URBAN TRAFFIC STATE ESTIMATION & PREDICTION

Master Thesis

Luuk de Vries
M-CEM
S1003712
Master Thesis
21th of October 2016

UNIVERSITY OF TWENTE.

Title: Urban traffic state estimation & prediction

Author: Luuk de Vries
University of Twente
Civil Engineering & Management
Traffic Engineering & Management
S1003712
l.o.devries@student.utwente.nl

Date: 21th of October 2016

Contacts: **Daily Supervisor**
Luc Wismans
University of Twente
Tel: 0570 - 666 840
Email: l.j.j.wismans@utwente.nl

UT Supervisor
Eric van Berkum
University of Twente
Tel: 053 - 489 1098
Email: e.c.vanberkum@utwente.nl

1 Table of Contents

2 Abstract (Dutch)

In de laatste jaren wordt informatietechnologie (IT) in het verkeer steeds vaker toegepast en onderzocht. Het gebruik van IT in zowel voertuigen als infrastructuur, valt onder de noemer van intelligente transportsystemen (ITS). Voordelen van het toepassen van ITS zijn divers en kunnen variëren van; een verhoogde verkeersveiligheid, verbeterde prestaties van het verkeersnetwerk, verbeterde mobiliteit, milieuvoordelen tot economische groei (Ezell, 2010). Typisch gebruik van ITS leunt op een belangrijk ingrediënt; een robuust, volledig en accuraat beeld van de verkeerstoestand in het netwerk. Dit beeld wordt gerealiseerd in een proces welke wordt omschreven als *traffic state estimation*. Het verlengde van dit proces, de *traffic state prediction*, voegt daar een extra stap aan toe door de verkeerstoestand in het netwerk te voorspellen voor de (nabije) toekomst.

In dit onderzoek met als onderwerp *urban traffic state estimation-* en *traffic state prediction*, worden twee algemene uitdagingen geïdentificeerd. Als eerste is er een mismatch tussen de hoeveelheid beschikbare meetgegevens en de te schatten en voorspellen verkeersstroomvariabelen. Een verkeersnetwerk wordt immers nooit volledig door verkeersinformatiebronnen gedekt, waardoor technieken nodig zijn om de data die wel beschikbaar is, zo goed mogelijk te extrapoleren en te gebruiken. De tweede uitdaging ontstaat door in dit onderzoek te focussen op een stedelijke omgeving. Een stedelijke omgeving verhoogt de moeilijkheidsgraad, als gevolg van bijvoorbeeld lagere verkeersvolumes, lagere maximumsnelheden, meer variatie in snelheden, een hogere dichtheid van kruispunten, verkeerslichten, rotondes, prioriteiten en dynamische interacties tussen vervoerswijzen, in tegenstelling tot een netwerk dat enkel uit snelwegen bestaat.

Van Hinsbergen et al. (2007) beschrijft dat er reeds een grote hoeveelheid methodes bestaan voor *traffic state estimation* en *traffic state prediction*. Er zijn naïeve modellen, parametrische modellen, niet-parametrische modellen en een hybride mengsel van twee of meer van deze categorieën. Van Lint (2011) benadrukt dat het vinden van een balans tussen geavanceerde en complexe modellen aan de ene kant en robuuste, snelle en algemeen toepasbare modellen aan de andere kant, het grootste probleem in de praktijk van *traffic state-estimation* en *–prediction* is.

De naïeve categorie vertegenwoordigt de modellen, waarbij verkeersgegevens alleen gebruikt worden om daaruit directe relaties te berekenen. Voorbeelden zijn *instantaneous travel time* of *historical averaging* modellen. Voordelen van deze modellen zijn de lage berekeningscomplexiteit en eenvoudige implementatie. Het nadeel is dat vanwege het weglaten van de verkeersstroomtheorie, de uitkomsten doorgaans onlogisch of onnauwkeurig zijn. De parametrische categorie vertegenwoordigt modellen waarin het principe van Lighthill-Whitham-Richards (Lighthill & Whitham, 1955) wordt toegepast. Klassieke voorbeelden zijn; *Newell's simplified kinematic wave model* (Newell, 1993) en *cell transmission models* (Daganzo, 1994). Het voordeel van deze modellen is dat de verkeersstroomtheorie wèl gebruikt wordt, maar met als nadeel dat veel kalibratie van parameters nodig is. Een ander nadeel is dat door de stedelijke omgeving van het netwerk in dit onderzoek, de fundamenten van de verkeersstroomtheorie niet noodzakelijk meer gelden, wat de betrouwbaarheid van deze modellen negatief beïnvloed. De niet-parametrische categorie omvat de modellen waarin relaties uit de verkeersgegevens worden beschouwd, maar geen parameters worden geschat. Voorbeelden van deze modellen zijn meestal gebaseerd op lineaire regressie. Deze modellen hebben gemeen dat ze een lage complexiteit hebben en daardoor gemakkelijk in real-time kunnen worden uitgevoerd. De nauwkeurigheid is volgens Van Lint (2011) echter meestal laag.

Theoretisch meer geschikt voor de stedelijke toepassing zijn hybride model types, welke elementen van niet-parametrische, parametrische en naïeve methoden combineren. Het meest bekende voorbeeld hiervan is de *Kalman Filtering* (Kalman & Bucy, 1961). Met als nadeel dat deze opgebouwd is met het in achting nemen van de verkeersstroomtheorie. Van alle andere hybride model types beschouwd in dit onderzoek, lijkt de hybride black-box benadering theoretisch, praktisch en intuïtief het meest geschikt voor gebruik in een stedelijke omgeving. Binnen deze subcategorie ontwikkelden Morita (2011), Esaway (2012) twee interessante kaders die patronen in historische verkeersgegevens van naburige wegsegmenten gebruiken als indicator voor de huidige en toekomstige verkeersituatie op de segmenten in een stedelijk netwerk.

Het doel van dit onderzoek is om een kader te ontwerpen voor *traffic state estimation* en *traffic state prediction,* in een stedelijke omgeving, welk gebruik maakt van zowel *floating car*- en tellus data om real-time en toekomstige uitspraken te doen over de snelheid, dichtheid en intensiteit op iedere link in het netwerk. Als uitgangspunt voor dit ontwerp worden de methoden van Morita (2011) en Esaway (2012) gebruikt. De nieuw ontworpen Neighbourhood Link Method (NLM) wordt vervolgens beoordeeld op de nauwkeurigheid van de schattingen en voorspellingen.

De gebruikte onderzoeksmethode bestaat uit drie afzonderlijke delen; 1) het simuleren van de z.g. *ground truth*, 2) de *traffic state estimation* en 3) de *traffic state prediction*. In het eerste deel van het onderzoek wordt een microsimulatie programma (PARAMICS) gebruikt om een 100% accuraat en 100% dekkende *ground truth* te genereren in het als case gekozen Sioux Falls verkeersnetwerk. Deze *ground truth* blijft verder ongewijzigd gedurende dit onderzoek. Het tweede deel van dit onderzoek beschrijft de totstandkoming van de NLM voor *traffic state estimation.* Dit raamwerk wordt vervolgens onderworpen aan een evaluatie met als doel om het ontwerp verder te verbeteren. Met de *ground truth* beschikbaar voor vergelijkingsdoeleinden, worden de schattingen beoordeeld op juistheid. In het derde deel wordt deze procedure herhaalt, om het NLM raamwerk uit te breiden voor *traffic state prediction*.

Het nieuw ontwikkelde NLM raamwerk, bestaat uit 7 stappen. Aanvankelijk worden gemeten verkeersgegevens ongewijzigd opgeslagen in een database. Vervolgens worden de links welke zich over de tijd hetzelfde gedragen aan elkaar gekoppeld (neighbours/buren) op basis van correlatie. De mest recente real-time beschikbare verkeersdata wordt vervolgens gebruikt om uitsluitend de verkeersdata van de buren, met behulp van lineaire regressie, om te zetten in een z.g. buurt-schatting van verkeersvariabelen voor een wegvak. Deze wordt vervolgens gefuseerd op basis van betrouwbaarheid, met de meest recente verkeersdata van de link zelf om de *traffic state estimation* te realiseren. Voor het voorspellen, wordt aan het raamwerk een tijdsdimensie gelijk aan de voorspelhorizon toegevoegd, om met de meest recente verkeersdata de toekomst te voorspellen.

De resultaten laten zien dat NLM bij 5% FCD; gemiddeld 60% van de stedelijke segmenten in het netwerk binnen 5km/uur (3 mi/h) van de waarheid kan schatten, gedurende de spits. Buiten de spits daalt dit percentage tot 50%. De dichtheid wordt voor 80% van de segmenten geschat binnen 7,5 vtg/km/rijstrook en voor de intensiteitsschattingen geldt dat voor 60% van de segmenten binnen 100 vtg/uur/rijstrook kunnen worden geschat. Met bijbehorende correlatie waarden van meer dan 0,91 voor snelheid; 0,93 voor dichtheid en 0,75 voor intensiteit. Het NLM raamwerk voor voorspelling laat bij 5% FCD en een voorspelhorizon van 5 minuten opnieuw veelbelovende resultaten zien, met slechts een 20% reductie in nauwkeurigheid van snelheidsvoorspellingen. Echter de dichtheid is tot 50% minder nauwkeurig, en intensiteitsvoorspellingen zijn zelfs tot 300% slechter. De correlatie van de snelheid en dichtheid voorspellingen daalt slechts marginaal, maar de correlatie van intensiteit daalt met 10% tot 15%. Voor een voorspelhorizon van 15 minuten neemt de zowel de nauwkeurigheid als de correlatie verder af tot voor de correlatie ver onder de 0,75.

Dit onderzoek laat zien dat het ontworpen NLM raamwerk een veelbelovende *traffic state estimation* en redelijke *traffic state prediction* weet te genereren in de gemodelleerde en gesimuleerde stedelijke omgeving van Sioux Falls. Vanwege de relatieve eenvoud, lage complexiteit is het bovendien gemakkelijk om NLM toe te passen in de echte wereld. Ook biedt NLM de mogelijkheid om andere verkeers-databronnen moeiteloos te integreren. Er zijn ook gebieden waarop verder onderzoek nodig is, vooral op het voorspellende vermogen van NLM, welke op dit moment nog onvoldoende is. Geavanceerde technieken zoals *bagging*, alsmede het vinden van een betere methode om de buren van een link te koppelen, kunnen extra nauwkeurigheid verschaffen. De tijd-lag vertraging welke gelieerd is aan de manier waarop NLM omgaat met het vinden van relaties is nog niet overwonnen in dit onderzoek. Een plug-in die historische verkeersgegevens op een andere manier gebruikt kan hier een oplossing bieden. Daarnaast is verder werk nodig om de huidige implementatie te testen en verbeteren voor complexere en realistischere stedelijke netwerken, omdat niet alle eigenschappen die kenmerkend zijn voor een stedelijke omgeving werden opgenomen in de case studie van dit onderzoek. Voorbeelden hiervan zijn; interactie tussen gebruikers, vervoerswijzen, een heterogene voertuigmix en dynamische verkeerslichten.

3 Abstract

The use of information technology (IT) in traffic systems is becoming a hot topic within the traffic research community. It is within this context that IT and traffic, blend together to create intelligent transportation systems (ITS). Benefits of employing ITS are ample and can range from; increased safety, improved operational performance, enhanced mobility, environmental benefits to boosts of productivity leading to economic and employment growth (Ezell, 2010). One of the practical outcomes can for example be that all actors of the transportation systems are allowed to enlighten themselves with information and make better informed decisions. This type of utilization of ITS leans on a key ingredient; a robust, complete and accurate picture of the traffic state in the network. This picture is generated in a process called traffic state estimation. It generally goes hand in hand with the process of traffic state prediction, which produces the future pictures of the traffic state.

In this research towards urban traffic state estimation and prediction, two challenges are introduced. Firstly there is a mismatch between the amount of data and the needed traffic flow variables. A traffic network is in practice never fully covered by traffic information sources, thus requiring techniques to extrapolate and utilize the traffic data that ís available. The second challenge is a result of choosing an urban environment as subject of study in this research. An urban environment which as opposed to a freeway only network, comes with a higher complexity due to e.g. lower traffic volumes, lower speed limits, more variability in velocities, a higher density of intersections, traffic signals, roundabouts, priority-junctions and dynamic interactions between other modes of transport.

Van Hinsbergen et al. (2007) describes that the vast amount of traffic estimation and predictions models used in literature can be fitted into four categories. There are naïve models, parametric models, non-parametric models and a hybrid blend of two or more of these categories. Van Lint (2011) emphasizes that the key difficulty for traffic state estimation and prediction is therefore to find a balance between sophisticated and complex models on one side and smooth, fast, general applicable models on the other side, to make valid estimations and forecasts given the data available.

The naïve categorization represents models, in which only traffic data is used from which direct relations are calculated. Examples are instantaneous travel time or historical averaging models. The advantage of these models is the favourable low computational complexity and easy to implementation. The downside is that because of the lack of traffic theory, results are usually illogical and inaccurate. The parametric categorization represents models in which the principle of the Lighthill–Whitham–Richards (Lighthill & Whitham, 1955) model are applied. Classical examples are; Newell's simplified kinematic wave model (Newell, 1993) and cell transmission models (Daganzo, 1994). The advantage of these models is that they implement logical real world traffic theory, but with the disadvantage of requiring vast calibration of parameters. Additionally due to the case being an urban network where the traffic flow fundamentals of for example flow conservation might not be applicable, accuracy is negatively affected. The non-parametric categorization represents the traffic models in which relations in traffic data are considered, but no traffic flow parameters are estimated. Examples of these models are mostly based on simple regression. These models have in common that while their complexity is low and therefore they can easily be run in real-time speed, their accuracy is according to Van Lint (2011) generally fairly low.

Theoretically more suitable for urban traffic state estimation and prediction are hybrid model types, which take elements from non-parametric, parametric and naïve methods to output more accurate estimations and predictions. The most famous example is Kalman Filtering (Kalman & Bucy, 1961), which again assumes all traffic flow fundamentals to hold. Of all the other hybrid models considered in this research, the hybrid black-box approaches seem to be theoretically, practically and intuitively, the best suitable in the urban environment chosen as subject of this study. Within these lines Morita (2011), Esaway (2012) developed two interesting frameworks which consider the use of patterns in historical traffic data which allow the current traffic state of links to be used as indicators for the traffic state on neighbouring links.

The goal of this research is to design a performing traffic state estimation and traffic state prediction framework, which by utilizing both floating car- and inductive loop detector- data, delivers real-time and future link -velocities, -densities and -flows within an urban traffic environment. Used as a starting point for this design are the previously methods of Morita (2011) and Esaway (2012), from which this newly developed neighbourhood link method (NLM) is created. Aimed is to answer the question on how to design a neighbour link framework which delivers both a traffic state estimation and state prediction of all relevant traffic flow variables within an urban network and assess both the performance and accuracy of the traffic states outputted by this NLM framework.

The research method applied is divided into three separate parts; 1) the urban traffic state ground truth, 2) the urban traffic state estimation and 3) the urban traffic state prediction part. In the first part of the research a microsimulation programme (PARAMICS) is used to generate a 100% accurate and 100% covered ground truth for the Sioux Falls network. This ground truth designed remains unchanged throughout this research. The second part of this research presents the extension of the ground truth framework with the NLM for traffic state estimation. Additionally the steps of performance assessment, evaluation and synthesis are included to complete the design cycle. With the ground truth available (for comparative purposes), the estimations are assessed on accuracy and correlation leading to the designing of the best performing NLM variant for both traffic state-estimation and -prediction.

The newly developed NLM framework can be described by the following 7 steps. Initially traffic data is stored in a database. Next from this database for each link it is determined which links behave the same and can be considered neighbours based on correlation in traffic data. Then newly arrived data is considered upon which an estimation from solely the traffic data of neighbouring links is generated using linear regression. Consequently this neighbourhood estimation is fused, weighted on reliability, with the traffic data from the link itself, generating the final traffic state estimation. For the prediction part, an extra time dimension is included, to incorporate the prediction horizon.

The results reveal that the NLM framework for estimation at 5% FCD, is able to estimate on average 60% of the urban links in the network within 3 mi/h of the ground truth during rush hour periods. In free flow traffic this percentage drops to 50%. Density estimations show 80% of all links to be estimated within 12 veh/mi/lane and of the flow estimations 60% of the links can be estimated within 100 veh/h/lane deviations in rush hour. With corresponding correlation values of over 0,91 for velocity; 0,93 for density and 0,75 for flow estimations. The NLM framework for prediction shows at 5% FCD and a prediction horizon of 5 minutes again promising results, with only a 20% drop in accuracy for velocity. However density predictions are up to 50% less accurate, and flow predictions are up to 300% worse. The correlation of velocity and density predictions only drop marginally, but the correlation of flow predictions lowers by 10% to 15%. For a prediction horizon of 15 minutes the degradation continues as the prediction accuracy decreases further and the correlation for density and flow drops well below 0,75.

This research reveals that the considered NLM framework can yield very reasonable traffic state estimation results in a modelled and simulated environment. Due to the fact it is simple in essence and algorithmically not very complex, NLM can also be easily transferred to a real world scenario. Additionally other traffic data sources can be effortlessly implemented in the process. There are however areas suitable for further research, especially as the predictive ability of NLM is currently unsatisfactory. Advanced bagging of historical traffic data can provide additional accuracy, as well as a different approach to finding the neighbourhood space for each link. The time-lag inherently apparent in NLM (because the first link that experiences congestion cannot be predicted by its neighbours) has not been overcome. A plug-in that incorporates historical traffic data differently might provide a solution here. Additionally further work is needed to improve and test the current implementation into more complex and realistic urban traffic networks as not all traits that typically describe an urban environment were included in the used case study (e.g. user-interaction, mode-interaction, a heterogeneous vehicle mix and dynamic traffic lights).

4 Introduction

4.1 Preface

Information technology (abbreviated as IT) is currently being applied on a global scale. In education, health, industries and in government the role of IT has become more and more important in the last few years. With the positive effects of IT not going unnoticed, IT use in traffic systems is also on the rise. For many countries, building new roads and other traffic infrastructure to keep up with the ever growing traffic demand, is no longer a viable solution. The scope is turning from asphalt, concrete and steel to more intelligent use of the traffic infrastructure currently in place. It is within this context that IT and traffic blend to the subject of intelligent transportation systems (ITS). ITS gives the actors of the transportation systems (both the controllers as the actual users) the power to enlighten themselves with information (in all kinds of formats) to make better informed decisions. Examples of these decisions are ample on the user side; e.g. route choice behaviour, departure time behaviour and mode choice behaviour, yet also on the controller side; e.g. automatic incident detection (AIDA) (Wang, 2005), optimizing traffic signals, improve current roads and provide better public transportation. ITS can deliver five key benefits; increasing safety, improving operational performance, enhancing mobility and convenience, delivering environmental benefits, and lastly boosting productivity leading to economic and employment growth (Ezell, 2010).

A key ingredient of ITS is real-time traffic information. It is therefore required to somehow get a complete image of the network's traffic state at current and near future times (Wang, 2005) to derive this traffic information from. The more accurate this image the more focussed ITS can be implemented for actors in the traffic network. The task of mathematically deriving an image of the network's traffic state from traffic data, are separated into the *traffic state estimation*- and *traffic state prediction* task. The goal of both is to deliver a robust, complete and accurate picture of the urban traffic state on all links in the network.

The *traffic state estimation* task refers in this context to real-time estimation of different traffic flow variables for a network within a specific time interval and spatial resolution. The *traffic state prediction* task refers to real-time prediction over a certain time horizon of different traffic flow variables for a network within a specific time interval and spatial resolution. Both the estimation and prediction are generated by applying mathematical techniques and concepts of traffic flow, upon available real-time traffic data.

Within the research field of traffic state estimation and prediction, a recent shift has occurred altering the scope of research from the relative comfort of freeway segments in highway networks to larger and more detailed urban road networks. These urban networks cover besides the upper network also the important primary and secondary urban arterials. A first example of such an urban study is the *Instrumented City Project* (Bell et al. 1992). Practical examples of recent urban traffic state studies in the Netherlands are Sensor City Assen (as from 2011) and the Praktijkproef Amsterdam (as from 2013). Within both these projects and in literature, the traffic flow variables of interest are; link wise space mean speeds, travel times, traffic densities and flows (e.g. Snelder & Calvert, 2015; Wang, 2005). Accurately estimating these traffic flow variables for different pre-defined network segments is the ultimate goal for the application of traffic state estimation techniques. Additionally predicting these traffic flow variables is the goal of the traffic prediction counterpart. As the number of unknown traffic flow variables to be estimated and predicted are generally larger than the number of variables that are measured, the relative complex task of a deriving an accurate real-time network traffic state for ITS purposes, becomes visible.

An exploratory research towards traffic state estimation by De Vries (undersigned) in 2015 revealed new ideas on how to improve the traffic state estimation accuracy in urban networks by using a learning database heuristics. This result initiated this research, which dives further into the subject of urban traffic state estimation and urban traffic state prediction.

