edges and networks are in semi-structured type and Rotterdam has fewer number of edges and nodes with square typed structure.

**Keywords:** Street network, Bayesian Network Model, Intelligent Transport System, Visualization, Sensors, Hill Climbing algorithm, Probability distribution

## ACKNOWLEDGEMENTS

Thanks to my mother, elder siblings and brothers-in-law who always believed in me and helped me in pursuit of my dreams. Their unconditional love keeps me growing in life.

Thanks to my first supervisor, Dr Frank Badu Osei, for the motivation and comments that he provided during my entire 2<sup>nd</sup> year of masters. His support and guidance during a crucial phase of this research helped to finish the study in the best way.

Thanks to my chair Prof. Dr IR. Alfred Stein and my second supervisor Dr Michael Yang for the critical questions during the proposal phase and midterm that helped me to think out of the box for the solutions.

Thanks to drs. J.P.G. Wan Bakx for his enormous support and being a mentor for my internship at United Nations.

Extended thanks to Dr.Elenna Dugundji for being a host supervisor and guiding me to complete 1<sup>st</sup> internship at CWI, Amsterdam. Secondly, Ayako Kagawa for providing me an opportunity to intern with GIS-Carto unit at United Nations headquarters.

Thanks to all the teaching faculty and staff at ITC for making my entire journey unforgettable.

Thanks to almighty God for giving me the confidence to pursue my passion, knowledge, and ability to finish the study. Finally, I would like to thank my therapist, who always inspired me to practice yoga and meditation for physical and mental well-being which helped me sustain during pandemic times.

And finally, I would like to thank all my friends at ITC who became an important part of my life in these 24 months of the journey. Thanks for all the special moments and memories you decided to share with me. Thanks for being there for me at tough times.

# TABLE OF CONTENTS

---

1.	Introduction.....	1
1.1.	Background.....	1
1.2.	Problem Statement.....	3
1.3.	Research Objectives.....	3
1.4.	Innovation and Novelty of the Research.....	3
1.5.	Thesis Structure.....	4
2.	Literature Review.....	5
2.1.	Probabilistic Graphical Method.....	6
2.1.1.	Markov Network.....	7
2.1.2.	Factor Graph.....	8
2.1.3.	Bayesian Network.....	9
3.	Study Area.....	11
3.1.	Study Area.....	11
3.2.	Amsterdam.....	12
3.3.	Rotterdam.....	13
4.	Data and Software.....	14
4.1.	Historical Traffic data.....	14
4.2.	Road Networks Data.....	14
4.3.	Scripts (R and Python).....	15
5.	Methodology.....	16
5.1.	Pre-Processing Data.....	16
5.2.	Method Explanation.....	19
5.3.	Variables.....	20
5.4.	Structure.....	22
5.5.	Parameter Learning.....	24
5.6.	Identifying major causes for increased ‘Traffic Intensity’ by measuring the odd ratio.....	25
5.7.	Evaluating model performance for the continuous variables using RMSE, MAE & MAPE.....	26
6.	Results & Discussions.....	27
6.1.	Model Authentication.....	28
6.2.	Parameter Estimation Results.....	29
6.3.	Analysis and Inference.....	35
6.3.1.	Influence of ‘Time of the Day’ on ‘Traffic Intensity’.....	36
6.3.2.	Influence on Traffic Intensity during ‘Weekdays & Weekends’.....	39
6.3.3.	Influence of ‘Vehicle Size’ on ‘Traffic Intensity’.....	42
6.3.4.	Influence of ‘Travel Time’ on ‘Traffic Intensity’.....	45
6.3.5.	Identifying the major causes of increased ‘Traffic Intensity’.....	48
6.3.6.	Traffic flow prediction.....	51
6.4.	Accuracy Assessment of the BN model through RMSE, MAE and MAPE.....	55
6.5.	Road Network characteristics of Amsterdam & Rotterdam.....	55
6.5.1.	Analysing street network.....	56
6.5.2.	Open Street Map.....	56
7.	Conclusion and recommendations.....	58
7.1.	Conclusion.....	58
7.2.	Research Questions: Answered.....	59

7.3. Recommendations..... 60



## LIST OF FIGURES

---

Figure 1(left): Comparison of congestion indices for Urban networks in six cities of Netherlands from 2011-2015 -----	11
Figure 2(right): Comparison of congestion indices between motorways and urban networks of six cities in Netherlands from 2012 to 2015 -----	11
Figure 3: Road network of Amsterdam with sensors (With base layer) -----	12
Figure 4: Road network of Amsterdam with sensors (Without base layer) -----	12
Figure 5: Road network of Rotterdam with sensors (Without base layer) -----	13
Figure 6: Road network of Rotterdam with sensors (With base layer)-----	13
Figure 7: Snapshot of the Road network (Amsterdam & Rotterdam)-----	15
Figure 8: The snapshot of the driving lane-----	17
Figure 9: Data Normalization of input variables – Intensity (A), Travel Time (B) and Speed (C) -----	19
Figure 10: Three types for Bayesian Network Model -----	22
Figure 11: Bayesian Network Structure using IAMB Algorithm -----	23
Figure 12: Bayesian Network Structure using HC Algorithm -----	24
Figure 13: Network score values of Bayesian Network structure -----	28
Figure 14: Representing Marginal Probability Distribution for 7 nodes in the designed BN mode <b>(Amsterdam, 2018)</b> . Node <b>A</b> - Intensity, <b>B</b> - Travel Time, <b>C</b> - Traffic Speed, <b>D</b> - Vehicle Categories, <b>E</b> - Weekend or Weekday, <b>F</b> - Day in week & <b>G</b> - Time of the day-----	29
Figure 15: Representing Marginal Probability Distribution for 7 nodes in the designed BN mode <b>(Amsterdam, 2018)</b> . Node <b>A</b> - Intensity, <b>B</b> - Travel Time, <b>C</b> - Traffic Speed, <b>D</b> - Vehicle Categories, <b>E</b> - Weekend or Weekday, <b>F</b> - Day in week & <b>G</b> - Time of the day-----	30
Figure 16: Representing Marginal Probability Distribution for 7 nodes in the designed BN mode <b>(Amsterdam, 2020)</b> . Node <b>A</b> - Intensity, <b>B</b> - Travel Time, <b>C</b> - Traffic Speed, <b>D</b> - Vehicle Categories, <b>E</b> - Weekend or Weekday, <b>F</b> - Day in week & <b>G</b> - Time of the day-----	31
Figure 17: Representing Marginal Probability Distribution for 7 nodes in the designed BN mode <b>(Amsterdam, 2020)</b> . Node <b>A</b> - Intensity, <b>B</b> - Travel Time, <b>C</b> - Traffic Speed, <b>D</b> - Vehicle Categories, <b>E</b> - Weekend or Weekday, <b>F</b> - Day in week & <b>G</b> - Time of the day-----	32
Figure 18: Representing Marginal Probability Distribution for 7 nodes in the designed BN mode <b>(Rotterdam, 2019)</b> . Node <b>A</b> - Intensity, <b>B</b> - Travel Time, <b>C</b> - Traffic Speed, <b>D</b> - Vehicle Categories, <b>E</b> - Weekend or Weekday, <b>F</b> - Day in week & <b>G</b> - Time of the day-----	33
Figure 19: Representing Marginal Probability Distribution for 7 nodes in the designed BN mode <b>(Rotterdam, 2020)</b> . Node <b>A</b> - Intensity, <b>B</b> - Travel Time, <b>C</b> - Traffic Speed, <b>D</b> - Vehicle Categories, <b>E</b> - Weekend or Weekday, <b>F</b> - Day in week & <b>G</b> - Time of the day-----	34
Figure 20: CPT table showing impact of ‘Time of the day’ on ‘Traffic Intensity’ – AMS, 2018-----	36
Figure 21: CPT table showing impact of ‘Time of the day’ on ‘Traffic Intensity’ – RTM, 2018 -----	36
Figure 22: CPT table showing impact of ‘Time of the day’ on ‘Traffic Intensity’ – AMS, 2019-----	37
Figure 23: CPT table showing impact of ‘Time of the day’ on ‘Traffic Intensity’ – RTM, 2019 -----	37
Figure 24: CPT table showing impact of ‘Time of the day’ on ‘Traffic Intensity’ – AMS, 2020-----	38
Figure 25: CPT table showing impact of ‘Time of the day’ on ‘Traffic Intensity’ – RTM, 2020 -----	38
Figure 26: CPT table showing impact on ‘Traffic Intensity’ during WD/WE – AMS, 2018 -----	39
Figure 27: CPT table showing impact on ‘Traffic Intensity’ during WD/WE – RTM, 2018 -----	39
Figure 28: CPT table showing impact on ‘Traffic Intensity’ during WD/WE – AMS, 2019 -----	40
Figure 29: CPT table showing impact on ‘Traffic Intensity’ during WD/WE – RTM, 2019 -----	40
Figure 30: CPT table showing impact on ‘Traffic Intensity’ during WD/WE – AMS, 2020 -----	41

Figure 31: CPT table showing impact on ‘Traffic Intensity’ during WD/WE – RTM, 2020 -----	41
Figure 32: CPT table showing the impact of Vehicles Size on ‘Traffic Intensity’ – AMS, 2018 -----	<b>Error!</b>
<b>Bookmark not defined.</b>	
Figure 33: CPT table showing the impact of Vehicles Size on ‘Traffic Intensity’ – RTM, 2018 -----	<b>Error!</b>
<b>Bookmark not defined.</b>	
Figure 34: CPT table showing the impact of vehicles size on ‘Traffic Intensity’ – AMS, 2019 -----	43
Figure 35: CPT table showing the impact of vehicles size on ‘Traffic Intensity’ – RTM, 2019 -----	43
Figure 36: CPT table showing the impact of vehicles size on ‘Traffic Intensity’ – AMS, 2020 -----	44
Figure 37: CPT table showing the impact of vehicles size on ‘Traffic Intensity’ – RTM, 2020 -----	44
Figure 38: CPT table showing the effect of Travel Time on ‘Traffic Intensity’ – AMS, 2018 -----	45
Figure 39: CPT table showing the effect of Travel Time on ‘Traffic Intensity’ – RTM, 2018 -----	45
Figure 40: CPT table showing the effect of Travel Time on ‘Traffic Intensity’ – AMS, 2019 -----	46
Figure 41: CPT table showing the effect of Travel Time on ‘Traffic Intensity’ – RTM, 2019 -----	46
Figure 42: CPT table showing the effect of Travel Time on ‘Traffic Intensity’ – AMS, 2020 -----	47
Figure 43: CPT table showing the effect of Travel Time on ‘Traffic Intensity’ – RTM, 2020 -----	47
Figure 44: OR between Contributing nodes and Traffic Intensity = “High”   AMS & RTM - 2018 -----	48
Figure 45: OR between Contributing nodes and Traffic Intensity = “High”   AMS & RTM - 2019 -----	49
Figure 46: OR between Contributing nodes and Traffic Intensity = “High”   AMS & RTM - 2020 -----	50
Figure 47: Traffic intensity forecasting (30 minutes interval)   AMS - 2018 -----	51
Figure 48: Traffic intensity forecasting (30 minutes interval)   RTM - 2018 -----	52
Figure 49: Traffic intensity forecasting (30 minutes interval)   AMS - 2018 -----	53
Figure 50: Traffic intensity forecasting (30 minutes interval)   RTM - 2018 -----	54
Figure 51: Monochrome images of Amsterdam (left) and Rotterdam(right) street network (One square mile) -----	56
Figure 52: Monochrome images of Amsterdam (left) and Rotterdam(right) street network (Node highlighted) -----	57

## LIST OF TABLES

---

Table 1: Vehicle Characteristics by length differentiated by NDW sensors.-----	16
Table 2: Area and a total count of sensors in the study area.-----	17
Table 4: Discretization of input data-----	20
Table 3: Description of input data-----	21
Table 5: K-fold Cross-validation results of two learning algorithm-----	28
Table 6: Comparison of RMSE, MAE & MAPE for Amsterdam & Rotterdam - 2018-----	55
Table 7: Comparison of RMSE, MAE & MAPE for Amsterdam & Rotterdam - 2020-----	55
Table 8: Descriptive statistics of 2 street networks in the Netherlands-----	57
Table 9: Table showing data coding for the node ‘Days of the week’ implemented in the BN model-----	64
Table 10: Table showing data coding for the node ‘Weekend/Weekday’ implemented in the BN model-----	64
Table 11: Table showing data coding for the node ‘Time of the day’ implemented in the BN model-----	64
Table 12: Range defined for the node ‘Speed’, ‘Intensity’ & ‘Travel time.’-----	64
Table 13: Vehicle categories-----	65



# 1. INTRODUCTION

## 1.1. Background

Across several industrial nations, there is an increasing volume of traffic, especially on the motorways, which is putting pressure on the existing road systems (Whittaker, Garside, & Lindveld, 1997). It is predicted by several academicians that a gradual increase in population and growth would exacerbate delays in traffic, increase energy consumptions, and would also lead to reduced safety standards (Pieters, 2015; Wang et al., 2018). Wegman, (2007) notes that at least 1.2 million people are killed in road accidents throughout the world. With such intensities on the rise, transportation systems mandate forecasts that give an indication of traffic conditions, both on a short term and a long-term basis, because in many cases the occurrence of unpredictable events like accidents or extreme weather events would hinder the management and flow of traffic (Vlahogianni, Karlaftis, & Golias, 2014). Predicting traffic aids in reducing public accidents by notifying traffic administrations, thereby aiding in formulating emergency response plans that can be deployed at the earliest (Zheng, Ismail, & Meng, 2014). The motivation for such forecasting becomes obvious then, that is, actions taken now would have a better chance to succeed if the future conditions of the system are already anticipated.

On parallel lines, it is important to note that real-time traffic flow state identification and traffic predictions have been quite integral within the Intelligent Transportation Systems (ITS) since the 1980s (Lu, Sun, & Qu, 2015). Based on present and past traffic information, ITS delves into predictions that are made from few seconds to few hours into the future (Vlahogianni et al., 2014). Much of the focus back then was on developing methods and models that could be used for predicting traffic characteristics like volume, travel time, density, and speed of traffic (Ahmed & Cook, 1979). However, this classical approach provided only a preliminary understanding of the traffic patterns. Hence over the years, the focus shifted to using data-driven approaches which involved the use of a variety of algorithms that emphasise on the use of real-time information. The use of real-time data is, therefore, of utmost importance for traffic prediction (Davarynejad et al., 2011). Continuous prediction and revision of the traffic states in terms of density and traffic behaviour, and congestion Wegman, (2007) is required for an effective traffic monitoring system.

Meanwhile, traffic data is vast and multisource, and this fact has been acknowledged by several academicians (Li et al., 2015). Hence, there arises a need to effectively manage and control such data. The continuous development in communication technologies has proven to bring about a dramatic shift in the management and storage of such large amounts of data.

Traffic congestion is a serious problem in many of the big cities in the world. Traffic congestions have resulted in extra energy costs worth \$160 million in at least 471 urban areas of the United States in 2014 (Schrank, Eisele, & Lomax, 2019). According to Algemene Nederlandse Wielrijders Bond (ANWB), there was a 17% rise in the volume of traffic jams on Netherlands in 2019 when compared to the previous years. These traffic jams are frequent during morning and evening rush hours. Several studies released by the state authorities argue that the present-day congestion levels in the Netherlands reel back to 2008, despite the continuous expansion of roads network in the country. Much of this can be attributed to a dramatic increase in the number of cars on the road. The largest traffic blackspot in the country is from Amsterdam to Apeldoorn, because this route is used by several millions of people to traverse to workplaces.

On the other hand, cyclists and pedestrians are also a major contribution to the traffic congestion in Amsterdam city itself. According to the 2011 study, human error played a vital role in most of the accidents. For example, majority of serious injuries were hurt by cyclists, just 45% of bicycles are provided with functioning lights, 23% of bikes users obey speed limits, and 80% of critical accidents occur on roads with 50 km/hr speed limits (Amsterdam, 2020). Geographically the four major cities Amsterdam, Rotterdam, The Hague and Utrecht are located on the western part of the Netherlands in an urban ring with a distance of 40-60 km between the four cities and the road networks are gridlocked structure. These geographical characteristics also lead to traffic congestions (Rietveld, 2004).

Furthermore, long durations of traffic congestion cause greater emissions of carbon dioxide and other greenhouse gases, thereby increasing the risk of air pollution and climate variability. Multiple factors contribute to traffic congestions, viz. rainfall, construction or infrastructural related work, double parking on lanes, closure of lanes due to repair works or other road-related works. Such factors lead to traffic congestion. On the other hand, traffic signals could be out of sync owing to malfunctioning of the computers, inadequate green time for traffic passage and the road system proves to be inadequate to control traffic congestions. Hence, in such instances short term traffic predictions of traffic congestions become very important to allow a smooth flow of traffic, thereby reducing additional energy costs and accidents.

However, while looking at traffic predictions, most of the prediction techniques make use of big data that are licensed and expensive. This tends to hinder many researchers for not all can avail such expensive data for conducting research. Furthermore, reliability on volunteered geographic data (VGA) is not considered as robust, because data is not captured by 'sound machines' and 'trained professional'. Using data from such sources tends to question the quality of the outputs (Papapesios et al., 2019). In order to tackle this, there are several agencies and governmental organisations that have started to make data open-sourced, especially for research purposes. However, in the past, traffic congestion was remained untouched due to its complexity of the framework proposed initially. The framework was built based on Advanced Traveller Information System (ATIS). In the historical framework, Origin-Destination data coupled with real-time data obtained from surveillance system were used as an input to predict the movement of traffic congestion. This helped to navigate traffic control and route guidance to alter the traffic congestions. As suggested by Huisken Giovanni, (2006) Dynamic Traffic Assignment (DTA) model, with 3-dimensional data matrix keeping time as a 3<sup>rd</sup> dimension which helps to predict the traffic congestion with good accuracy. As an alternative, the two approaches, such as Least square Estimation and Kalman filtering procedure, were suggested (Dougherty & Cobbett, 1997). The two-wheeler data were gathered in the Netherlands from sensor-based induction loops which collect data such as Volume, Speed, and occupancy at lane level. Later, these data were aggregated at single minute using field expert knowledge the traffic congestion was predicted. These predictions were between 10 to 30 minutes, a single method titled "Multi-Layer Feedforward Artificial Neural Networks". However, these predictions yield reasonable results but were not accurate enough to all occasions. On the other hand, using the same data which was aggregated at 10-30 minutes was used for time series analysis and contrasted with ANN models. The resultant error was classified as False alarm (False prediction) and if there were errors which resulted in no predictions and the results were measured in percentage along with time. According to Huisken Giovanni, (2006), the ANN yielded better accuracy predictions compared to Time series analysis. To draw the conclusion about short term traffic congestion predictions is deficient with the existing techniques. The bottleneck of all the traffic predictions are model-driven. The performance of the model can be scaled based on the data provided.

## 1.2. Problem Statement

This study is a two fold approach which attempts to address traffic congestion at two levels. Firstly, making use of open-sourced data and the concept of Bayesian Network analysis approach to model the probabilistic dependency structure, which leads to the causes of traffic congestion on each of the cities, namely Amsterdam and Rotterdam. The data is downloaded at an interval of 30 mins & 60mins for a period of three months (1st March to 31st March 2018), (1st March to 31st March 2019) and (1st March to 31st March 2020). The Bayesian approach is used to build the joint probability distribution over a set of random variables. With the random set of variables, the probabilistic model is built to predict traffic congestion. To validate the prediction results, observed data for the study period will be used to check the accuracy of the prediction model. The results will then be critically examined for the robustness of the model. Secondly, by making use of OSM data to visually inspect and analyse the structure of road networks.

## 1.3. Research Objectives

The objective of the current study involves the identification and prediction of real-time traffic congestion in Amsterdam using the probabilistic graphical method. Several reports and articles have suggested that of all the cities in the Netherlands. Rotterdam and Amsterdam score badly in terms of traffic flow (Pieters, 2015). For this reason, the present study attempts to explore the possibilities of using the probabilistic graphical method or short-term prediction of real-time traffic congestion in major metropolitan cities by using the cases of road traffic networks in Amsterdam and Rotterdam, Netherlands

### 1.3.1 Sub-research objectives

1. To investigate the characteristics of road segments of Amsterdam and Rotterdam and study their impacts on the traffic prediction mechanisms of the proposed model.
2. To inquire regarding the different methods that can be used to calculate traffic density and understand how the traffic density moves from one node to another node.
3. To examine the accuracy and reliability of the short-term traffic prediction outcomes using Bayesian Networks.

### 1.3.2 Research Questions

1. What are the characteristics of road segments of Amsterdam and Rotterdam road networks?
2. What are the techniques to identify and examine the conditional independence of each node in the road networks, and how are the short-term predictions made?
3. How accurate and reliable are the short-term traffic predictions made using Bayesian Networks?

## 1.4. Innovation and Novelty of the Research

Much of the studies around traffic predictions have made use of data that are collected or obtained through expensive sources. Furthermore, while reviewing literature, many of the scholars use complex fluid dynamic principles to make these predictions. However, the present study aims at bringing innovation in using open sources data, i.e. from the National Databank Wegverkeersgegevens (NDW), for short term traffic predictions. The study employs a combination of this data source with probabilistic graphical methods to predict traffic congestions, which is a field that has minimal studies done.

## 1.5. Thesis Structure

The thesis consists of seven chapters. Chapter 1 briefly discusses Background, Problem statement and novelty of the proposed research that will be addressed in the further chapters. Chapter 2 reviews the previous studies related to the currently proposed research. Chapter 3 describes the selected study area chosen in this research. Chapter 4 describes the data and the software's used. Chapter 5 delineates the methodology used to build the prediction model using Bayesian Network, while Chapter 6 & 7 explains the results obtained from the prediction model, talks about the conditional probability table, limitations, strength, and weakness of the methods. Finally, Chapter 8 gives a short conclusion on this research along with the future recommendations.



## 2. LITERATURE REVIEW

There are several authors who have predicted traffic process using different means. Whittaker et al., (1997) utilized **dynamic state-space models** to track and predict a network traffic process. Using a skeletal traffic model, the authors seek to ascertain that **dynamic state-space model** are a reliable way for monitoring and predicting, thereby rendering as a useful tool for traffic management. In order to build these dynamic state-space models, the authors make use of conditional independence relationships and Bayesian methodology. To estimate the optimal state, Kalman filters were used. The outputs of this model were compared with predictions from the Naïve predictors, thereby testing the accuracy and reliability of the predictions made. The results seem favourably good; wherein, predictions were accurate at lower magnitudes. As the magnitudes increased, the predictions were not in sync but were in close alignment with the observed traffic process. Acknowledging that the model holds several issues, the authors suggest using decomposition techniques that tend to simplify traffic networks or more specifically traffic models in which path choice becomes an important determinant for prediction.

The spectrum of traffic modelling has also paved the way for considering the nuances of uncertainties-in terms of speed, flow, density, and volume-that are present in traffic systems. Probabilistic traffic models are one such kind of models that represent traffic flow accurately while considering the traffic certainties and stochastic traffic flow fluctuations (Calvert et al. 2017). Apart from this, there are several other scholars who have attempted to study traffic flow state and traffic predictions. Based on the Kalman-filter method, designed an estimator of the traffic flow. Similarly, Canaud et al. (2013), used a model based on the probability hypothesis density filtering to predict traffic flow. Using different models, these scholars have proposed and suggested ways to predict traffic flows.

According to the author's Yu & Cho, (2008), who proposed a short-term traffic prediction model using Bayesian Network, a short-term meaning is forecasting a condition of the traffic in near time future, which ranges from 05 mins to 60 minutes. In general, the author's Sun, Zhang & Yu, (2006) has adopted the concept of Bayesian Network model, a well-known model in PGMs (Probabilistic Graphical Models) to predict the flow of traffic at given road segment and its neighbouring road segment using historical data. The real-time traffic information, such as the average speed of the current flow of vehicles is used to predict the average speed of the vehicles in the near future. To build the Bayesian network model the author has used GMM (Gaussian Mixture Model) with two or more components as a distribution of random variables and utilized Expectation Maximum (EM) algorithm to extract parameters of Joint probability density function of GMM (Williams & Hoel, 2003). According to Yu & Cho, (2008) who has implemented the entire process in two different steps: Firstly, design the Bayesian network by inputting the data of both streams, which are upstream and downstream with respective historical data. Secondly, by applying the current traffic flows to each of the links to compute the final predictions presuming the occurrence of accident incidences which might affect the accuracy of the Bayesian network model. To estimate the accuracy of the implemented model, the author carried out several iterations with two sorts of measures: RMSE and travel time of the real-time vehicles flow. As a result, the RMSE of all the road links were slightly increased as the prediction time increased from 5<sup>th</sup> minute to 40<sup>th</sup> minute. When the prediction was carried out at 60<sup>th</sup> minute, the RMSE was less than 8. In order to reach from any source to the destination, the travel time was computed using numerous paths. Out of which four paths were considered; namely, static

shortest path, dynamic shortest path, and real shortest path (Maroto et al., 2006). After several iterations, the author concluded that the predicted travel time acquired from the dynamic shortest path as a stronger correlation to the real shortest path compare to the other methods.

According to authors Sun et al., (2015), the Dynamic Bayesian Network model (DBN) was implemented to predict the traffic crashes which occurs in highways/expressways. The DBN model was proposed to estimate the correlation between crash incident and dynamic speed condition data. As additional input to the DBN, the traffic congestion at the crash site were observed and fed into the model. Apart from crash incident data, several other data such as traffic flow data (Volume, Speed and Occupancy), road geometry alignment and other external environmental data were used in the development of the crash prediction model. However, due to the inconsistency in the floating car and GPS data, the traffic flow data which are collected from the sensors were used to fill these gaps. Although the several studies (M. M. Ahmed & Abdel-Aty, 2012; Li, Hao, Gan, & Chen, 2013) have accepted speed only data to build the prediction model, here the author's Sun et al., (2015) used secondary data such as 'speed condition data' with certain limitations collected from the same sensors were used to build the prediction model. The author depicts that the state of traffic is very crucial before the crash incident occurs; this is an important factor in building robust crash prediction models. To make the DBN more complex and robust - Initially, the speed condition data collected at several intervals coupled with congestion level data of both upstream and downstream of the crash location were considered as explanatory variables for the prediction model. The author also uses a static Bayesian network model for comparative analysis to estimate the accuracy of the DBN by inputting data which is collected by sensors. As previously mentioned, the focus of the SBN model is to compare with the DBN model, which includes time-dependent and crash occurrence variables; these variables are evaluated in both the models. As similar to any other classification models, the original data, which consists of crash and non-crash information with respective traffic condition data, was randomly assigned for training and evaluation of BN models. To minimize the error caused by random sampling, ten combinations with various training and testing datasets were generated. The lower RMSE value of the model depicts the performance of the model. The DBN model comparatively yields better performance than SBN model for both crash prediction accuracy and false alarm rate. Another advantage of DBN over SBN is that the DBN can manage the time series data very well before crash incidence and detects the state transition at the first place when the same results were compared with SBN, the DBN model was more suitable to predict the real-time crashes considering time dependency across a different time interval.