4.2 Problem definition

Research into urban traffic state estimation and prediction comes with two challenges. Firstly there is the previously mentioned mismatch between the amounts of *big data* and the needed traffic flow variables. The urban network is in practice never fully covered by traffic information sources, leading to gaps and errors when no techniques are applied to the traffic data that ís available. The second challenge is related to the urban nature of the traffic network. While research on traffic state estimation on freeways is vast (Nantes, 2015), literature on traffic state estimation in an urban environment is less common. It is however not less interesting, as negative effects of traffic affect a larger population in urban areas than in rural areas. The ante is upped because of the fact that an urban environment provides additional challenges as opposed to a freeway only network, due to for example; lower traffic volumes, lower speed limits, more variability in velocities, a higher density of intersections, traffic signals, roundabouts, priority-junctions and lastly dynamic interactions between other modes of transport. Generally the state estimation and state prediction are part of the traffic control cycle (presented below), in which according to the graphic of Van Lint et al. (2015) a traffic flow model is located at the centre. It must be noted that this traffic flow model can be replaced with other mathematical models and methods which do not use the traffic flow fundamentals. This research discusses these other types later on.

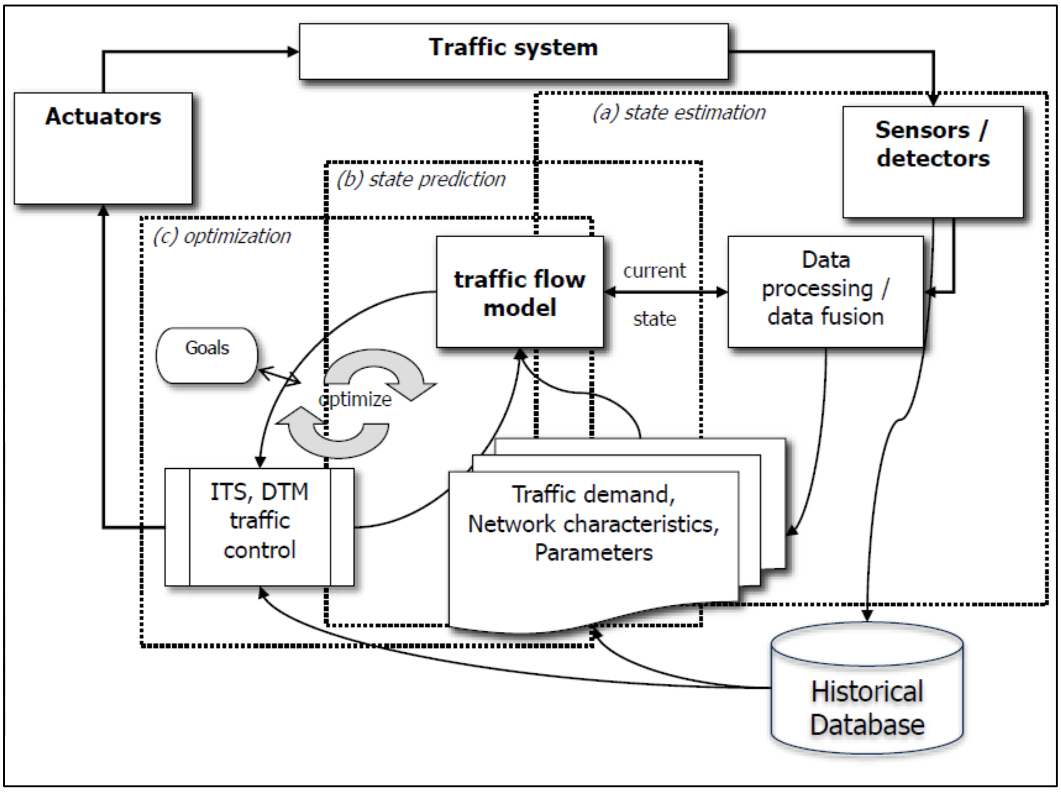

Figure 1: The traffic control cycle. Source: Van Lint, J. (2015)

4.2.1 Traffic data

Regarding the traffic information and traffic data collection, the common practice is to rely solely on a network of roadside sensors responsible for generating solid detector data (SDD) (Tao et al., 2012). These most commonly equipped sensors are inductive loop detectors (ILD), which – if dually equipped – give information about all three main macroscopic traffic flow variables (flow, occupancy and speed) at fixed positions on the road network. Inductive loop detector data (ILDD) comes with two major drawbacks, firstly it is prone to errors (Herrera, 2010) and secondly the broader spatial representativeness of measured traffic flow variables is questionable (Tao, 2012). As in urban networks the coverage rate of ILDD is generally low, the measurements from a limited amount of ILD, do not suffice to provide the complete traffic information needed (Wang et al., 2011).

Radio-frequency identification (RFID) or Bluetooth identification are alternatives used to obtain individual travel times based on vehicle identification and re-identification (Herrera, 2010). Yet come at high cost, privacy concerns, low coverage and can only measure travel times between set locations. For License plate recognition (LPR) and other video image techniques, the same disadvantages apply.

A more promising data source is Floating Car Data (FCD). FCD consists of reported vehicle positions, direction of driving and velocities for timestamps with a predefined temporal spacing from dedicated vehicle probes. These vehicles are equipped with a form of GPS and a communication link for transferring this data. FCD has the advantage opposed to SDD to be able to determine a representative mean speed for a whole road segment. The downsides of FCD are the wobbly representativeness of FCD due to the penetration rate, resolution (Cayford & Johnson, 2003) and GPS accuracy as a result of mapping issues due to tall buildings and complex networks within an urban environment (Li et al., 2013). The use of FCD has been researched in literature quite extensively in both real-world practice and simulations. Focus lies for example on determining the percentage of FCD required for an accurate speed estimation. Srinivasan & Jovanis (1996) showed that for an urban traffic network in Sacramento (California) a bare minimum of 5% of the total vehicle population is required for a reasonable travel time estimation. Cheu et al. (2002) concluded that to achieve a standard deviation of maximum of 5 km/h in space mean speed (for 95% of the links in the network), with a 10 minute resolution, 4 to 5 percent FCD is needed. A sample size of less than 10 cars might however, not be adequate enough in a given interval. Cheu (2002) also added that beyond 15% FCD the additional profit diminishes in its urban case study of Singapore. Herrera (2010) revealed in its Mobile Century field experiment a penetration rate of 2% to 4% is sufficient for a freeway stretch with current GPS technology. De Vries (2015) concluded that in a microscopic simulation of an urban network 10% FCD is needed to provide accurate (57% of the links in the network with a deviation of less than 5 km/h) speed estimations, though with a much higher resolution of only 1 minute as opposed to the 5 and 10 minute sample rates within the other researches. Even at this high resolution the profit of having more than 12% of FCD does not yield much better results. The focus of FCD should therefore not necessarily be solely on quantity, but on quality and coverage.

More recently, a new data source is compiled by mining of traffic jam reactions through online social sensors (Georgiou et al., 2015). These social media sensors (SMS) gather data by crawling through posts on regular social media (e.g. Twitter) or on more specialized traffic apps (e.g. Waze). SMS offers a fast and low cost way to understand what is happening in the physical world, although the noisy nature of the data makes quality still lacking. Deng & Zhou (2011) argue that each type of information has its own advantages for traffic state estimation. The table below summarizes the different traffic data sources available, its' respective quality and costs.

Information Source	Traffic Data	Data Quality	Costs & Concerns
Point detector data (ILDD)	Vehicle counts & Point speeds	High accuracy but low reliability	High maintenance cost.
Automatic vehicle identification	Point to point travel times and volumes of tagged vehicles	Accuracy dependent on penetration rate	High installation costs.
Video Image Processing	Continuous path trajectory for vehicles	Accuracy depends on machine vision algorithms	High investment and communication costs.
Floating Car Data (FCD)	Continuous path trajectories and travel times on traversed links	Accuracy dependent on penetration rate, no direct information on density and flow	Trade-off between utility and privacy. Big data sizes. Owned by corporate companies.
Social Media Sensors (SMS)	Congestive/Incident traffic state reports	Very fast! Noisy, lack of thorough understanding, vast sizes	Publically available, cheap

Table 1: Summary of available data sources. Expanded, actualized and adjusted from Deng (2011).

To make the most of all available traffic data, the method or algorithm used for both traffic state estimation and traffic state prediction, calls for fusion of heterogeneous data sources to maximize the utility of the available information (Treiber & Kesting, 2012). The gathering and application of historical traffic data is already extensively used in the real world for traffic state estimation and prediction (e.g. Snelder, 2015; Wang, 2005). It is within this context that another challenge arises regarding the subject of big data. Whereas the common single freeway practice is to rely solely on a selection of roadside sensors in the form of loop detectors, smart cities equip and utilize a full network of data and wireless sensors to map each activity in the city. Data might be gathered from e.g. point detectors (ILD), automatic vehicle identification sensors, video image processing sensors, floating cars (FC) and social media sensors (SMS). The traffic data gathered from these sources is generally separately processed and stored into traffic databases. The question rises on how to deal with this ever-growing amount of *big data* and how to integrate difference sources of traffic data as to achieve its full potential. But still then the usefulness of this vast amount of *big data* depends on the quality of the recorded data and used extrapolation technique (Wang, 2005).

4.2.2 Urban environment

An urban environment increases the complexity of the state estimation and prediction task as opposed to a freeway environment due to several reasons. The general theme of an urban environment is very dynamic. Different modes of transport interact and also the degrees of freedom a vehicle or person have are greater than on a stretch of freeway. Interactions between users interrupt steady flows of traffic and generate different incident patterns. For example a crosswalk creates a localized temporary disturbance. Signalized and priority junctions have an even larger effect, making it very challenging to separate an incident, from a (regular) queue of cars (Tampère et al., 2012). Urban road segments additionally provide on-street parking, branches and parking garages. Whereas the law of flow conservation generally applies on freeway network, it does not apply in urban networks (Gosh & Smith, 2015). On the data-collection side an urban environment makes it harder to separate FCD from all the data that pedestrians, cars and cyclists generate. Van Lint (2011) states that the complexity of the urban road network makes the choice for any type (heuristic/algorithm) of traffic flow model for traffic state estimation/prediction an important one.

4.2.3 UTSE & UTSP

As previously mentioned, *urban traffic state estimation* (UTSE) refers to estimating relevant traffic flow variables such as flows, densities, speed and travel times for links in an urban road network with a certain temporal and spatial resolution based on traffic data available (Wang, 2008). *Urban traffic state prediction* (UTSP) refers to predicting the same traffic flow variables using the most current traffic data with a predefined prediction horizon (generally up to 30 minutes). Figure 2 shows the generalized form of a UTSE and UTSP model.

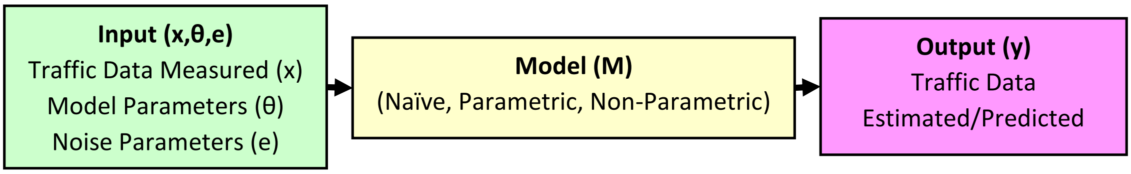

Figure 2: General form of any traffic state estimation and prediction models. Derived from Van Lint (2011)

Snelder (2015) gives an inexhaustible list of estimation and prediction model types used in literature being; statistical, dynamical, microscopic, macroscopic, offline, online, data driven, model driven and deterministic models. Because each type of model has its own advantages and disadvantages there is currently no model available which outperforms them all, in every context. For literature review purposes, the categorization proposed in Van Hinsbergen et al. (2007) is adopted in which four categories describe the differences between traffic estimation models. There are naive models, parametric models, non-parametric models and a hybrid blend of two or all of these categories. Van Lint (2011) adds that the key difficulty for traffic state estimation and prediction is therefore to find a balance between sophisticated and complex models on one side and smooth, fast, general applicable models on the other side, to make valid estimations and forecasts given the data available. A more detailed look into these model types is provided next.

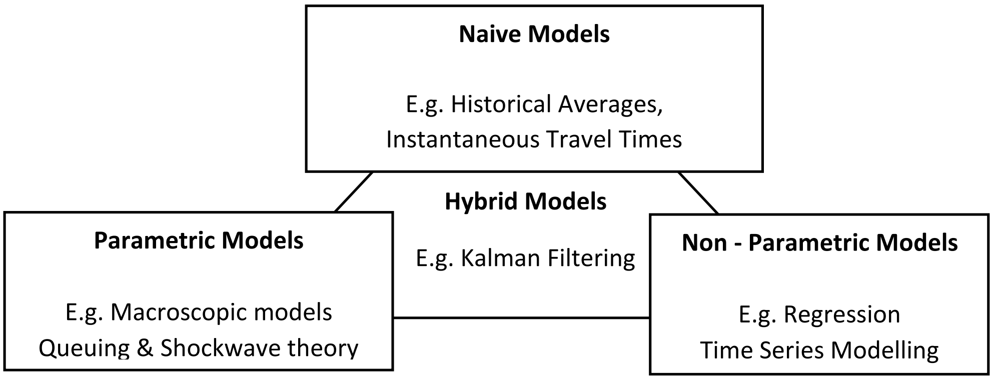

Figure 3: The three model types of traffic state estimation. Edited from Van Hinsbergen et al. (2007).

The naïve categorization represents the traffic models in which only the traffic data is used and direct relations are calculated. No model structure or parameters are inputted, which results in favourable low computational complexity and very easy implementation. The naive method Snelder (2015) uses in the earlier mentioned Praktijkproef Amsterdam, is based the assumption that in short term the traffic situation on the network does not change. Other examples of naïve methods are based on measured instantaneous travel times or historical averages. It can be argued that because of the lack of traffic theory, the results are usually illogical and inaccurate (Van Lint, 2011). As the traffic situation in an urban network can be considered to be even more dynamic, results are likely to become more illogical.

The parametric categorization represents models in which the principles behind the Lighthill–Whitham–Richards (Lighthill & Whitham, 1955) model, also known as a first order traffic flow and kinematic wave theory model are used. The two principles of LWR are that traffic is modelled and simulated conform: a fundamental diagram and the traffic flow conservation law. These models are therefore based on plausible theoretical assumptions on traffic behaviour in time (Van Lint, 2011). The inputted data consists of e.g. flows, OD-pairs and/or turn rates, for which the models determine macroscopic parameters (e.g. link capacities or parameters related to a fundamental diagram) or microscopic parameters (e.g. car-following behaviour or lane change behaviour). As these models try to incorporate real world car traffic theory such as e.g. queueing theory, car following theory and/or shockwave theory (Van Lint, 2011), it becomes inevitable that vast calibration of parameters is required as to assure that the model results comply with the real-world. Examples of these model types are; Newell's simplified kinematic wave model (Newell, 1993), cell transmission models (Daganzo, 1994), the variational kinematic wave theory (Daganzo, 2005), link transmission models (Yperman, 2007) and more recently a Lagrangian based approach (Laval & Leclercq, 2013). These parametric based traffic estimation models do come with some limitations. Firstly traffic flow theory represents abstracted versions of real-world phenomena, as for example a fundamental diagram can only be approximated (Seo, 2015). This is due to individual driving behaviour e.g. differences in desired acceleration and deceleration, vehicle types, vehicle lengths, platooning, lane changing and changing traffic states. Determination of a realistic fundamental diagram for each segment of the network is therefore difficult, especially in an urban environment. The second reason is related to this urban environment, as already mentioned, traffic flow conservation law does not necessarily hold, making these models relatively unsuitable for urban traffic state estimation.

The non-parametric categorization represents the traffic models in which relations in traffic data are considered, but no traffic flow parameters are estimated. They estimate traffic state based on real-time traffic data with some relation abstracted from historical traffic data. This relationship can be spatial and/or temporal. The examples of these type of models are again ample. E.g. a simple regression approach tries to approximate the output by using weighted combinations of input data and is generally classified as an autoregressive–moving-average (ARMA) model. Expansions of ARMA are again possible by e.g. considering locally weighted regression in which each data point is weighted proportionally to its proximity to the investigated data point (which is also in basis used in the Adaptive Smoothing Method (Treiber, 2012), considering a linear combination of historical and current states (ATHENA; Aron & Danech-Pajouh (1991)) or even considering recent measurements to be weighted more heavily as opposed to the more historical measurements (SETAR; Watson et al. (1992)). These models have in common that while their complexity is low and therefore they can easily be run in real-time speed, their accuracy is according to Van Lint (2011) fairly low.

The hybrid categorization represents the traffic models which take elements from the different categories. Examples of popular parametric models which also apply some non-parametric techniques are mostly based on Kalman Filtering (KF)(Kalman & Bucy, 1961) e.g. EKF, UFK, PF, DEKF (Van Lint et al., 2008). These models are already quite well performing (e.g. Wan & Merwe, 2000; Ristic et al., 2004; Wang, 2005), though Tampère (2011) argues that for KF to keep performing in urban networks their complexity must be raised. It is due to the influence and dominance of capacity restrictions at intersections, the second order traffic flow phenomena (instability, capacity drop, stop and go traffic) are disturbed as well. The cell transmission models (Daganzo, 1994) enriched with link transmission models and node models have the advantage that they do not consider these traffic flow phenomena. The advantages of hybrid parametrized models are numerous as these traffic models allow decision support, scenario analysis and real-time traffic control (Van Lint, 2011). The limitations are related to the designed complexity as demands, turning rates, on-street parking rates, route choice patterns, traffic signal cycles and the vast amount of parameters to be calibrated. And as this calibration requires the outputted (real) traffic states as main ingredient, a vicious circle is potentially designed. Additionally computation complexity is increased and therefore there models might forfeit the real-time prediction ability (Van Lint, 2008; Snelder, 2015) making them less suitable for urban traffic state estimation.

The examples mentioned in the previous chapter fall all within the white box hybrid model category. When the processing phase of the model becomes vaguer and abstracter, a black box approach is adopted. As a basis these black box approaches use for example artificial Intelligence (AI) or Artificial Neural Networks (ANN). These machine based learning models are expansions of non-parametric models. Van Lint (2011) argues that with the traffic data inputted and outputted being noisy and the relationships being multivariate and not necessarily linear, the problem of traffic estimation and prediction is intuitively suitable for an AI approach. The most commonly used ANN models all share the same basis consisting of (at least) two weight layers. The idea of these models is to find weights such that the deviation from expected output is minimized (also known as the *least squares problem*). As studies on ANN show varying results in terms of accuracy more sophisticated models and expansions are available. Mention worthy examples are modular neural networks (MNN), which process input in sub-networks and make calculations much faster. Radial basis frequency networks (RBFNN) add another layer to cope with clustering the input space. Conjugate gradient algorithms (CGA) use higher order information to determine the magnitude and vector of weight changes. Additionally wavelet functions (sinuses) instead of the sigmoid functions (s-shape) and Fourier transformations in generalized neural networks (GNN) can be used to improve accuracy even further. Additionally there are Bayesian Believe Network (BBN) methods, which according to Van Lint (2011) work already well within urban environments. A BNN is defined as a directed graphical model in which conditional dependencies between variables are included with the goal of selecting the most probable outcomes. An example of such a type of method, is the k-nearest neighbour method. In this method a historical database is scanned for the most similar entries to the current situation. A dynamic weighting is then applied to yield a most probable final estimation or prediction. Smith et al. (2002) concludes that this method outperforms the more naïve approaches and might even improve further when the sizes of the historical databases is expanded, or weightings are spatially dependent. Another branch of BNN suggests the ability of memorization captured in evolutionary learning networks, in which previous outputs are stored in a hidden layer such that previous outputs and patterns can be reused. Van Lint (2011) underlines that the area of a data driven BNN's is deemed very promising.

It becomes apparent from above analysis that the number of methods reviewed and used in literature within each model domain is ample. It is the different way of method effectiveness determination or accuracy within each study that makes comparison on an absolute scale difficult. It is for this reason that in most studies only a relative conclusion is drawn. The accuracy of the developed model is e.g. compared to a predecessor or in quantities which does not allow for comparison between models. More extremely, commercial reasons might even disallow comparison completely. Regarding the way of measurement for travel times, the most common expressions are MSE (mean square error), RMSE (root mean square error) and MAPE (mean absolute percentage error) of which the MSE and RMSE would be preferred in terms of used expression, because it harbours the same quantity of measurement (Van Lint, 2011). The bottom-line is that there is no method which outperforms one another in every situation and many of the mentioned methods seem at least theoretically to be less suitable for traffic state estimation and prediction within a detailed urban network as opposed to freeway stretches. The choice for a certain model should therefore be completely context-dependent.

Of all the mentioned models, the hybrid *black-box* approaches seems to be at least theoretically, the best suitable for the dynamic urban environment chosen as subject of this study. They can perform in real-time, are simple yet complex enough to cope with the heterogeneous urban traffic network and yield both traffic state estimations and predictions. By including conditional dependencies (BNN) and memorization of patterns in previous outputs (ANN) naive approaches are outperformed. As a prerequisite this does require some existing intuitive relationship as to give direction to the weight layers and a starting point for further research. At a first thought, an obvious relationship is that of travel time correlation between neighbouring links. This relationship has already been researched and confirmed in literature (e.g. Gajewski & Rilett, 2003; Sen et al., 1997). It comes with no surprise that research on correlation of velocities for neighbouring links reveals already two separate frameworks developed, by Morita (2009, 2010, 2011) and Esawey (2012), with both showing promising results. A more elaborate relationship is used by INRIX Traffic (2016), in which data from adjacent links are considered informative for the current and future state of other links. Due to obvious commercial reasons no performance accuracy is given, but as this approach is already commercially in use it shows there is merit in further research. This approach is classifiable within the hybrid category where it combines naïve, non-parametric and perhaps even slight hints of parametric ideas. These methods do hold some disadvantages. Firstly there is (a lot of) historical data required for finding the relationship. Secondly these relationships only apply during the normal traffic situations present in the traffic databases. Though this latter disadvantage means that extraordinary traffic situations can be detected when comparing the estimated traffic states with the true traffic states. It is within this line of reasoning, that the idea of correlation of traffic flow variables between neighbouring links is the most deem worthy area to be further explored within this urban traffic state estimation and prediction research.

4.2.4 The neighbourhood link method (NLM)

This thesis proposes to further research the vitality of using the current and future situation of neighbouring links as indicators for the both a traffic state estimation/prediction. The previously mentioned frameworks of Morita (2011) and Esaway (2012) are used a starting point. Currently they only output estimations along the dimension of space and for traffic state prediction the dimension of time needs to be included and completely redesigned. Additionally the frameworks differ in some detailed approaches and mathematical techniques, where comparison is required to select the best way to continue. Lastly the frameworks only output velocity/travel time, which means extension and designing is required to include intensities and densities as output.

The backbone of both neighbour link methods (NLM) is based on weighting of neighbour links to improve accuracy of the final state estimation/prediction and to fill in the gaps of links with no data by using the data from its neighbours. The global structure of these NLMs is depicted in the figure below. Without going into details here (a complete step-for-step explanation is provided later on in this thesis), it stores newly received traffic data (x) in a traffic database (A). Then the neuron layer (w) finds for each neighbour link, a linear weighting (w) such that this weighting (w) multiplied with the input layer (x) becomes the outputted traffic data (y) of that particular link.

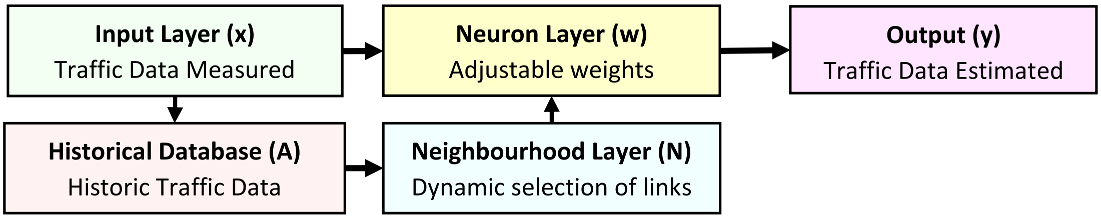

Figure 4: Global structure of the NLM framework

The first big difference between Morita (2011) and Esaway (2012) is the way the neighbourhood space is calculated. Obviously some kind of threshold should be applied as to achieve a reasonable trade-off between accuracy and number of neighbours (which directly influences calculation times. Esaway (2012) makes its selection of neighbourhood links based on correlation in traffic data. The links that show correlating behaviour in the past are assumed to behave the same in the future. Morita (2011) relaxes this assumption and defines the neighbourhood space as just the directly adjacent edges of a link in the network. However, in an urban network it is not necessarily true that directly adjacent links show correlation with the link in question as for example near (signalized) intersections the merging, diverging and yielding of traffic plays a significant role (Tampère, 2011). The approach taken by Esaway (2012) is considered to be intuitively more suitable than the approach of Morita (2011).

The second difference between Morita (2011) and Esaway (2012) is the way of weighting applied to the links in the neighbourhood space. Upon selecting of the neighbourhood space, Esaway (2012) weights the traffic data of the links, by their respective variance. That is; the less reliable a measurement, the less it is weighted (just as in KF). Subsequently a normalized weighting of these links then yields a final estimation which forcefully lays in between the highest and lowest measurement on the neighbouring links. Morita (2011) however, weights the traffic data of the links, by using linear regression. By using the traffic data from the link itself as a target, it tries to find any linear weighting of the traffic data on neighbouring links, such that for all times in the traffic database the target is aimed for. This regression approach therefore weights traffic data of neighbour links that correlate most, relatively higher than neighbour links that correlate less. It can therefore be considered in line with the filtering consequence of the neighbourhood space correlation approach. As both approaches have merit, combining them and developing them further, is adopted to be a suitable area for further research.

To make NLM suitable for traffic state predictions, both the neighbourhood space determination and the linear regression parts of NLM must be modified. The neighbourhood space determination can be relatively easily adjusted by determining the correlation between a link and its neighbours not at a set time t, but at a time in the past, at an exact *distance* equal to the forecast horizon. The linear regression part can be modified along the same lines, whereas the target of the weighting is not at a time t, but at a time in the future, at an exact 'distance' equal to the forecast horizon. With these modifications, the historic traffic database is searched for exactly the idea behind NLM; namely finding which links are indicators for the future traffic state of a link itself.

Lastly the NLM must be made suitable to output flows and densities as well. Essentially the correlated relationship in velocities also applies to densities. Lower velocities imply higher densities and high velocities imply lower densities. Using the multiplicative relationship between velocity and density gives the flow required. However, density and flow cannot be reasonably be estimated from solely FCD (Seo, 2015) as the fundamental diagram is inherently inaccurate and hard to obtain. For this the fusing of data from FCD and ILDD is to be explored for implantation in the NLM.

For accuracy determination, the model outputs (expected) should be compared with the real world situation (observed). However, the traffic data used to determine this observed situation are inherently noisy and thus not necessarily correct (Snelder, 2015). Treiber (2012) argues that it can only be correct if the full data set serving as reference is so dense that it can be regarded as representing the ground truth. Proposed within this thesis is to use a microscopic traffic model in which an urban network is modelled. This will provide a 100% accurate ground truth and therefore allows unbiased and fair comparison of the performance of the neighbourhood link correlation method. Obviously the downside of this choice is that real world traffic behaviour might be different than the behaviour programmed in the simulation model. For example as the network changes or flow is disturbed, traffic data is noisy (Snelder, 2015), users might be only boundedly rational (e.g. Mahmassani & Chang, 1987).

4.3 Thesis outline

In this fourth chapter, the research subject, context and the challenges that come with urban traffic state estimation and prediction are introduced. In the next chapter, the research objective, research question and scope of this research are presented upon which in the third chapter, the used research methodology is presented.

The subsequent section of this research is divided into three parts. The first part introduces the case study of this research in detail. This first part then yields the ground truth which is used to assess the performance of the NLM. The second part considers the urban traffic state estimation share of this research. In this part the NLM is applied upon the traffic data derived from the case study scenario and subsequently its results are assessed and evaluated. The third part considers the urban traffic state prediction share of this research. Here the modifications to make NLM suitable for prediction are discussed in more detail, after which its' results are again presented and discussed.

In the last chapter, this research is finished by discussing this research's methodology, results and remarks. Additionally conclusions are drawn from the findings in the estimation and prediction parts of this research and the research question is answered. This research is then finalized with research implications and recommendations for further research.

5 Research objective, research question and scope

In this chapter the research objective, the research question and scope of this research are presented.

5.1 Research objective

The main research objective of this master thesis is:

> *"Design of a traffic state estimation and traffic state prediction method and framework, which by utilizing both floating car- and inductive loop detector data, delivers real-time link-velocities, -densities and -flows within an urban traffic environment."*

From the previously presented literature review, the theoretically most likely model to perform in an urban network is an improved and redesigned version of the neighbourhood link correlation methods of Morita (2011) and Esaway (2012). These methods are therefore selected as a starting point for further development of a neighbourhood link method (NLM) framework that fulfils the stated objective.

The first part of the research therefore focusses on designing the NLM framework as to allow enrichment of floating car data considered, by including loop detector data through the process of data fusion. Additionally it needs to be researched how every traffic flow variable can be deduced from the limited availability of both FCD and ILDD. Another objective of research is how the framework should cope with changing traffic conditions within the network. By using a historical database for correlation and weight determination, within day and between day traffic variations provide an additional challenge. The last part assesses and improves the accuracy of the traffic state estimation outputted by analysing sensitivity towards certain choices and finding quick wins to implement in the improved version of the framework.

The second part of the research focuses on the development of the NLM framework towards traffic state prediction. It requires the designed framework from the first part of this research in which the correlation between links is not determined at the same times, but over a certain time horizon (e.g. 5 and 15 minutes). The objective of this second part is then again to assess and improve the accuracy of the traffic state predictions outputted.

5.2 Research question

The main research question of this master thesis is:

> *"How can a NLM framework best be designed as to deliver both a traffic state estimation and state prediction of all relevant traffic flow variables within an urban network, and how well does it perform in delivering an accurate estimation and prediction for different traffic conditions within the urban network?"*

5.3 Scope

The scope of this research is fully aimed towards development and redesign, of the as starting point presented frameworks of Morita (2011) and Esaway (2012). In its current form these are limited in number of output variables (travel time/velocity), cannot predict and make choices that limit their respective performance. For this extent the results and recommendations from previous research on these methods are taken into account (i.e. De Vries (2015), Esaway (2012), Morita (2011)).

Proposed in this research is to work with three design guidelines to frame and limit the direction of the research, to most suitable area for further research identified earlier. The design guidelines proposed are:

1) Every part of the new NLM framework is developed within line of the idea that; *patterns in historical traffic data can be used to allow the current traffic state of links to serve as indicators for the current and future traffic state on neighbouring links*
2) The developed NLM framework remains classifiable as a (hybrid) non-parametric model;
3) The real-time processing power is not forfeited.

The case study traffic data used for performance assessment in this research, is generated by implementing the urban road network of the *Enriched Sioux Falls Scenario* (Chakirov & Fourie, 2014) within the microsimulation software environment of PARAMICS. It will output 100% accurate and 100% true real-time provided – or at least as close to real-time as possible – loop detector data and floating car data aggregated on a resolution of 60 seconds. For simplicity, no subcategorizations of different types of motorized traffic are made within the modelled environment. Simulated are three different two hour intervals, capturing morning rush hour, a period between morning and evening rush hour and an evening rush hour period within the *Enriched Sioux Falls Scenario's* working day.

6 Research methodology

In this chapter, the research method within this research is introduced. The research method can be divided into three separate parts being; 1) the urban traffic state ground truth, 2) the urban traffic state estimation and 3) the urban traffic state prediction part.

6.1 Part 1: Urban traffic state ground truth

This first part of the research method describes how the ground truth for this research is generated. This designed ground truth remains unchanged throughout this research. It is used to quantify the quality and accuracy of the traffic state estimations and predictions outputted in the second and third part of the research method. The ground truth is – in this context – a database with one minute aggregated velocities, densities and flows on every link in the urban network. A structural overview of the steps executed that lead to the determination of this ground truth is displayed below.

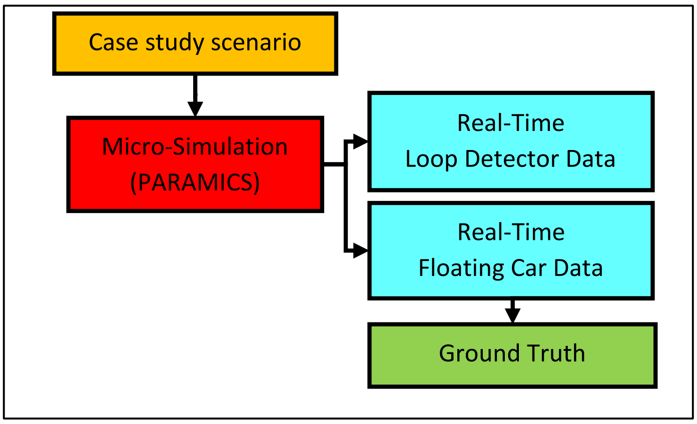
Case study scenario
Micro-Simulation (PARAMICS)
Real-Time Loop Detector Data
Real-Time Floating Car Data
Ground Truth

Figure 5: Ground truth traffic state framework

The first step in the ground truth determination process is regarding the choice and set-up of the case study scenario. That is, the network layout, vehicle characteristics, demand and trip distribution are determined. Consecutively it is implemented into the microscopic simulation environment of PARAMICS, upon which the software is run. Each run then yields a full report on real-time loop detector and floating car data. As the floating car data is both 100% accurate and fully covers all the vehicles in the network, the last step is then to aggregate the floating car data into one minute intervals and filling up the ground truth database with average link velocities, densities and flows.

6.2 Part 2: Urban traffic state estimation

The second part of this research method covers the traffic state estimation research method. It is an extension of the previously mentioned ground truth framework with one additional modification. Whereas the ground truth is derived from a 100% penetration rate, the traffic state estimation relies on a lower penetration rate. Dai et al. (2003) conclude that 3% should be a bare minimum for freeway roads and 5% a minimum for so called surface roads. De Vries (2015) shows that with a 5% penetration rate the learning database interpolation already reaches up to 50%-60% of its maximum potential, continuing up to 95% of its max potential at 10% penetration rate. Therefore a 5% floating car rate is adopted as a starting point. Additionally, the steps of performance results and evaluation and synthesis are included to complete the design cycle. The traffic state estimation outputted is – in this context – a database with one minute aggregated velocities, densities and flow estimations for every link in the previously defined urban network. With the ground truth available (for solely comparative purposes), the state estimations outputted is assessed on accuracy and correlation. Evaluation of these results and more in depth synthesis are then executed to find and design the best performing NLM variant upon the case study scenario. A structural overview of the whole traffic state estimation framework is presented on the next page.

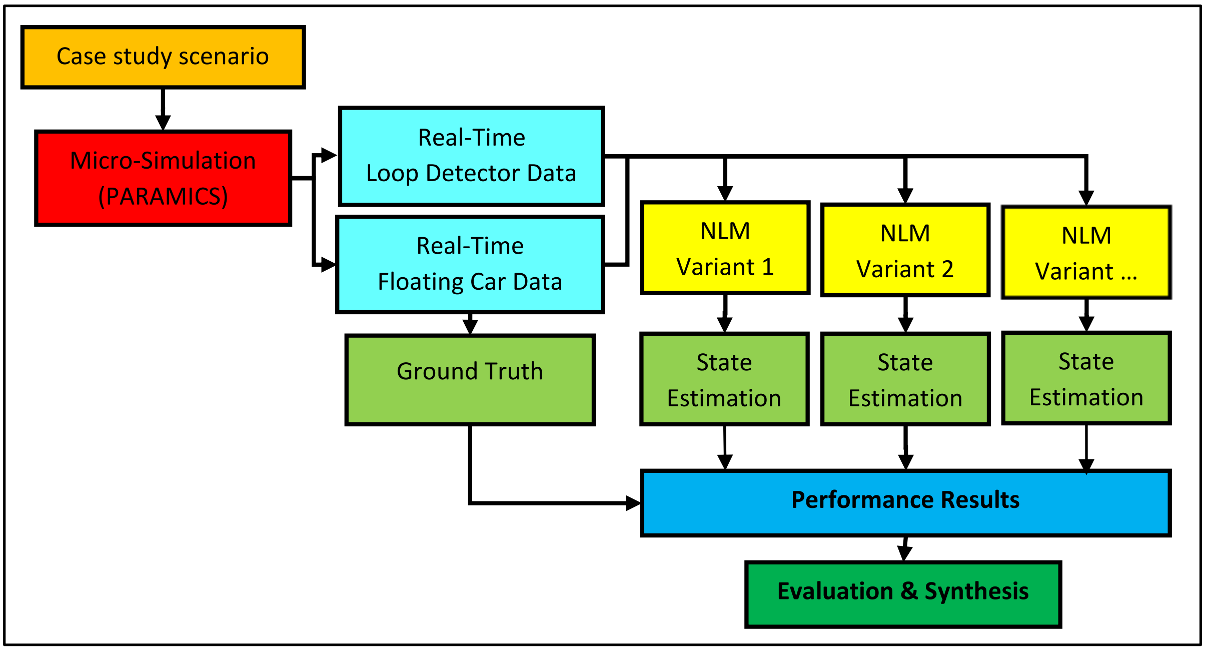

Figure 6. Traffic state estimation framework. Refined expansion of Tao (2012).

The additional steps taken to derive the traffic state estimations thus start with the filtering of the floating car traffic data as a reduction from 100% coverage to 5% is simulated. Additionally another copy is made with 10% penetration rate for review purposes. This data is then inputted in pre-defined different variants of the NLM upon which the state estimations are outputted.

6.3 Part 3: Urban traffic state prediction

The prediction part of this thesis is developed within the more or less same framework as within traffic state estimation. It does however imply some modifications needed to the NLM, because the traffic state is outputted for a time in the future, beyond the most recent available traffic data. For completeness the modified framework is presented below.

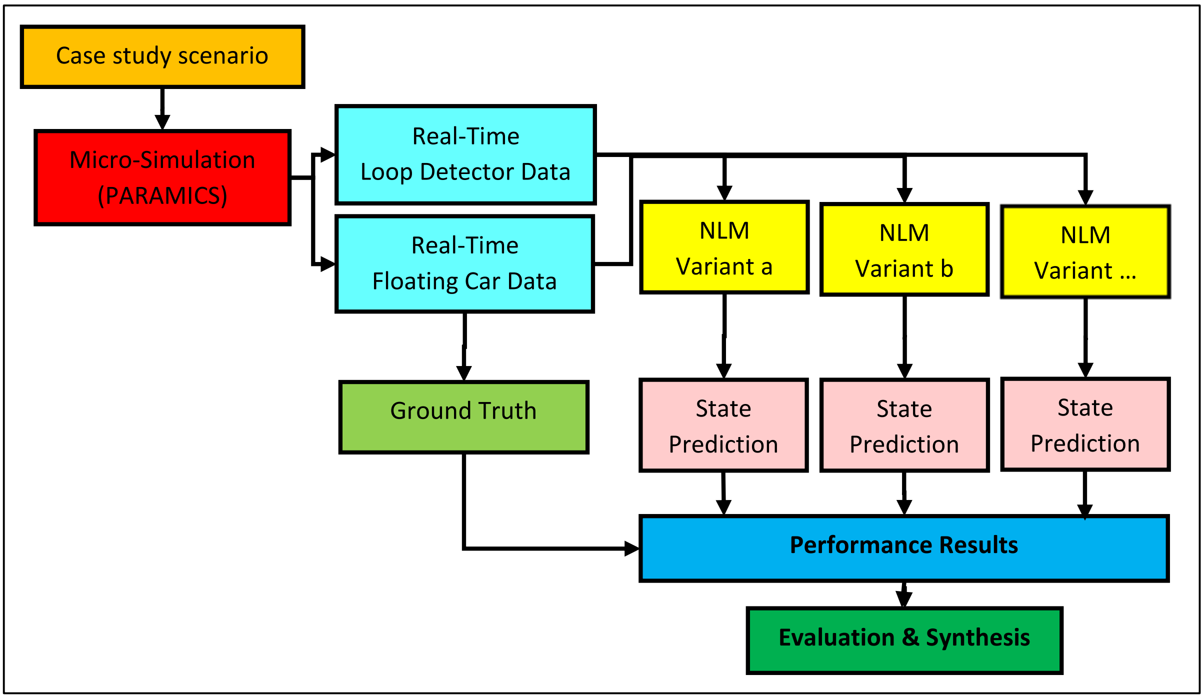

Figure 7. Traffic state prediction framework. Refined expansion of Tao (2012).

PART I

URBAN TRAFFIC STATE CASE

7 The case: Enriched Sioux Falls Scenario

The urban road network chosen for this research is the *Enriched Sioux Falls Scenario* introduced firstly by Morlok et al. (1973) as a traffic equilibrium network. It has been adopted as a benchmark and test scenario for many publications. Professor Hillel Bar-Gera from Ben-Gurion University of the Negev supplies the open source data for this network (Bar-Gera, 2014). In this chapter the reasoning for choosing this network, as well as the specific network layout, network demand and trip distribution are discussed.

7.1 Introduction

The considerations leading to the choice for the *Enriched Sioux Falls Scenario* are based on the following requirements. For this research a small-scale scenario with realistic demand and a high level of disaggregated information is needed to test and demonstrate the urban traffic state approach. Therefore it is required to find a scenario that is computationally manageable (in this case study setting, without access to optimized soft- and hardware). An additional requirement is that it should be compiled out of open source data in order to ensure both comparability and free public access. Furthermore, the network should resemble a real world scenario, but without necessarily exactly mimicking a particular place. Recently Chakirov (2014) developed an enriched version of the Sioux Falls Scenario with the purpose of mimicking socio-demographic characteristics and spatial distributions. This leads to the development of a scenario which serves for users and developers of agent-based simulation tools as a convenient test-case to serve as a test bed for new software extensions. It therefore aligns with the aim of this thesis. The test scenario used in this thesis is therefore the original Sioux Falls test scenario (Morlok, 1973) expanded in some areas with the *Enriched Sioux Falls Scenario* (Chakirov, 2014) to make it suitable for agent-based simulation.