## **2.1. Probabilistic Graphical Method**

The Graph theory presents a unified and intuitive modelling framework that aids in expressing multivariate dependencies and causal relationships, coupled with a natural structure for the design of a model algorithm. Whittaker et al., (1997) states that Probabilistic Graphical Model (PGM) is a model that identifies the probability relationships between random variables. PGM is an amalgamation of graph theory and probability theory. It has a strict theoretical foundation making the joint probability distribution concise and compact. Such an application relies on the independent variable relationship, thereby simplifying uncertain and complex system problems, ensuring simplistic knowledge acquisition and domain modelling. Today the concepts of PGM are applied in artificial intelligence, machine learning and data mining fields.

Delving deeper into PGM, it can be ascertained that PGM presents a modelling framework for processing the uncertainties, which makes the probability-based relationships to be expressed and calculated with accuracy. Currently, many of the statistical probability models can be understood as special cases of PGM. In 1988, Pearl, J. (2014) proposed a detailed elaboration of two GMPs, viz. directed graph (Bayesian networks) and undirected graph (Markov networks) and indicated their buildout in processing uncertain

information. Following the development of PG theory, a novel PGM factor graph was also introduced in 2001 that showcases a uniform representing methods which shows a greater efficiency the fields of data mining, machine learning, pattern recognition, and other important areas.

In PGM there are mainly 3 types:

1. Markov Network
2. Factor Graph
3. Bayesian Network

**2.1.1. Markov Network**

Based on the undirected graph, Markov Network (MN) is a probabilistic graphical model which displays the independent relationships among variables. It attempts to delineate the joint probability distribution of random variables of Markov’s property which indicates that the probability of state transition is directly proportional to its adjacent state, furthermore, the values may be discrete or continuous. Markov networks predominantly aid in analysing the spatial bonding of certain physical occurrences. It has several applications, such as it is helpful in making statistical inferences and in pattern recognition, to name a few. While exploring different networks, it can be said that the Markov network is slightly different from the Bayesian Network (BN), wherein the MN aids in solving symmetric relationships and dependencies. It is also capable of showcasing certain dependencies like the cycles dependence which BN is not capable of. In this way, MN holds applicable for a wider range of representations (Li et al., 2013).

The Markov network is illustrated as (UGM) undirected graphical model, which is represented by two elements, namely, ‘V’ and ‘E’. For any given undirected graph ‘G’. The Markov model is given as

$G = \{V, E\}$  |  $V$  = Set of nodes depicts random variables |  $E$  = Set of edges shows dependency among variables.

The Joint Probability Distribution (JPD)<sup>1</sup> is described over the set of possible functions, and it is nothing but the normalization of the product of all possible functions. The following is the form

$$P(V = v) = \frac{1}{Z} \prod_{c \in C} \varphi_c(v_c) \dots\dots\dots (1)$$

$\varphi$  = Non-negative possible function expressed over the maximal clique<sup>2</sup>  
 $V = \{V1, V2, \dots, Vn\}$ ,  $v = \{v1, v2, \dots, vn\}$  is state value of  $V$   
 $Z$  = normalization factor

---

<sup>1</sup> Joint probability distribution = It is a phenomenon in which events A and B are occurring together at the same time which is denoted as ‘probability distribution’. The joint probability distribution denotes the events which occurs are between two or more random variables.

<sup>2</sup> Maximal Clique = In the graphical representation, maximal clique represents adjacent vertex which cannot be extended. In other words, the adjacent vertex is not subset of larger clique.

The applications of the Bayesian network and inference algorithms can be implemented into a Markov network. When it comes to the exact inference, the Markov network faces NP-hard problem instead approximate inference can be estimated like Markov chain Monte Carlo algorithm and propagation algorithm (Li et al., 2013).

### 2.1.1.1. Practical Implications of the Markov Network

Markov Networks are generally taking ‘discrete probability distribution’ into consideration. In the practical implications, the most used models are Markov Logic Network (MLN), Markov random field (MRF) and Relative Markov Network (RMN) & Conditional Random Field (CRF) in the regime of statistical learning.

Most of the MRF models are cyclic graphs. Moreover, a Markov network is used as a network structure for building a model. However, the theoretical concept remains the same, but the approach towards problem-solving might differ. In the field of **Image processing**, MRF models are used in the lower level of image processing for image restoration, surface reconstruction, edge detection and image segmentation analysis. As per Li et al., (2013) who developed MRF model used Markov joint distribution and the performance of MN was like Bayesian framework these aids in solving vision aid problems using mathematical computational methods.

In the regime of NLP (Natural language processing), CRF models are widely used; this is due to its nature, the graphs are usually represented in cyclic, which adopts MN network structure. This probability model used for predictions and classification of sequential structured data.

### 2.1.2. Factor Graph

Being a bidirectional graph, Factor graph delineates the manner in which multivariate global function is decomposed into local functions by breaking down complex computation from a global to a local level—being an intuitive and generic method that eases up complexities. Factor graph aids in the interpretation of several iterative processes are making it widely useful in statistical studies. Majority of PGM complexities can be unified by a factor graph.

Factor graphs are binary graph which breakdowns global function into a series of local function which contains two nodes, a) Variable and b) Factor nodes.

Factor graph can be written as

$$g(x_1, x_2, \dots, x_n) = f_A(x_1)f_B(x_2)f_C(x_1, x_2, x_3)f_D(x_3, x_4)f_E(x_3, x_5) \dots \dots \dots (2)$$

With respect to modelling, Factor graphs consist of topological structure modelling and local function modelling with the matching factor nodes. In the practical applications, the use of expert knowledge is to develop the initial prototype and later ‘sample-based modelling’ can be used to improve the prototype. The main aim of sum product algorithm is to calculate the marginal function of the factor graphs. Here, the assumption of factor graphs is interpreted as to calculate marginal functions of global functions using iterative message passing mechanism. With respect to cyclic factor graphs, one can obtain exact inference using sum product algorithm on the other side, for acyclic factor graphs, approximate inference is used to calculate marginal functions which results in obtaining approximate results.

The sum product algorithm is used in acyclic graphs which consists of series of “addition” and “multiplication” operations in the execution of communication takes place between variable nodes and local functional nodes. Here, the ‘communication’ refers to a function which are either continuous or discrete variables which are calculated in the inference processes. During every iteration, every single variable nodes and factor nodes as to communicate between adjacent nodes and pass on the information to obtain accurate information.

The factor graphs are still emerging in the modern years. Factor graphs are simplified and generalized due the nature of factor graphs it is applied in many different areas such as Signal processing and channel coding. The complex representation of majority of PGM can be unified by factor graph. While comparing factor graph with BN and MN, it can be said that factor graph showcases better representation of the factorised forms of probabilistic distribution with a better clarity.

### 2.1.3. Bayesian Network

Bayesian Network is a type of probabilistic graphical model which depicts probabilistic relationships between a set of variables through a directed acyclic graph (DAG). It is a causal model that represents conditional independencies between a set of random variables (Sun et al., 2006). It is a combination of probability theory and graph theory, thereby acting as a tool to deal with two problems- uncertainty and complexity (Jordan et al., 1999). Therefore, this study makes use of Bayesian Network model to predict traffic congestions.

A Bayesian Network comprises of a set of nodes and arcs, wherein nodes represent random variables and arc connecting the pair of nodes, thereby representing direct dependencies amidst variables. Constructing a Bayesian Model entails three steps: i) Defining variables (nodes); ii) describing the structure (arcs); iii) identifying and describing parameters that involves specifying a conditional probability distribution for each node.

Here, defining conditional independence becomes important because it plays an important role in using the probabilistic model for pattern recognition. Conditional independence properties attempt to simplify i) the structure of the model and ii) the computations needed to perform and learn within that model (M. Jordan et al., 2006). The conditional independence relationship aids in representing joint probability distribution more compactly and conveniently, especially for a larger network, like the traffic networks. S. Sun et al., (2006) define conditional independence relationships in simple terms: a node is considered independent of its ancestors given its parents, wherein the ancestors/parent relationship is with respect to the fixed topological ordering of the nodes.

Within a Bayesian network that comprises of  $n$  nodes ( $x_1, x_2, x_3 \dots x_n$ ) the joint probability distribution can then be represented as:

$$p(x_1, x_2, \dots, x_n) = \prod_{i=1}^n p(x_i | x_{P_i}) \dots \dots \dots (3)$$

Wherein  $p(x_i | x_{P_i})$  is the local conditional probability distribution of node  $i$ , while  $P_i$  is the set of indices labelling the parents of node  $i$ . (Sometimes  $P_i$  can be empty if there are no parents). To derive the joint probability distribution between the input and output, there are several distribution models that aid in doing this approximation. However, one of the robust methods to do so, that will be used in this study are Hill Climbing (HC) and Incremental Association Markov Blanket (IAMB) which is a weighted combination of various normal distribution functions (S. Sun et al., 2006).

In this way, the Bayesian Network gives a clear representation of the joint probability distribution over all its nodes, which in turn enables the user to compute the probability of each state of the node that is conditioned over any other subset of other variables (S. Sun et al., 2006).

This process of probabilistic inference aids in diagnosing and predicting traffic congestion, thereby making Bayesian Network a powerful tool for traffic prediction.

Summarizing the three basic models of PGM which are BN, MN and factor graphs which includes understanding the theoretical concepts, inferences, and applications. The relations between distinction between all these three models are represented in each of the sections. In general, Bayesian Network is commonly used in modelling the casual relationship which are represented by a directed acyclic graph. Markov Network are commonly used in modelling the non-causal relationship by undirected acyclic graph. Factor graphs are generally an iterative process applied in the problem solving and represented in bidirectional graph. The three models of PGM which are discussed in the above sections are widely applied in numerous domains due to its high performance, and the results are trustable.

### 3. STUDY AREA

#### 3.1. Study Area

With over 130,000km of public roads, the Netherlands has one of the densest road networks. Taale & Wilmink, (2016) highlight that traffic congestions in the Netherlands have increased due to the rising economy that was complemented with more infrastructure and more automobiles, which culminated into more traffic congestions<sup>3</sup>. In 2016, an increase in traffic jams by 32% was reported in the region of Amsterdam<sup>4</sup>. Figure 1 attempts to illustrate the increase in traffic congestions in Amsterdam, Rotterdam, Hague, Utrecht, Groningen, and Eindhoven. From the figure, it is quite evident that traffic congestions in these cities have increased from 2011. While looking deeper into which kind of networks face more congestions, Figure 2 illustrates that congestions are higher on the main road networks around the city. From these initial inquiries, the cities of Amsterdam and Rotterdam are selected as study areas for this study, since these cities are within the Randstad area-- forming the main economic areas of the Netherlands-- and both these cities have experienced a considerable increase in traffic. Hence it becomes important to not just look at the causes of congestions, but also to look at ways of predicting traffic which would aid in better monitoring and planning with these commercially important areas.

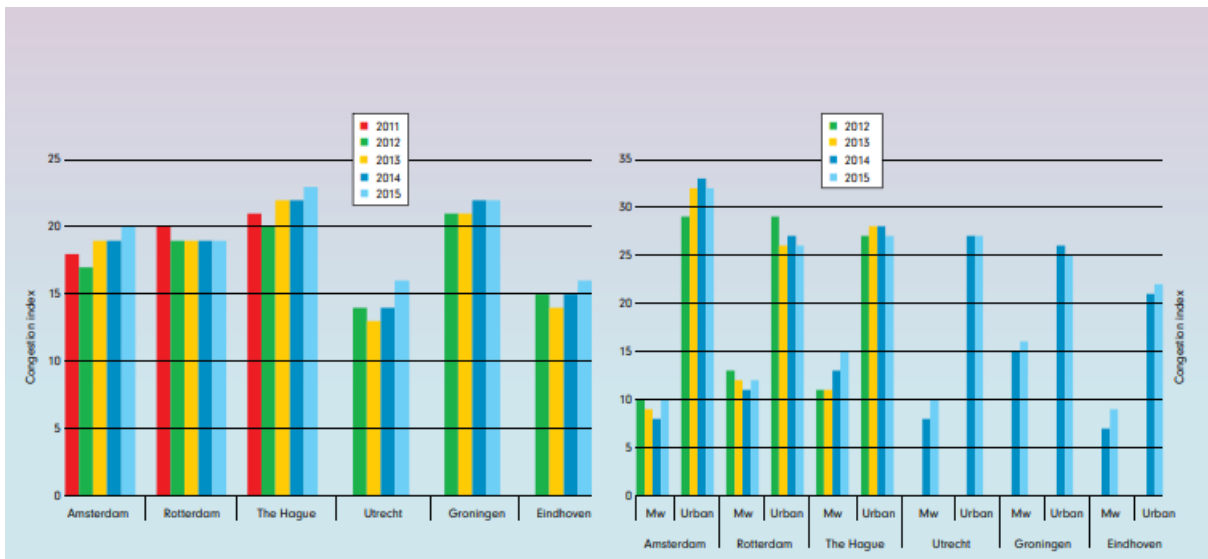


Figure 1(left): Comparison of congestion indices for Urban networks in six cities of Netherlands from 2011-2015

Figure 2(right): Comparison of congestion indices between motorways and urban networks of six cities in Netherlands from 2012 to 2015

(Source: Taale & Wilmink, 2016)

<sup>3</sup> Congestion here refers as when Traffic congestion occurs when travel demand exceeds the existing road system capacity. (Rosenbloom, 1978)

<sup>4</sup> Practical Trial Amsterdam (2016) Overall final report In Car. 16 February 2016.

### 3.2. Amsterdam

Located in the Northern province of the Netherlands, Amsterdam houses nearly one million inhabitants, as of 2018 (Thomas B, 2020). Studies by Melnikov et al., (2016), suggests that the urban area of Amsterdam comprises of 118577 nodes and 207577 links that belong to the urban road networks of the city. Such a huge density makes it necessary and important to predict traffic, for it is the most populous city of the Netherlands that has been facing severe traffic issues for years now, as illustrated in Figure 1.

Figure 2(right) illustrates the comparison of congestion indices among motor vehicles and urban network from 2012 to 2015. The below map (Figure 3 & Figure 4) illustrates the distribution of sensors across the area of Amsterdam.

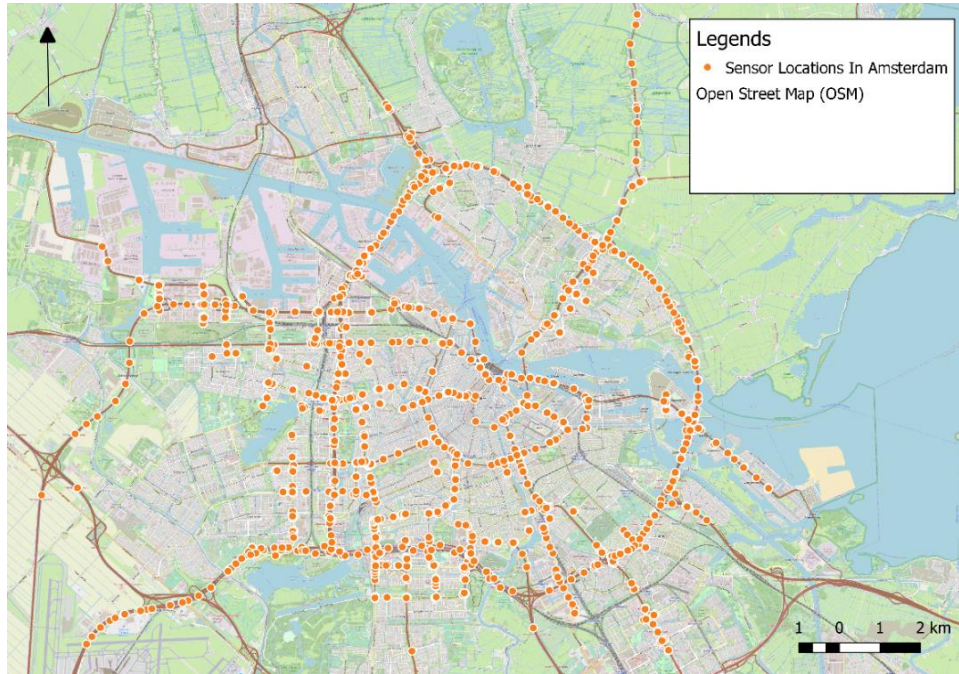


Figure 3: Road network of Amsterdam with sensors (With base layer)

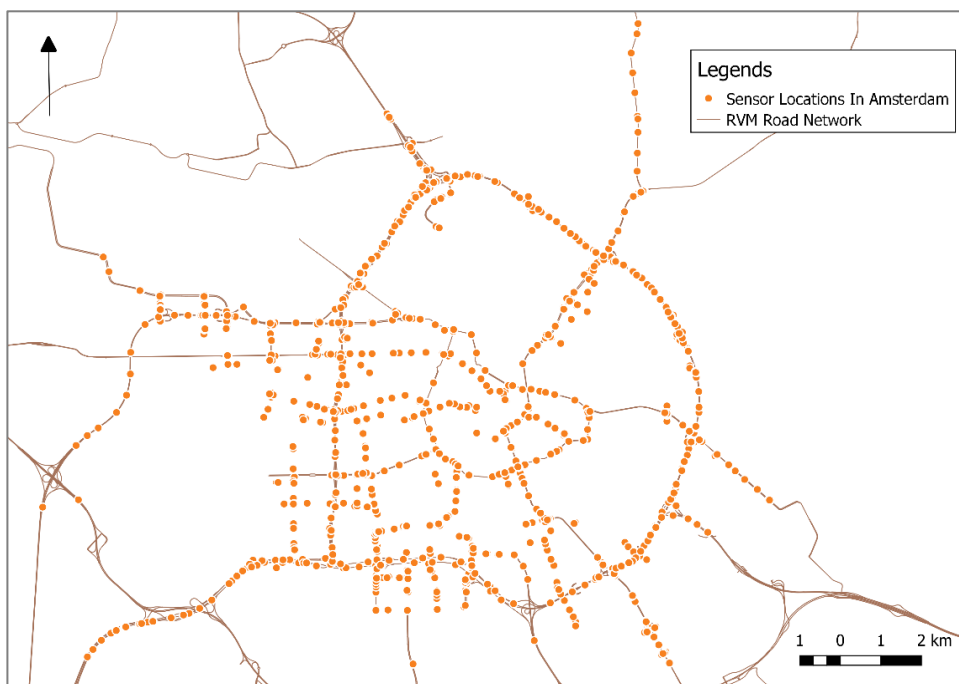


Figure 4: Road network of Amsterdam with sensors (Without base layer)



### 3.3. Rotterdam

After Amsterdam, Rotterdam is the second-largest city with a population less than a million. The city evolved from a small fishing village in 1328 to a large commercial and trade area that it is now (Rotterdam Population, 2019). This growth was coupled with expansions of road networks. Therefore, the location of Amsterdam and Rotterdam with the economic region of the county makes it more important to look at these areas for predicting traffic. The below map (Figure 5 & Figure 6) illustrates the distribution of sensors across the area of Amsterdam.

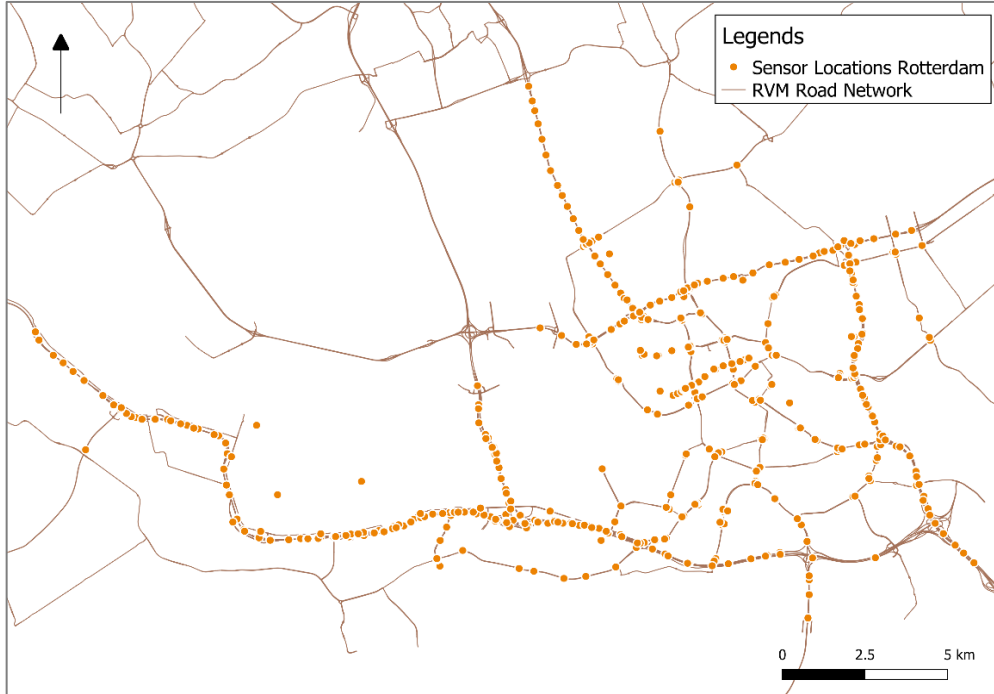


Figure 5: Road network of Rotterdam with sensors (Without base layer)

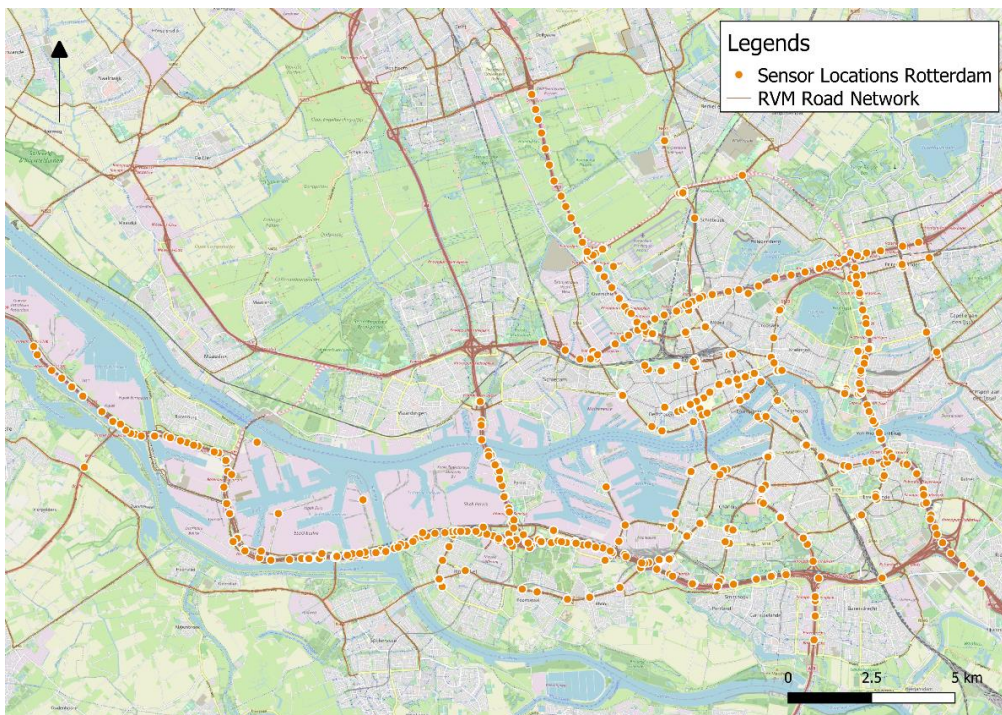


Figure 6: Road network of Rotterdam with sensors (With base layer)

## 4. DATA AND SOFTWARE

The data for the present study is extracted from the National Databank Wegverkeersgegevens (NDW), The Netherlands.

### 4.1. Historical Traffic data

The historical dataset was extracted from NDW (DATEX II Dutch Profile, 2015) which is an open-access database which attempts to collect, process, store and distribute all the traffic-related data. This data was downloaded in the form of a .csv file. About 27,000 sensors of the NDW collect the entire road network data of the Netherlands. Amidst these close to 500-600 sensors collect the traffic and road network data of Amsterdam and Rotterdam.

Over the last decade, traffic in the Netherlands has increased in a manner has created severe and dense congestions in the road networks(Kim & Wang, 2016) In 2008, the National Data Warehouse (NDW) was formed for collecting and providing road traffic data in terms of speed, intensity, travel time, vehicle characteristics, the total number of lanes, and specific lanes information. The information regarding different parameters (location, data measured, accuracy) extracted from the sensor is stored in separate files, i.e. each parameter is stored in a separate folder thereby containing all the characteristics of the parameter.

The NDW sensor data is provided in two ways: -

- **Real-time data** can be downloaded from the public File transfer protocol server which is saved as XML files with a refresh rate of 1-minute interval.
- **Historical data** provided by a web service interface by NDW, which as the ability to filter the data and aggregate according to the user choices. The aggregation ranges from 1 minute to 1 month. The interface enables the user to select area based on the drop-down list or the by drawing rectangle on the map. The robust and efficient interface enables the user to download the data at ease.

### 4.2. Road Networks Data

As mentioned above, the road network data was extracted from NDW (DATEX II Dutch Profile, 2015). The files of road networks have been downloaded in the form of .shp files. Figure 7 illustrates the snapshot of the road network of Amsterdam and Rotterdam. The road network data was just used for visualization purposes. However, the road network information such as length of the road and area data were not used in the Bayesian network model.