7.2 Network layout & vehicle characteristics

The *Enriched Sioux Falls Scenario* network in this thesis consists out of the original 24 zones with 24 nodes and 76 links. As the maximum area to be simulated by PARAMICS Discovery v15.0, the microscopic modelling environment available is capped by license (available for this study) at 10 km² and the original network is of size 17 km², a 25% reduction in node distances is applied (new size = 9,6 km²). The original zone placement on the actual nodes is changed to be on the links directly adjacent to the node, allowing (random) disappearance and appearance of vehicles on links itself, mimicking real world behaviour. The length of all links is set to be equal to the Euclidian distance between nodes, yielding an average link length of 1290 meters. A full map of the network, depicting the nodes, links and road types is displayed on the next page. For the setting of physical road parameters such as road length, number of lanes and a legal speed limit, the two types of road links defined by Chakirov (2014) are copied to PARAMICS. These are highways (2x3 lanes (width: 11m) with an imposed speed limit of 70 mi/h) and urban roads (2x2 lanes (width: 7,3m) with an imposed speed limit of 30 mi/h).

Nodes between highway and urban roads are modelled as simple priority junctions in PARAMICS. The nodes between urban roads are modelled as signalized junctions, with a very simple non-actuated cycle of 30s green time for each opposing direction (allowing left turns in conflict). The cycle times are therefore independent of the actual traffic demand. For node 10 the intersection is a node between five links, and consequently has a total traffic light cycle time of 1m30s. Node 4 is modelled as a signalized junction without any designated priorities. Additionally every link leading to a signalized junction is equipped with dual inductive loop detectors.

Regarding vehicle characteristics, it is chosen (for simplicity) to model a unimodal vehicle mix consisting of passenger cars only. For the actual modelling in PARAMICS, the default values for behavioural parameters are kept e.g. awareness (norm. distr. at lvl 5), aggressiveness (norm. distr. at lvl 5), minimum headway (2.0m), mean headway (1.0s) and a simulation time steps of 0.5s.

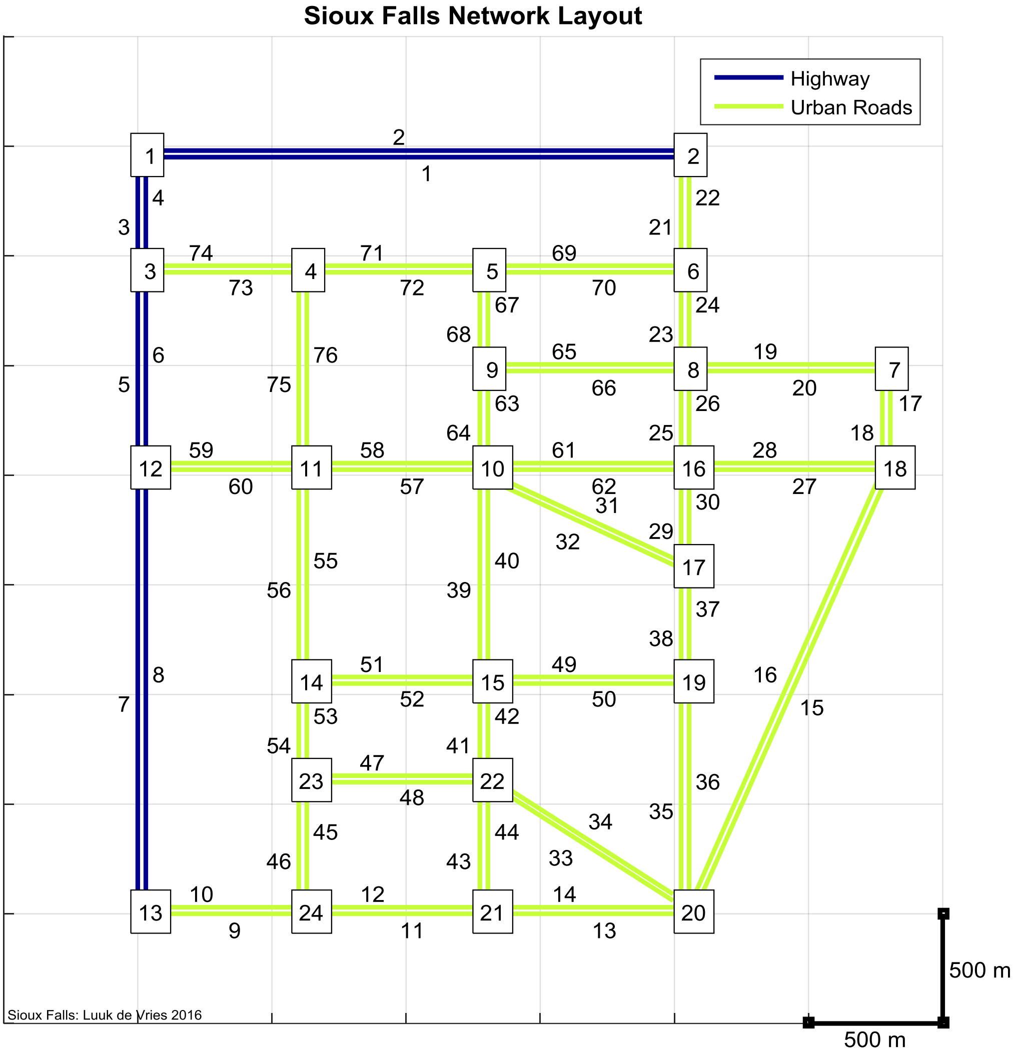

Figure 8: Sioux Falls Scenario network layout used in this research (block size is 0,5 km x 0,5 km)

7.3 Network demand

The starting point for the network demand is an OD-Matrix presented by LeBlanc *et al.* (1975). This contains a total of 360.600 trips. Chakirov (2014) re-estimates this OD-matrix by designing a household structure with a demographic profile and income distribution using data from census districts in and around the City of Sioux Falls which results in approximately half the total number of trips (d≈0,5). However with the 168.220 trips, the large number of vehicles in the network produces a drop in space-mean flow that persists from 7:30h until 12:00h and therefore yields 4,5h of congestion (Chakirov, 2014). A trial run in PARAMICS with this parameter yielded a grid lock from which the network does not recover. A more modest figure is suggested by Josefsson & Patriksson (2007) and Bar-Gera (2013) by multiplying the OD-matrix from Leblanc (1975) by 0.11 (d=0.11). A trial run in PARAMICS with this parameter still yielded some congestion and even allows the network to recover from peek hour within a more reasonable time. The network demand is therefore set at a total of 12% (due to rounding) of the trips from the OD-Matrix presented by LeBlanc (1975).

7.4 Trip distribution

On the demand side of this *Enriched Sioux Falls Scenario*, two simple activity chains are introduced to enhance the degree of realism. Instead of a flat uniform trip distribution which will assign the OD-table over time, a more complex trip distribution from Chakirov (2014) is included. The basis for this trip distribution is the two different trip activity chains. The first trip chain is: *home – work – home*. The second trip chain is: *home – other – home*. Each trip chain can then be further divided into the trip from home and the return trip.

Chakirov (2014) then adds additional constraints to each separate activity lenient on the characteristics of: opening hours, work hours and activity durations. Using simple assumptions on normally distributed departure times for work trips and uniformly distributed departure times for secondary activity trips, the network trip distribution for this research is designed with the characteristics of the trip chains are displayed the table below.

Primary (Work)	**Distribution**	**Mean**	**St. Deviation**
1. Home → Work	Normal	07:30	15 minutes
2. Work → Home	Normal	17:30	15 minutes

Table 2: Primary Activity Trip Distribution Characteristics

Secondary (Other)	**Distribution**	**Constraints**
3. Home → Activity	Uniform	From 07:45 – 19:45
4. Activity → Home	Uniform	From 09:00 – 21:00

Table 3: Secondary Activity Trip Distribution Characteristics

For practical implementation in PARAMICS, it is required to determine the profile of distribution, by defining how many of the actual trips are started within time bins of 5 minutes. Because in PARAMICS the OD-matrix implies only go trips (and not return trips), the OD-matrix is split into two tables, each therefore responsible for half the number of trips. The first table represents the first stage of network trip activities (from home). The second table represents the second stage of network trip activities (return trip). Using the same key for primary and secondary trip distribution as Chakirov (2014) uses, the ratio between 'work' and 'other' trip activities is set to 2:1. The profile inputted into PARAMICS is displayed in the figure below in terms of percentage of total trips. These values are to be treated as indicators as the decision to start the trip in PARAMICS is a random decision.

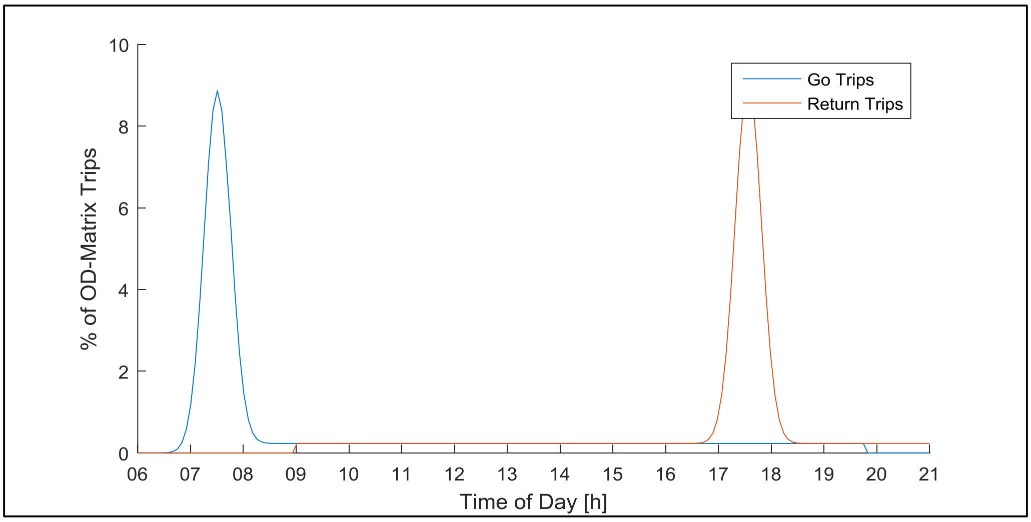

Figure 9: Network Trip Distribution as input for PARAMICS

8 Microscopic simulation

The microsimulation software used for this research is PARAMICS Discovery v15.0. It is within this software that the context defined in the previous chapter is drawn up. As previously stated the license available for this research comes with two limits; (1) the maximum traffic area of simulation is maxed at 10 km² and (2) the maximum simulation time is limited to two hours. In this chapter, the simulation characteristics are discussed resulting into 100% correct logs of FCD and ILDD for use in the performance assessment of the interpolation method.

8.1 Simulation Timeframes

The first step concerns choosing one or more interesting timeframes of the simulation. Keeping in mind that the goal of this research is to analyse the performance of the NLM framework in different traffic conditions, the simulation timeframes are chosen such that the most interesting traffic flow characterizations are covered (i.e. traffic breakdown, congestion, recovery). Additionally a timeframe in which solely free-flow traffic conditions are experienced, is added to allow assessment of model performance in free flow conditions and serve as a basis for comparison. Multiple runs of these timeframes with different starting seeds are proposed to mimic between day differences in all simulated periods (not with the intention to find e.g. an average rush hour period). As simulations are time costly and also generate a sizable amount of data, selecting three timeframes of two hour length, and running each simulations a total of 5 times for each of the timeframes is suggested to yield a reasonable trade-off between accuracy and research time.

The rush hour simulation timeframes are set around the mean time of primary trip activities to capture the traffic breakdown, congestion and recovery phases. The two hours of simulation time available, are distributed evenly along the time at which the maximum number of trips are suspected to take place. For the third timeframe an arbitrarily period between morning and evening rush hour is chosen. All periods are considered separately in this research. The simulated timeframes are:

(1) Morning Rush Hour	From: 06.30h – 08.30h	Peak expected at: 07:30
(2) Evening Rush Hour	From: 16.30h – 18.30h	Peak expected at: 17:30
(3) Outside (of) Rush Hour	From: 12:30h – 14:30h	

8.2 Simulation Characteristics

Each period up for simulation is then run a total of five times, to mimic daily variation in traffic due to PARAMICS' random decision process regarding the release of vehicles, vehicle parameters and choice processes. Each simulation therefore has a unique random generated seed. The output of each simulation are the log files which contain the FCD and ILDD. These make up the 'traffic data' used throughout this research. While the simulation itself ran quite fast, averaging 60 times the real time speed, the time consumption for the transferring and saving of the generated traffic data required up to 3 hours per run. A summary of the simulation characteristics and corresponding data sizes are shown in the table below.

Simulation Type	Timeframe	Runs (Days)	$\sum$ **Size of data**
Morning Rush Hour	06:30 – 08:30	5	6,89GB
Evening Rush Hour	16:30 – 18:30	5	9,71GB
Low Hours	12:30 – 14:30	5	0,35GB

Table 4: Simulation characteristics

For notation and reference throughout this research, while mimicking consecutive days by running each simulation, the word 'run' in combination with the timeframe is used to pinpoint to one of the five simulated days of the simulation. E.g. M=3 refers to the 3rd simulation of a timeframe.

9 Simulated traffic data

The result of the preceding chapter can be described as a huge amount of raw data and can be called (while generally even larger sizes of data apply) the big data available in this research. The big data consists of non-aggregated FCD per time step of 0.5s and aggregated ILDD per interval of 60s. In this chapter the mathematical techniques and reasoning behind the narrow selection of useful data from the big data is discussed. The resulting data is stored in either the so called (historical) database, which is used for the interpolation method later on in this research and for determination of the ground truth in the next chapter. This chapter starts off by defining the notation and consecutively the traffic flow variables used in this research.

9.1 Basic Notation

The list below describes the notation of commonly used variables in this research. This list of variables will be expanded as new notations are used and explained.

$U(X,J)$ =	The urban network used in this research	[-]
$x \in X$ =	Link number in the urban network (edge)	[-]
$j \in J$ =	Junction number in the urban network (node)	[-]
$r(x)$ =	Lanes on link x	[-]
$l(x)$ =	Length of link x	[mi]
$v_{max}(x)$ =	Speed limit on link x	[mi/h]
n =	Number of vehicles	[-]
a=	Unique vehicle ID	[-]
tt_a =	Travel time of vehicle a	[h]
s_a =	Distance travelled by vehicle a	[mi]
$\bar{v}_a$ =	Mean velocity of vehicle a	[mi/h]
t =	Timestamp of the time interval	[min]
Δt_d =	Default time interval length (1/60h)	[h]
M=	Run number (between 1 and 5)	[-]
A =	Historical traffic database	[-]

9.2 Traffic flow variables

The goal within this research is to provide an as accurately as possible reconstruction of the traffic state in the previously mentioned urban network at any given time. This however creates a problem as there are both an infinite amount of locations and an infinite amount of times in the simulated network. It is infeasible to compare the output of the neighbour link method with the ground truth without discretizing the traffic flow data both spatially and temporally. A practical spatial aggregation per link is proposed. Additionally a temporal one minute interval is adopted as real-life ILDD and FCD are aggregated for one minute intervals. This means that this research is aimed towards reconstructing (as accurately as possible) the traffic state on each link in the network in intervals of one minute.

The traffic flow variables of interest therefore should be dependent on both space (length of the link) and time (an interval of 60 seconds). The traffic flow variables of interest are the traffic speed, traffic flow and traffic density. Notably in other research the travel time is deemed of interest as well (e.g. Esaway, 2012) but as it can be directly derived from the traffic speed it is omitted in this research. In general these traffic flow variables are defined solely for either space or time (i.e. Treiber, 2012). This means that in this research the traffic flow variables are expanded as to write the traffic-state estimators as a function of space and time. For this purpose spatiotemporal aggregation is required.

This process of spatiotemporal expansion of the traffic flow variables is explained by using the figure below adopted from Treiber (p.9, 2012) in which trajectories of vehicles are plotted. Identified can be 2 links ((x_1, x_2): [0m-100m] and [100m-200m]) and 2 time intervals (($\Delta t_1, \Delta t_2$): ([0s-60s] and [60s-120s], which are assumed to represent the intervals and sections of interest.

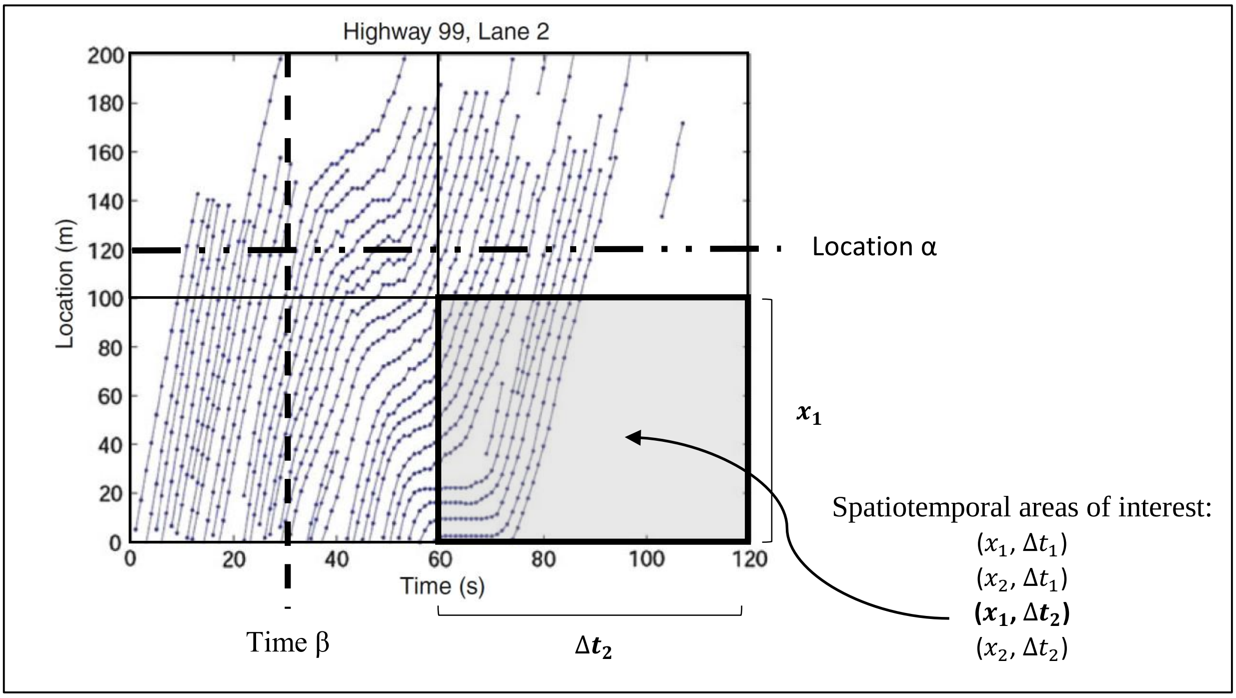

Figure 10: Trajectories from vehicles on highway 99 (Treiber, 2011, p.9)

The first expansion that has to be made, affects the mean speed. The time mean speed describes the arithmetic mean of speeds of vehicles passing a point (e.g. location α). The space mean speed describes the average speeds over a length of roadway (e.g. at time β). A combination of both is required to find the spatiotemporal mean speed. Secondly the density is generally expressed as the number of vehicles over a stretch of roadway (e.g. at time β), which has to be expanded to a time interval. Lastly the flow is generally expressed as the rate at which vehicles pass a fixed point (e.g. at location α). Therefore expansion over a length of road is required. Bottom-line is that that where normally these variables are only integrated over time or space (to describe the situation at either a point (α) or at a time (β)), they are to be integrated for both in this research, to reflect the required spatiotemporal areas. The three modified traffic flow variables are presented in the table below.

Spatiotemporal Traffic flow variable	Description	Unit
$\boldsymbol{v(x,t)}$	Spatiotemporal mean speed	[mi/h]
$\boldsymbol{p(x,t)}$	Spatiotemporal mean density	[veh/mi]
$\boldsymbol{q(x,t)}$	Spatiotemporal mean flow	[veh/h]

Table 5: Modified traffic flow variables.

9.2.1 The spatiotemporal mean speed

The mean speed in [mi/h] on link x at interval t: $v(x,t)$ is defined as follows:

$$v(x,t) = \frac{\sum_{a=1}^{n(x,t)} s_a(x,t)}{\sum_{a=1}^{n(x,t)} tt_a(x,t)} \qquad [1]$$

In which:

$n(x,t)$ =	Number of Vehicles measured within the interval t on link x	[-]
$s_a(x,t)$ =	Distance travelled by the vehicle a within the interval t on link x	[mi]
$tt_a(x,t)$ =	Time spent by the vehicle a within the interval t on link x	[h]

! Note: where both x and t are actually referring to a location interval (Δx) and a time interval (Δt) respectively, the delta notation is omitted in this research.