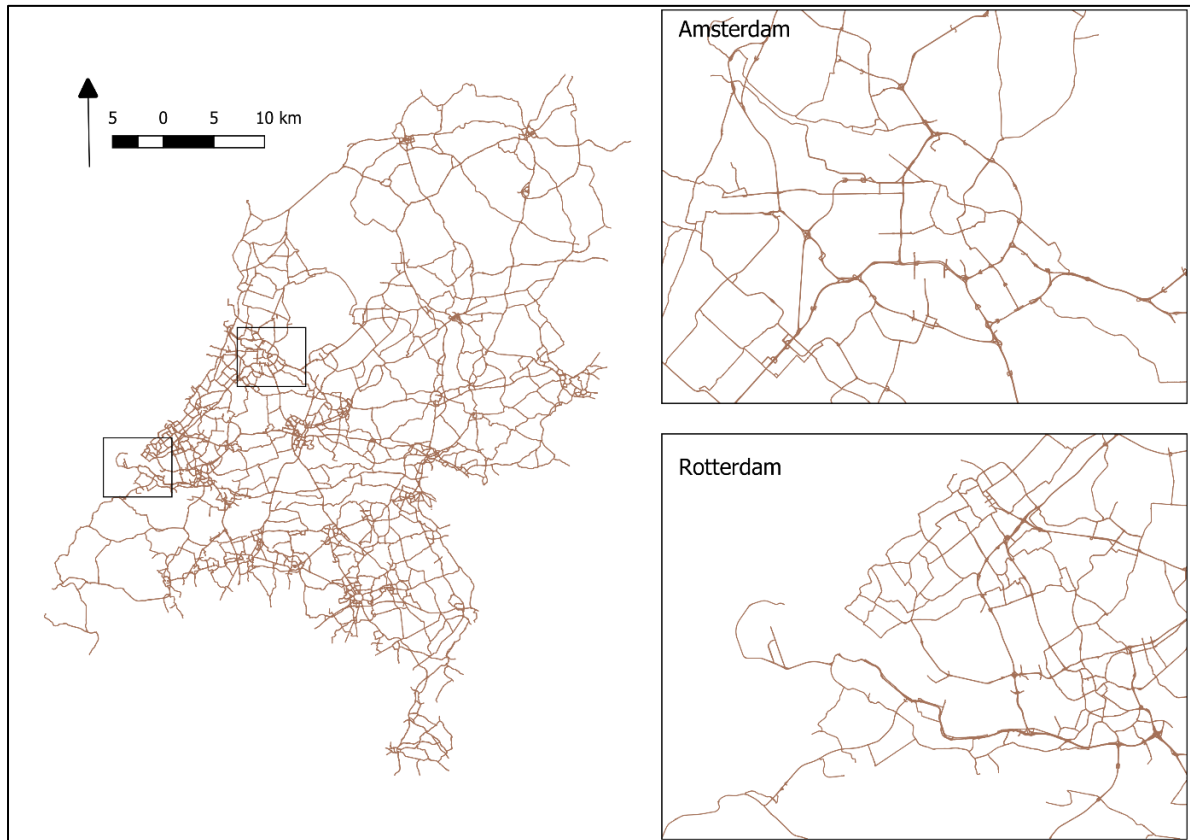


Figure 7: Snapshot of the Road network (Amsterdam & Rotterdam)

#### 4.3. Scripts (R and Python)

The scripts for the entire process were done by using a library called “BNLearn – An R package for Bayesian network Learning and Inference” (Bnlearn, 2020). An open-source library built to simplify the complexity of the Bayesian Network concept and lastly ‘Excel bar charts’ were used for visualization of the results.

**Road Network Data:** Analysing and processing of road network data were done on Jupyter Notebook (Project Jupyter, 2020) an open-source web platform which enables users to create and share the code and visualize the results on the go.

**OSMnx Plugin:** An open-source tool built on top of OSM (Boeing, 2017). With these tools, one can download spatial information such as building footprints, spatial boundary, and street networks. This tool helps users to analyse and visualize complex street networks. Several functionalities help the network to be built based on a walking path or driving path. Other functionalities include node elevation calculations and street grades.

## 5. METHODOLOGY

This section includes the key process entangled in this research. The section is divided into seven sub-sections:

### 5.1 Pre-Processing Data

### 5.2 Method Explanation

### 5.3 Variables

### 5.4 Structure

### 5.5 Parameter Learning

### 5.6 Identifying major causing nodes for increased ‘Traffic Intensity’ using ‘Odd Ratio’

### 5.7 Evaluating model performance for the continuous variables using RMSE, MAE & MAPE

#### 5.1. Pre-Processing Data

For the study, the data from NDW were downloaded at an interval of 60 minutes and 30 minutes for a period of three months (1st March to 31st March, 2018), (1st March to 31st March, 2019) and (1st March to 31st March, 2020) The downloaded CSV files consists of two main tables. One table consists of data for ‘Travel Speed & Intensity’ and the second table consists of ‘Travel Time’. Other information such as vehicle characteristics, latitude & longitude, date and time, sensor information and road-specific information was downloaded. The detail explanation of each data variables are explained as follows:

##### 5.1.1. Vehicle Characteristics

The measurement points that are equipped with instruments of higher accuracy aid in differentiating between small and very small vehicles, i.e. it allows the distinction between cars and motorcycles coupled with aiding in a clear distinction between trucks and busses. Such differentiation is based on the length of vehicular types (DATEX II Dutch Profile, 2015).






Category	Description	Length Range (In Mts)
Cat 01	Motorcycle, Moped	$\geq 1.85$ & $\leq 2.40$ 
Cat 02	Car/Van	$> 2.40$ & $\leq 5.60$ 
Cat 03	Non-articulated Truck	$> 5.60$ & $\leq 11.50$ 
Cat 04	Non-articulated bus	$> 11.50$ & $\leq 12.20$ 
Cat 05	Articulated Truck	$> 12.20$ & $25$ 

Table 1: Vehicle characteristics by length differentiated by NDW sensors.

5.1.2. NDW Sensor Table

Place Name	Sensor Count	Area (In Sq. KM)
Amsterdam	638	219.40
Rotterdam	412	325.80

Table 2: Area and a total count of sensors in the study area.

**5.1.3. Travel Time (S):** Estimated as current or estimated travel time, that is calculated in one driving direction recorded by sensors in seconds. The definition for Travel Time(S) was followed as per the NDW guidelines (DATEX II Dutch Profile, 2015).

**5.1.4. Intensity (Vehicles per hour):** Intensity was calculated based on the number of motor vehicles that pass at a given point during a given period calculated in one direction of travel. The definition of intensity was followed as per the NDW guidelines (DATEX II Dutch Profile, 2015).

**5.1.5. Speed (km/h):** The speed is estimated using the harmonic mean speed of vehicles that pass through a point or a turn in a unit of time, which is calculated in one direction of travel. The definition for Speed was followed as per the NDW guidelines (DATEX II Dutch Profile, 2015).

**5.1.6 Driving Lane (DL):** Driving lane depicts different categories of vehicles which passes by the sensors in the respective lane. In the major roads of the Netherlands comprises of 6 lanes, i.e. Lane01, Lane02, Lane03, Lane04, Lane05 & Lane06. Figure 8 depicts the snapshot of the Driving Lane.

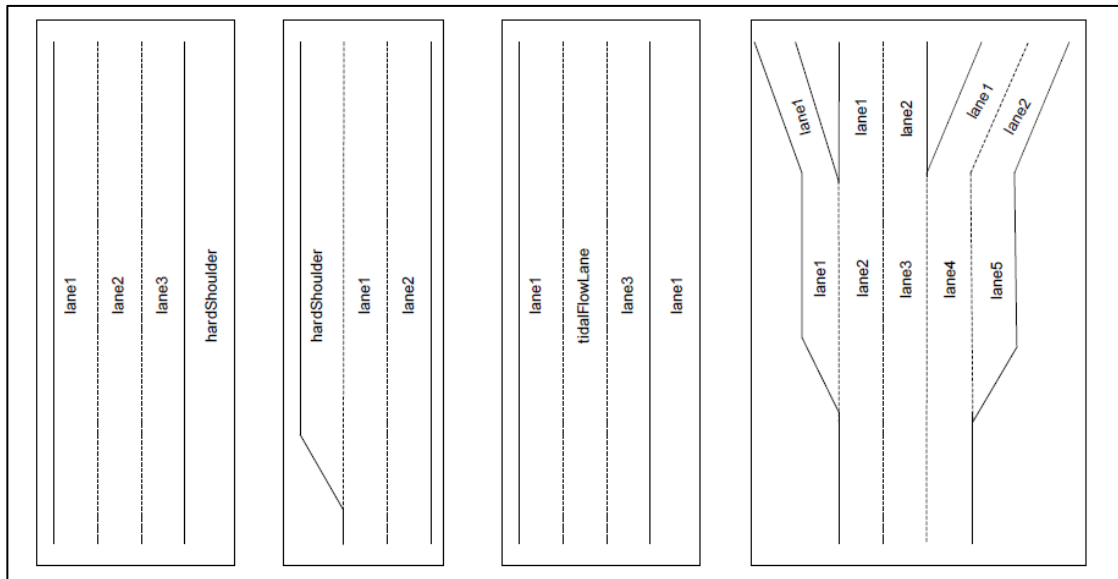


Figure 8: The snapshot of the driving lane

The following guidelines were used from NDW for naming the lanes: -

- Naming conventions for all the lanes except bus lanes, tidal flow lanes and emergency lanes does follow based on the road orientation. For example, left-most lane observed by a driver is named as '**lane1**'.
- In certain cases, a new lane is added to the leftmost side of the road called '**Rush hour lane**'.
- The lane which carries heavy vehicles is called as '**Bus lane**'.
- The lane which carries the traffic on both sides is called '**Tidal flow lane**'.
- In emergency cases, the rush-hour lane is considered as an emergency lane in those cases the respective lane is called '**Hard shoulder**'.
- Carriage ways are titled as 'all lanes complete carriageway' these names supports in making it as unique and traceable which differentiates from the rest of the lane. Nevertheless, the same does not apply the left lane since the numbering is not done. In general, to preserve the lane numbering of the carriageway sometimes it becomes difficult to provide additional names. In those cases, '**Rush hour lane**' is used. This eventually avoids the problem of numbering road lanes.

**5.1.7. Specific Lane (SL):** Specific lane is based on the functional road classification, which is an indication of the importance of the road segment. This category has 5 subcategories: -

- **Connecting Carriageway:** representing the principal road at a motorway junction
- **Entry Slip road:** representing the entrance slip road
- **Exit Slip road:** representing the exit slip road
- **Main Carriageway:** representing the main carriageway
- **Parallel Carriageway:** representing the parallel carriageway

Also, in many instances, the measurement locations (sensor location) might not be present on the main carriageway. Due to this, the supplement information about the carriageway must be provided in order to detect the measurements accurately. Therefore, if the measurement location is situated at the exit or entry of the 'Slip road' or in 'Parallel Carriageway', the carriageway element name will be affected. In that case, 'Carriageway' and 'lanes' should be used.

**5.1.8. Time of Day (ToD):** 'Timestamp' information obtained from the sensors were categorised into five discrete states {Early morning, morning peak hours, off-peak hours, evening peak hours and night}.

**5.1.9. Day of Week (DW):** 'Day of week' data obtained from sensors were categorised into 7 discrete states {Sunday, Monday, Tuesday, Wednesday, Thursday, Friday and Saturday}

**5.1.10. Weekday or Weekend (WD/WE):** This node includes the two states, namely, Weekday or Weekend. The categorization of data was based on the date & time stamp information obtained from sensor data.

### 5.2. Method Explanation

**Structuring a BN model for Traffic congestion and prediction:** This study considers a BN model that represents the dependency structure of link-level measures. For a given link, a BN model is designed that describes relationships between link performance measures (e.g., flow, intensity, and speed) and external factors that affect the target link (e.g., time of day and incident). The goal of this model is to assess the effects of external factors on traffic conditions on a targeted link. The spatial component of the flow direction of traffic is not considered in this model. The BN model in this study is considered to be static; the model represents a time-independent knowledge of dependency relationships between variables (that is, long-term average patterns).

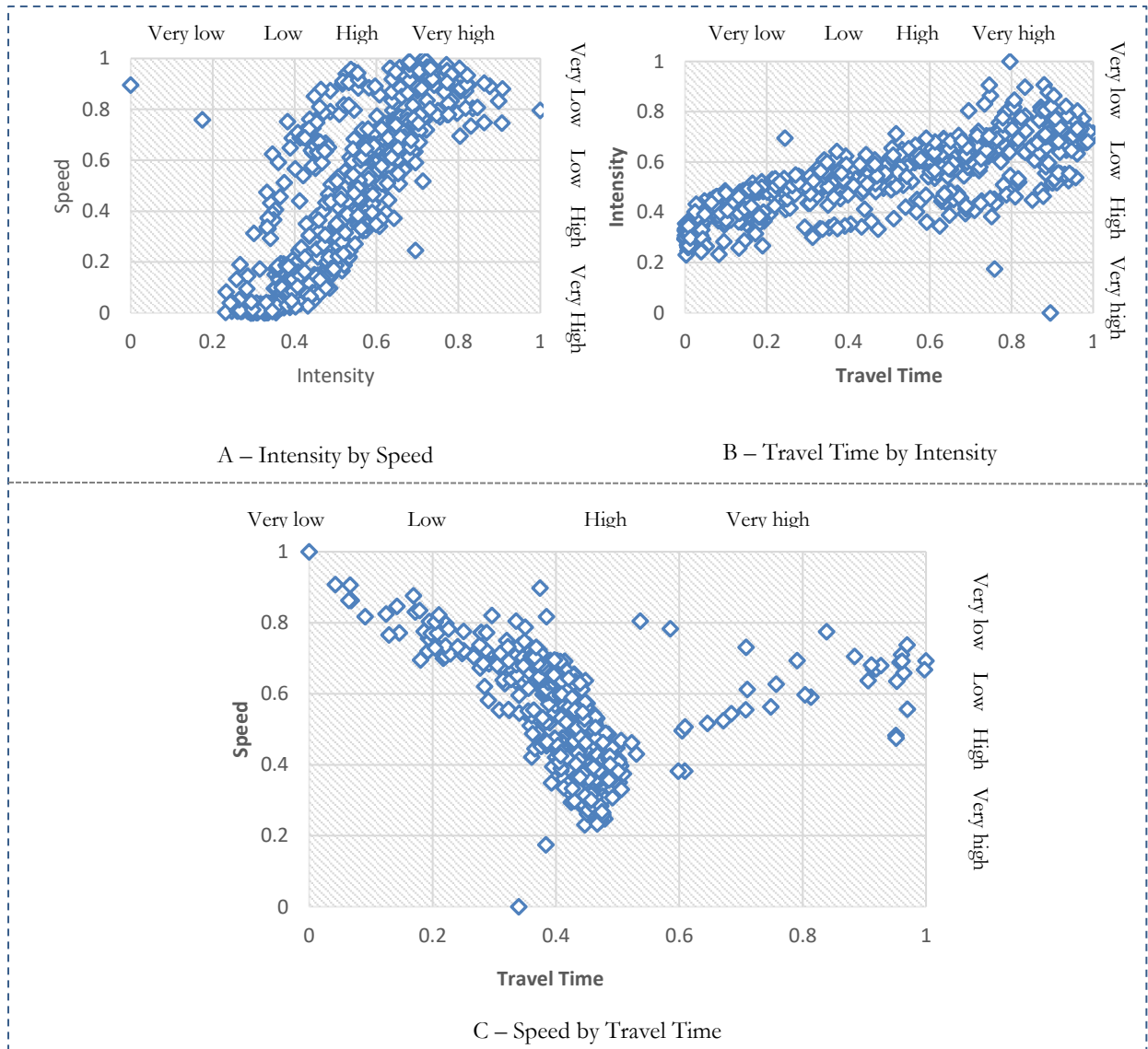


Figure 9: Data normalization of 'Group 2' input variables – Intensity (A), Travel Time (B) and Speed (C)

**5.3. Variables**

The variables used in the preferred Bayesian Network model are given in **Error! Reference source not found.**. A total of 7 variables were chosen. In the study, all the continuous data were converted into discrete variables which indicates that each node takes discrete values. The process of discretization involves two steps,

1. **Normalization of continuous variables:** According to the Patro & sahu, (2015) the term normalization is defined as a scaling strategy or mapping technique which usually carried during the pre-processing stage. The aim of this technique is to derive a new range from the existing range and to maintain the consistency between the values; these techniques are useful in the study of prediction or forecasting modelling. Well-known normalization techniques such as Min-Max normalization, Z-score normalization, and Decimal scaling normalization techniques. In the current study, simple Min-Max normalization technique is adopted.

$$Y = \frac{A - \text{Minimum value of } A}{\text{Maximum value of } A - \text{Minimum values of } A} * 1 \dots\dots\dots (4)$$

Where,

Y = Normalized data

A = Original data

2. **Categorising based on the distribution of normalised variables:** As shown in Figure 9, For each node, the illustration of Speed, Travel Time & Intensity associations are plotted. Furthermore, the range of input data is equally distributed into four parts, as shown in Table 3.

Table 3: Discretization of input data

Discrete Variables	Range of continuous variables
<b>Very low</b>	0.00 – 0.25
<b>Low</b>	0.25 – 0.50
<b>High</b>	0.50 – 0.75
<b>Very high</b>	0.75 – 1.00

The input data is classified into two groups: Group 01 and Group 02. Each of the variables in the group is described as follows:

**Group 01:** The variables in the group indicates characteristics of the vehicles, Time of the day, Weekend or Weekday, Day of the week and lastly the type of road, single lane or multiple lanes and a total number of lanes. The vehicle categories variable takes five values {Categories: 01 - 05}; Day in a Week variable takes (DW) takes seven values {Sunday, Monday, Tuesday, Wednesday, Thursday, Friday and Saturday}; Time of the day variable takes five values {Early morning, morning peak hours, off-peak hours, evening peak hours and night}. Table 1 shows the input data combination for different variables across different years across.

**Group 02:** The variables in the group denotes the traffic situation on each sensor. The group contains three variables, namely Speed (AS), Intensity (AI) and Travel Time (TT). Each of these variables contains four discrete values {Very low, Low, High, and Very high} as stated, before the continuous data were converted into discrete variables. Before discretization, the data values were normalised, which ranges between 0 to 1 and the distribution of the normalised data is shown in Figure 9.



<b>Input Data Explanation</b>					
<b>Amsterdam</b>					
<b>Years</b>	<b>Group 01</b>	<b>Group 02</b>	<b>Time Period</b>	<b>Nodes Count</b>	<b>Total No. of entries</b>
2018	VC, ToD, DW, WD/WE	AT, AI, AS	1 <sup>st</sup> – 31 <sup>st</sup> Mar	7	1,635,898
2019	VC, ToD, DW, WD/WE	AT, AI, AS	1 <sup>st</sup> – 31 <sup>st</sup> Mar	7	1,873,848
2020	VC, ToD, DW, WD/WE	AT, AI, AS	1 <sup>st</sup> – 31 <sup>st</sup> Mar	7	1,797,634
<b>Rotterdam</b>					
2018	VC, ToD, DW, WD/WE	AT, AI, AS	1 <sup>st</sup> – 31 <sup>st</sup> Mar	7	1,345,266
2019	VC, ToD, DW, WD/WE	AT, AI, AS	1 <sup>st</sup> – 31 <sup>st</sup> Mar	7	1,463,546
2020	VC, ToD, DW, WD/WE	AT, AI, AS	1 <sup>st</sup> – 31 <sup>st</sup> Mar	7	1,563,324

Table 4: Description of input data

#### 5.4. Structure

Right after the Bayesian Network is stated, the following step is to establish the relationships between the variables. In this study, the structure of the Bayesian Network is learned from the data. Several **Constraint-Based** learning algorithms are available, such as Grow-Shrink Markov Blanket (GS) algorithm (Tsamardinos, Aliferis, & Statnikov, 2003), practical constraint-based structure learning algorithm (PC) (Kim & Wang, 2016) and Incremental Association (iamb) algorithm (Tsamardinos et al., 2003). Similarly, in **Score Based**, several learning algorithms are available such as Hill Climbing (HC) and Tabu search (tabu). In this study will be choosing one learning algorithm from constraint-based and score-based, later evaluate and use the optimum learning algorithm for the study. To build the model systematically firstly the relationship between the group nodes should be established and then establish the relationship between the individual nodes. For example, From the user knowledge or from the data definitions, if the user assumes that group 01 affects group 02. The node is directed from group 01 to group 02. Similarly, the following assumptions were made:

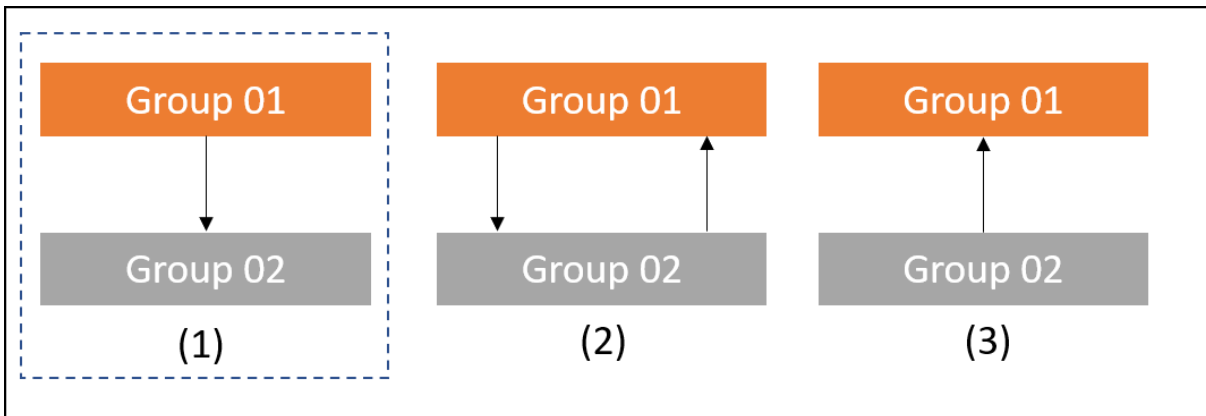


Figure 10: Three possible configuration of Bayesian Network Model

On the foundation of these assumptions, a total of three possible types were defined as depicted in Figure 12. Out of these three types, the type (1) is considered as a study model. The reason behind selecting the type (1) is that the nodes in the **Group 01** such as vehicle characteristics, time of the day, weekday/weekend and day of the week are directly influencing the variables such as speed, intensity, and travel time in **Group 02**.

**5.5.1 IAMB (Incremental Association Markov blanket) Algorithm Description:** It is an algorithm which locates the Markov blanket to the assigned variable. This receives input of a whole dataset D as a training and the assigned variable. This algorithm is categorized as forwarding strategic algorithms. This is due to the nature of the algorithm meaning, which creates the empty sets initially and start assigning the nodes one by one. The accuracy of the algorithm is judged based on the following assumptions, the learning dataset D is not dependent and identically distributed sample from a Probability distribution P in combination with the direct acyclic graph (DAG) G and that tests the conditional independence of each node and resulting values are assumed to be correct (Zhang, Zhang, Liu, & Qian, 2010).

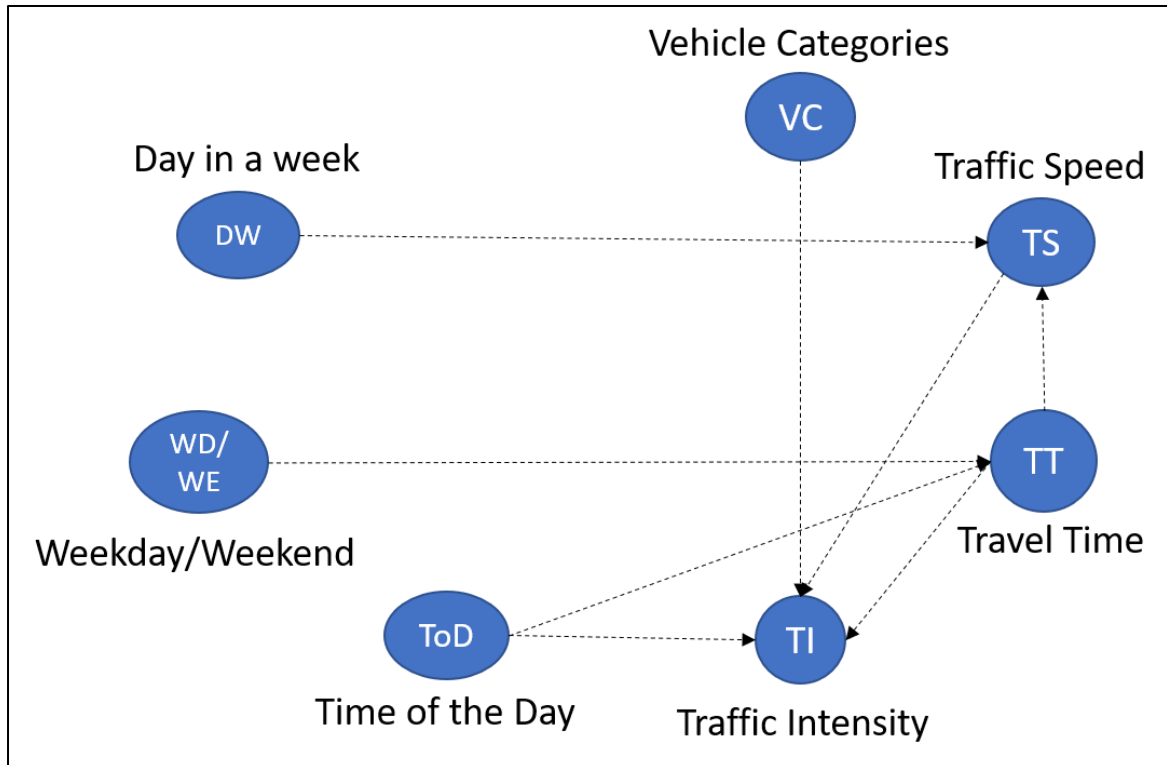


Figure 11: Bayesian Network Structure using IAMB Algorithm

**5.5.2 HC (Hill Climbing) Algorithm Description:** This algorithm is a trial and error search which is used to solve mathematical optimization problems in the field of machine learning. The main advantage of Hill climbing is given that the large input datasets with trail & error function, it will find a reasonable solution. But however, the major drawback of Hill climbing is that the resultant output is not optimal. The definition of Trial & Error (aka Heuristic Function) is a function that will list all possible combinations based on the given data. This function helps the algorithm to choose the best combination. The characteristics of HC are:

**a.) The variant of generate & Test algorithm:** Produce all possible solution and validate for the expected results if the expected results are reasonably quiet. To simplify the complexity, the variant of generating creates feedback from test results and generator uses this feedback to decide the next move while searching.

b.) **Utilization of a greedy approach:** At any given point, the greedy approach moves in a particular direction to optimize the cost function with the expectation of finding the optimal solution in the end (Bhavek, 2019).

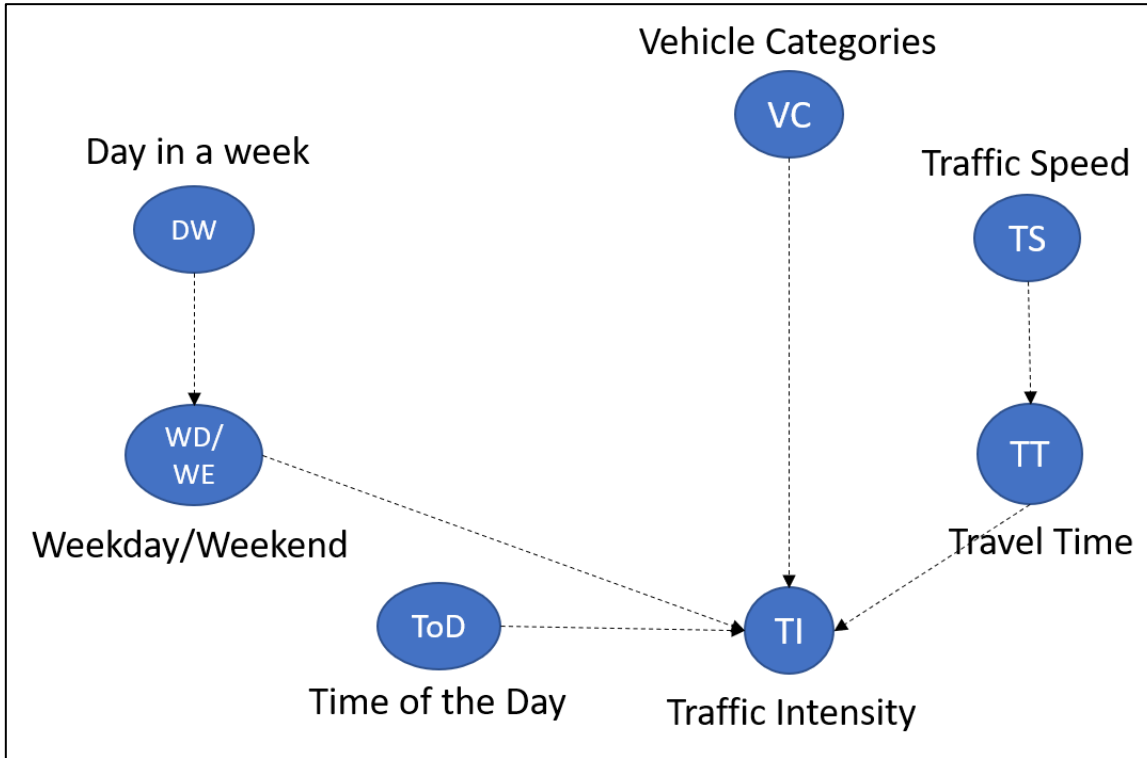


Figure 12: Bayesian Network Structure using HC Algorithm

### 5.5. Parameter Learning

After the structure of Bayesian Network is established, the following step is to measure the relationships between connected nodes. This is accomplished by calculating a conditional probability at an individual node. The nodes which do not have parent node will have a marginal distribution of the node independently. The nodes which have parents, a conditional probability table will be generated for each node. For example, From Figure 11, For the node traffic intensity (TI) the CPT is calculated as  $P(TI = t \mid VC = v, TT = a, TS = s, ToD = d)$ . The child node which as many parents' node the CPT will be too large and leads to mathematical complexities. To avoid manual calculations, several machine learning techniques can be used (Kim & Wang, 2016). In chapter 6, the different machine learning techniques which have been adapted in the study are explained in detail.

**5.6. Identifying major causes for increased ‘Traffic Intensity’ by measuring the odd ratio**

To investigate the association between parent nodes (contributing nodes) and the target node, further, the association can be quantified by measuring using ‘**odds ratio**’. The odds of an event happening is the ratio of the probability of the event will happen to the probability that the event does not happen. According (Persoskie & Ferrer, 2017) the odds ratio is defined as “The odds of an outcome event in the first category is divided by the odds of the second category”. Odds ratio (ORs) is termed as the ratios of odds which are not same as compare to probabilities. In other words, the concept of odds ratios is unintuitive. The concept of the odds ratio is majorly applied in case-control studies, where an outcome of the incident is completely unknown. When the odd ratio is greater than one ( $OR > 1$ ) states that chances of event occurrence are very high. When the odd ratio is lesser than one ( $OR < 1$ ) states that the chances of event occurrence are very low. To authenticate the outcome of odds ratio will rely on the values of Confidence Interval (CI) and P-value for statistical significance.

For example, if we consider the event of “Traffic Intensity” compared to the two groups of ‘Travel Time,’ i.e., (More Travel Time and Less Travel Time).

$$OR = \frac{p(\text{high intensity}|\text{more traveltime}) / p(\text{low intensity}|\text{more traveltime})}{p(\text{high intensity}|\text{less traveltime}) / p(\text{low intensity}|\text{less traveltime})}$$

When the odds ratio is greater than 1 depicts that the event is likely to happen in the first group. With respect to the above example, when ( $OR < 1$ ) depicts that Traffic Intensity will be higher and takes more time meaning the traffic congestion is more likely to occur when compared to low traffic intensity which takes lesser travel time meaning that there is no traffic congestion.

$$OR = \frac{\frac{P(T|C)}{P(\sim T|C)}}{\frac{P(T|\sim C)}{P(\sim T|\sim C)}} = \frac{P(T, C). P(\sim T, \sim C)}{P(T, \sim C). P(\sim T, \sim C)} \dots\dots\dots (5)$$

- Where,
- $P(T)$  = probability of traffic intensity T is high [e.g.,  $P(T = High)$  ]
- $P(\sim T)$  = probability of traffic intensity T is not high [e.g.,  $P(T \neq High)$  ]
- $P(C)$  = probability of Contributing nodes<sup>5</sup> C is occurring [e.g.,  $P(Travel\ Time = More)$  ]
- $P(\sim C)$  = probability of Contributing nodes C is not occurring [e.g.,  $P(Travel\ Time \neq More)$  ]

The OR can also be called a relative joint probability distribution. As explained in the equation ..... (5), the formula is the product of the probability of both T and C takes place, and the probability that both T and C does not takes place is divided by the product of the probabilities that one might take place.

---

<sup>5</sup> Here Parent nodes are called as contributing nodes. In this case contributing nodes are (Time of the day (ToD), Weekday/Weekend (WD/WE), Vehicle Categories (VC) and Travel Time (TT))

**5.7. Evaluating model performance for the continuous variables using RMSE, MAE & MAPE**

According to Jia, Wu & Xu, (2017) the evaluation of the Bayesian model is crucial. Three well-known performance measurements are selected to validate the model. Namely, mean absolute error (MAE), the mean absolute percentage error (MAPE), and the root mean square error (RMSE), mentioned below

$$MAE = \frac{1}{N} \sum_{i=1}^N |x_i - y_i|$$

$$MAPE = \frac{1}{N} \sum_{i=1}^N \left| \frac{x_i - y_i}{x_i} \right|$$

..... (6)

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - y_i)^2}$$

Where  $x_i$  and  $y_i$  are predicted and observed traffic intensity data.  $i$  is an interval time interval, and  $N$  is the sample size. The units of RMSE and MAE are vehicles per hour.

## 6. RESULTS & DISCUSSIONS

In this section, the results of the Bayesian Network model are discussed in detail for both Amsterdam and Rotterdam city for three time periods (2018, 2019 and 2020). The results chapter is broadly divided into four main sections. Each of the section is furthermore divided into subsections.

### **6.1 Model authentication**

### **6.2 Parameter Estimation Results**

### **6.3 Analysis and Inference**

### **6.4 Accuracy Assessment of the model through RMSE, MAE and MAPE**

### 6.1. Model Authentication

According to Sinharay, (2006) the model validation of any BN is not straight forward. This is due to the large number of input variables, which will eventually lead to a huge combination of output variables. Thus, assessing the goodness of fit of the model is essential. As stated by Kim & Wang, (2016) to assess any BN model, the two standard tests are performed.

1. Comparing the network structure using a score function.
2. To choose the optimal network structure by validating the k-fold cross-validation.

To choose the best learning algorithm for assessing goodness of fit for the Bayesian Network model. Various score functions were assigned. Familiar score functions such as Akaike’s information criterion (AIC), Bayesian information criterion (BIC) and Bayesian Dirichlet likelihood-equivalence (Bde) Carvalho, (2020) were used. To choose the optimal network structure “k-fold cross validation” function was incorporated.

k-fold cross-validation for Bayesian networks		
Target learning algorithm	IAMB	Hill-Climbing
Number of folds	10	10
Loss function	Log-Likelihood Loss (disc.)	Log-Likelihood Loss (disc.)
Expected loss	<b>7.234212</b>	<b>6.804658</b>

Table 5: K-fold Cross-validation results of two learning algorithm

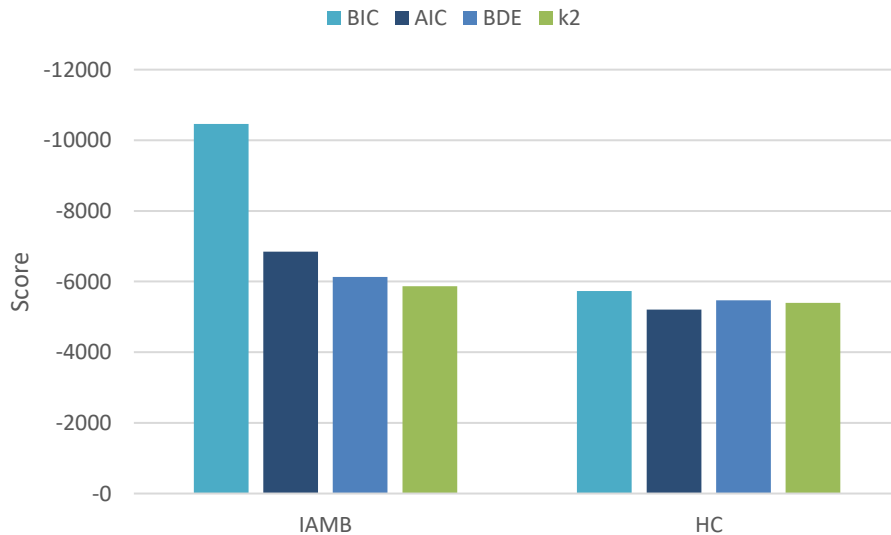


Figure 13: Network score values of Bayesian Network structure

**Table 5** shows k-fold cross-validation results for both the learning algorithm. The lower the expected loss, the better the model performs. From the table, it is very clear that the ‘HC’ algorithm with the expected loss 6.8 is expected to perform better compared to the ‘IAMB’ algorithm with the expected loss of 7.2. Figure 13 depicts the Network Score for both the learning algorithms. The network score for ‘HC’ algorithm is lesser compare to ‘IAMB’ which indicates that the ‘HC’ algorithm is comparatively better.



## 6.2. Parameter Estimation Results

In 6.2 subsection, marginal probability distribution of each node are calculated. The results are represented in the bar charts for Amsterdam and Rotterdam across three periods (2018, 2019 & 2020).

6.2.1 Marginal probability distribution table for Amsterdam - 2018, 2019 & 2020

6.2.2 Marginal probability distribution table for Rotterdam - 2018, 2019 & 2020

### 6.2.1.1 Marginal probability distribution table for Amsterdam | 2018

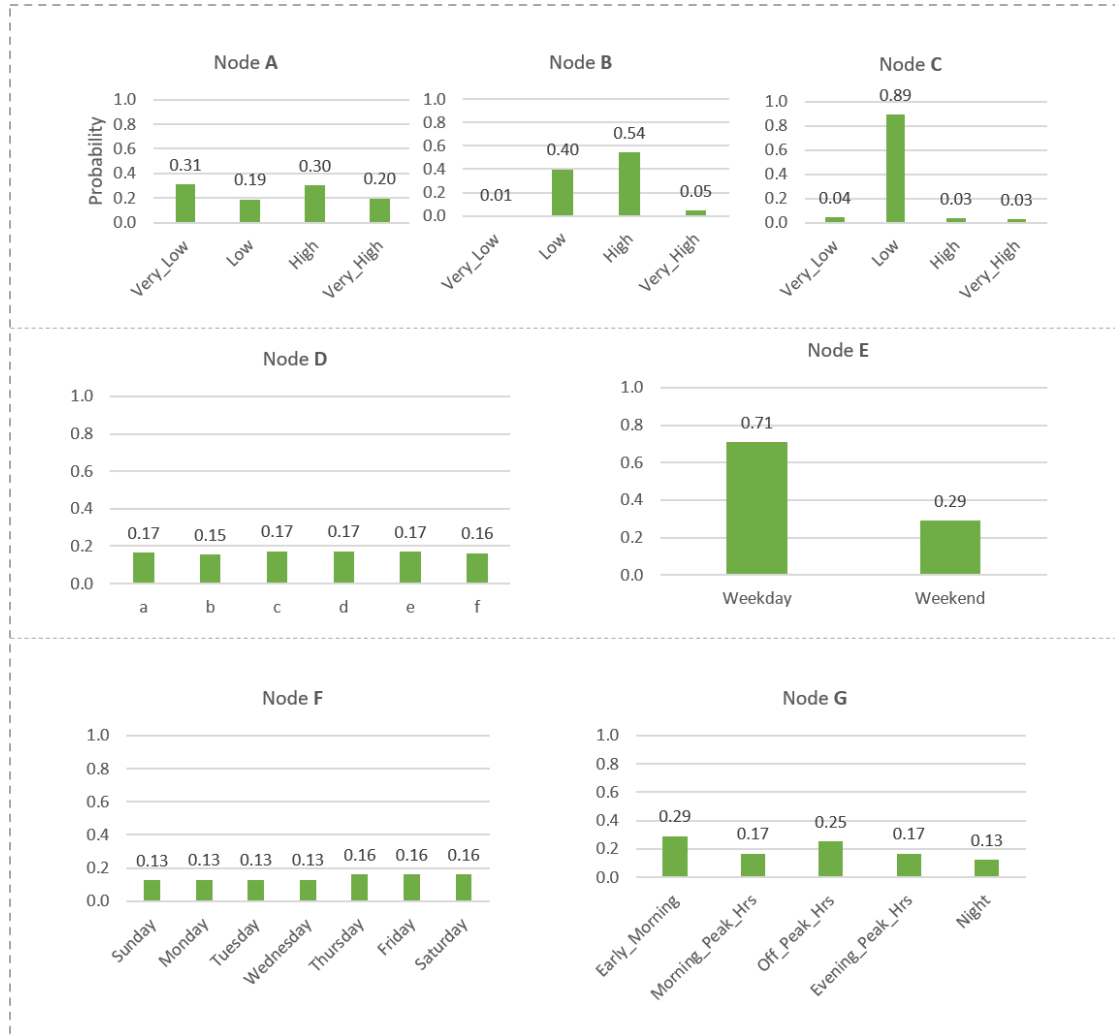


Figure 14: Representing marginal probability distribution for 7 nodes in the designed BN model (Amsterdam, 2018). Node A- Intensity, B- Travel Time, C- Traffic Speed, D- Vehicle Categories, E- Weekend or Weekday, F- Day in week & G- Time of the day

Figure 14 represents the results of the marginal probability distribution of each node for the year 2018, Amsterdam city area. The distributions of nodes such as Traffic Intensity (Node A), Vehicle Categories (Node D) & Time of the day (Node G) are consistent. The probability of Speed (Node C) being Low is 89% which is contrasting with Travel time (Node B) being high, which is 54.00%. The values from the MPD table states that, as the ‘Travel Time’ is increasing the probability of ‘Traffic speed’ is gradually getting lower. From a day in a week (Node F), it is clearly seen that the ‘Traffic Intensity’ are usually higher in the ‘Early Morning’ (29.00%) compare to other times in the day.

6.2.1.2 Marginal probability distribution table for Amsterdam | 2019

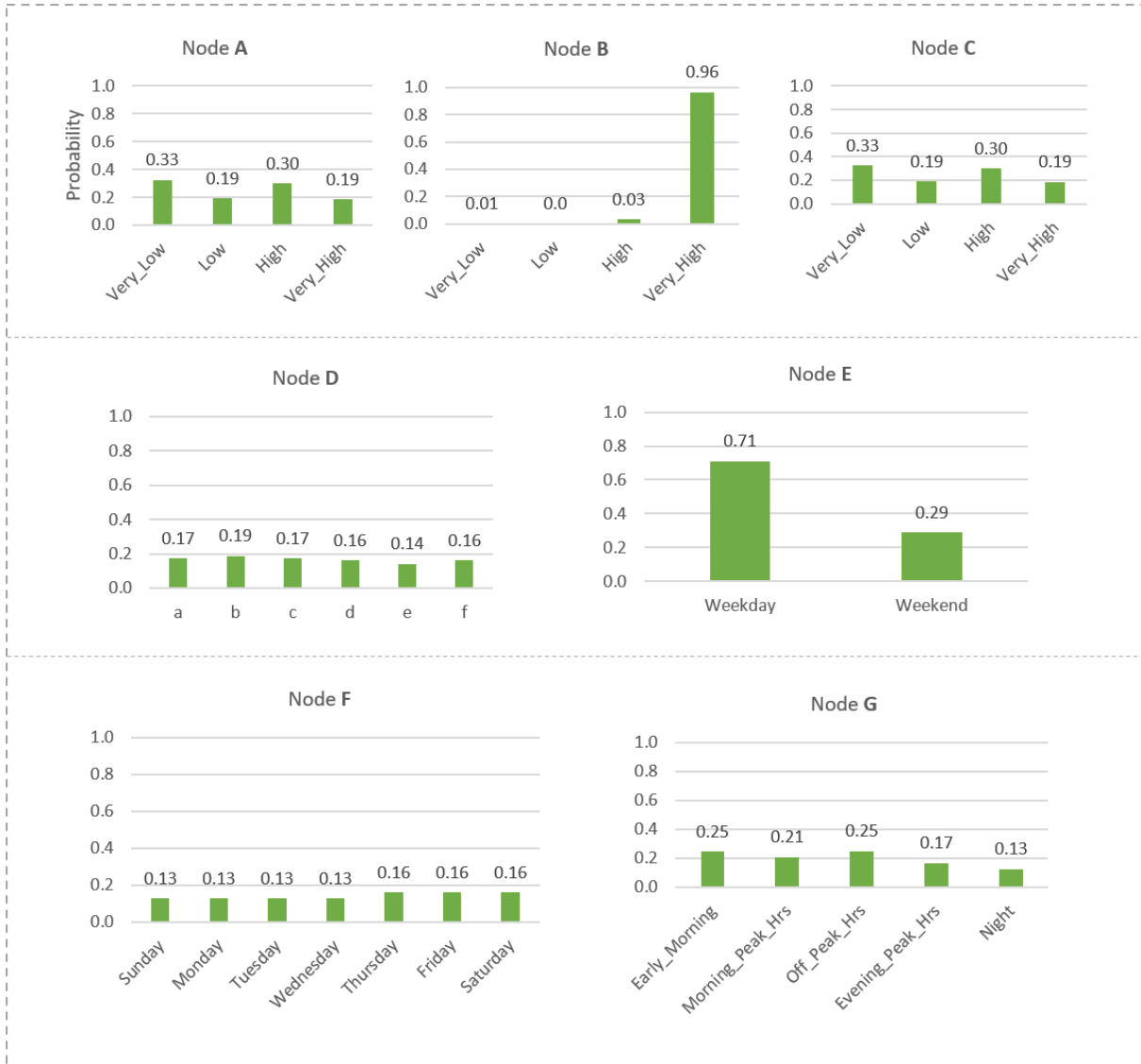


Figure 15: Representing marginal probability distribution for 7 nodes in the designed BN model (**Amsterdam, 2019**). Node **A**- Intensity, **B**- Travel Time, **C**- Traffic Speed, **D**- Vehicle Categories, **E**- Weekend or Weekday, **F**- Day in week & **G**- Time of the day

**Figure 15** represents the results of marginal distribution calculated for each defined parameters of the designed BN model. These calculations are performed for the year 2019, Amsterdam city area. The distributions for each node such as ‘Traffic Intensity’ (Node A), ‘Vehicle Categories’ (Node D) & ‘Weekend/Weekday’ are consistent which are very much evident from the bar charts. The probability of ‘Travel Time’ (Node B) being ‘Very High’ is 96.37% which is contrasting with ‘Weekdays/Weekend’ (Node E) which states that 71.06% of activities are more on ‘weekdays’ than compare to the ‘weekend’ with the probability of 28.94%. The probability of ‘Traffic Speed’ being ‘Very low’ is 32.57% compare to Traffic speed being ‘Very High’ which is 29.88%. From the Node F, ‘**Traffic intensity**’ is moderately higher in the ‘Early Morning’ (24.90%) and in ‘Off-Peak Hours’ (25.03%) compare to other times in the day.

6.2.1.3 Marginal probability distribution table for Amsterdam | 2020

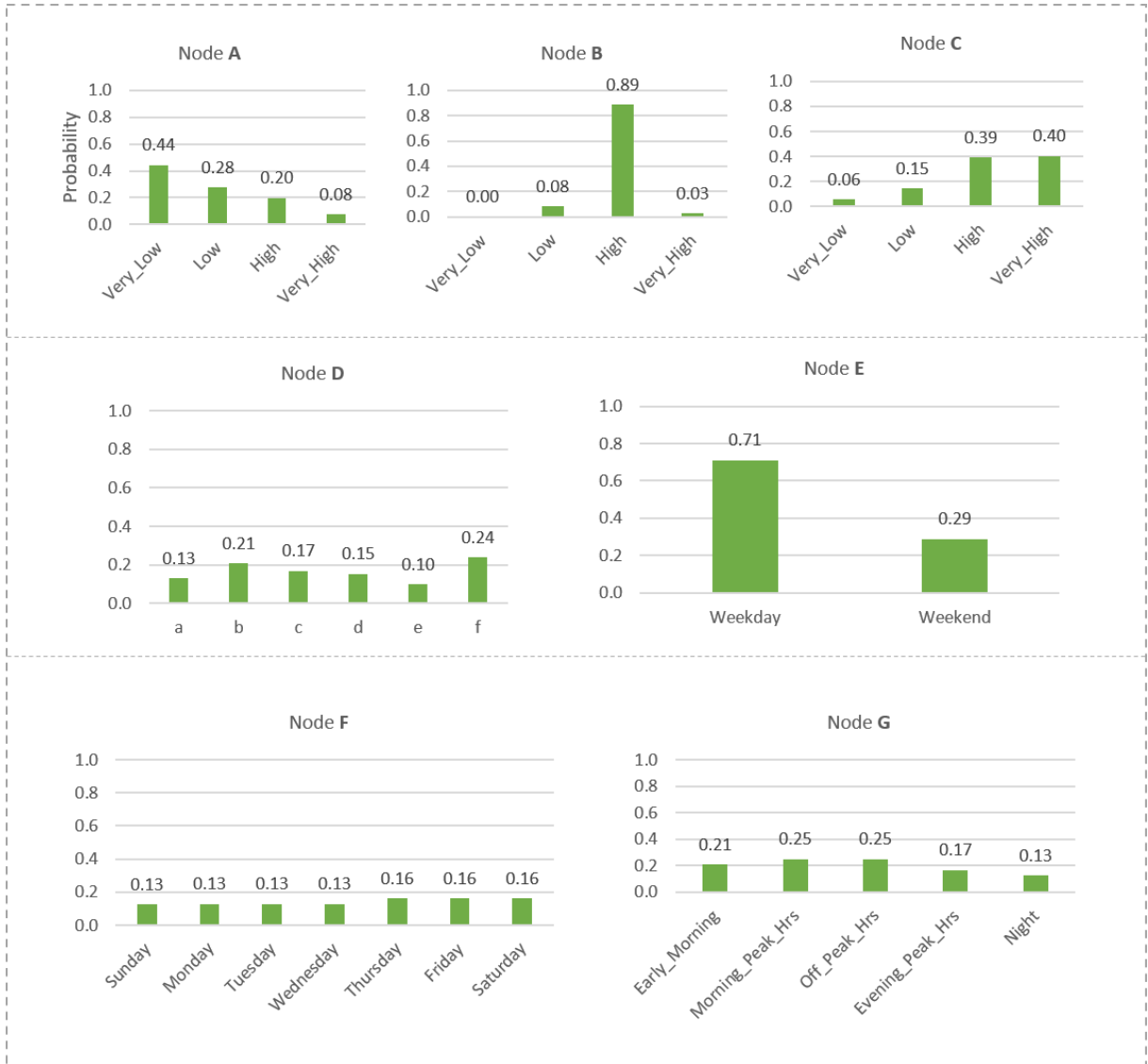


Figure 16: Representing Marginal probability distribution for 7 nodes in the designed BN model (Amsterdam, 2020). Node A- Intensity, B- Travel Time, C- Traffic Speed, D- Vehicle Categories, E- Weekend or Weekday, F- Day in week & G- Time of the day

Figure 16 represents the results of a marginal probability distribution for each node, calculated for the year 2020, Amsterdam city area. The marginal distributions of nodes such as Traffic Intensity (Node A), Vehicle Categories (Node D) & Weekend/Weekday are consistent, which are evident from the bar charts. The probability of Travel Time (Node B) being ‘High’ is 88.56% which is contrasting with Weekdays/Weekend (Node E) which states that 71.06% of activities are more on ‘weekdays’ than compare to ‘weekend’ with the probability of just 28.94%. The travel time is increasing; on the other hand, the probability of Traffic Speed is gradually getting lower. From Node F, the traffic intensity is moderately higher in the Early Morning (24.90%) and ‘Off-Peak Hours’ (25.03%) compare to other times in the day.

6.2.2.1 Marginal Probability Distribution table for Rotterdam | 2018

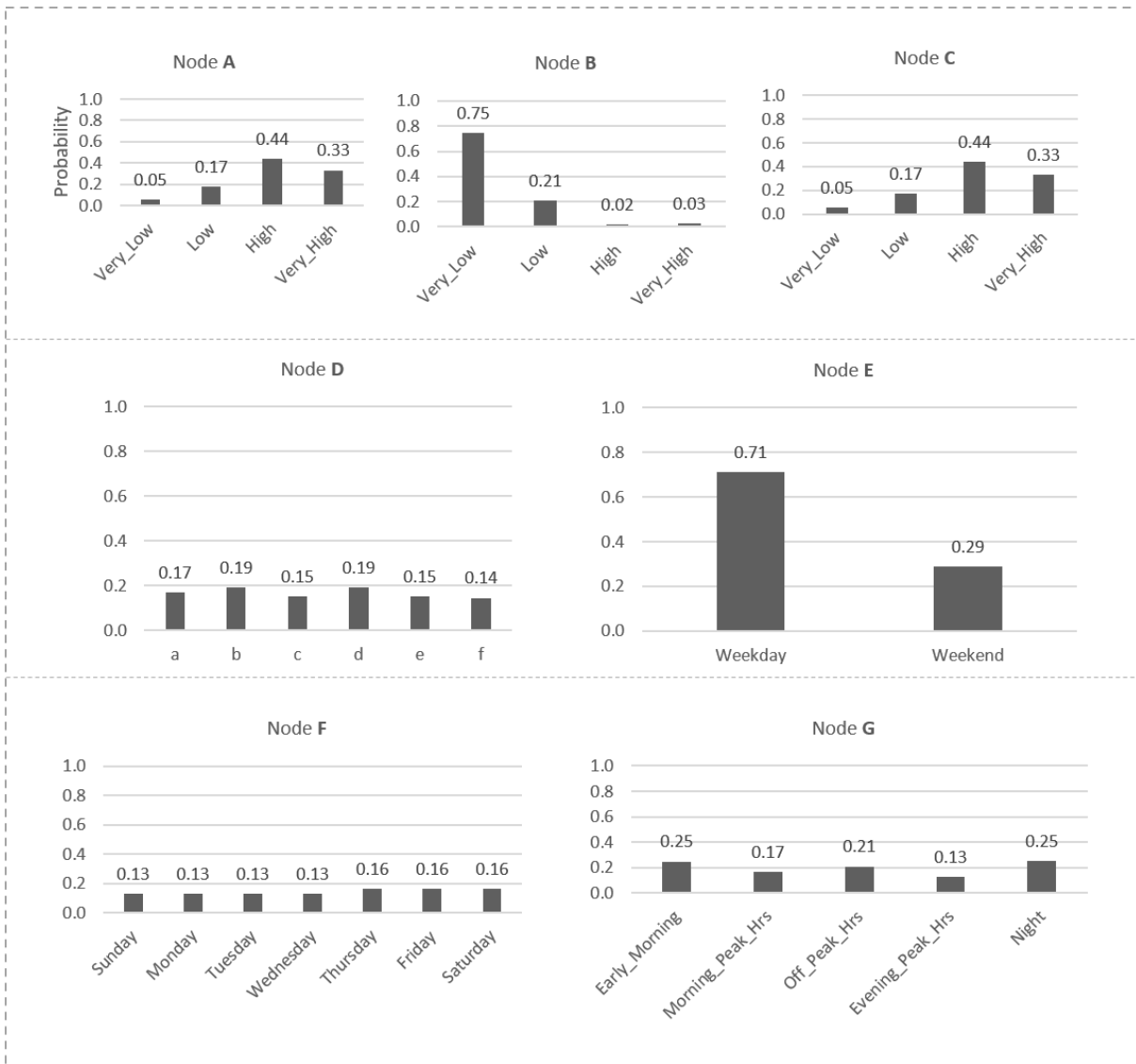


Figure 17: Representing marginal probability distribution for 7 nodes in the designed BN model (Rotterdam, 2020). Node A- Intensity, B- Travel Time, C- Traffic Speed, D- Vehicle Categories, E- Weekend or Weekday, F- Day in week & G- Time of the day

Figure 17 shows the marginal distributions calculated for each defined parameters of the designed BN model. These calculations were performed for the year 2018 of the Rotterdam area. From the graph, the probability of Travel Time (Node B) being ‘High’ is 75.00%. From ‘Weekend or Weekday’ (Node E), travel activities are more on weekdays with a probability of 71.06% compare to ‘weekend’ with the probability of 28.94% and the probability of Travel Time (Node B) and Traffic Speed (Node C) are inversely proportional.