9.2.2 The spatiotemporal mean density

The mean density [veh/mi] on link x at interval t: $p(x,t)$ is defined as follows:

$$p(x,t) = \frac{\sum_{a=1}^{n(x,t)} tt_a(x,t)}{\Delta t_d * l(x)} \qquad [2]$$

In which:

$n(x,t)$ =	Number of Vehicles measured within the interval t on link x	[-]
$l(x)$ =	Length of link x	[mi]
Δt_d =	Default time interval length (1/60 h)	[h]
$tt_a(x,t)$ =	Time spent by the vehicle a within the interval t on link x	[h]

! Note: In this research the density is edited to be in format of [veh/mi/lane] by dividing the result by r(x) to allow fair comparison between different types of links.

9.2.3 The spatiotemporal mean flow

The mean flow [veh/h] on link x at interval t: $q(x,t)$ is defined using the relationship between speed, density and flow, as follows:

$$q(x,t) = p(x,t) * v(x,t) \qquad [3]$$

In which:

$q(x,t) =$	Average flow on link x, during time interval t	[veh/h/lane]
$p(x,t) =$	Average density on link x, during time interval t	[veh/mi/lane]
$v(x,t) =$	Average speed on link x, during time interval t	[mi/h]

! Note: In this research the flow is also edited to be in format of [veh/h/lane] to allow again fair comparison between different link types.

9.3 Floating car data

The floating car data (FCD) used in this research consists of reported vehicle positions, direction of driving, velocities and times from dedicated vehicle probes. As all vehicles in the PARAMICS simulation are theoretically equipped with GPS and a communication link for transferring this data, the FCD generated is a 100% accurate and therefore a full reflection of the traffic states in the urban network. For the rush hour periods a single run contains up to 22 million of logged data entries.

To show how this data is treated and transferred into one minute aggregated intervals and stored into a traffic database, an example is provided. The table below shows 20 seconds of the (sorted) FCD output of a vehicle traversing on link ’20:21’, and coming to a full stop at 9 meters before the end of the link, because of a red traffic light.

[Timestamp]	[Link]	[Tag Number]	[PosX]	[PosY]	[Acceleration]	[Speed]	[D_Linkend]
23.481.000	'20:21	1	-672.32	-1994.52	0.000	32.358	51.98
23.481.500	'20:21	1	-679.55	-1994.52	0.000	32.358	44.75
23.482.000	'20:21	1	-686.11	-1994.52	-2.685	29.355	38.19
23.482.500	'20:21	1	-692.00	-1994.52	-2.712	26.322	32.30
23.483.000	'20:21	1	-697.20	-1994.52	-2.737	23.261	27.10
23.483.500	'20:21	1	-701.71	-1994.52	-2.758	20.177	22.59
23.484.000	'20:21	1	-705.52	-1994.52	-2.771	17.078	18.78
23.484.500	'20:21	1	-708.65	-1994.52	-2.769	13.980	15.65
23.485.000	'20:21	1	-711.09	-1994.52	-2.740	10.916	13.21
23.485.500	'20:21	1	-712.86	-1994.52	-2.656	7.946	11.44
23.486.000	'20:21	1	-714.03	-1994.52	-2.460	5.194	10.27
23.486.500	'20:21	1	-714.67	-1994.52	-2.052	2.898	9.63
23.487.000	'20:21	1	-715.14	-1994.52	-0.738	2.072	9.16
23.487.500	'20:21	1	-715.25	-1994.52	-1.853	0.000	9.05
23.488.000	'20:21	1	-715.25	-1994.52	0.000	0.000	9.05
23.488.500	'20:21	1	-715.25	-1994.52	0.000	0.000	9.05
23.489.000	'20:21	1	-715.25	-1994.52	0.000	0.000	9.05

Table 6: Example of some lines of raw FCD logs

The variables mentioned in the headers of the above table are:

[Timestamp]	Number. Indicates the time of measurement since 0h00 in 0.5s increments.
[Link]	String. Indicator of the link traversed on.
[Tag Number]	Number. Unique number connected to a vehicle (license plate).
[PosX]	Number. X coordinate of vehicle position at that measurement in meters.
[PosY]	Number. Y coordinate of vehicle position at that measurement in meters.
[Acceleration]	Number. Indicates the acceleration in the previous 0.5s. In mi/s².
[Speed]	Number. Indicates the average speed of the previous 0.5s. In mi/h.
[D_Linkend]	Number. Indicates the remaining distance to the end of the link in meters.

The result of the aggregation of FCD to one minute intervals, should allow for future (easy) calculation of the; space mean speed, space mean density and space mean flow per link, per time-interval. The prerequisite is that the FCD is aggregated such that these traffic flow variables can be calculated. This means that a total of 5 relevant variables are selected for storage, which must answer the following questions.

1) Variable: [Tag Number]	What is the vehicles’ unique id?
2) Variable: [Timestamp]	In which time-interval(s) did the vehicle ‘drive’?
3) Variable: [Link]	Which link(s) were visited by the vehicle?
4) Variable: [Distance]	What distance did it travel on the link within(s) the time-interval(s)?
5) Variable: [Time]	How long did it travel on the link(s) within the time-interval(s)?

While some of these questions can be answered (more or less) directly from the raw FCD, others require some intermediate steps (e.g. distance travelled is the sum product of *speed* x *time*). The pseudo-code displayed below – with complexity $(O = nxt)$ – summarizes this process.

```
Load 'FCD'
For each unique vehicle id (a)
    For each time-interval (t)
        For each link (x)
            Add into the database [x,t]:
                the vehicles id (a)
                distance travelled (s)
                time travelled (tt)
            End
        End
    End
End
```

Figure 11: Pseudo-Code of the aggregation process of FCD

9.4 Inductive loop detector data

In this case study inductive loop detector data (ILDD) is introduced as a second data source. It is assumed that on all urban roads in the network a double inductive loop detector is present. The data from these loops is aggregated for a 60 second interval. The exact location is displayed by [x.1] in figure 12. It must be noted however, that in common real-world practice these assumptions with regards to location, aggregation and equipment are likely to not satisfied.

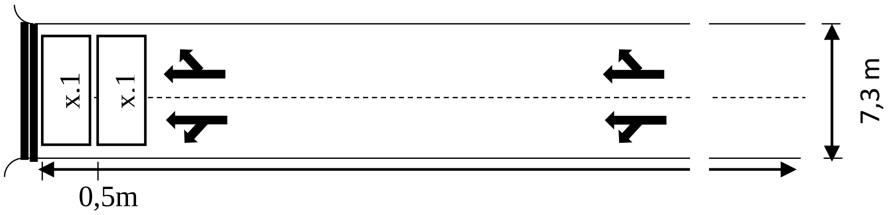

Figure 12: Dual inductive loop detector location

The usability of ILDD is limited due to the nature of the spatiotemporal traffic flow variables (Tao, 2011). Whereas ILDD can be considered perfectly able to deliver aggregated point measurements, the location of the loop detector makes it very questionable if the measured speeds at a stop-and-go location are representative for the mean traffic variables on a whole link. Other factors of influence are; cycle time of the link, the link section length and exact detector location. A preliminary investigation is executed to verify which variables of ILD's are suitable for further use.

This investigation starts with a comparison of the mean speed from FCD (considered the ground truth) and the time mean velocity from ILD. Assumed is that the installed detectors aggregate within one minute intervals. Arguably 5 minute aggregation yields a different comparison. Plotting the one minute aggregated detector speed (point speed) and the one minute aggregated ground truth for all urban links (with the average length of 1.300 meters) during 5 runs of off-peak hours, generates the figure presented on the next page. While the real mean speed of a link generally stays above 15mi/h, the time mean speed at the detector location is seriously biased to underestimation. The spread of the data away from the trend line, as well as the trend line's position render the ILD's speed measurements unusable for data fusion. Further research on how to make ILDD more representative for the space mean speed determination, lies outside the scope of this research.

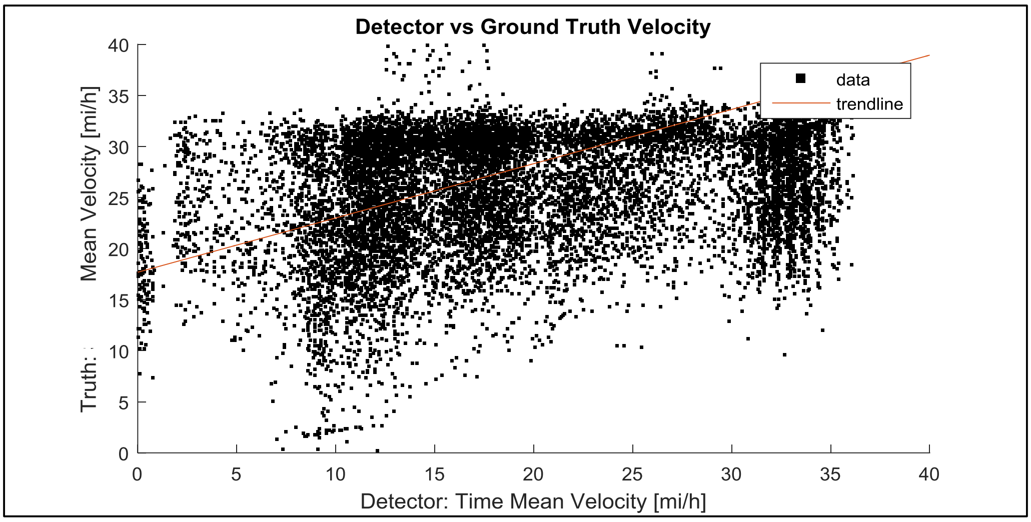

Figure 13: Time mean velocity and true mean velocity scatter plot, using FCD and ILDD

Additionally it is trivial that the true mean density and mean density (derived from occupancy) will be unrepresentative as well. The location of the ILD at the front of the queue is obviously not representative for the (average) occupancy of the full road segment.

The investigation continues with a comparison of the mean flow from FCD (again the ground truth) and the local point flow/intensity/volume from the ILD. A same approach can be taken such that the aggregated detector flow (point flow) and ground truth are plotted and displayed in the figure below. This time the trendline starts around the origin, but reveals a serious bias towards overestimation with a factor of approximately 1.5. Four reasons are identified to be explanatory of this bias; first of all, due to rounding and sampling the ILD flows are multiples of 60, and are more capricious then flow determined from the FCD. Secondly the principle of flow conservation on the urban links, does not hold in this network as in the zone placement in PARAMICS, mimics the on-street-parking-behaviour. It is very plausible that vehicles join and disappear halfway the link, and get fully counted (or not at all) at the ILD's location. Thirdly the length of the link itself influences the bias. Lastly the cycle time of the traffic lights plays an important role, as time mean flow is zero during the red light phase, yet traffic may still be able to flow (up to the queue). Again it is concluded that ILDD's flow measurements are not useful for mean flow determination.

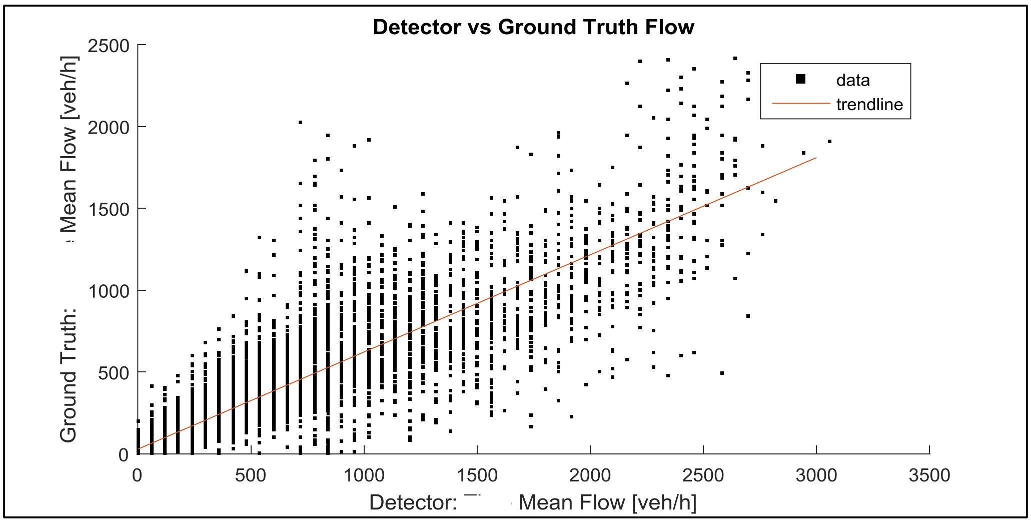

Figure 14: Detector mean flow and true mean flow scatter plot, using both FCD and ILDD

There is however one relationship between FCD and ILDD available which will be used. Comparison between FCD and ILDD point flow counts for exactly the ILD's location, makes it possible to calculate the *local FCD penetration rate* $\leftthreetimes= n/\tilde{n}$ on that link. This rate is needed as obviously with the coverage rate of FCD is less than 100%, the number of vehicles $\tilde{n} \leq n$ equipped with GPS sensors is smaller than the actual vehicles on the link. Using this rate to compensate yields the mean density. Therefore another operation on the FCD is added, by checking whether the vehicle leaves the link within a time step. The updated pseudo-code becomes:

```
Load 'FCD'
For each unique vehicle id (a)
        For each time-interval (t)
                For each link (x)
                        Add into the database [x,t]:
                                the vehicles id (a)
                                distance travelled (s)
                                time travelled (tt)
                                did vehicle pass detector (q)
                        End
                End
        End
End
```

Figure 15: Final pseudo-code of the aggregation process of FCD and ILDD

9.5 Traffic database

The result of running the code presented in the previous paragraph, is a 3D-database containing both 60s aggregated FCD and ILDD. For each 60s time interval of the simulation (t) and for each link (x) the relevant data from the FCD and ILD belonging to that specific link within that time step can be found. Whenever a vehicle is detected at the time interval, on that specific link, it adds an array consisting of the four (6 minus *time* and *link*) previously mentioned variables in depth of the database. The depth of each cell of the database is therefore equal to the number of vehicles on the link x during the interval t. An illustrative example of a selection of links in the traffic database (A) is shown below.

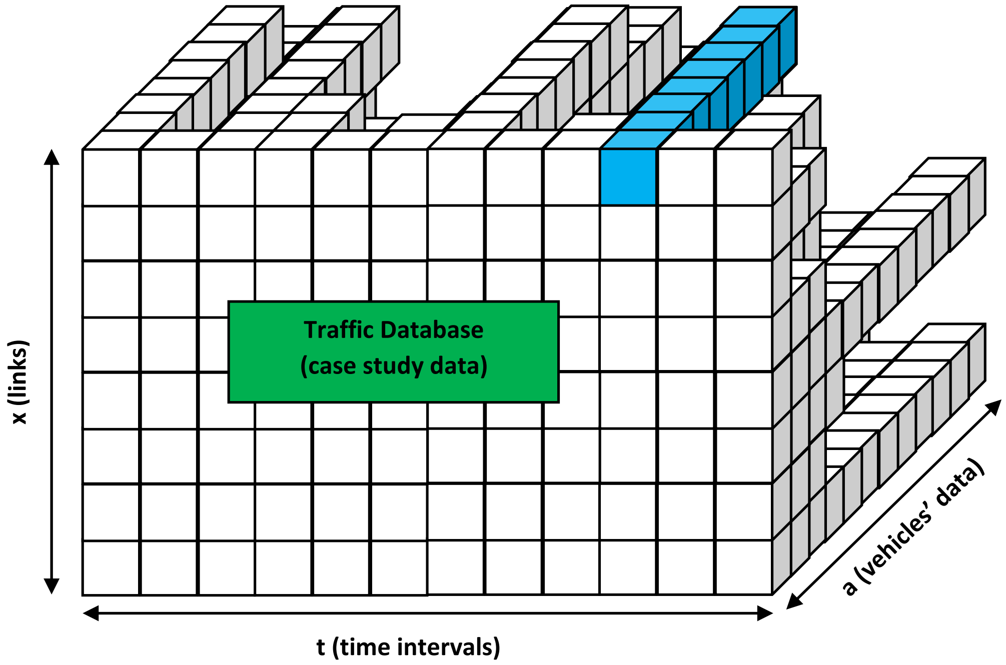

Figure 16: Illustrative example of a section of the traffic database

An example of the array within the marked cells of the illustrative traffic database presented on the previous page, where headers are added for convenience, is shown below.

a	s_a	tt_a	q_end_a
12465	0,0771	0,0028	1
12500	0,1279	0,0047	0
11675	0,2877	0,0093	0
12545	0,1159	0,0043	1
12558	0,0888	0,0042	1
11551	0,0101	0,0010	1
11960	0,0088	0,0006	0
11865	0,0121	0,0007	0
12673	0,1148	0,0040	0
12706	0,0355	0,0015	1
11227	0,0113	0,0007	0

Table 7: Example of the traffic database for x=16 at time interval t=58

The variables stored are:
a = Unique vehicle ID [-]
$s_a(x,t)$= Distance travelled on link x, within time interval t [mi]
$tt_a(x,t)$ = Time spent on link x, within time interval t [h]
q_end_a= Does vehicle (a) on link x, pass the detector within time t [-]

! When no vehicles travel on the link x at time interval t, the (sub-)database is empty.

Additionally three copies of this traffic database are made, as morning rush hour, evening rush hour and outside rush hour are stored into separate databases. The unmodified database serves as the ground truth database for the next chapter. Additionally two database copies are made in which 90% and 95% of the vehicles are randomly removed (by randomly drawing a value between 0 and 1 and removing the vehicle if this value is >0,10 or >0,05) respectively. The latter two databases are used to assess the performance of the interpolation method. The resulting selection is presented below. Assumed for this selection is that there is no sample bias present in this urban network, which might not be true for real world conditions. Examples of possible reasons for bias in real world conditions are; (1) Some road types might be more frequently travelled by GPS equipped vehicles; (2) longer trips might be logged more; (3) certain destinations might attract more GPS equipped vehicles and (4) at certain times the coverage rate is more than average, while at other times less.

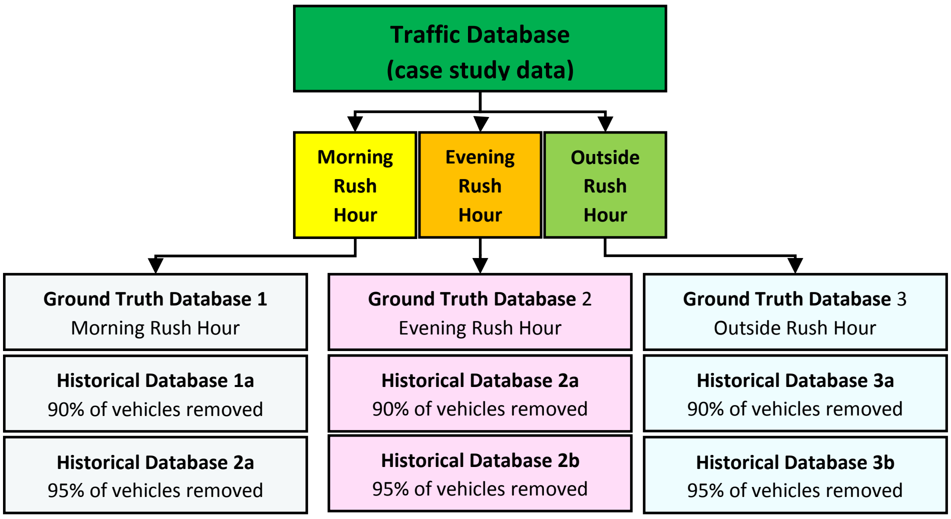

Figure 17: The 9 databases with traffic data, used within this research.

10 The ground truth

Now that a fully aggregated traffic database is set-up, it is time to derive the ground truth of this research. This means calculating from the previously made ground truth database the complete set traffic flow variables for each time interval describing the average traffic state on a road link. It therefore consists of the true mean speed (v), mean density (p) and mean flow (q), per link (x), per time interval (t) of 60 seconds. This paragraph focuses on applying the previously defined formulas to the unfiltered ground truth database.

10.1 Mean speed

The mean speed is determined by applying [1] on the data. The form is trivially distance travelled divided by time travelled. For completeness [1] is repeated:

$$v_{GT}(x,t) = \frac{\sum_{a=1}^{n(x,t)} s_a(x,t)}{\sum_{a=1}^{n(x,t)} tt_a(x,t)} \qquad [4]$$

In which:

$v_{GT}(x,t) =$ Ground truth mean speed on link x, at time interval t [mi/h]
$n(x,t) =$ Number of vehicles, traversing on link x within time interval t [-]
$s_a(x,t)$= Distance travelled on link x, within time interval t [mi]
$tt_a(x,t) =$ Time spent on link x, within time interval t [h]

! When there are no vehicles, i.e. $\sum_{a=1}^{n} tt_a(x,t) = 0$*, then* $v(x,t) = v_{max}(x)$*.*

An example of the ground truth velocity traffic state while in the midst of evening rush hour is depicted in Figure 18.