6.2.2.2 Marginal Probability Distribution table for Rotterdam | 2019

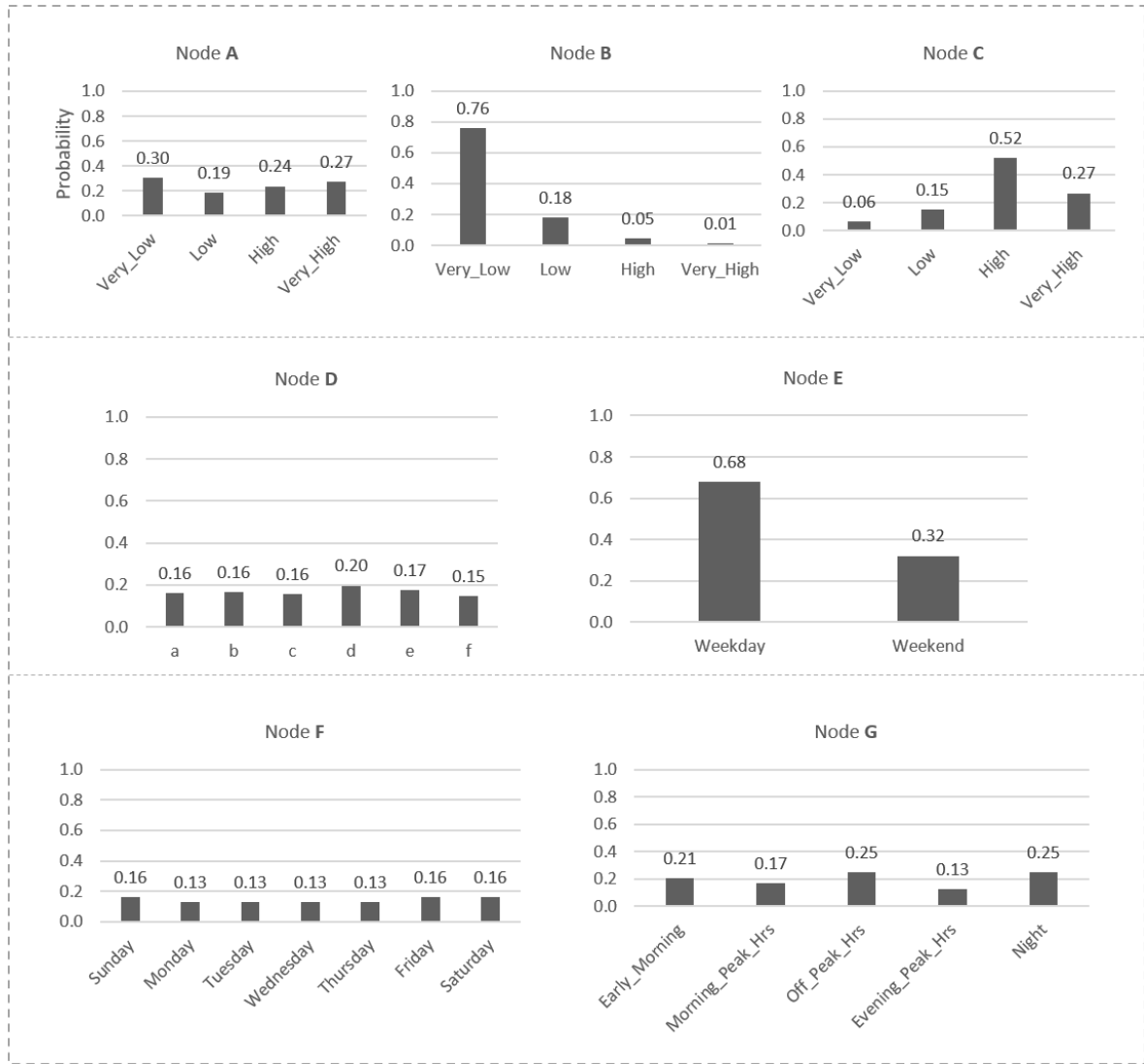


Figure 18: Representing Marginal Probability Distribution for 7 nodes in the designed BN model (Rotterdam, 2019). Node A- Intensity, B- Travel Time, C- Traffic Speed, D- Vehicle Categories, E- Weekend or Weekday, F- Day in week & G- Time of the day

Figure 18, shows the marginal distributions calculated for each defined parameters of the designed BN model for 2019, Rotterdam. The probability of travel time (node B) being 'low' is 75.91%, and the probability of traffic speed (node C) is high with 52.22%. These association between two nodes shows the inverse relationship between 2 nodes. From the (node G) it is clear that the traffic intensity is moderately higher in the 'Off-peak hours' & in 'Night' with probabilities of 25.30% and moderately higher during 'early mornings' compare to other times in the day.

6.2.2.3 Marginal Probability Distribution table for Rotterdam | 2020

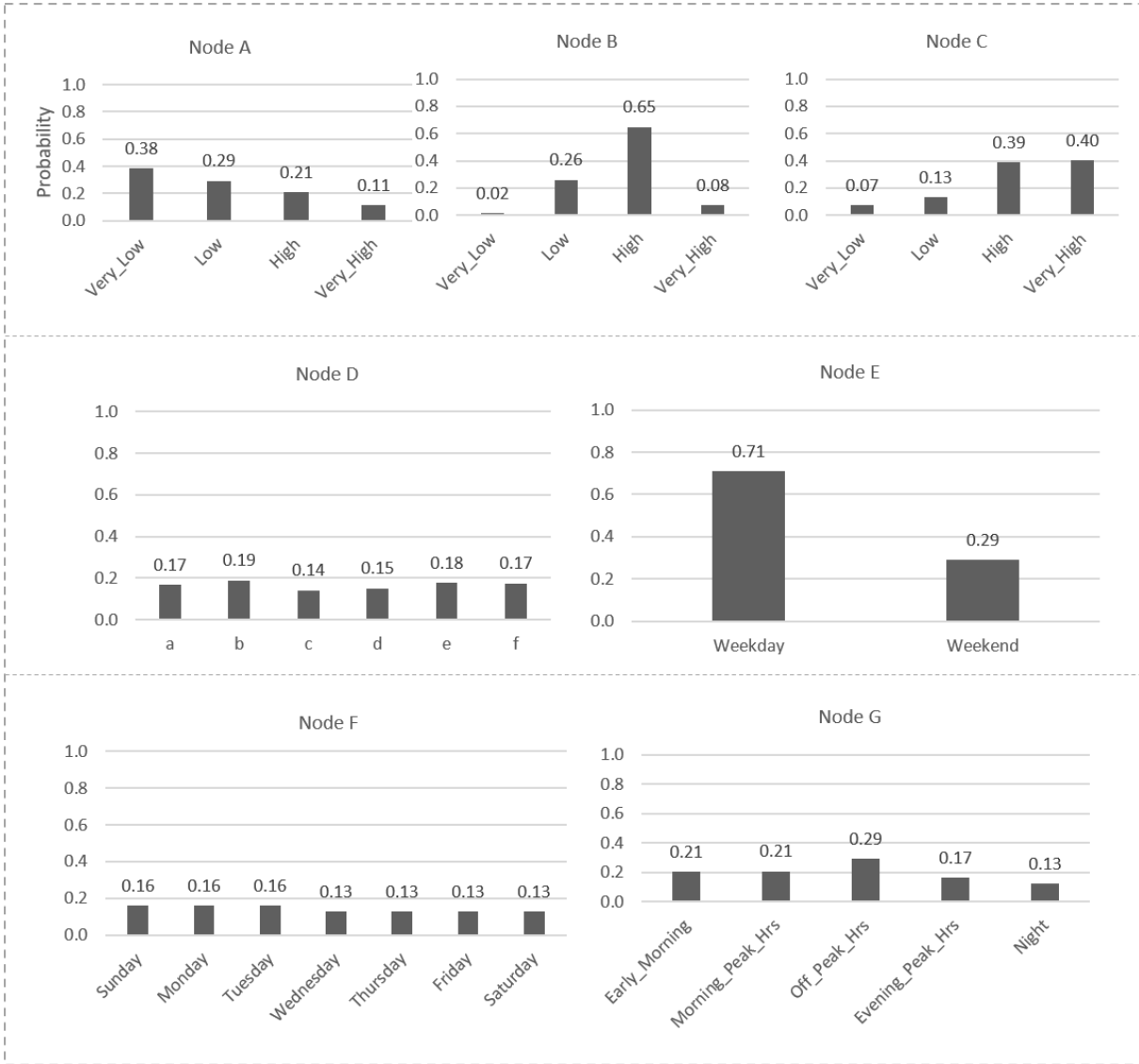


Figure 19: Representing Marginal Probability Distribution for 7 nodes in the designed BN model (Rotterdam, 2020). Node A- Intensity, B- Travel Time, C- Traffic Speed, D- Vehicle Categories, E- Weekend or Weekday, F- Day in week & G- Time of the day

Figure 19 shows the marginal distribution calculations for each defined parameters of the designed BN model. These calculations are performed for the year 2020 of the Rotterdam area. The marginal distributions values of nodes such as traffic intensity (Node A), vehicle categories (Node D) & weekend/weekday (Node E) are consistent, which are evident from the bar charts. The probabilities of travel time (Node B) being ‘High’ is 65.00% and traffic speed (Node C) being ‘very high’ is 40.00%. The association between these two nodes are linear. On the other hand, Weekdays/Weekend (Node E) states that 71.06% of activities are more on ‘weekdays’ than compare to ‘weekend’ with the probability of just 28.94%. From the Node F it is clear that the traffic intensity is moderately higher in the ‘Off-peak hours’ (29.00%) and with ‘Early mornings’ (21.03%) compare to other times in the day.

### 6.3. Analysis and Inference

In the current subsection, will start inferring from the designed BN model. The analysis/results are more focused on identifying contributing nodes that lead to the **'Traffic intensity'**. In the given study the target node is considered as **'Traffic intensity'** and will examine the results using 'cpquery', a command in a bn-learn package which identifies the association between two nodes through a probability value. From the **(Figure 11)** we can observe that the contributing nodes such as [Time of the day (ToD), weekday/weekend (WD/WE), vehicle categories (VC) and travel time (TT)] are contributing nodes for the target node 'traffic intensity (II)'.

In the following sub-section, using a BN model – one can perform diagnostic reasoning to identify the causes, which means, finding the leading causes for increased traffic intensity. In each of the given subsection, results of a posterior probability distribution of each causing node (also called as contributing nodes) are discussed.

- 6.3.1 Influence of 'Time of the Day' on 'Traffic Intensity'**  
Amsterdam & Rotterdam - 2018, 2019 & 2020
- 6.3.2 Influence on Traffic Intensity during 'Weekdays & Weekends'**  
Amsterdam & Rotterdam - 2018, 2019 & 2020
- 6.3.3 Influence of 'Vehicle Size' on 'Traffic Intensity'**  
Amsterdam & Rotterdam - 2018, 2019 & 2020
- 6.3.4 Influence of 'Travel Time' on 'Traffic Intensity'**  
Amsterdam & Rotterdam - 2018, 2019 & 2020
- 6.3.5 Identifying major causes for increased 'Traffic Intensity'**  
Amsterdam & Rotterdam - 2018, 2019 & 2020
- 6.4.5 Traffic flow prediction using continuous variables**  
Amsterdam & Rotterdam - 2018, 2019 & 2020

6.3.1. Influence of ‘Time of the Day’ on ‘Traffic Intensity’

Amsterdam, 2018

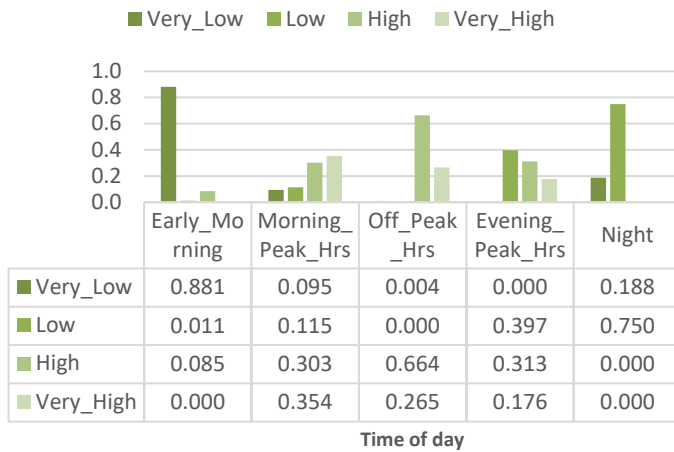


Figure 20: CPT table showing impact of ‘Time of the day’ on ‘Traffic Intensity’ – AMS, 2018

The probability of traffic intensity being ‘high’ in the ‘off-peak hours’ is 67.99%, 31.07% during ‘morning peak hours’ & 32.03% during ‘evening peak hours’.

From Figure 20,  
 The probability that a traffic intensity being ‘very low’ in the ‘early morning’ is 90.23%; during ‘night’, the traffic intensity being ‘very low’ is 19.23%.  
 The probability that traffic intensity being ‘very high’ is 36.24% during ‘morning peak hours’; 27.00% during ‘off-peak hours’ & 18.06% during ‘evening peak hours’.  
 The probability of ‘low’ traffic intensity during ‘night’ is 76.81%, this percentage depicts that there is a higher chance of traffic being ‘low’ in the night compared to other times of the day.

Rotterdam, 2018

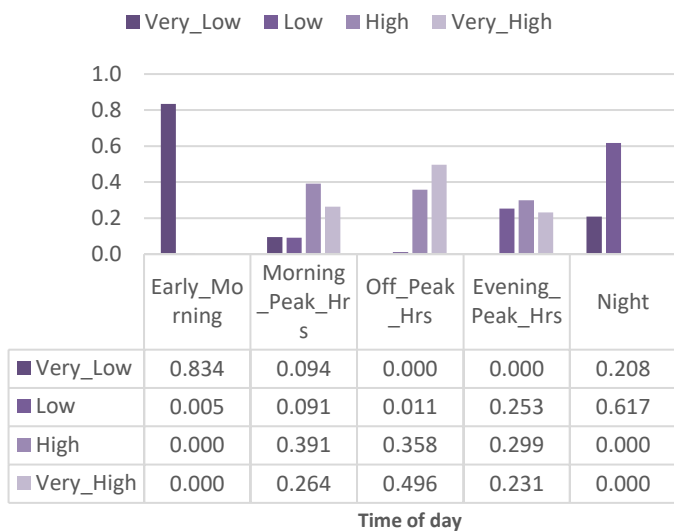


Figure 21: CPT table showing impact of ‘Time of the day’ on ‘Traffic Intensity’ – RTM, 2018

From Figure 21,  
 The probability that a traffic intensity being ‘very low’ in the ‘early morning’ is 99.43% & during ‘night’, the traffic intensity being ‘low’ is 74.23%.  
 The probability that traffic intensity being ‘very high’ is 57.35% during ‘off-peak hours’; 31.45% during ‘morning peak hours’ & 29.53% during ‘Evening Peak Hours’.  
 The probability of ‘low’ traffic intensity during ‘night’ is 74.75%; this percentage depicts that there is a higher chance of low traffic in the night compared to other times of the day.



**Amsterdam, 2019**

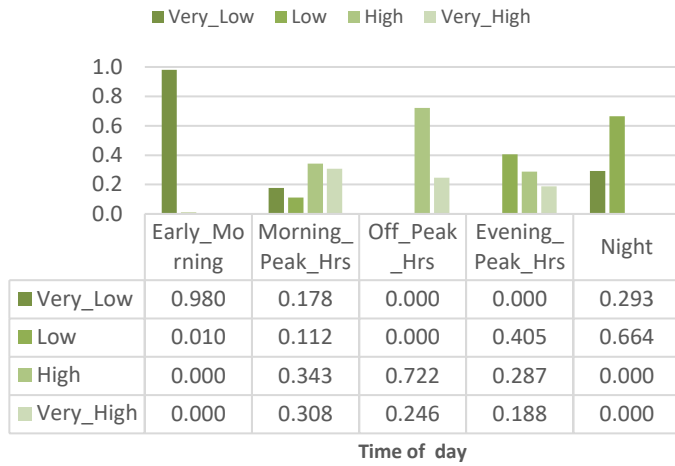


Figure 22: CPT table showing impact of ‘Time of the day’ on ‘Traffic Intensity’ – AMS, 2019

day.

The probability of traffic intensity being ‘high’ in the ‘off-peak hours’ is 74.56% & 36.52% during ‘morning peak hours’ & 32.64% during ‘evening peak hours’.

From Figure 22,

The probability that traffic intensity is ‘very low’ in the ‘early morning’ is 99.03%; during ‘night’, the intensity of traffic gradually drops down to 30.62%.

The probability of traffic intensity is ‘High’ is 36.52% during ‘morning peak hours’ & 74.56% during ‘off-peak hours’ & 32.64% during ‘evening peak hours’.

The probability of ‘low’ traffic intensity during ‘night’ is 69.38%, this percentage depicts that there is a higher chance of traffic being ‘low’ in the night compared to other times of the

**Rotterdam, 2019**

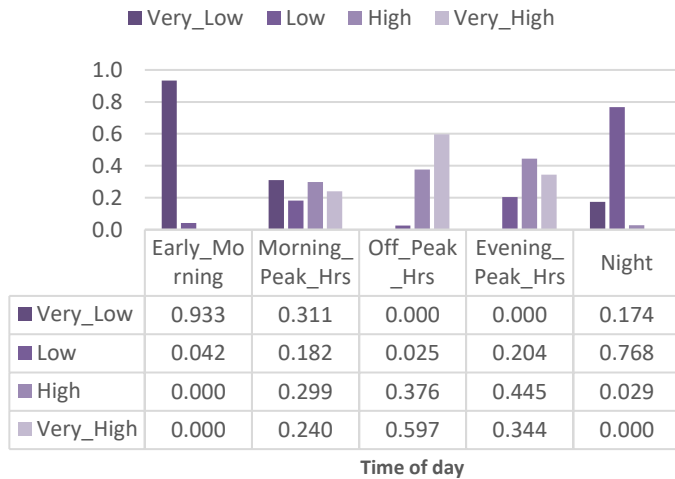


Figure 23: CPT table showing impact of ‘Time of the day’ on ‘Traffic Intensity’ – RTM, 2019

The probability of traffic intensity being ‘high’ in the ‘off-peak hours’ is 59.79%; 23.29% during ‘morning peak hours’ & 34.59% during ‘evening peak hours’.

From (Figure 22),

The probability that the traffic intensity is ‘very low’ in the ‘early morning’ is 95.69%, during ‘night’, the intensity of traffic gradually drops down to 17.93%.

The probability of traffic intensity being ‘high’ during ‘evening peak hours’ is 44.83%; during ‘off-peak hours’ reduces to 37.70%.

The probability of ‘low’ traffic intensity during ‘night’ is 79.11%, this percentage depicts that there is a higher chance of traffic being ‘low’ in the night compared to other times of the day.

The above evidence depicts the presence of a stronger correlation between “traffic intensity” node and “Time of the Day” node.

**Amsterdam, 2020**

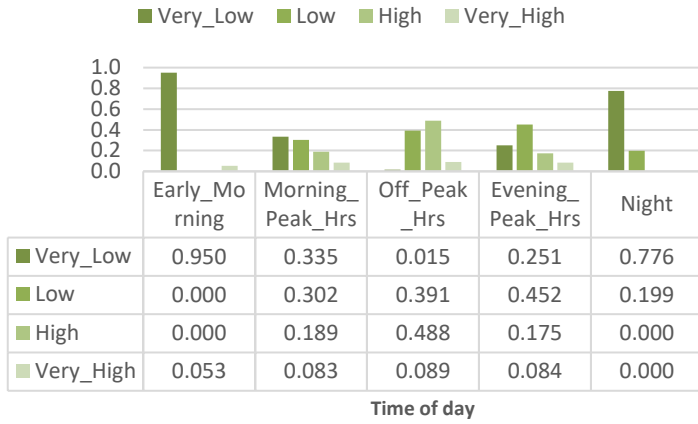


Figure 24: CPT table showing impact of ‘Time of the day’ on ‘Traffic Intensity’ – AMS, 2020

The probability of ‘very low’ traffic intensity during ‘Night’ is relatively higher. This percentage depicts that there is a higher chance of traffic being “low” in the night compared to other times of the day.

**Rotterdam, 2020**

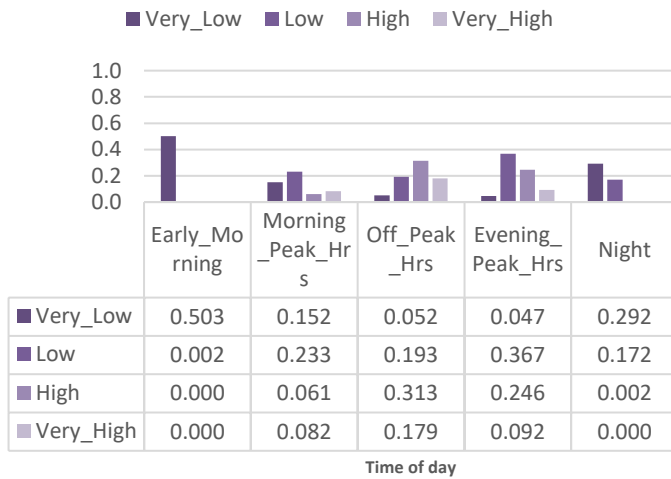


Figure 25: CPT table showing impact of ‘Time of the day’ on ‘Traffic Intensity’ – RTM, 2020

The probability of ‘low’ traffic intensity is observed in the overall times of the day, which are evident from the CPT values.

The probability of traffic intensity being ‘High’ during ‘off-peak hours’ is 49.60% & ‘very high’ is 12.85% during ‘Morning peak hours’ & 8.69% during ‘Evening peak hours’.

The above evidence depicts that, there is an increase in probability values in ‘time of the day’ when it is correlated with the ‘Traffic intensity’ for all the three periods. Earlier, from the marginal distribution table, the values of the node ‘ToD’ were comparatively lower.

From Figure 24,

The probability that a traffic intensity being ‘very low’ in the ‘early morning’ is 87.23% & 79.55% in the ‘night’. Compare to 2018 and 2019, the probability value of traffic intensity being ‘very high’ is lesser than 12.85% during all the times of the day.

Also, the probability of traffic intensity being ‘high’ is 20.82% during ‘morning peak hours’ & 49.60% during ‘off-peak Hours’ & 18.19% during ‘evening peak hours’.

From Figure 25,

The probability that the traffic intensity being ‘very low’ in the ‘early morning’ is 99.66%, 62.67% during ‘night’. Similarly, for ‘very low’, it is 79.55%. The values depict that the ‘traffic intensity’ shows a major peak on ‘early morning’ and in the ‘late evenings’. The probability of traffic intensity being ‘high’ is 42.45% during ‘off-peak hours’ & 32.67% during ‘evening peak hours’ & 11.15% during ‘morning peak hours’.

The probability of ‘low’ traffic intensity

6.3.2. Influence on Traffic Intensity during ‘Weekdays & Weekends’

Amsterdam, 2018

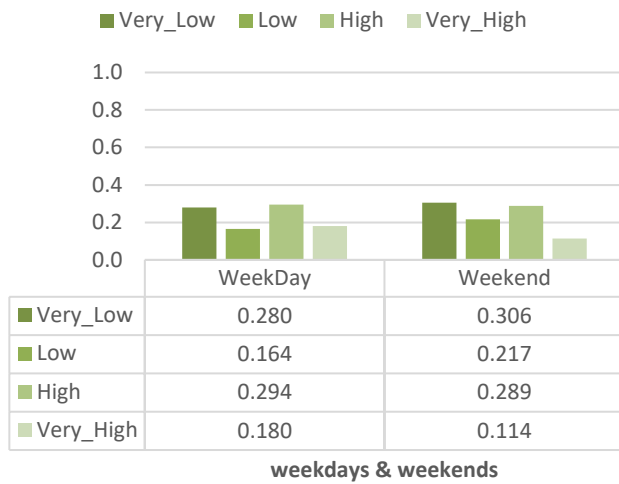


Figure 26: CPT table showing impact on ‘Traffic Intensity’ during WD/WE – AMS, 2018

In the current sub-section, the association between ‘traffic intensity’ and ‘weekdays/weekends’ are explained in detail.

Figure 26, shows the probability of traffic intensity being ‘very low’ is 30.04% on weekdays & 33.04% during weekends. This percentage represents that the traffic intensity is comparatively lower on weekends and higher on weekdays.

The probability of traffic intensity being ‘very high’ is 19.6% on weekdays when it is comparing with weekends which is just 12.34%. This percentage shows a stronger correlation between the two nodes.