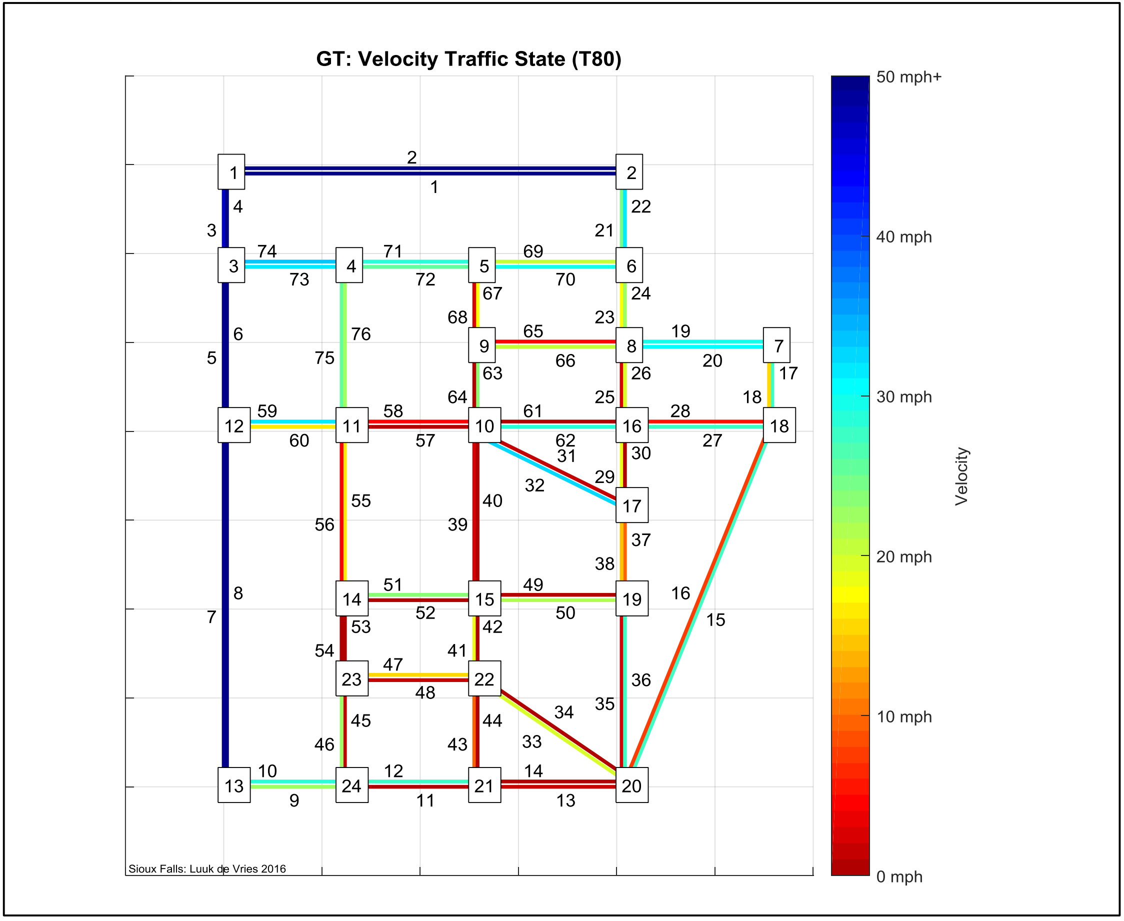

Figure 18: Velocity traffic states in evening rush hour

10.2 Mean density per lane

The mean density is determined by only modifying [2] as to include the lane divisor r(x):

$$p_{GT}(x,t) = \frac{1}{\Delta t_d * \ l(x) * r(x)} \sum\nolimits_{a=1}^{n(x,t)} tt_a(x,t) \qquad [5]$$

In which:

p_{GT} =	Ground truth mean density	[veh/mi/lane]
$n(x,t)$ =	Number of Vehicles measured within the interval t on link x	[-]
$l(x)$ =	Length of link x	[mi]
$tt_a(x,t)$ =	Time spent by the vehicle a within the interval t on link x	[h]
Δt_d =	Default time interval length (1/60 h)	[h]
$r(x)$ =	Number of lanes on link (x)	[-]

! When there are no vehicles, i.e. $n(x,t) = 0$ *then* $p_{GT}(x,t) = 0$

An example of the ground truth density traffic state in the same evening rush hour is depicted in Figure 19.

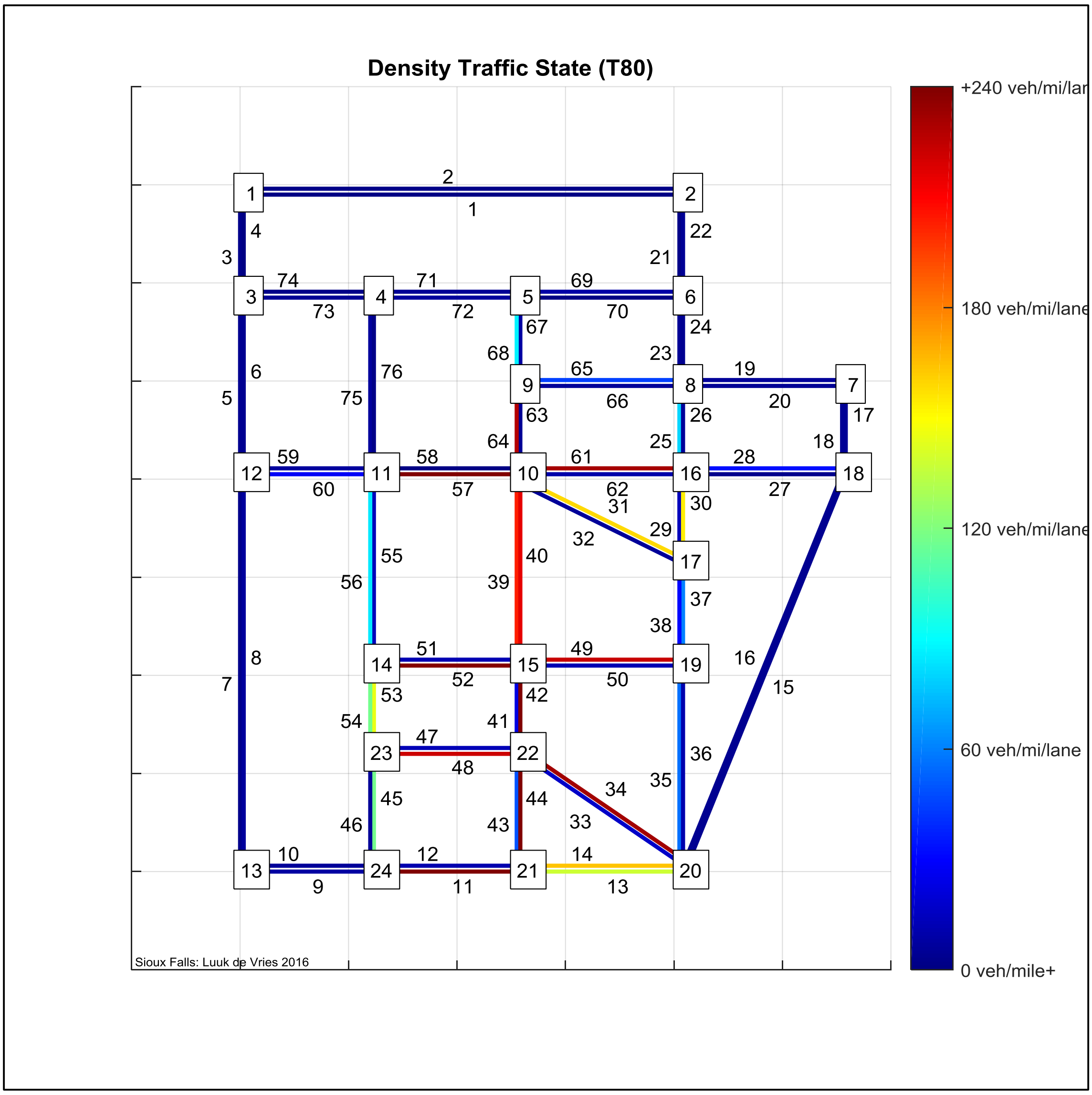

Figure 19: Density traffic states in evening rush hour.

10.3 Mean flow per lane

The mean flow in [veh/h/lane] can then be determined by just applying formula [3].

$$q_{GT}(x,t) = p_{GT}(x,t) \times v_{GT}(x,t) \qquad [6]$$

In which:

$q_{GT}(x,t) =$	Ground truth average flow on link x, during time interval t	[veh/h/lane]
$p_{GT}(x,t) =$	Ground truth average density on link x, during time interval t	[veh/mi/lane]
$v_{GT}(x,t) =$	Ground truth average speed on link x, during time interval t	[mi/h]

An example of the ground truth flow traffic state in the same evening rush hour is depicted in Figure 20.

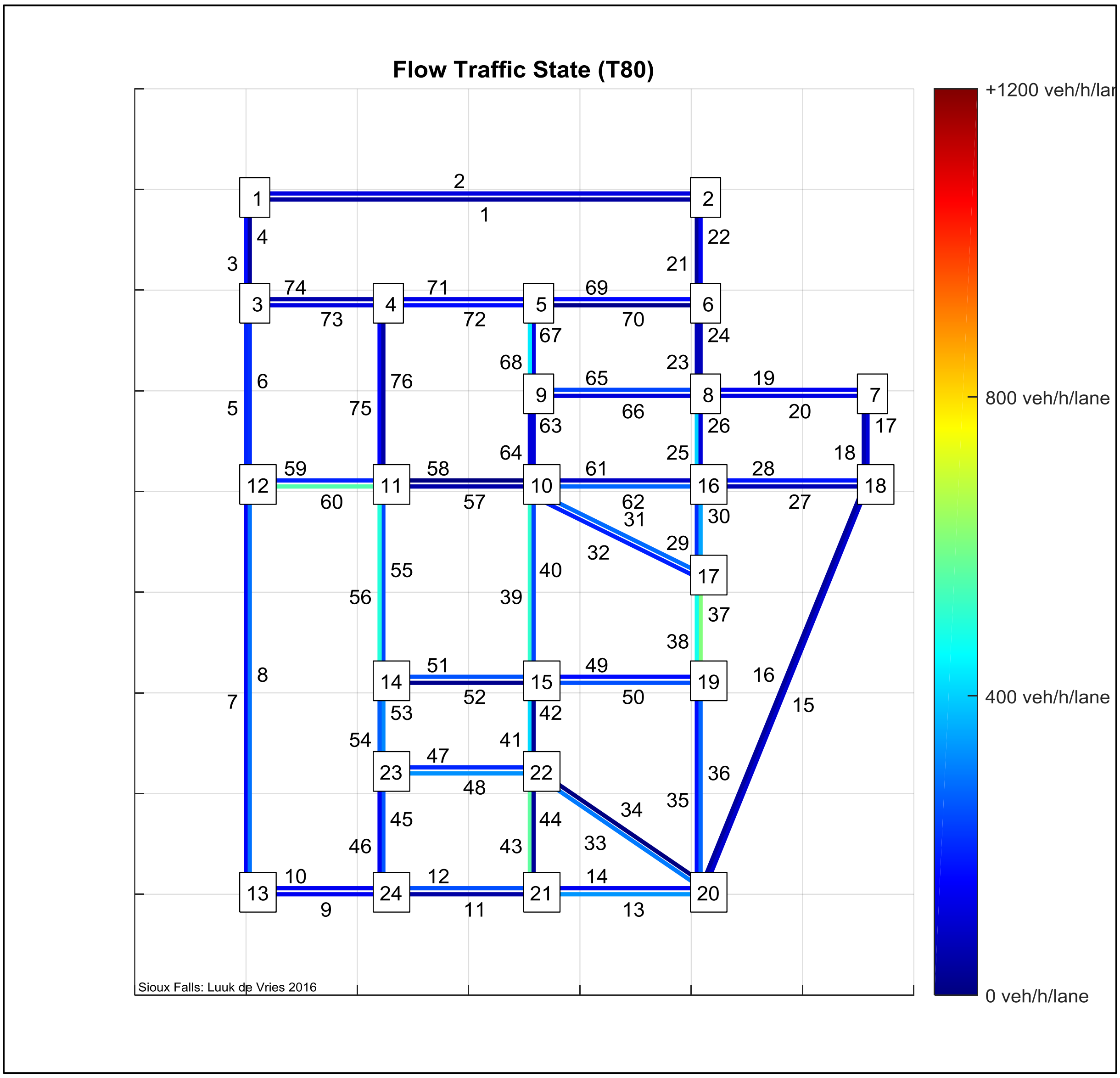

Figure 20: Flow traffic states in evening rush hour.

10.4 Ground truth example

To check whether the output presented in the three plots of the velocity, density and flow traffic states within the urban network is correct, an example is presented in this paragraph to show how the ground truth is calculated from the FCD which draws up the ground truth database. For practical reasons, this is done for link 17 (x=17), which can be found at the far right of the network, just under node 7.

10.4.1 Raw printout of the ground truth database

The (ground truth) database, reveals for x=17 at time t=80, the following raw printout of 'recorded' traffic data (displayed in transposed form, due to its size) in the table below.

$\boldsymbol{a}$	**14013**	**14003**	**13028**	**13654**	**14095**	**14167**	**13440**	**12494**	**14206**	**14212**	**14259**
$\boldsymbol{s_a(17,80)}$	0,065	0,060	0,010	0,289	0,116	0,174	0,010	0,009	0,174	0,130	0,090
$\boldsymbol{tt_a(17,80)}$	0,002	0,002	0,001	0,009	0,004	0,006	0,001	0,001	0,006	0,005	0,003
$\boldsymbol{q_{end_a}(17,80)}$	1	1	0	1	1	1	0	0	1	1	0

Table 8: Raw printout of the ground truth traffic database on link 17 at time 80 (during evening rush hour).

The variables presented in this printout are:

a =	Vehicle id /Tag number	[-]
$s_a(17{,}80)$ =	Distance travelled on link 17, within time interval 80	[mi]
$tt_a(17{,}80)$ =	Time spent on link 17, within time interval 80	[h]
$\boldsymbol{q_{end_a}}(17{,}80)$ =	Link end passed, within time interval 80	[-]

10.4.2 Mean speed, density and flow calculation

The mean speed, density and flow can be straightforwardly calculating by applying [4], [5] and [3] respectively, yielding the following results. These confirm with their given colours:

$$v_{GT}(x,t) = \frac{\sum_{a=1}^{n} s_a(x,t)}{\sum_{a=1}^{n} tt_a(x,t)} = \frac{1{,}002}{0{,}036} =$$ **27,833 mi/h**

$$p_{GT}(x,t) = \frac{\sum_{a=1}^{n} tt_a(x,t)}{\Delta t_d * l(x) * r(x)} = \frac{0{,}036}{\frac{1}{60} * 0{,}2796 * 2} =$$ **3,863 veh/mi/lane**

$$q_{GT}(x,t) = \bar{\rho}(x,t) \times \bar{v}(x,t) = 3{,}863 * 27{,}833 =$$ **107,519 veh/h/lane**

10.5 Ground truth storyboard

This chapter is concluded by presenting the storyboard which shows the development of traffic and congestion in the simulated Sioux Falls scenario. The figures below show the density traffic state per link for the time intervals; 20, 40, 60, 80, 100 and 120 minutes in the morning rush hour simulation respectively. It is at around t=60 that congestion appears on link 40, below the centre node 10, which then spreads further downstream through the network. At t=100 recovery is visible.

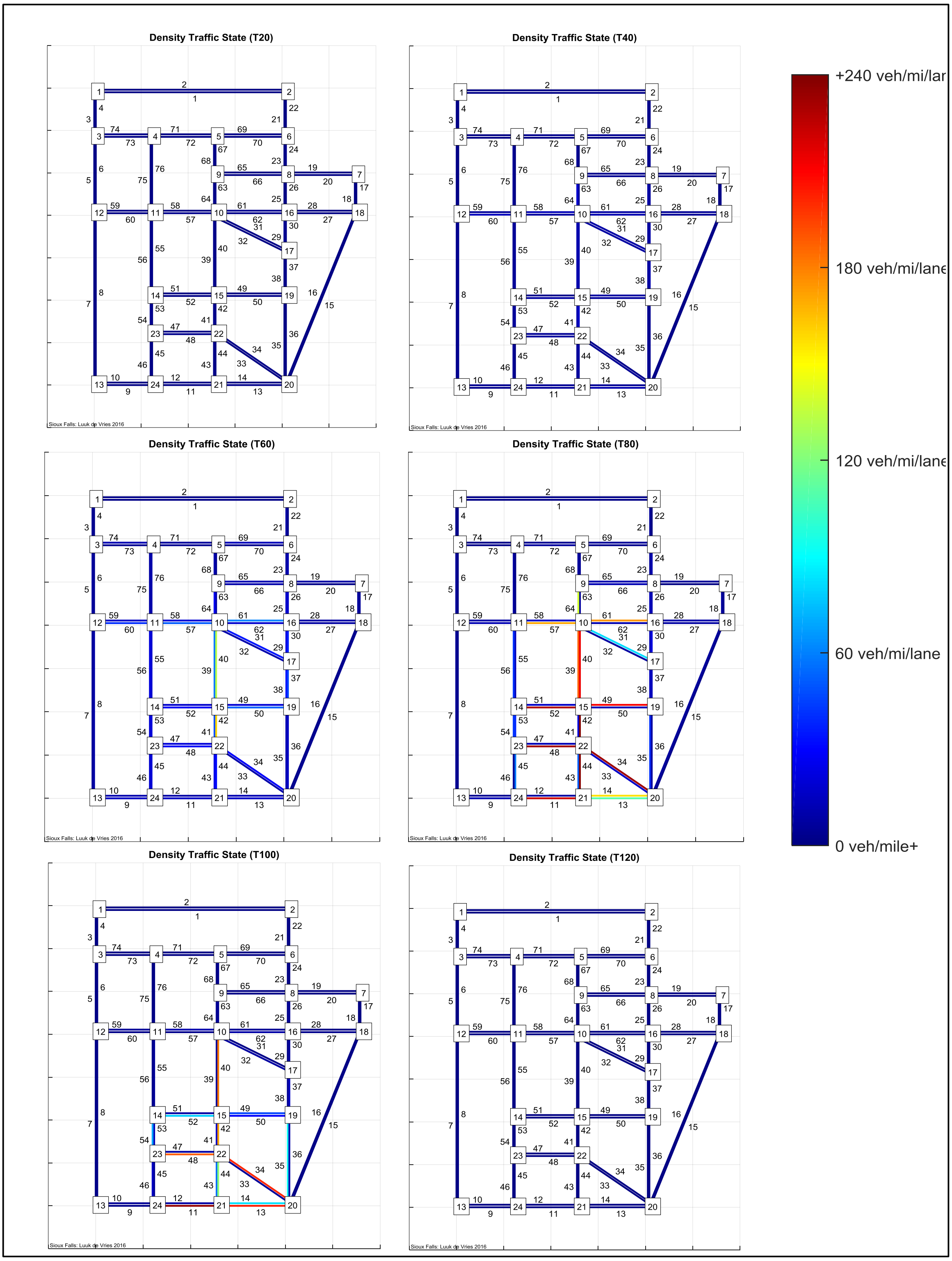

Figure 21: Density traffic state storyboard of morning rush hour (for 20 minute increments).

The evolution of traffic in the evening rush hour is shown in the next figure. As the traffic demand is higher than in morning rush hour, the spread of congestion from node 10 and link 40 is much larger. For clarity the storyboard for the evening rush hour is added below. As in outside-rush hour little to none congestion appears (apart from the temporary queues for the traffic lights), the outside-rush hour storyboard is omitted.

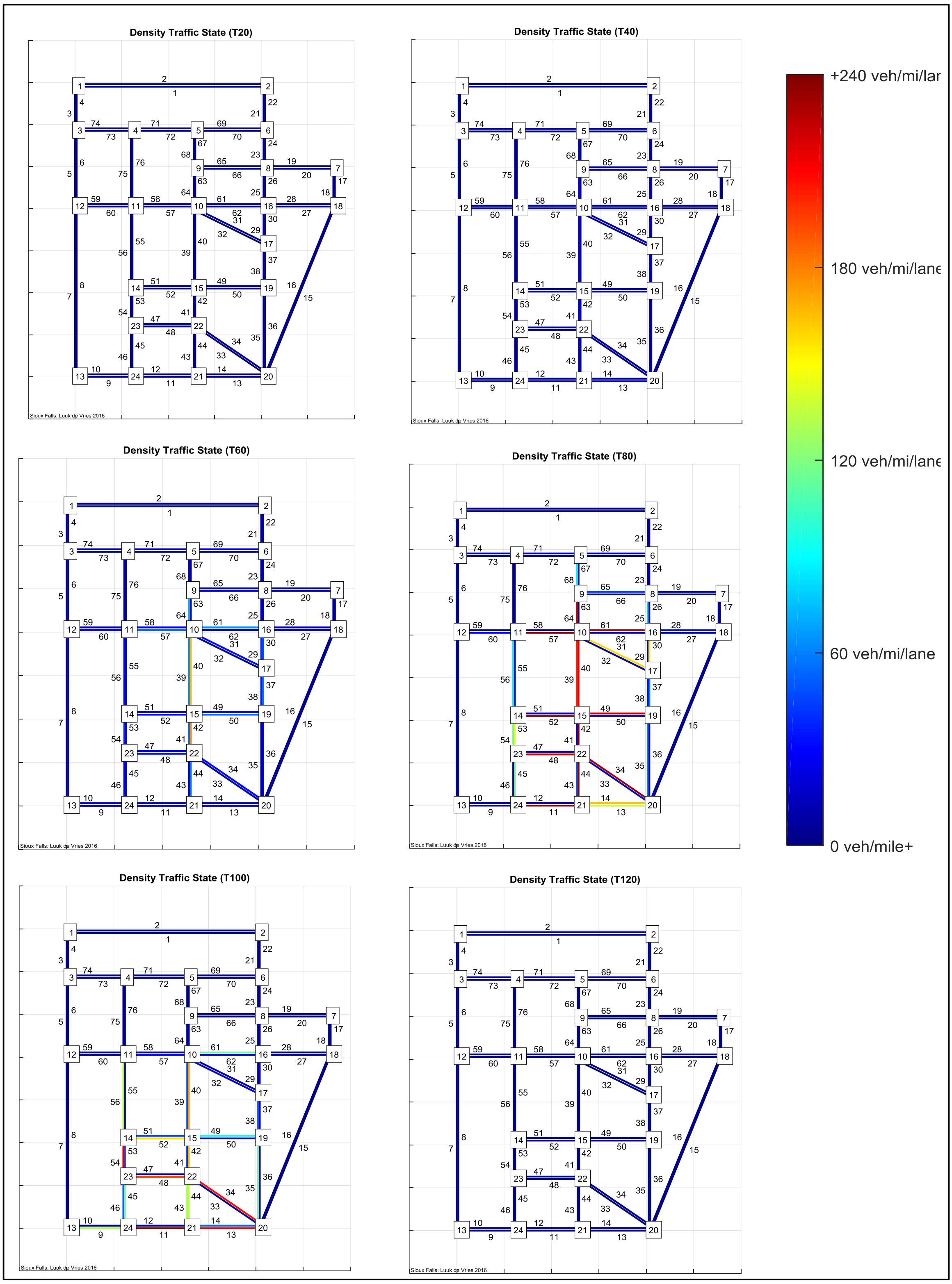

Figure 22: Density traffic state storyboard of evening rush hour (for 20 minute increments).

●▶

PART II

URBAN TRAFFIC STATE ESTIMATION

11 NLM estimation framework

In this chapter, the interpolation method used to derive the traffic state estimation is discussed in more detail. It's branded the "Neighbour Link Method" (NLM) as is based on the idea that by using patterns in historical traffic data, the current traffic state of links can be used as indicators for the traffic state on neighbouring links. However, these segments are not necessarily easily found (Esawey, 2012). They can be the simple adjacent preceding and succeeding segments, parallel segments, intersecting segments or even segments on the opposite side of the network.

11.1 NLM structure

The goal of the Neighbour Link Method is therefore to not only find the appropriate neighbourhood for each link, but also to derive the robust, complete and accurate picture of the urban traffic state on all links in the network. It does so by fusing and enriching both historical traffic data and neighbouring link data to compensate for the less accurate estimations derived from a limited sample rate of FCD and the limited or even absent coverage of ILDD.

The NLM therefore takes the approach of a 'black box' with an artificial neural network and regression model at its core. The artificial neural network aims to find not only the neighbouring links of each link, but also a suitable weighting to express the solidity of the relationship. The global structure of the NLM is therefore based on two of these dynamic artificial layers. The first layer is the neighbourhood layer, which identifies the links which are most likely showing the same traffic behaviour over time. The second layer is the weight layer, which optimizes weights for the neighbouring links such that the relation of copying behaviour between links is quantified. The global structure of the NLM is for completeness again displayed in the figure below.

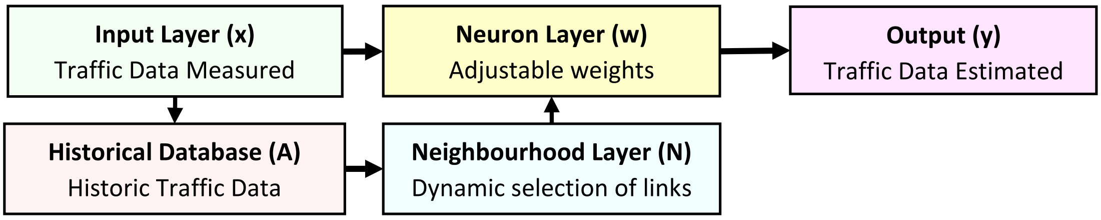

Figure 23: Global structure of the NLM framework

This chapter serves as the guide-book, in which all subsequent steps taken to arrive at the output (y), are explained in more detail. For that purpose a more detailed step-by-step division of the NLM is presented in the figure below. The (mathematical) content of each step is explained in the next paragraphs. Additionally in the next chapter a numerical example is provided for clarity.

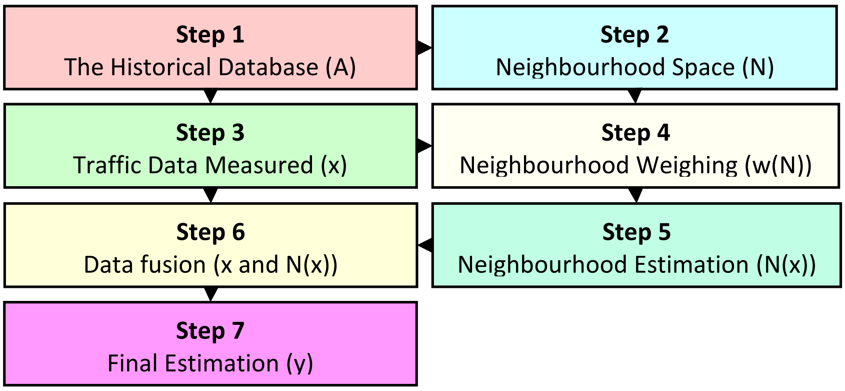
Step 1
The Historical Database (A)
Step 2
Neighbourhood Space (N)
Step 3
Traffic Data Measured (x)
Step 4
Neighbourhood Weighing (w(N))
Step 6
Data fusion (x and N(x))
Step 5
Neighbourhood Estimation (N(x))
Step 7
Final Estimation (y)

Figure 24: Step by step structure of the NLM framework for estimation

11.1.1 Step 1: The historical database

As opposed to the ground truth database, the historical databases considered in the Neighbour Link Method contains only a filtered number $\tilde{n} \leq n$ of vehicles which provide the data for the actual state estimation. Obviously each of the modelled time periods (morning rush hour, evening rush hour, outside rush hour) is considered separately. Additionally it handles the FCD and ILDD in the historical databases in a slightly different way as opposed to the formulas used for determination of the ground truth. This paragraph presents the modified formulas used to derive the traffic flow variables from the remaining data.

Mean speed estimation

The estimation for mean speed is calculated from the vehicles ($\tilde{n} \leq n$) that remain equipped with GPS and thus yield FCD. Formula [4] is therefore modified slightly.

$$\bar{v}(x,t) = \frac{\sum_{a=1}^{\tilde{n}(x,t)} s_a(x,t)}{\sum_{a=1}^{\tilde{n}(x,t)} tt_a(x,t)} \qquad [7]$$

In which:

$\bar{v}(x,t) =$	Mean speed estimation on link x, at time interval t	[mi/h]
$\tilde{n}(x,t) =$	Number of vehicles equipped with GPS on link x within time interval t	[-]
$s_a(x,t)$=	Distance travelled on link x, within time interval t	[mi]
$tt_a(x,t) =$	Time spent on link x, within time interval t	[h]

Mean density estimation

The estimation for mean density is slightly more challenging. As the subset of vehicles $\tilde{n} \sqsubseteq n$ is smaller then the actual vehicles on the link, using formula [f.5] makes little sense as it would always yield a lower density then the true density. Suggested is to incorporate the local floating car penetration rate $\leftthreetimes \approx n/\tilde{n}$ to compensate for that. While this rate converges globally to either $\bar{\leftthreetimes}$=20 (at 5% of FCD) or $\bar{\leftthreetimes}$=10 (at 10% of FCD), it can of course differ from location to location and from time to time. The factor can be derived by comparing the ILDD with the FCD, at the point of measurement. For the links without this data (highway links and the links around node 4) the global average of $\bar{\leftthreetimes}\ (t)$ is used as an estimator $\bar{\leftthreetimes}\ (t) = \sum n(x,t)\ /\ \sum \tilde{n}\ (x,t)$.

! When $\tilde{n}(x,t)$ = 0, there is no floating car that passes the detector on the link, $\leftthreetimes$ is then defined to be 0 as well. This assures that $\leftthreetimes\ (x,t) \in \mathbb{R}^{+}$ and thus the outcome remains feasible.

The old formula for density estimation is [5] edited to include the local penetration rate:

$$\bar{\rho}(x,t) = \frac{60 * \leftthreetimes\ (x,t)}{l(x) * r(x)} \sum_{a=1}^{\tilde{n}(x,t)} tt_a(x,t) \qquad [8]$$

In which:

$\bar{\rho}(x,t) =$	Density estimation on link x, at time interval t	[veh/mi/lane]
$l(x)$ =	Length of link x	[mi]
$tt_a(x,t)$ =	Time spent by the vehicle a within the interval t on link x	[h]
$r(x)$ =	Number of lanes on link (x)	[-]
$\leftthreetimes\ (x,t)$ =	Local penetration rate of link (x) at time (t)	[-]

Mean flow estimation

The formula used to determine the mean flow, remains unmodified. However flow is not estimated in this intermediate step as it is no input of the NLM as it is solely a direct output of step 7; the final estimation.

11.1.2 Step 2: The neighbourhood space

The first *real* step in this method is to determine the neighbourhood space ($N_x(t)$) for $\forall x \in X$. The neighbourhood space contains the links which can be used as indicators for a current (time t) traffic state on the link in question. The exact size of the neighbourhood space is set to 4 links, which seems a reasonable cut-off value without losing estimation accuracy (Esaway, 2012). The link itself is never included in the neighbourhood space of itself though it is used later on when fusing the neighbourhood link data with the most current traffic data on the link itself. The mathematical description of the neighbourhood space becomes:

$$N_x(\mathrm{t}) = \{x_1, x_2, x_3, x_4\} \subset X \backslash x$$

The selection procedure for links to be included in this neighbourhood space is determined by calculating the Pearson' correlation coefficient of the traffic data on the link in question to the traffic data on every other link (Esayway, 2015). The neighbourhood space (N_x) of link x is then filled with the 4 links (x_1, x_2, x_3, x_4) which give the highest correlation. The condition for determination of the 4 links in the neighbourhood space becomes mathematically:

$$corr\left(A_x(t), A_j(t)\right) \geq corr\left(A_x(t), A_k(t)\right) \quad | \, \forall \mathrm{j} \in N_x(\mathrm{t}) \ , \ \ \forall \mathrm{k} \in X \backslash N_x(\mathrm{t})$$

In which:
$A_x(t)$ = The historical database of velocities on link x for times: $t \in (t_0 \, , t_{max})$

As this operation can become quite time consuming – the complexity ($\mathcal{O} = X^2 T$) increases due to the increasing size of the database due to continuously storing of traffic data – suggested is to follow the recommendation of Esaway (2012) to not bluntly refresh the neighbourhood space every minute. Instead it is chosen to refresh the neighbourhood space at the beginning of each run M. The consequences of this choice shall be discussed later.

11.1.3 Step 3: Considering newly arrived data

The next step is to consider the traffic data measured at a time (t) at which a state estimation is required. This traffic data comprises for each link (x) out of aggregated FCD and ILDD which are used to calculate the speed $\bar{v}(x, t)$ and density $\bar{p}(x, t)$ estimations using formula [6] and [7].

These data are then added to the historical database, upon which the neighbourhood space can be refreshed. It is also used in the next steps, as to derive the traffic state estimation at time (t).

11.1.4 Step 4: The neighbourhood space weighting

With the data from the previous steps, the goal of this neighbourhood space algorithm is to find for each link (x) a linear weighting of the links in the neighbourhood space ($w_v(N_x)$) such that the indicator of velocity ($\bar{v}(x, t)$) is reached as closely as possible for all times (t). Additionally a second, completely separate weighting scheme ($w_p(N_x)$) is required, which aims towards reaching the indicator of density $\bar{p}(x, t)$ as closely as possible for all times (t).

The optimization problem for finding these weights is a special case of the *least squares problem* (Morita, 2012). In formula form, the following two problems are to be solved*:*

$$\min_{w_v(N_x)} \left|\left| A_{N_{v,x}}(t) * w_v(N_x) - \bar{v}(x, t) \right|\right|_2^2. \qquad [9a]$$

$$\min_{w_p(N_x)} \left|\left| A_{N_{p,x}}(t) * w_p(N_x) - \bar{p}(x, t) \right|\right|_2^2. \qquad [9b]$$

In which:
$A_{N_{v,x}}(t) \in \mathbb{R}_{+}^{|N_i(t)|\times|T|}$ = The historical database of velocities on neighbourhood links N_x. [mi/h]
$A_{N_{p,x}}(t) \in \mathbb{R}_{+}^{|N_i(t)|\times|T|}$ = The historical database of densities on neighbourhood links N_x. [veh/mi]

This step is then repeated for each link (x), such that for each link two (unique) weight layers: $w_v(N_x) \in \mathbb{R}^{|N_i(t)|}$ and $w_p(N_x) \in \mathbb{R}^{|N_i(t)|}$ are calculated.

11.1.5 Step 5: The neighbourhood estimation

The neighbourhood estimation of the velocity and density can be calculated by using the weighting from the previous step by multiplying the indicators for velocity/density with this weighting. In formula form:

$$\acute{v}(N_x, t) = \sum_{i=1}^{|N_x|} \bar{v}\left(N_{x,i}, t\right) * w\left(N_{x,i}, t\right) \quad [10a]$$

$$\acute{p}(N_x, t) = \sum_{i=1}^{|N_x|} \bar{v}\left(N_{x,i}, t\right) * w\left(N_{x,i}, t\right) \quad [10b]$$

In which:
$\acute{v}(N_x, t)$= Neighbourhood velocity estimation of link x at time t [mi/h]
$\acute{p}(N_x, t)$= Neighbourhood density estimation of link x at time t [veh/mi]

11.1.6 Step 6: Data fusion

The intermediate data available up until this step are considered to be two datasets. The first set is the previously determined neighbourhood estimations: $\acute{v}(N_x, t)$ and $\acute{p}(N_x, t)$. The second set is the data from the link itself: $\bar{v}(x, t)$ and $\bar{p}(x, t)$. A weighting scheme is required to assure that the final estimation for the link maximizes the utility of both data sources.

Esaway (2012) describes that there are ample examples of weighting schemes which can be used here; e.g. straight average, coefficients of determination, correlation coefficients, model variance, exponent of correlation, exponent of model variance and an Empirical Bayes (EB) method. Esaway concludes firstly that all of the schemes yield the same accuracy (the maximal difference between all schemes performance is <0,20% of the RRSE between estimation and truth). Secondly, Esaway concludes that not all schemes can be applied due to data unavailability. Chosen in this research is to weigh the velocity estimations using the respective variances of the estimations, to account for the reliability of the estimations. Whereas for density no variances are available (they are unknown), a simple straight average is proposed. Trial runs with different weighting schemes did not disprove the conclusions of Esaway (2012), though using both the data from the link itself and the neighbourhood yielded better results than using solely the data from either sources.

For velocity the variance in the mean velocity of the all traffic data $s^2{}_{\bar{v}}(x, t)$ is therefore calculated on each link, using the standard statistical formula for variance:

$$s^2{}_{\bar{v}}(x, t) = \frac{1}{n(x,t)-1} \sum_{a=1}^{\tilde{n}(x,t)} (v_a(x, t) - \bar{v}(x, t))^2 \quad [11]$$

! When $\tilde{n}(x, t) \in (0,1)$, $s^2{}_{\bar{v}}(x, t)$ is defined to be $v_{max}(x)$.

The variance of the neighbourhood estimation ($s^2{}_{\acute{v}}(N_x, t)$) can be derived using the normalized weights (w_{norm}) from step 4:

$$s^2{}_{\acute{v}}(N_x, t) = \sum_{i=1}^{|N_x|} s^2{}_{\acute{v}}\left(N_{x,i}, t\right) * \left(w_{norm}\left(N_{x,i}, t\right)\right)^2 \quad [12]$$

With now ${s^2}_{\bar{v}}(N_x, t)$ and ${s^2}_{\bar{v}}(x, t)$ known, the final estimation is calculated by minimizing the variance of both estimations and thus weighting each with their inverse variance. In formula form:

$$\ddot{w}\big(\acute{v}(N_x, t)\big) = \frac{\frac{1}{{s^2}_{\acute{v}}(N_x,t)}}{\frac{1}{{s^2}_{\acute{v}}(N_x,t)}+\frac{1}{{s^2}_{\bar{v}}(x,t)}} \text{ and } \ddot{w}\big(\bar{v}(x, t)\big) = 1 - \ddot{w}\big(\acute{v}(N_x, t)\big) \qquad \text{[13a][13b]}$$

For density the straight average is proposed, which yields the weights:

$$\ddot{w}\big(\bar{p}(N_x, t)\big) = 1 - \ddot{w}\big(\bar{p}(x, t)\big) = 0{,}5 \qquad \text{[13c]}$$

! Note that currently only 2 'data sources' are weighted and fused. Obviously if a representative ILD would be available, the data from it can be easily plugged into both equations.

11.1.7 Step 7: Final estimation

As the last step of the NLM the final estimation can then be calculated for all links:

$$\hat{v}(x, t) = \acute{v}(N_x, t) * \ddot{w}\big(\acute{v}(N_x, t)\big) + \bar{v}(x, t) * \ddot{w}(\bar{v}(x, t)) \qquad \text{[14a]}$$

$$\hat{p}(x, t) = \acute{p}(N_x, t) * \ddot{w}\big(\acute{p}(N_x, t)\big) + \bar{p}(x, t) * \ddot{w}\big(\bar{p}(x, t)\big) \qquad \text{[14b]}$$

$$\hat{q}(x, t) = \hat{v}(x, t) * \hat{p}(x, t) \qquad \text{[14c]}$$

11.2 NLM Example

For clarity and completeness a numerical example of NLM is provided in this chapter. Assumed is that with a historical database comprising out of 59 intervals of the morning rush hour and a FCD coverage rate of 5% the real time traffic state of link 26 (x=26) at time interval 60 (t=60) is to be calculated. This paragraph applies the steps mentioned in the previous chapter to derive a final estimation by applying the NLM.

11.2.1 Step 1: The historical database

The historical database of this example initially starts with 59 (times) $*|X|=76$ (links) cells filled with arrays of $\bar{v}(x,t)$ and $\bar{\rho}(x,t)$ determined from previously recorded traffic data.

11.2.2 Step 2: The neighbourhood space

Let's assume that in this example, the neighbourhood space is refreshed at this time interval. Using Pearson's correlation coefficient, for the whole historical database the correlation between velocities from link 26 and the other links is then determined. The 4 most correlating links are then selected. In this example, the result are; link 37 with a correlation of 0,5739, link 42 with a correlation of 0,5537, link 50 with a correlation of 0,4951 and link 10 with a correlation of 0,4434.

The result is: $N_{26} = (37{,}42{,}50{,}10)$.

11.2.3 Step 3: Considering newly arrived data

At time 60, new (real-time) traffic data arrives for all links in the network. As this example is limited to providing a state estimation on link 26. Only the data from relevant links is displayed in the table below. Again formulas [7] and [8] are applied to derive these initial estimations. Additionally all link data is added to the historical database. Obviously the correlation between link 26 and link 26 is 1.