Rotterdam, 2018

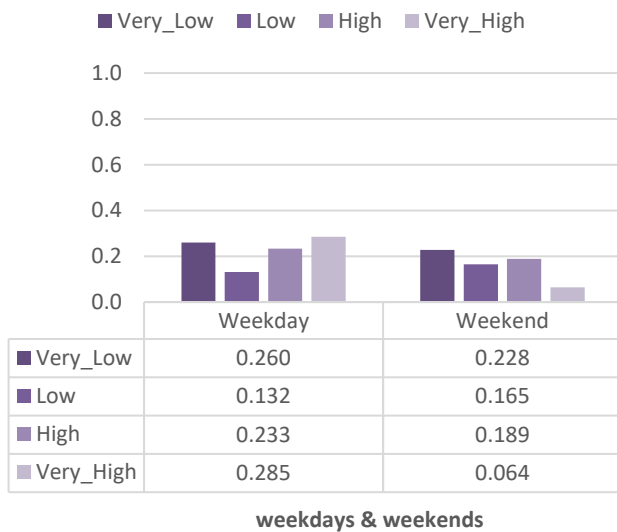


Figure 27: CPT table showing impact on ‘Traffic Intensity’ during WD/WE – RTM, 2018

Figure 27. shows the probability values of traffic intensity being ‘very low’ is 28.54% during weekdays & 35.30% during weekends. This percentage represents that the traffic intensity is slightly lower on weekends.

The probability of traffic intensity being ‘very high’ 31.34% on weekdays when it is comparing with weekends, is 9.924%.

**Amsterdam, 2019**



Figure 28: CPT table showing impact on ‘Traffic Intensity’ during WD/WE – AMS, 2019

From Figure 28, the probability of traffic intensity being ‘very low’ is 32.77% during weekdays & 36.04% during weekends. This percentage represents that the traffic intensity is comparatively lower on weekdays and higher on weekends similar to 2018.

The probability of traffic intensity being ‘high’ is 31.23% on weekdays when it is compared with weekends which is 29.96%.

**Rotterdam, 2019**

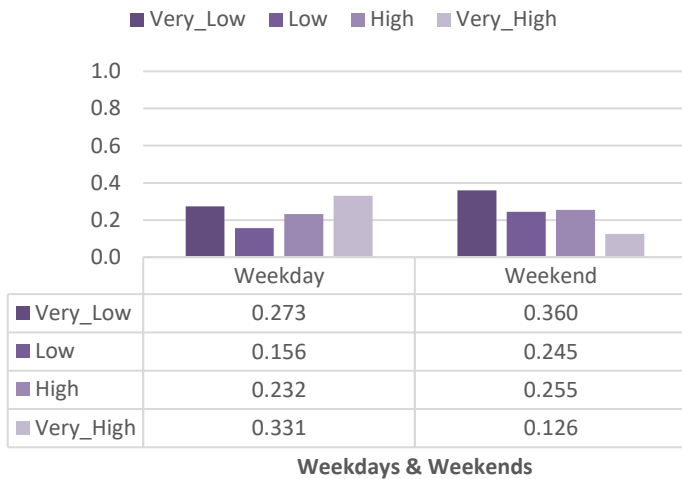


Figure 29: CPT table showing impact on ‘Traffic Intensity’ during WD/WE – RTM, 2019

From Figure 29, the probability of traffic intensity being ‘very low’ is 27.53% during weekdays & 36.04% on weekends. This percentage represents that the traffic intensity gradually reduces over the weekend and increases on weekdays.

The probability of traffic intensity being ‘high’ is 33.35% on weekdays when it is compared with weekends which is 12.80%. This percentage shows a stronger correlation between the two nodes.

**Amsterdam, 2020**

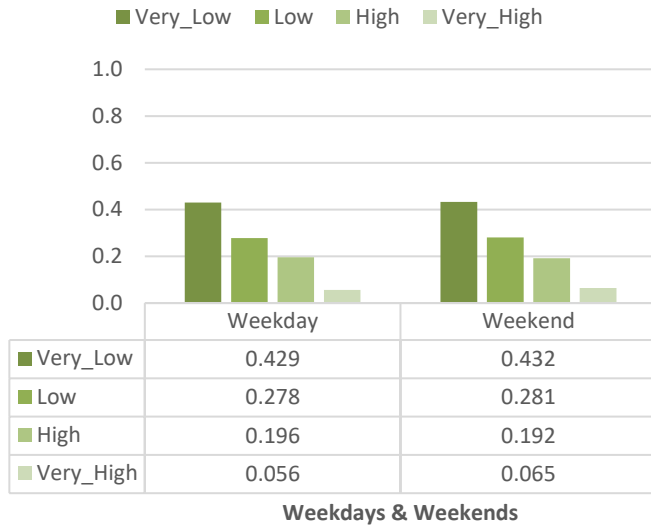


Figure 30: CPT table showing impact on ‘Traffic Intensity’ during WD/WE – AMS, 2020

From Figure 30,

The probability of traffic intensity being ‘very low’ is 44.74% during weekdays. Traffic intensity being ‘very low’ is 44.56% on weekends. This percentage represents that the traffic intensity is comparatively lower on weekdays and weekends.

The probability of traffic intensity being ‘high’ is 5.81% on weekdays when it is compared with the weekend, which is 6.70%. This percentage shows a stronger correlation between the two nodes.

**Rotterdam, 2020**

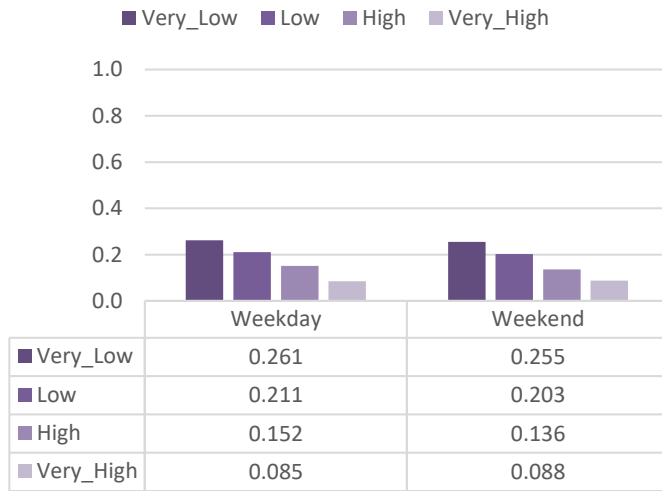


Figure 31: CPT table showing impact on ‘Traffic Intensity’ during WD/WE – RTM, 2020

Figure 31 depicts,

The probability of traffic intensity being ‘very low’ is 36.85% during weekdays, and the values remain the same on the weekends. These shows fewer vehicles on the road. This percentage represents that the traffic intensity is the same on weekdays and weekends.

The probability of traffic intensity being ‘high’ is 12.03% during weekends & 12.41% during weekdays.

The CPT values of the node ‘WD/WE’ remains almost similar in all the three-time periods.

The above evidence depicts that, there is no increase in the probability values on ‘weekend/weekday’ node when correlated with the ‘traffic intensity’ for all the three periods. Earlier, from the marginal distribution table, the values of the node ‘WE/WD’ remains almost the same. These association states that the traffic intensity is not majorly affected by the parent node.

6.3.3. Influence of ‘Vehicle Size’ on ‘Traffic Intensity’

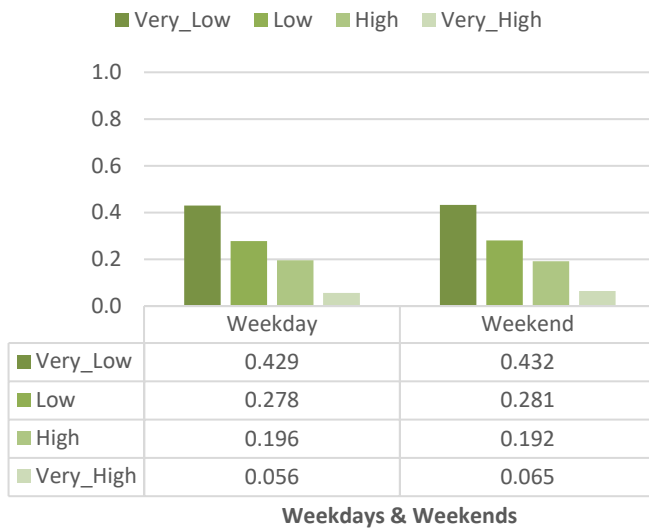


Figure 32: CPT table showing impact on ‘Traffic Intensity’ during WD/WE – AMS, 2020

**Amsterdam, 2020**

From Figure 30, The probability of traffic intensity being ‘very low’ is 44.74% during weekdays. Traffic intensity being ‘very low’ is 44.56% on weekends. This percentage represents that the traffic intensity is comparatively lower on weekdays and weekends.

The probability of traffic intensity being ‘high’ is 5.81% on weekdays when it is compared with the weekend, which is 6.70%. This percentage shows a stronger correlation between the two nodes.

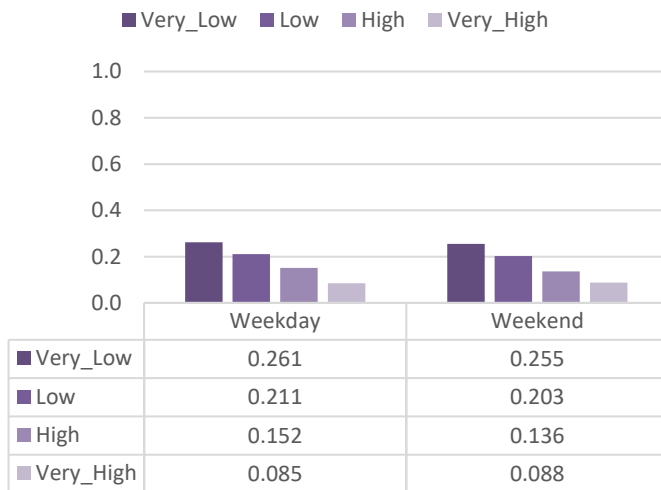


Figure 33: CPT table showing impact on ‘Traffic Intensity’ during WD/WE – RTM, 2020

**Rotterdam, 2020**

Figure 31 depicts, The probability of traffic intensity being ‘very low’ is 36.85% during weekdays, and the values remain the same on the weekends. These shows fewer vehicles on the road. This percentage represents that the traffic intensity is the same on weekdays and weekends.

The probability of traffic intensity being ‘high’ is 12.03% during weekends & 12.41% during weekdays.

The CPT values of the node ‘WD/WE’ remains almost similar in all the three-time periods.

The above evidence depicts that, there is no increase in the probability values on ‘weekend/weekday’ node when correlated with the ‘traffic intensity’ for all the three periods. Earlier, from the marginal distribution table, the values of the node ‘WE/WD’ remains almost the same. These association states that the traffic intensity is not majorly affected by the parent node.

**Amsterdam, 2019**

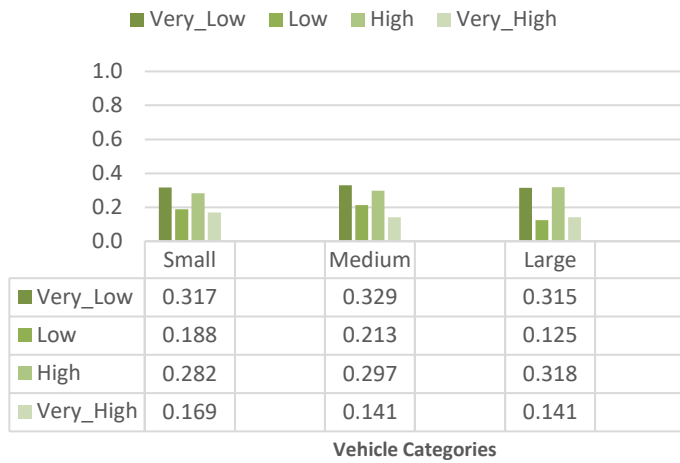


Figure 34: CPT table showing the impact of vehicles size on ‘Traffic Intensity’ – AMS, 2019

on traffic intensity is relatively high, with a probability of 29.50%.

**Rotterdam, 2019**

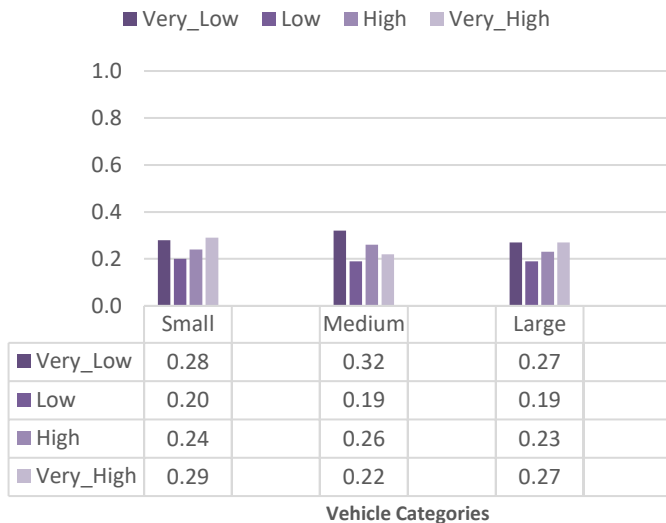


Figure 35: CPT table showing the impact of vehicles size on ‘Traffic Intensity’ – RTM, 2019

on traffic intensity is relatively high, with a probability of 32.32%.

Figure 34 states the impact of different length of vehicles, causing an effect on traffic intensity. The impact of small to medium scale vehicles on traffic intensity is relatively similar (small size = 33.16%) compared to the medium size vehicles (medium size = 33.57%).

**[Large Vehicles]** The probability of large size vehicles, causing an effect on traffic intensity is high with a probability of 35.3%.

**[Medium Vehicles]** The probability of medium size vehicles, causing an effect on traffic intensity is moderate, with a probability of 30.3%.

**[Small Vehicles]** The probability of small size vehicles, causing an effect

Figure 35 states the impact of vehicle sizes on traffic intensity. The impact of small to medium scale vehicles on traffic intensity is relatively lower (small size = 27.72%) compared to the medium size vehicles (medium size = 32.32%).

**[Large vehicles]** The probability of large size vehicles causing an effect on traffic intensity is high with a probability of 28.13%.

**[Medium vehicles]** The probability of medium size vehicles causing the effect on traffic intensity is moderate with a probability of 30.30%.

**[Small vehicles]** The probability of medium size vehicles, causing an effect

**Amsterdam, 2020**

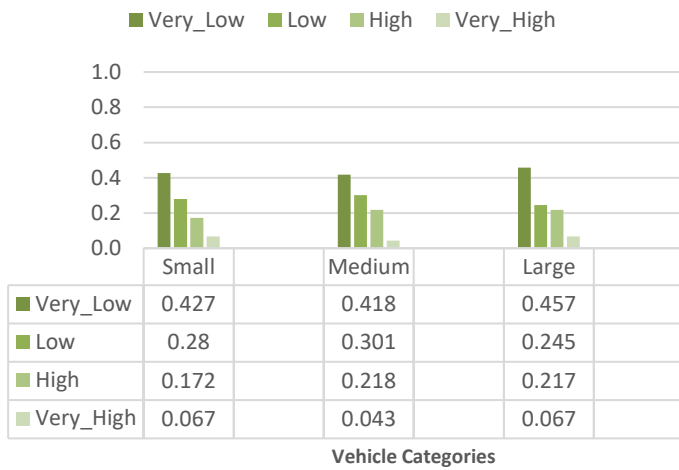


Figure 36: CPT table showing the impact of vehicles size on ‘Traffic Intensity’ – AMS, 2020

Figure 36 states the impact of vehicle sizes on traffic intensity. The impact of small to medium scale vehicles on traffic intensity being ‘low’ is relatively lower (small size = 45.13%) when it is compared to the medium size vehicles (medium size = 42.65%).

**[Large vehicles]** The probability of larger size vehicles, causing an effect on traffic intensity is high with a probability of 22.00%.

**[Medium vehicles]** The probability of medium size vehicles, causing an effect on traffic intensity is moderate, with a probability of 22.24%.

**[Small vehicles]** The probability of small size vehicles, causing an effect on traffic intensity is comparatively high with a probability of 29.50%.

The above evidence depicts the correlation between the Vehicle Categories, and traffic intensity does not have stronger correlation.

**Rotterdam, 2020**

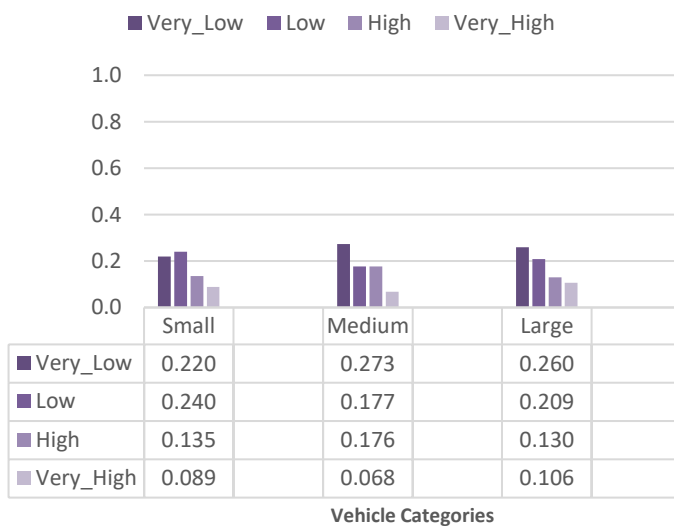


Figure 37: CPT table showing the impact of vehicles size on ‘Traffic Intensity’ – RTM, 2020

Figure 37 states the impact of different size vehicles on traffic intensity. The impact of small to medium scale vehicles on traffic intensity is relatively lower (small size = 32.16%) when it is compared to the medium size vehicles (medium size = 39.34%).

**[Large vehicles]** The probability of large size vehicles causing an effect on traffic intensity is low with a probability of 15.04%.

**[Medium vehicles]** The probability of medium size vehicles, causing an effect on traffic intensity is very low with a probability of 39.34%.

**[Small vehicles]** The probability of small size vehicles, causing an effect on traffic intensity is comparatively high with a probability of 29.50%.

The above evidence depicts that, there is a slight significant increase in the probability values in ‘vehicle categories’ node when correlated with the ‘traffic intensity’ for all the three periods. Earlier, from the marginal distribution table, the values for the node ‘vehicle categories’ relatively lower. These association states that the traffic intensity is affected by the parent node.



6.3.4. Influence of ‘Travel Time’ on ‘Traffic Intensity’

Amsterdam, 2018

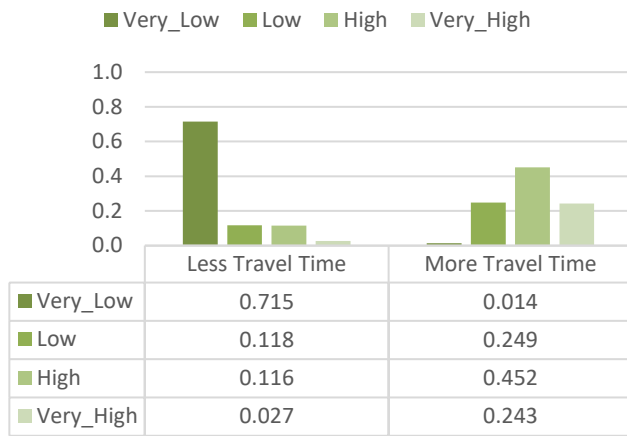


Figure 38: CPT table showing the effect of Travel Time on ‘Traffic Intensity’ – AMS, 2018

Figure 38 shows the probability distribution between travel time and traffic intensity for Amsterdam, 2018.

**[Less Travel Time]**. With a probability of 73.23%, traffic intensity being ‘very low’ is due to less travel time. In other words, the travel intensity is relatively less, meaning the flow of traffic is smooth and takes less time to reach from point A to point B.

**[More Travel Time]**. The probability of traffic intensity being more is contributed more to the travel time. Thus, the travel time becomes higher with the probability of 47.18%.

The above evidence depicts the correlation between ‘travel time’ and ‘traffic intensity’ does have a stronger correlation.

Rotterdam, 2018

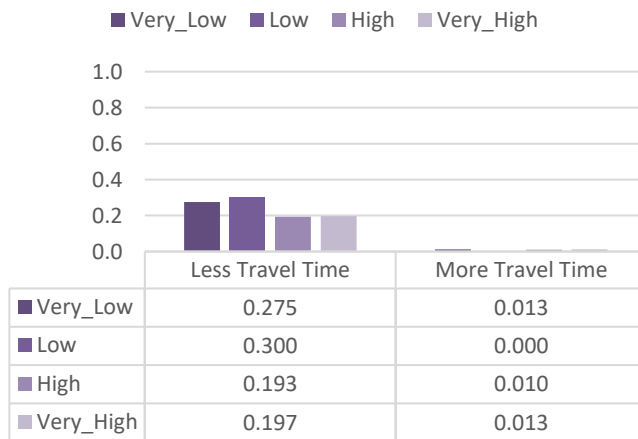


Figure 39: CPT table showing the effect of Travel Time on ‘Traffic Intensity’ – RTM, 2018

Figure 39 shows the probability distribution between travel time and traffic intensity for Rotterdam, 2018.

**[Less Travel Time]**. With the probability of 31.10% that the traffic intensity being low is due to less travel time. In simple words, the travel intensity is relatively low meaning the flow of traffic is smooth.

**[More Travel Time]** The probability of traffic intensity being ‘high’ is contributed more to the travel time. Thus, the travel time becomes higher with the probability of 35.53%.

The above evidence depicts the stronger correlation between the ‘Travel Time’ and ‘Traffic Intensity’ node.

**Amsterdam, 2019**

Figure 40 & Figure 41 shows the probability of distribution between travel time and traffic intensity for Amsterdam & Rotterdam, 2019.



Figure 40: CPT table showing the effect of Travel Time on ‘Traffic Intensity’ – AMS, 2019

**[Less Travel Time].** With the probability of 54.25% that the traffic intensity being ‘very low’ is due to less travel time. In simple words, the travel intensity is relatively low meaning the flow of traffic is smooth and takes less time to reach from point A to point B.

**[More Travel Time].** The probability of traffic intensity being ‘high’ is contributed more to the travel time. Thus, the travel time becomes higher with the probability of 57.31%.

**Rotterdam, 2019**

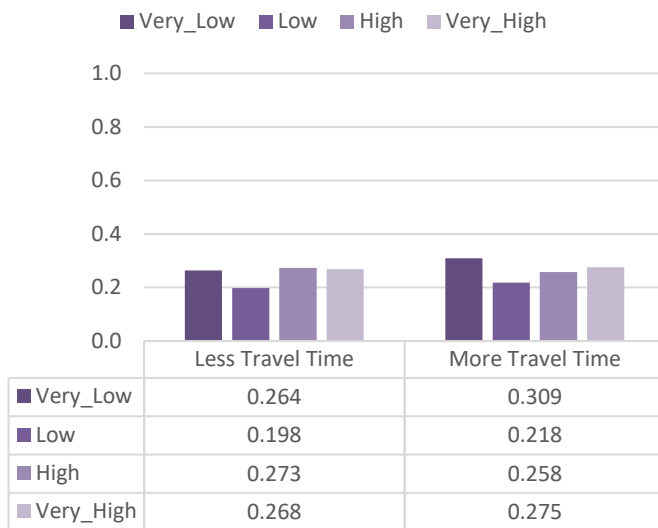


Figure 41: CPT table showing the effect of Travel Time on ‘Traffic Intensity’ – RTM, 2019

**[Less Travel Time].** With the probability of 26.30% that the traffic intensity is very low is due to less travel time. In other words, the travel intensity is relatively low meaning the flow of traffic is smooth and takes less time to reach from point A to point B.

**[More Travel Time].** The probability of traffic intensity being ‘high’ is contributed more to the travel time. Thus, the travel time becomes higher with the probability of 37.54%.

**Amsterdam, 2020**

Figure 42 & Figure 43 shows the probability of distribution between travel time and traffic intensity for Amsterdam & Rotterdam, 2020.

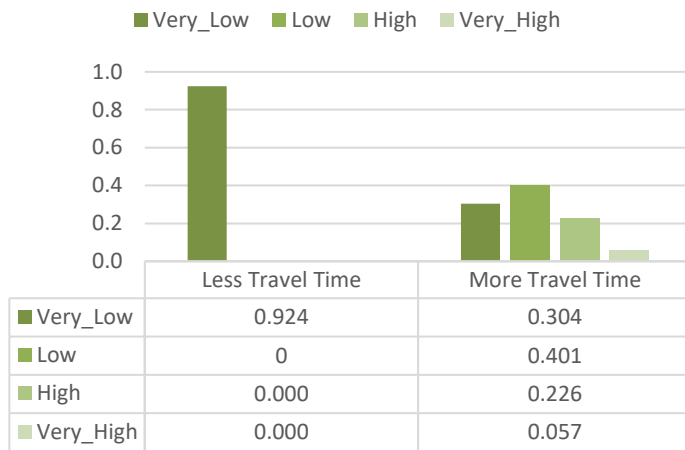


Figure 42: CPT table showing the effect of Travel Time on ‘Traffic Intensity’ – AMS, 2020

**Rotterdam, 2020**



Figure 43: CPT table showing the effect of Travel Time on ‘Traffic Intensity’ – RTM, 2020

**[Less Travel Time]** With the probability of 95% that the traffic intensity being ‘very low’ is due to less travel time. In simple words, the travel intensity is relatively low meaning the flow of traffic is smooth and takes less time to reach from point A to point B.

**[More Travel Time]** The probability of traffic intensity being ‘high’ is contributed more to the travel time. Thus, the travel time becomes higher with the probability of 5.81%.

**[Less Travel Time]** With the probability of 95% that the traffic intensity being ‘very low’ is due to less travel time. In other words, the travel intensity is relatively low meaning the flow of traffic is smooth and takes less time to reach from point A to point B.

**[More Travel Time]** The probability of traffic intensity being ‘high’ is contributed more to the travel time. Thus, the travel time becomes higher with the probability of 5.81%.

From Figure 42 & Figure 43, the lowest travel time was recorded during the month of Mar-2020. The CPT numbers can be believed that the values showing are less due to the current Covid-19 pandemic. Due to the travel restrictions, fewer vehicles are observed on the road.

6.3.5. Identifying the major causes of increased ‘Traffic Intensity’

Amsterdam & Rotterdam - 2018

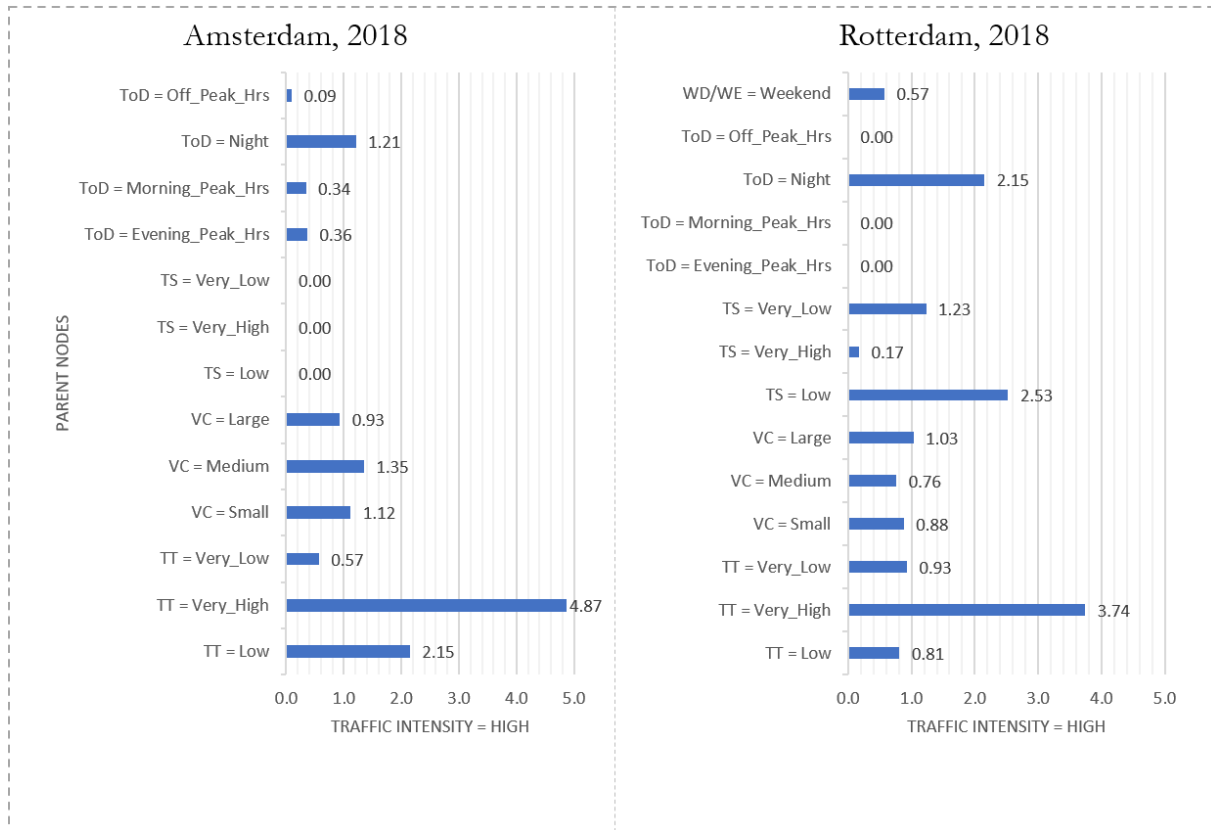


Figure 44: OR between Contributing nodes and Traffic Intensity = “High” | AMS & RTM - 2018

From the definition of the odds ratio, explained in chapter 5.6 & equation ..... (5, the input values for measuring odds ratios are the probability values obtained from the BN model. The results of odds ratios for Amsterdam Figure 44(left) and Rotterdam Figure 44(right) are shown.

For Amsterdam, the highest correlation can be found between ‘travel time’ and ‘traffic intensity’; which is 4.87, which indicates that the chances of both events occurring together are high. The next highest correlation can be found in the event, vehicle categories (VC = Medium & VC = Small); it is 1.35 and 1.12, respectively. The odds ratio of other variables such as Speed = Low, Speed = High and Time of the day = ‘off-peak hours’ are lesser than 1 stating that these events are less likely to takes place together with traffic intensity event.

For Rotterdam, the highest correlation can be found between ‘travel time’ and ‘traffic intensity’, which is 3.74. This value indicates that the chance of both events occurring together is likely to happen. The next highest correlation is traffic speed (TS) = 2.53, ToD = 2.15. Similarly, other events such as vehicle categories (VC = medium; VC = high; TT = very low) are less than one stating that these events are less likely to takes place together when the traffic intensity is ‘high’.

Amsterdam & Rotterdam - 2019

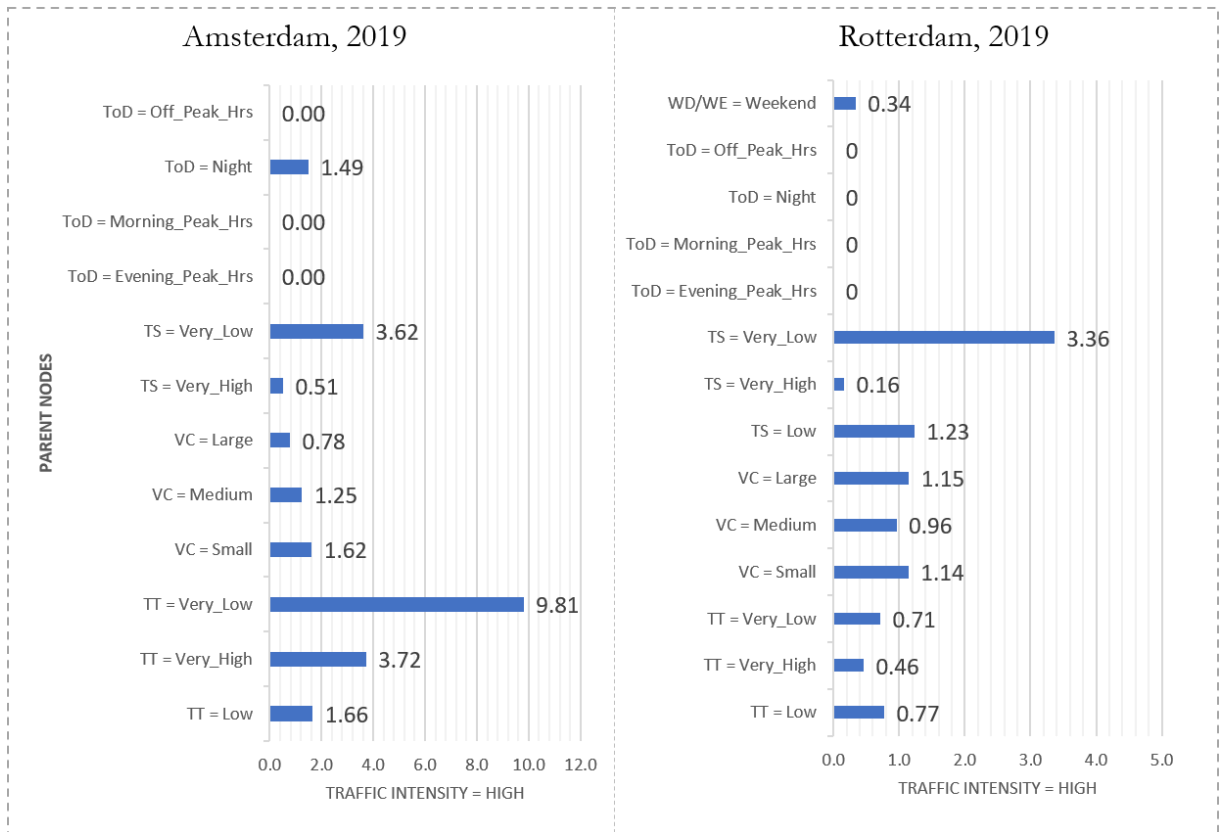


Figure 45: OR between Contributing nodes and Traffic Intensity = “High” | AMS & RTM - 2019

From the definition of the odds ratio, explained in chapter 5.6 & equation ..... (5, the input values for measuring odds ratios are the probability values obtained from the BN model. The results of odds ratios for Amsterdam Figure 45(left) and Rotterdam Figure 45(right) are shown.

For Amsterdam, the highest correlation can be found between ‘travel time’ and ‘traffic intensity’; which is 9.81, which indicates that the chances of both events occurring together are very likely. The next highest correlation can be found in the event, vehicle categories (VC = Medium & VC = Small); it is 1.25 and 1.62, respectively. The odds ratio of other variables such as speed =VH, speed = high and time of the day = ‘off-peak hours’ are lesser than 1 stating that these events are less likely to takes place together with traffic intensity event.

For Rotterdam, the highest correlation can be found between ‘traffic speed’ and ‘traffic intensity’, which is 3.36. This value indicates that the chance of both events occurring together is likely to happen. The next highest correlation is traffic speed (TS = low) = 1.23. Similarly, other events such as vehicle categories (VC = medium; VC = high; TT = very low) are less than one stating that these events are less likely to takes place together when the traffic intensity is ‘high’.

Amsterdam & Rotterdam – 2020

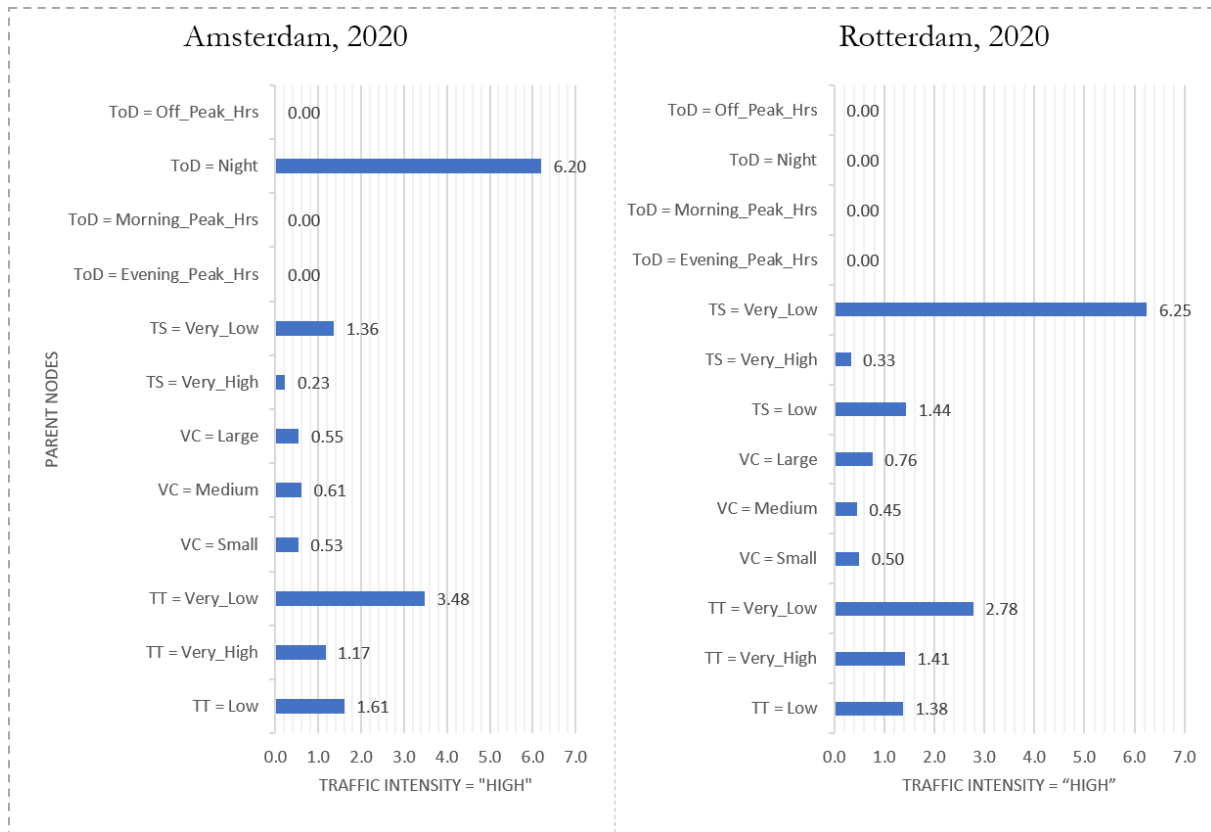


Figure 46: OR between Contributing nodes and Traffic Intensity = “High” | AMS & RTM - 2020

From the definition of the odds ratio, explained in chapter 5.6 & equation ..... (5, the input values for measuring odds ratios are the probability values obtained from the BN model. The results of odds ratios for Amsterdam Figure 46(left) and Rotterdam Figure 46(right) are shown.

For Amsterdam, like 2018 and 2019, the highest correlation can be found between ‘travel time’ and ‘traffic intensity’; which is ToD, which indicates that the chances of both events occurring together are likely to occur. The next highest correlation can be found in the event ‘traffic speed’, which is 1.36. The odds ratio of other variables such as traffic speed =VH, Speed = High and Time of the day = ‘off-peak hours’ are lesser than 1 stating that these events are less likely to takes place together with traffic intensity event.

For Rotterdam, the highest correlation can be found between ‘travel speed and ‘traffic intensity’, which are 6.25 & 2.78. This value indicates that the chance of both events occurring together is likely to happen. The next highest correlation is traffic speed (TS = low) = 1.44. Similarly, other events such as vehicle categories (VC = medium; VC = high; TT = very low) are less than one stating that these events are less likely to takes place together when the traffic intensity is ‘high’.

**6.3.6. Traffic flow prediction**

So far, the association between contributing nodes and the target node were discussed mainly focusing on the individual contributing node and its impact on the target node in the previous sections. In this subsection will consider continuous variables and input the continuous variables into the Bayesian model and analyse the predicted results from the observed results. At last, validate the BN model accuracy based on RMSE, MAPE and MAE for the target node ‘traffic intensity’.

The continuous variables of the input data were classified into training samples, and testing samples with the ratio of 70% is to 30%. The below **Error! Reference source not found.** shows the traffic intensity forecast for the sample of 300 data points for Amsterdam, 2018. The MAE, MAPE and RMSE between observed and predicted data are shown in Table 6. While looking at the RMSE values, it is very evident that the RMSE is much lower for the 30-minutes than compared to a 60-minutes prediction for overall data.

**Amsterdam, 2018**

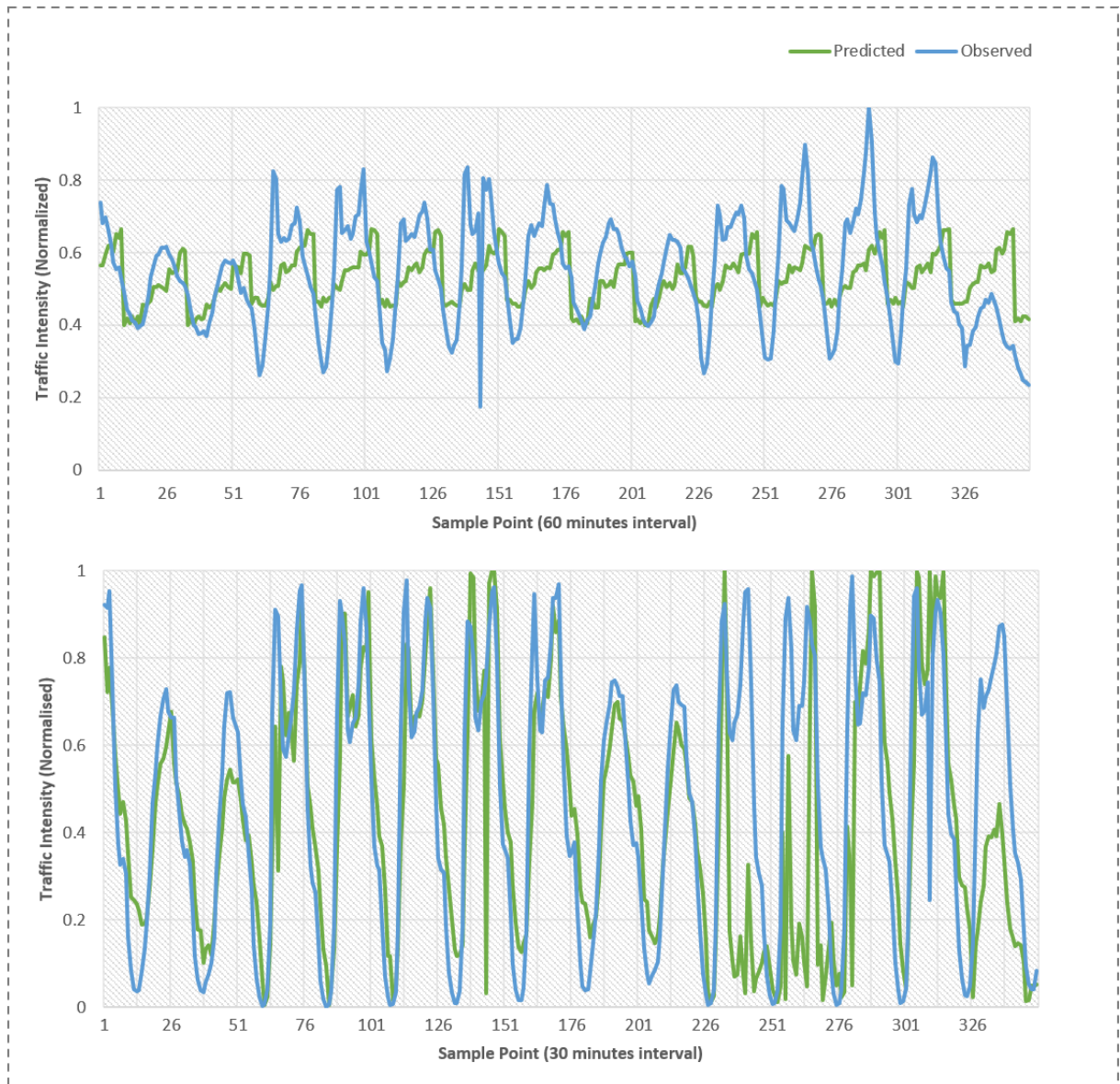


Figure 47: Traffic intensity forecasting (30 minutes interval) | AMS - 2018

Rotterdam, 2018

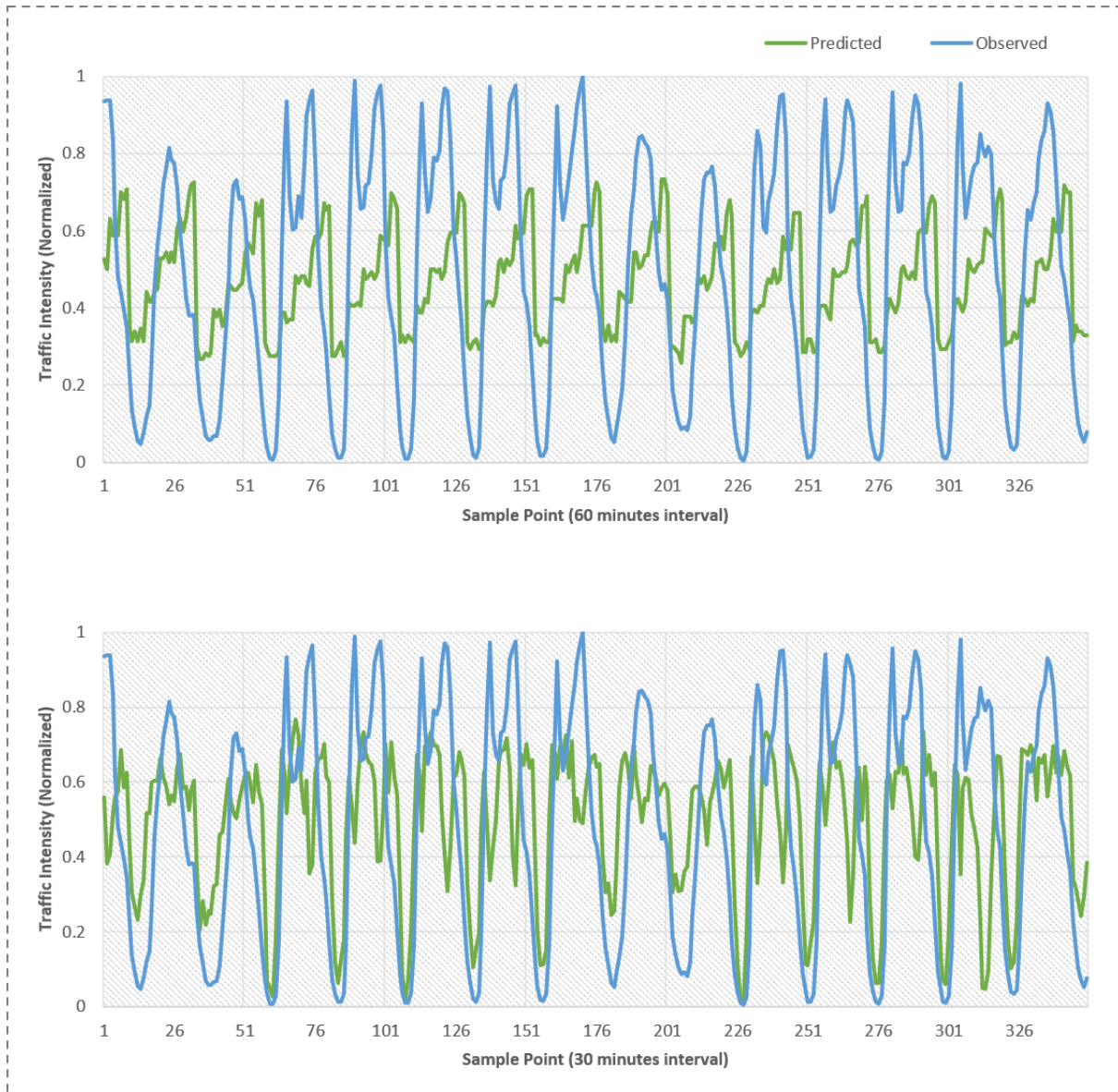


Figure 48: Traffic intensity forecasting (60 & 30 minutes interval) | RTM - 2018

Figure 48 depicts the traffic intensity forecast for the sample of 300 data points for Rotterdam, 2018. The MAE, MAPE and RMSE between observed and predicted data are shown in Table 6. While looking at the RMSE values, it is very evident that the RMSE is much lower for the 30-minutes than compared to a 60-minutes prediction for overall data.



## Amsterdam, 2020

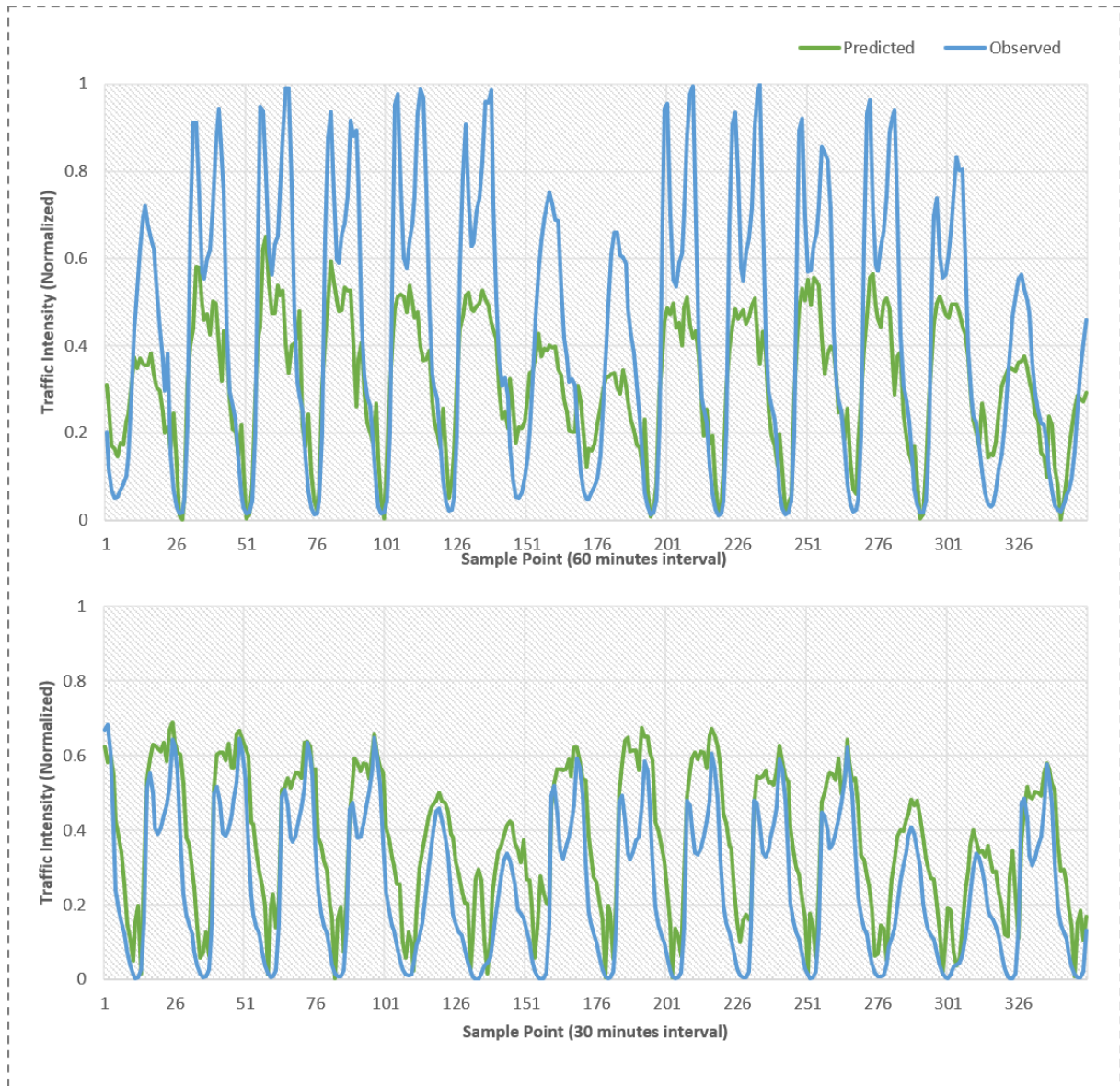


Figure 49: Traffic intensity forecasting (60 &amp; 30-minutes interval) | AMS - 2020

Figure 49 depicts the traffic intensity forecast for the sample of 300 data points for Amsterdam, 2020. The MAE, MAPE and RMSE between observed and predicted data are shown in Table 6. While looking at the RMSE values, it is very evident that the RMSE is much lower for the 30-minutes than compared to a 60-minutes prediction for overall data.

Rotterdam, 2020

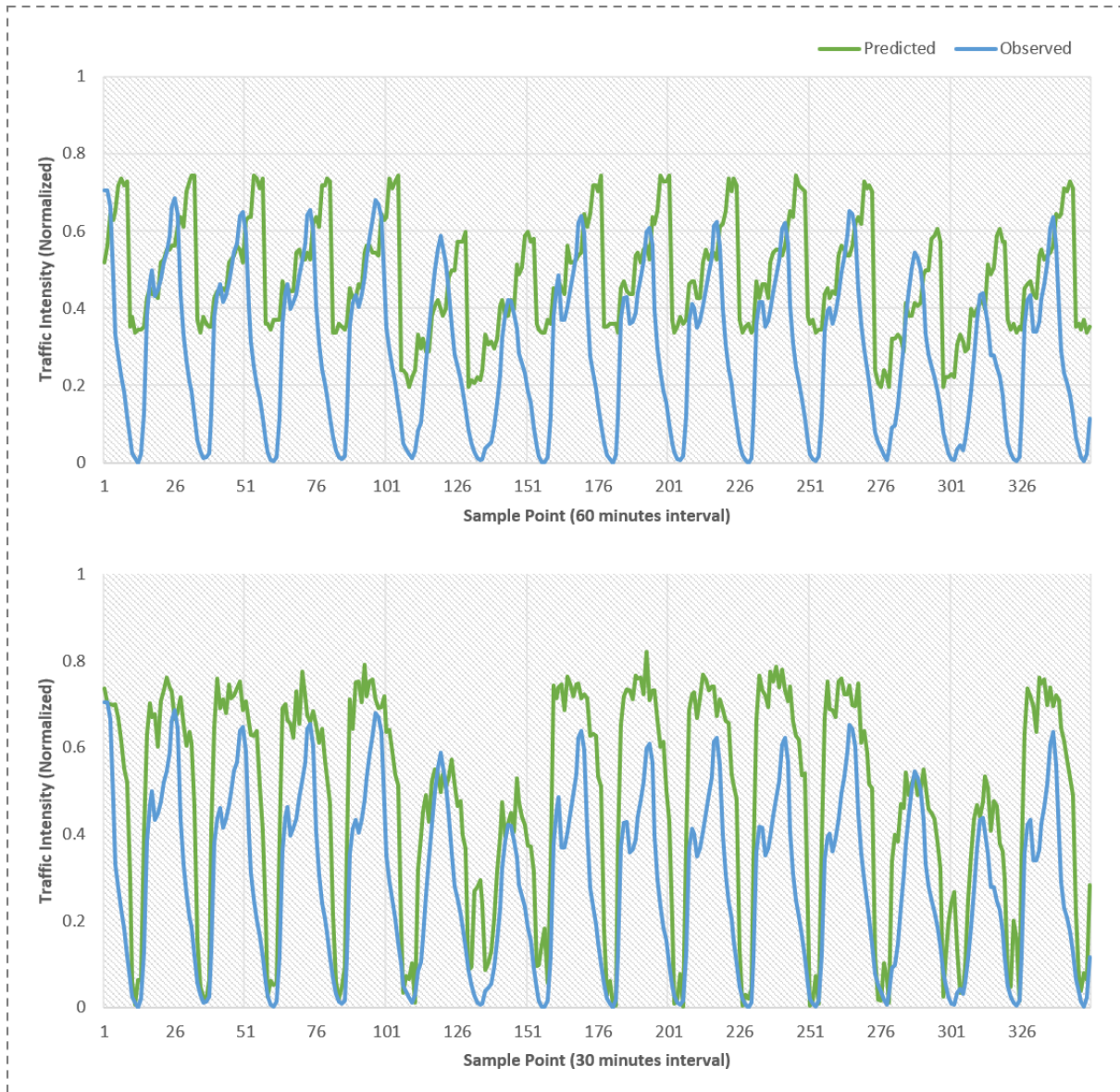


Figure 50: Traffic intensity forecasting (60 & 30 minutes interval) | RTM - 2020

Figure 49 depicts the traffic intensity forecast for the sample of 300 data points for Rotterdam, 2020. The MAE, MAPE and RMSE between observed and predicted data are shown in Table 6. While looking at the RMSE values, it is very evident that the RMSE is much lower for the 30-minutes than compared to a 60-minutes prediction for overall data.

#### 6.4. Accuracy Assessment of the BN model through RMSE, MAE and MAPE

The MAE, MAPE and RMSE values between observed and actual data are shown in Table 6 & Table 7. Given the model, the predictions for 30 minutes gives better performance results compared to 60 minutes of predictions.

Table 6: Comparison of RMSE, MAE & MAPE for Amsterdam & Rotterdam - 2018

Model	Measurement	30 Minutes		60 Minutes	
		Amsterdam	Rotterdam	Amsterdam	Rotterdam
HC	RMSE	0.784	0.116	0.920	0.346
	MAE	0.586	0.866	0.609	0.947
	MAPE	58.19	23.44	349.79	41.85

Table 7: Comparison of RMSE, MAE & MAPE for Amsterdam & Rotterdam - 2020

Model	Measurement	30 Minutes		60 Minutes	
		Amsterdam	Rotterdam	Amsterdam	Rotterdam
HC	RMSE	0.779	0.090	0.901	0.137
	MAE	0.543	0.739	0.670	0.992
	MAPE	39.04	98.62	58.72	206.89

#### 6.5. Road Network characteristics of Amsterdam & Rotterdam

To understand the characteristics of any road segments, one has to study road network formation and its connecting junctions. From the graph theory point of view, any road network is nothing but a connection of nodes and vertices. According to the author Boeing, (2017) the study of any road segments includes various approaches, the formation of road network based on geographical region, transportation demand, topological structure, and historical impact on the current road network. Street network analysis is an integral part of network science since its beginning. The foundation of any street network analysis is on graph theory since its origin from the 18<sup>th</sup> century. The concept of graph theory became popular through the 'seven bridges of Königsberg problem'. Graph theory is an illustrative representation of a combination of elements (Vertices/Nodes) & connections between them (links/edges) (Powell & Hopkins, 2015).

Any spatial network can be a planar or non-planar which can be shown in two dimensions. In real-time instance, most of the street networks are non-planar. For example, the road network consists of overpass, bridges and tunnels. To ease the mathematical computation, street networks are transformed and represented in a planar form (Buhl et al., 2006; Barthélemy & Flammini, 2008). However, this extreme situation is uncommon in most of the places, and thus one can still estimate the network as planar.

### 6.5.1. Analysing street network

As per the definition of Newman, (2010) any street network is considered as primal, planar/non-planar with cyclic loops, and these characteristics of roads can be done through visual inspection. Later can be quantified by metric and topological measures. One common measure performed in any road networks is measuring the area and length of roads (Cervero & Kockelman, 1997). By measuring average street length (measured in mts), depicts the texture of network. On the other hand, the topological measures tell the configuration, efficiency, and characteristics of the network. The indicators include average node degree and average street per node.

### 6.5.2. Open Street Map

A successful collaborative project which aims at providing a free and editable map of the world for the public use. Recently OSM has come out as a large player by providing mapping support and spatial data (Jokar Arsanjani et al., 2015; Corcoran, Mooney, & Bertolotto, 2013)

**OSMnx Plugin:** An open-source tool built on top of OSM. With these tools, one can download spatial information such as building footprints, spatial boundary, and street networks. This tool helps users to analyse and visualize complex street networks. Several functionalities help the network to build based on a walking path or driving path. Other functionalities include node elevation calculations and street grades.

This study makes use of a street network analysis concept coupled with OSMnx plugin, which provides network statistics of Amsterdam and Rotterdam. The network statistics include various measures such as average node degree, average street per node, total edge length, node density etc.



Figure 51: Monochrome images of Amsterdam (left) and Rotterdam (right) street network (One square mile)

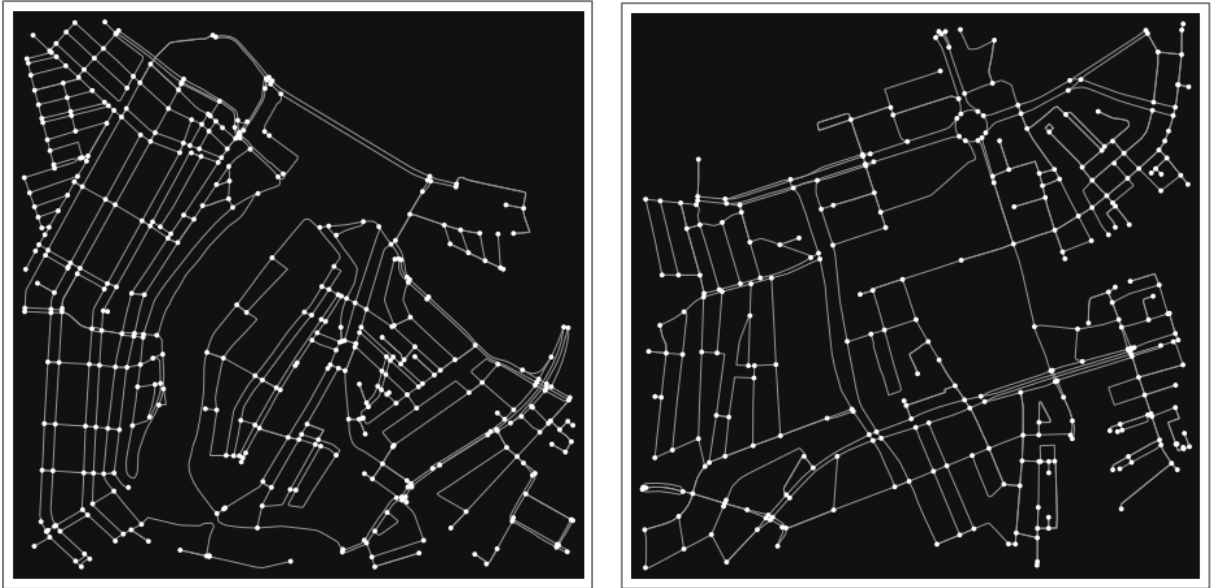


Figure 52: Monochrome images of Amsterdam (left) and Rotterdam(right) street network (Node highlighted)

Table 8: Descriptive statistics of 2 street networks in the Netherlands

Indicators	Amsterdam	Rotterdam
Area covered	193.5 sq km	161.6 sq km
Number of nodes in a network (n)	11520	11245
Number of edges in a network (m)	26580	25387
Average node degree	4.6	4.5
Intersection count	9995	9682
Intersection density (per km <sup>2</sup> )	51.64	59.88
Average streets per node	2.93	2.91
Average street length(m)	102.58	97.7

The various histories and design pattern of the network tell us the historical events, planning authorities, livelihood, design patterns and topographies(Guzowski, 1990). Figure 51Figure 52 depict the chunk of the street network of Amsterdam and Rotterdam. The quantitative measures from Table 8 show the differences between the two cities. Amsterdam has 51.64 intersection density /km<sup>2</sup>, and Rotterdam has 59.88. Average street length segment for Amsterdam is 102 mts, and for Rotterdam, it is 98 mts. From the visual inspection, we can observe the Amsterdam as semi-circular fashion with the clear symmetrical structure of streets. In Rotterdam, the street networks are square typed, and street ends are connected to the main roads. The descriptive statistics adds more weightage to visual inspection.

## 7. CONCLUSION AND RECOMMENDATIONS

### 7.1. Conclusion

This section deals with the advantages, drawbacks of the method implemented and the evaluation of the results which were acquired and presented in the previous section.

The study has combined discrete and continuous data and provided a framework using Bayesian network for estimating the traffic congestion by using a data-driven approach. The study coupled the historical data which are obtained from traffic sensors installed across the Amsterdam and Rotterdam area. The traffic-related data is further enriched by integrating other forms of data such as ‘vehicle categories’ obtained from NDW. In the current scenario, NDW is the single reliable source which as a potential to provide a good quantity of data because of its open-source policy promoting public researchers to access the data. So, this study mainly focused on a single source which enables enriching the data obtained from the traffic sensors. The data includes the information regarding vehicle speed, density of vehicles, road width, vehicle length etc. Among this information provided by inductive loop, speed of vehicle and density of the vehicles have been selected for building the congestion model. The data transformation was performed using the normalization technique. In order to reduce the data redundancy, the missing data were coded as ‘N/A’ and was eventually removed. Normalization technique helped in categorising the data into four different classes. Different learning algorithm such as Hill climbing and IAMB were compared and based on the performance; the HC learning algorithm was used for creating a Bayesian network. The accuracy achieved by the HC algorithm was better with the expected loss of 6.8 compare to IAMB with an expected loss of 7.2. Later traffic congestion levels were defined as low, medium and high. Using the Bayesian model, In the first half of the analysis, the diagnostic reasoning was performed by identifying the leading causes of traffic congestion. Using posterior probability distribution, the association each of nodes were derived. Later by using continuous data with the same BN network structure, prediction of the traffic congestion was performed with the interval of 30minutes and 60minutes. At last, the evaluation of the performance of the predictions was done by measuring RMSE, MAE and MAPE.

The prior and posterior distribution of each variable are visualised in a tabular form using bar charts. The major audience is traffic monitoring professional and general audience. This study will help one to understand what are the causing factors which are leading to traffic congestion. As mentioned in the related work section, previous studies focused more on evaluating the relative effect of various reasons for traffic congestion which includes other data sources such as weather data, traffic incident data and road work-related data coupled with statistical analysis such as regression modelling. The traditional approach has a major drawback which restricts in capturing complex dependencies between random variables and uncertainties with external data sources in the urban networks.

The BN model introduced in the current study helps in modelling the probabilistic dependencies between causing nodes, and the target node is given different scenarios variables. The joint probability distribution was derived across different variables which are represented as a factor influencing the traffic congestion.

The proposed BN model which are configured can be used in two ways: -

(a) Quantifying the contribution of each node or the combination of multiple nodes causing an effect on traffic congestion, which helps in investigating the leading causes for the purpose of congestion diagnosis.

(b) Predict future traffic flow based on historical data. The aim of the study is to build a BN model using historical data and perform traffic analysis using the Bayesian network.

## 7.2. Research Questions: Answered

The research questions mentioned in 1.3 are answered in this section briefly.

1. What are the characteristics of road segments of Amsterdam and Rotterdam road networks?

Answer: After reviewing various literature related to the road network, it has been observed that the spatial pattern of streets, street orientation and edges connecting to the junctions, all these characteristics collectively form the travel behaviour of vehicles on the road. As discussed in 6.5.2, the road network can be assessed in two different forms. Firstly, by visually looking at the formation of roads and its corresponding connection. But however, with a visual inspection, one can only observe the texture of the road network, secondly, by making use of network statistics. The combination of quantitative analysis and qualitative inspection adds more significance to assess the network structure.

2. What are the techniques to identify and examine the probabilistic dependency of nodes in the road networks, and how are the short-term predictions made?

Answer: As mentioned in 5.4 & 6.1, in the study two learning algorithms named “Hill Climbing” and “IAMB” were introduced and compared between them, it was observed from the structure that HC algorithm performance better compare to ‘IAMB’. To validate the same, various score functions such as BIC, AIC, BDE and K2 were measured to assess the goodness of fit. The same network structure was applied to the continuous variables to perform short-term predictions. The datasets of 30 minutes and 60 minutes were downloaded separately and compared between both. The results depict that 30 minutes predictions are more reliable compare to 60 minutes.

3. How accurate and reliable are the short-term traffic predictions made using Bayesian Networks?

Answer: Accuracy and reliability are measured through RMSE MAE and MAPE values. As discussed in 6.4, the model performs better with 30 minutes interval data compared to 60 minutes. By taking an instance of Amsterdam, the RMSE values for 30-minutes is 673.2; 60-minutes is 586.1.

### 7.3. Recommendations

Road traffic analysis and prediction is a vast subject and requires subject knowledge of what causes traffic congestions. Also, it is one of the major issues in countries like The Netherlands. The future of the scope of this study includes:

- The study made use of 4-5 parameters of the data from the sensor data that are Speed, Intensity, Travel Time and Vehicle characteristics. To make the BN model more robust, which yields good results, other variables such as specific lane, driving lane, the direction of the road and bicycle data can be included.
- The BN model completely depended on single-source data. To enhance the model, other sources like rainfall and temperature date etc. can be incorporated. Data from various sources will have an impact on the prediction model.
- The current BN model is a static & non-spatial one. In future, the entire can be made dynamic by adding time data and spatial components.



---

## LIST OF REFERENCES

---

- Ahmed, M. M., & Abdel-Aty, M. A. (2012). The viability of using automatic vehicle identification data for real-time crash prediction. *IEEE Transactions on Intelligent Transportation Systems*, 13(2), 459–468. <https://doi.org/10.1109/ITITS.2011.2171052>
- Ahmed, M. S. (n.d.). *Analysis of Freeway Traffic Time-Series Data by Using Box-Jenkins Techniques*.
- Amsterdam (Municipality, Noord-Holland, Netherlands) - Population Statistics, Charts, Map and Location. (n.d.). Retrieved October 14, 2019, from <http://www.citypopulation.de/php/netherlands-admin.php%3Fadm2id%3D0363>
- Arsanjani, J. J., Zipf, A., Mooney, P., & Helbich, M. (n.d.). *Lecture Notes in Geoinformation and Cartography*. Retrieved from <http://www.springer.com/series/7418>
- Barthélemy, M., & Flammini, A. (2008). *Modeling Urban Street Patterns*. <https://doi.org/10.1103/PhysRevLett.100.138702>
- Belgium & Netherlands have worst traffic in Europe; Rotterdam scores badly | NL Times. (n.d.). Retrieved June 23, 2020, from <https://nltimes.nl/2015/08/25/belgium-netherlands-worst-traffic-europe-rotterdam-scores-badly>
- bnlearn - Bayesian network structure learning. (n.d.). Retrieved June 1, 2020, from <https://www.bnlearn.com/>
- Boeing, G. (2017). OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. *Computers, Environment and Urban Systems*, 65, 126–139. <https://doi.org/10.1016/j.compenvurbsys.2017.05.004>
- Buhl, J., Gautrais, J., Reeves, N., Solé, R. v, Valverde, S., Kuntz, P., & Theraulaz, G. (2006). Topological patterns in street networks of self-organized urban settlements. *Eur. Phys. J. B*, 49(4), 513–522. <https://doi.org/10.1140/epjb/e2006-00085-1>
- Calvert, S. C., Rypkema, J., Holleman, B., Azulay, D., & de Jong, A. (2017). Visualisation of uncertainty in probabilistic traffic models for policy and operations. *Transportation*, 44(4), 701–729. <https://doi.org/10.1007/s11116-015-9673-3>
- Canaud, M., Mihaylova, L., Sau, J., & el Faouzi, N. E. (2013). Probability hypothesis density filtering for real-time traffic state estimation and prediction. *Networks and Heterogeneous Media*, 8(3), 825–842. <https://doi.org/10.3934/nhm.2013.8.825>
- Carvalho, A. M. (n.d.). *Scoring functions for learning Bayesian networks*.
- Cervero, R., & Kockelman, K. (1997). Travel demand and the 3Ds: Density, diversity, and design. *Transportation Research Part D: Transport and Environment*, 2(3), 199–219. [https://doi.org/10.1016/S1361-9209\(97\)00009-6](https://doi.org/10.1016/S1361-9209(97)00009-6)
- Corcoran, P., Mooney, P., & Bertolotto, M. (2013). Analysing the growth of OpenStreetMap networks. *Spatial Statistics*, 3, 21–32. <https://doi.org/10.1016/j.spasta.2013.01.002>
- DATEX II Dutch Profile 2015-2a. (n.d.).
- Davarynejad, M., Wang, Y., Vrancken, J., & van den Berg, J. (2011). Multi-phase time series models for motorway flow forecasting. *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*, 2033–2038. <https://doi.org/10.1109/ITSC.2011.6082839>
- Dougherty, M. S., & Cobbett, M. R. (1997). Short-term inter-urban traffic forecasts using neural networks. *International Journal of Forecasting*, 13(1), 21–31. [https://doi.org/10.1016/S0169-2070\(96\)00697-8](https://doi.org/10.1016/S0169-2070(96)00697-8)
- Huisken Giovanni. (n.d.). *Inter-Urban Short-TermTraffic Congestion Prediction*.
- Introduction to Hill Climbing | Artificial Intelligence. (n.d.). Retrieved June 4, 2020, from <https://medium.com/@bhavek.mahyavanshi50/introduction-to-hill-climbing-artificial-intelligence-a3714ed2d8d8>

- Jia, Y., Wu, J., & Xu, M. (2017). *Traffic Flow Prediction with Rainfall Impact Using a Deep Learning Method*.  
<https://doi.org/10.1155/2017/6575947>
- Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., & Saul, L. K. (1999). Introduction to variational methods for graphical models. *Machine Learning*, 37(2), 183–233. <https://doi.org/10.1023/A:1007665907178>
- Jordan, M., Kleinberg, J., & Schölkopf, B. (n.d.). *Pattern Recognition and Machine Learning*.
- Kim, J., & Wang, G. (2016). Diagnosis and Prediction of Traffic Congestion on Urban Road Networks Using Bayesian Networks. *Transportation Research Record: Journal of the Transportation Research Board*, 2595(1), 108–118. <https://doi.org/10.3141/2595-12>
- Life, M. N.-A., & 2012, undefined. (n.d.). *Networks: An Introduction*. 2010: Oxford University Press.
- Li, H., Hao, W., Gan, W., & Chen, G. (2013). Survey of probabilistic graphical models. *Proceedings - 2013 10th Web Information System and Application Conference, WISA 2013*, 275–280.  
<https://doi.org/10.1109/WISA.2013.59>
- Maroto, J., Delso, E., Félez, J., & Cabanellas, J. M. (2006). Real-time traffic simulation with a microscopic model. *IEEE Transactions on Intelligent Transportation Systems*, 7(4), 513–526.  
<https://doi.org/10.1109/TITS.2006.883937>
- Melnikov, V. R., Krzhizhanovskaya, V. v., Lees, M. H., & Boukhanovsky, A. v. (2016). Data-driven travel demand modelling and agent-based traffic simulation in Amsterdam urban area. *Procedia Computer Science*, 80, 2030–2041. <https://doi.org/10.1016/j.procs.2016.05.523>
- Papapesios, N., Ellul, C., Shakir, A., & Hart, G. (2019). Exploring the use of crowdsourced geographic information in defence: challenges and opportunities. *Journal of Geographical Systems*, 21(1), 133–160.  
<https://doi.org/10.1007/s10109-018-0282-5>
- Patro, S. G. K., & sahu, K. K. (2015). Normalization: A Preprocessing Stage. *LARJSET*, 20–22.  
<https://doi.org/10.17148/iarjset.2015.2305>
- Persoskie, A., & Ferrer, R. A. (2017). A Most Odd Ratio:: Interpreting and Describing Odds Ratios. *American Journal of Preventive Medicine*, 52(2), 224–228. <https://doi.org/10.1016/j.amepre.2016.07.030>
- Policy: Traffic safety - City of Amsterdam. (n.d.). Retrieved June 11, 2020, from  
<https://www.amsterdam.nl/en/policy/policy-traffic/policy-traffic/>
- Portland's Olmsted vision (1897-1915): A study of the public landscapes designed by Emanuel T. Mische in Portland, Oregon. (n.d.). Retrieved July 15, 2020, from  
<https://www.elibrary.ru/item.asp?id=6759397>
- Powell, J., & Hopkins, M. (2015). Graphs in theory. In *A Librarian's Guide to Graphs, Data and the Semantic Web* (pp. 1–6). <https://doi.org/10.1016/b978-1-84334-753-8.00001-4>
- Project Jupyter | Home. (n.d.). Retrieved July 23, 2020, from <https://jupyter.org/>
- Rietveld, P. (n.d.). *Urban Transport Policies*. Retrieved from <http://www.tinbergen.nl>.
- Rotterdam Population 2019 (Demographics, Maps, Graphs). (n.d.). Retrieved October 14, 2019, from  
<http://worldpopulationreview.com/world-cities/rotterdam-population/>
- Schrank, D., Eisele, B., & Lomax, T. (2019). *2019 Urban Mobility Report*.
- Sinharay, S. (n.d.). *Model Diagnostics for Bayesian Networks*.
- Sun, J., & Sun, J. (2015). A dynamic Bayesian network model for real-time crash prediction using traffic speed conditions data. *Transportation Research Part C: Emerging Technologies*, 54, 176–186.  
<https://doi.org/10.1016/j.trc.2015.03.006>
- Sun, S., Zhang, C., & Yu, G. (2006). A Bayesian network approach to traffic flow forecasting. *IEEE Transactions on Intelligent Transportation Systems*, 7(1), 124–133.  
<https://doi.org/10.1109/TITS.2006.869623>
- Taale, H., & Wilmink, I. (n.d.). *Traffic in the Netherlands 2016*.

- Tsamardinos, I., Aliferis, C. F., & Statnikov, A. (2003). *Algorithms for Large Scale Markov Blanket Discovery*. Retrieved from [www.aaai.org](http://www.aaai.org)
- Vlahogianni, E. I., Karlaftis, M. G., & Golias, J. C. (2014). Short-term traffic forecasting: Where we are and where we're going. *Transportation Research Part C: Emerging Technologies*, 43, 3–19. <https://doi.org/10.1016/j.trc.2014.01.005>
- Wegman, F. (2007). *Road traffic in the Netherlands: Relatively safe but not safe enough Sustainable Safety View project Road traffic in the Netherlands: Relatively safe but not safe enough!* Retrieved from <https://www.researchgate.net/publication/255578028>
- Whittaker, J., Garside, S., & Lindveld, K. (1997). Tracking and predicting a network traffic process. *International Journal of Forecasting*, 13(1), 51–61. [https://doi.org/10.1016/S0169-2070\(96\)00700-5](https://doi.org/10.1016/S0169-2070(96)00700-5)
- Williams, B. M., & Hoel, L. A. (2003). Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: Theoretical basis and empirical results. *Journal of Transportation Engineering*, 129(6), 664–672. [https://doi.org/10.1061/\(ASCE\)0733-947X\(2003\)129:6\(664\)](https://doi.org/10.1061/(ASCE)0733-947X(2003)129:6(664))
- Yu, Y. J., & Cho, M. G. (2008). A short-term prediction model for forecasting traffic information using Bayesian network. *Proceedings - 3rd International Conference on Convergence and Hybrid Information Technology, ICCIT 2008*, 1, 242–247. <https://doi.org/10.1109/ICCIT.2008.355>
- Zhang, Y., Zhang, Z., Liu, K., & Qian, G. (2010). An improved IAMB algorithm for Markov blanket discovery. *Journal of Computers*, 5(11), 1755–1761. <https://doi.org/10.4304/jcp.5.11.1755-1761>
- Zheng, L., Ismail, K., & Meng, X. (2014). Traffic conflict techniques for road safety analysis: Open questions and some insights. *Canadian Journal of Civil Engineering*, 41(7), 633–641. <https://doi.org/10.1139/cjce-2013-0558>

## APPENDIX - 1

Table 9: Table showing data coding for the node 'Days of the week' implemented in the BN model

Days of the week	Numeric coding	Character coding
Sunday	1	A
Monday	2	B
Tuesday	3	C
Wednesday	4	D
Thursday	5	E
Friday	6	F
Saturday	7	G

Table 10: Table showing data coding for the node 'Weekend/Weekday' implemented in the BN model

Weekend/Weekday	Numeric coding	Character coding
Weekday	1	A
Weekend	2	B

Table 11: Table showing data coding for the node 'Time of the day' implemented in the BN model

Time of the day	Numeric coding	Character coding
Early morning (12 am – 5 am)	1	A
Morning peak hours (5am – 10am)	2	B
Off peak hours (10am – 4pm)	3	C
Evening peak hours (4 pm – 8 pm)	4	D
Night (8pm – 12 am)	5	E

Table 12: Range defined for the node 'Speed', 'Intensity' &amp; 'Travel time.'

Range	States
0.00 -2.50	Very low
0.25 – 0.50	Low
0.50 – 0.75	High
0.75 – 1.00	Very high

Table 13: Vehicle categories

<b>Vehicle length (mts)</b>	<b>Categories</b>	<b>Character coding</b>	<b>Numeric coding</b>
<b>Greater than 2.40 &amp; less than or equal to 5.60</b>	Light vehicles	A	1
<b>Greater than 5.60 and less than or equal to 12.20</b>	Medium vehicles	B	2
<b>Greater than 12.20</b>	Heavy vehicles	C	3