Link #	**37**	**42**	**50**	**10**	**26**
r	0,5739	0,5537	0,4951	0,4434	1,00
$\bar{\boldsymbol{v}}(\boldsymbol{x},\boldsymbol{t})$	21,8400	5,4231	8,5300	23,7695	23,5703
$\bar{\boldsymbol{\rho}}(\boldsymbol{x},\boldsymbol{t})$	15,2217	61,1079	16,4153	8,45465	42,8413

Table 9: Overview of used traffic data

11.2.4 Step 4: The neighbourhood space weighting

Solving the least squares problems for both the velocity and density, with the historical database of size 60 (time) x 76 (links) yields the resulting two weight layers:

$$w_v(N_{26}) = [0.4365 \quad 0.2694 \quad 0.0232 \quad 0.2497]$$
$$w_p(N_{26}) = [0.1977 \quad 0.0198 \quad 0.1231 \quad 0.5612]$$

11.2.5 Step 5: The neighbourhood estimation

Using the weighting from above, and the corresponding traffic data, the neighbourhood estimation of velocity and density are calculated:

$$\bar{v}(N_{26}) = [21{,}8400 \quad 5{,}4231 \quad 8{,}5300 \quad 23{,}7695] \times \begin{bmatrix} 0.4365 \\ 0.2694 \\ 0.0232 \\ 0.2497 \end{bmatrix} = 17{,}129 \text{ mi/h}$$

$$\bar{p}(N_{26}) = [15{,}2217 \quad 61{,}1079 \quad 16{,}4153 \quad 8{,}45465] \times \begin{bmatrix} 0.1977 \\ 0.0198 \\ 0.1231 \\ 0.5612 \end{bmatrix} = 10{,}848 \text{ veh/mi/lane}$$

11.2.6 Step 6: Data fusion

Firstly the variances of the mean velocities are calculated, yielding the following array:

Link #	**37**	**42**	**50**	**10**	**26**
$\boldsymbol{s^2_{\bar{v}}(x,t)}$	81,8247	6,8840	125,0706	138,9276	43,8665

Table 10: Variance in mean link velocity

The variance of velocity on link 26 is:

$$s^2_{\bar{v}}(26) = 43{,}8665 \text{ mi/h}$$

The normalized weighting of the velocities in the neighbourhood space becomes:

$$w_{norm}(N_{26}) = \frac{w(N_{26})}{\sum|w(N_{26})|} = \frac{[0.4365 \quad 0.2694 \quad 0.0232 \quad 0.2497]}{0{,}9788} = [0.4460\ 0.2753\ 0.0237\ 0.2551]$$

Yielding a pooled variance of the neighbourhood velocity estimation:

$$s^2_{\bar{v}}(N_{26}) = [81{,}8247 \quad 6{,}8840 \quad 125{,}0706 \quad 138{,}9276] \times \begin{bmatrix} 0{,}4460^2 \\ 0{,}2753^2 \\ 0{,}0237^2 \\ 0{,}2551^2 \end{bmatrix} = 17{,}642 \text{ mi/h}$$

As $s^2_{\bar{v}}(N_{26}) = 17{,}642$ and $s^2_{\bar{v}}(26) = 43{,}8665$, the fusion of data is performed using the inverse variance weighting. This yields:

$$\ddot{w}\big(\bar{v}(N_{26})\big) = \frac{\frac{1}{17{,}642}}{\frac{1}{17{,}642}+\frac{1}{43{,}8665}} = 0{,}7131 \quad \text{and} \quad \ddot{w}\big(\bar{v}(26)\big) = 1 - 0{,}7131 = 0{,}2868$$

11.2.7 Step 7: Final Estimation

The final traffic state estimation for link 26 at time interval 60 is:

$$\hat{v}(26{,}60) = 17{,}129 * 0{,}7131 + 23{,}5703 * 0{,}2868 = \quad \textbf{19,0 mi/h}$$
$$\hat{p}(26{,}60) = 10{,}848 * 0{,}5 + 42{,}8413 * 0{,}5 = \quad \textbf{26,8 veh/mi/lane}$$
$$\hat{q}(26{,}60) = 18{,}98 * 26{,}845 = \quad \textbf{509,5 veh/h/lane}$$

For comparison, the actual ground truth traffic flow variables for this link are given:

$$v_{GT}(26{,}60) = \quad \textbf{19,492 mi/h}$$
$$p_{GT}(26{,}60) = \quad \textbf{17,330 veh/mi/lane}$$
$$q_{GT}(26{,}60) = \quad \textbf{337,545 veh/h/lane}$$

11.3 NLM performance evaluation

Whereas the previous chapters describe the detailed application of the neighbour link method, this chapter regarding the performance evaluation, serves as the last step of this framework. The indicator adopted for comparing the traffic state estimations to the actual observed ground truth, is the MSE (Mean Square Error). The MSE of each outputted state estimation can be calculated using the following general formula:

$$MSE_v(x,t) = \frac{1}{|X|}(\hat{v}(x,t) - v_{GT}(x,t))^2 \qquad [15]$$

In which:
$\hat{v}(x,t)$ = Velocity estimation for link x at time t [mi/h]
$v_{GT}(x,t)$ = Velocity ground truth for link x at time t [mi/h]

Additionally the average MSE of a full run M is the sum over all links x, and all times $t \in M$

$$MSE_v(M) = \frac{1}{T}\sum_{t=t_{min}}^{\mathrm{t}_{max}} \sum_{x=1}^{|X|} MSE_v(x,t) \qquad [16]$$

In which:

$MSE_v(x,t)$ =	Mean square error of estimation on link x at time t	[-]
T =	Run length	[h]
$MSE_v(M)$ =	Average run MSE	[-]

The average MSE over multiple runs: $\overline{\mathrm{MSE}}$ is calculated using the arithmetic mean of all runs M:

$$\overline{\mathrm{MSE}}_v = \frac{1}{|M|}\sum MSE_v(M) \qquad [17]$$

In which:

$MSE_v(M)$ =	Average run MSE	[-]
$\lvert M\rvert$ =	Number of runs	[-]
$\overline{\mathrm{MSE}}_v =$	Average MSE over all runs	[-]

Obviously the lower the mean square error, the better the estimations are.

Another way to assess the quality of the results is to calculate the correlation between estimation and ground truth. The degree of quality is expressed in Pearson's correlation coefficient between estimation and ground truth. E.g. for velocity:

$$r = corr(\{\hat{v}\},\{v_{GT}\}) \qquad [18]$$

In which:

r =	Pearson's correlation coefficient	[-]
$\{\hat{v}\}$ =	Set of all estimated velocities	[mi/h]
$\{v_{GT}\}$ =	Set of all ground truth velocities	[mi/h]

For this correlation coefficient, the closer the value to +1.00, the higher the correlation between the estimations and the ground truth.

! The MSE and r are calculated for all traffic flow variables; velocity in the formulas above can be arbitrarily replaced by density and flow as the ground truth is available for all.

12 NLM estimation variants

The next step in this research is to apply the previous chapters on the whole set of case study data as to assess the state estimation performance of the neighbour link method throughout the different modelled and simulated periods. By defining variants in which initially chosen settings are changed, the sensitivity of the neighbour link method is assessed as well. Four variants are defined for all simulation periods and two additional experimental variants are defined for solely the evening rush hour. Besides assessing sensitivity, the secondary goal is to find the best settings for this case study to use for the next part of this research regarding traffic state prediction.

12.1 Variant #1: baseline

Overview characteristics variant #1:

FCD coverage:	5%
Neighbourhood Refresh:	Every run (M)
Database Size:	Expanding
Run order:	[1,2,3,4,5]

The baseline variant compares the ground truth with the estimations from the NLM for all three simulation periods separately. The database is initially filled with the traffic data from run #1 of the respective period. Subsequently the simulated runs are iteratively added to the databases and for each of these runs (2,3,4,5) the minute to minute estimations are evaluated on accuracy. The baseline variant uses 5% of FCD and refreshes the neighbourhood space at the beginning of every run M. The historical database is iteratively expanding with each run. The goal of this baseline variant is to provide an initial basic run to be used as a comparative basis for the additional variants. The hypothesis is that this variant performs comparable to the learning database interpolation method used in De Vries (2015).

12.2 Variant #2: 10% FCD

Overview characteristics variant #2:

FCD coverage:	**10%**
Neighbourhood Refresh:	Every run (M)
Database Size:	Expanding
Run order:	[1,2,3,4,5]

The second variant changes the settings of the FCD coverage rate to 10% as opposed to the baseline variant. Again the minute to minute estimations are evaluated on accuracy. The goal of this variant is to analyse the sensitivity of the performance towards the average FCD coverage rate. The hypothesis is that this variant performs around 50% better than in variant 1, regarding velocity estimations, as seen in De Vries (2015). For the other traffic flow variables an unknown improvement is to be expected.

12.3 Variant #3: neighbourhood space refreshing

Overview characteristics variant #3:

FCD coverage:	5%
Neighbourhood Refresh:	**Every minute**
Database Size:	Expanding
Run order:	[1,2,3,4,5]

The third variant again uses 5% FCD but refreshes the neighbourhood space excessively, every minute as opposed to daily (once per run) in the baseline variant. The neighbourhood space is determined upon still analysing the whole traffic database. The goal of this variant is to analyse the sensitivity of the performance towards a continuously being updated neighbourhood space. The hypothesis of this variant is that no improvement is expected as argued by Esaway (2012).

12.4 Variant #4: maximum sized database

Overview characteristics variant #4:	
FCD coverage:	5%
Neighbourhood Refresh:	Every run (M)
Database Size:	**Limited**
Run order:	[1,2,3,4,5]

The fourth variant only varies compared to the baseline variant by defining a maximum size of the database. At the start of each run, the database size is checked and if larger or equal to two runs, all runs but the most recent are dropped. For example at the start of run 4, the historical database will only contain the data from run 3, instead of run 1,2 and 3 and is then throughout M=4 updated until it is at maximum size at the end of the run again (containing all data from M=3 and M=4). This therefore effects both the regression problem (the least squares minimization problem) and the neighbourhood space determination (the correlation problem) as the dropped data from the traffic database is not considered anymore. The goal of this variant is to analyse the sensitivity of the performance towards the database size. The hypothesis is that this variant reveals the trade-off between unwanted smoothing with more data and too little smoothing due to less data.

12.5 Variant #5: reversed ordering

Overview characteristics variant #5:	
FCD coverage:	5%
Neighbourhood Refresh:	Every run (M)
Database Size:	Expanding
Run order:	**[5,4,3,2,1]**

This 5th variant considers the variance in results by reversing the run order. Instead of M=1, 2, 3, 4, 5 in the baseline, the order of runs considered becomes M=5, 4, 3, 2, 1. The goal of this experimental variant is to analyse the sensitivity of the performance towards a different starting seed of the historical database. Assessment of the results takes place by considering the performance difference in average MSE over all runs. Hypothesis is that this variant will show some difference in performance as obviously the database seed affects both the selection of the neighbourhood space and next the easiness of finding a good weighting of these links. The observed performance difference can only be applied to designate the influence of database seed upon the performance within this research' context.

12.6 Variant #6: bagging

Overview characteristics variant #6:	
FCD coverage:	5%
Neighbourhood Refresh:	Every run (M)
Database Size:	Expanding
Run order:	[1,2,3,4,5]
Bins:	**2**

The last experimental variant varies from the baseline by partitioning the evening rush hour into two bins instead of the one bin in the previous variants. This data bagging or bucketing is a data pre-processing technique which is applied as it is suggested to be plausible that link correlation in the traffic breakdown phase does not necessarily also apply in the traffic recovery phases of a link. Separating the database and applying the traffic estimation framework on the created two partitions separately intuitively improves the result as now instead of smoothing due to finding the on average best correlating links throughout both traffic phases, the best links can be found in both phases separately. To uncover an easy applicable and suitable partitioning time of the traffic data, suggested is to look at both the average network velocity and the number of vehicles in the network. These both peek on average at t=90 yielding two bins of traffic data at (1) 1 to 90 minutes and (2) 91 to 120 minutes. The effects of this initial choosing are then assessed by using the weighted average of the MSE for each bin (¾,¼) and comparing the result with the MSE of the baseline. The hypothesis is that even with this rudimentary approach, accuracy of all traffic flow variables will improve.

13 NLM estimation results

In this chapter the raw results of each of the variants are displayed. The results are presented separately for each simulation interval and for each run of that respective interval. Additionally the mean MSE ($\overline{\text{MSE}}$) and correlation (r) of estimation versus ground truth are shown. A more thorough look at the individual results, comparison between results, discussion and synthesis is saved for the next chapter.

13.1 Variant #1: baseline

	Morning Rush Hour			**Evening Rush Hour**			**Outside Rush Hour**		
	v	p	q	v	p	q	v	p	q
MSE of Run #2	56,3	679	25518	59,2	1785	28925	72,9	1,5	924
MSE of Run #3	58,3	896	23981	61,2	1611	31066	74,1	1,5	906
MSE of Run #4	61,3	818	23466	66,5	896	27056	74,8	1,5	916
MSE of Run #5	59,6	839	24645	71,8	1255	29189	79,4	1,5	928
$\overline{\mathbf{MSE}}$	58,9	808	24402	64,7	1387	29059	75,3	1,5	919
r	0,851	0,766	0,687	0,849	0,751	0,685	0,821	0,294	0,268

Table 11: MSE per run, mean MSE and correlation for baseline variant

13.2 Variant #2: 10% FCD

	Morning Rush Hour			**Evening Rush Hour**			**Outside Rush Hour**		
	v	p	q	v	p	q	v	p	q
MSE of Run #2	47,9	411	18003	51,4	1174	23056	67,2	1,3	843
MSE of Run #3	49,9	478	18327	60,1	826	23045	67,7	1,3	823
MSE of Run #4	55,1	454	19186	59,0	478	20080	68,8	1,3	840
MSE of Run #5	55,2	423	19611	70,6	601	22184	72,7	1,3	846
$\overline{\mathbf{MSE}}$	52,0	442	18782	60,3	770	22091	69,1	1,3	838
r	0,851	0,866	0,772	0,844	0,854	0,761	0,824	0,412	0,372

Table 12: MSE per run, mean MSE and correlation for variant 2

13.3 Variant #3: neighbourhood space refreshing

	Morning Rush Hour			**Evening Rush Hour**			**Outside Rush Hour**		
	v	p	q	v	p	q	v	p	q
MSE of Run #2	57,4	700	24960	59,3	1774	30153	73,1	1,5	924
MSE of Run #3	59,8	906	24201	67,2	1590	31461	73,9	1,5	907
MSE of Run #4	61,7	814	24229	69,7	880	28205	74,7	1,5	916
MSE of Run #5	59,7	833	25092	72,1	1273	29715	79,3	1,5	928
$\overline{\mathbf{MSE}}$	59,7	813	24620	67,1	1379	29883	75,3	1.5	919
r	0,848	0,764	0,681	0,845	0,753	0,673	0,822	0,294	0,268

Table 13: MSE per run, mean MSE and correlation for variant 3

13.4 Variant #4: maximum sized database

	Morning Rush Hour			Evening Rush Hour			Outside Rush Hour		
	v	p	q	v	p	q	v	p	q
MSE of Run #2	56,3	679	25518	59,2	1785	28925	72,9	1,5	924
MSE of Run #3	49,5	916	22997	56,6	1603	30643	73,9	1,5	911
MSE of Run #4	52,8	826	23913	55,6	1009	26666	74,9	1,5	912
MSE of Run #5	50,3	836	24497	56,6	1318	28341	79,1	1,5	932
$\overline{\mathbf{MSE}}$	52,2	814	24231	57,0	1429	28644	75,2	1,5	920
r	0,885	0,760	0,691	0,887	0,739	0,688	0,824	0,288	0,263

Table 14: MSE per run, mean MSE and correlation for variant 4

13.5 Variant #5: reversed ordering

	Morning Rush Hour			Evening Rush Hour			Outside Rush Hour		
	v	p	q	v	p	q	v	p	q
MSE of Run #2	53,3	873	23717	58,5	937	26413	75,3	1,5	915
MSE of Run #3	55,2	928	23949	60,3	1580	31263	74,0	1,5	905
MSE of Run #4	58,3	635	24921	64,9	1701	31561	72,9	1,5	921
MSE of Run #5	55,0	882	22909	68,4	1644	30676	75,5	1,4	863
$\overline{\mathbf{MSE}}$	55,5	830	23874	63,0	1466	29978	74,4	1,5	901
r	0,866	0,772	0,694	0,852	0,736	0,669	0,824	0,299	0,272

Table 15: MSE per run, mean MSE and correlation for variant 5

13.6 Variant #6: bagging

	Morning Rush Hour			Evening Rush Hour			Outside Rush Hour		
	v	p	q	v	p	q	v	p	q
MSE of Run #2	56,6	693	23218	61,4	1767	31314	73,4	1,5	956
MSE of Run #3	58,7	1021	25672	62,1	1588	32820	74,0	1,5	901
MSE of Run #4	62,1	858	24589	66,2	955	29082	75,2	1,5	955
MSE of Run #5	59,5	859	25471	71,4	1296	30188	79,6	1,5	943
$\overline{\mathbf{MSE}}$	59,2	858	24738	65,3	1402	30851	75,6	1,5	939
r	0,812	0,741	0,655	0,823	0,712	0,648	0,821	0,291	0,2567

Table 16: MSE per run, mean MSE and correlation for variant 6

14 NLM estimation synthesis

With the results in, this chapter focusses on the synthesis of the results. First in this chapter, the practical meaning of a certain value for the MSE is quantified by deriving from the results the average chances that a link is estimated (or predicted) within a certain absolute distance of the ground truth.

Subsequently a global analysis of the results is performed. This analysis is aimed towards assessing the results of each of the designed variants, throughout the simulated periods of morning rush hour, evening rush hour and outside rush hour. The modelled periods are therefore divided into five types of traffic phases, based on their global characteristics in a fundamental diagram: free-flow, bounded and congestion. An additional distinction is made for the latter phase, separating the traffic breakdown and traffic recovery phase. Each variants average performance throughout each of these phases (if occurring) is then discussed separately.

In the next phase, the results of the variants are discussed in more depth. Whereas up until now only the average MSE has been discussed, it is in this analysis that the performance of variants is assessed in more depth, with the goal of finding improvements that can be applied to the NLM.

In the final phase, the final variant is designed and the performance of this final variant is presented again on average MSE but also on two representative individual links in the network.

14.1 Quantifying the MSE

In the previous chapter, the results show that the MSE of velocity lays between 47,9 and 79,6. The average MSE of density lays between 1 and 1785 and the average MSE of flow lays between 823 and 32820. These numbers allow for comparison between individual variants and other state estimation studies. However, practically the question regarding how bad or how good an average MSE of e.g. 60 is, remains unanswered. This paragraph gives the mean chance (C) that a result will lie within a certain absolute distance (d) from the ground truth at a given MSE value.

These threshold values (d) are determined next. For velocity the thresholds from De Vries (2015) are adopted such that d_{v1} =3, d_{v2} =6 and d_{v3} =9 mi/h. For density the intervals are chosen at 5%, 10% and 15% from the theoretical max per lane. This maximum density experienced in the ground truth is 250 veh/mi/lane. Therefore the density intervals are: d_{p1} =12, d_{p2} =25 and d_{p3} =38 veh/mi/lane. The flow reliability intervals are chosen d_{q1} =100, d_{q2} =200 and d_{q3} $=$ 300 veh/h/lane.

Next for each MSE increments of respectively 5 (velocity), 500 (density) and 5000 (flow), 5 one-minute simulation intervals are selected upon which the MSE at that time is close to the selected increment. Then within each of the selected times all individual link estimations are compared with their ground truth e.g. $\left(\hat{v}(x,t) - v_{GT}(x,t)\right)$ and counted if below the predefined absolute threshold (d). Then division by total number of links considered ($5|X|$) yields the mean chance that a link lays within the absolute distance (d). In formula form for e.g. velocity:

$$C(MSE, d) = 100\% * \frac{\sum_{a=1}^{5|X|}\left(\hat{v}(x,t) - v_{GT}(x,t)\right) \leq d}{5|X|} \mid \vec{t} = \{t1, t2, t3, t4, t5\} \subset T \qquad [19]$$

The following paragraphs show the resulting graphs for multiple increments. It must be noted that the distributions are scenario specific (e.g. the distribution of MSE along links for the used *Enriched Sioux Falls Scenario* is likely to be different in other scenarios), only the average MSE can be transferred and used to compare other research with.

14.1.1 Velocity MSE distribution

In the graph below, the velocity reliability intervals for the share of links in the network are shown. A MSE of 5 shows for example that on average, 78% of the links are estimated within 3 mi/h accuracy.

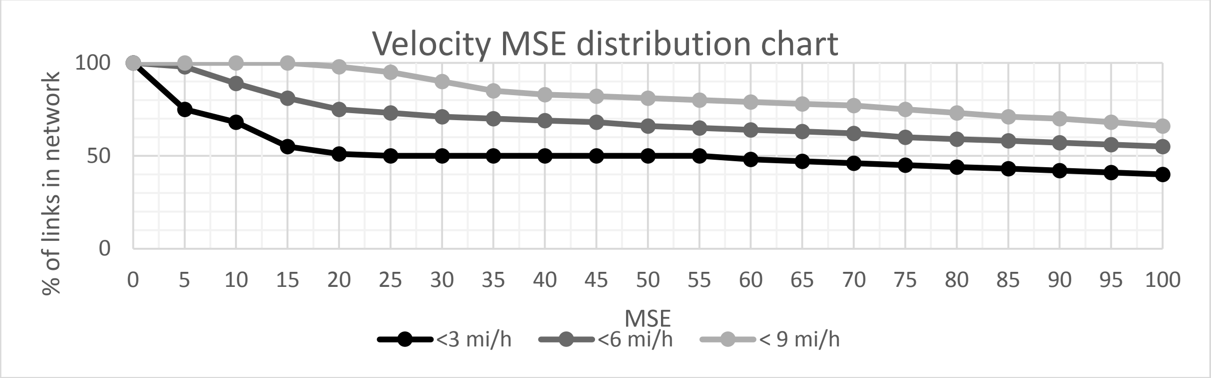

Figure 25: Velocity MSE distribution chart

14.1.2 Density MSE distribution

In the graph below, the density reliability intervals for the share of links in the network are shown. A MSE of 1000 means that on average 75% of the links are estimated within 12 veh/mi/lane, on average 88% are estimated within 25 veh/mi/lane and 91% are estimated within 38 veh/mi/lane from the ground truth.

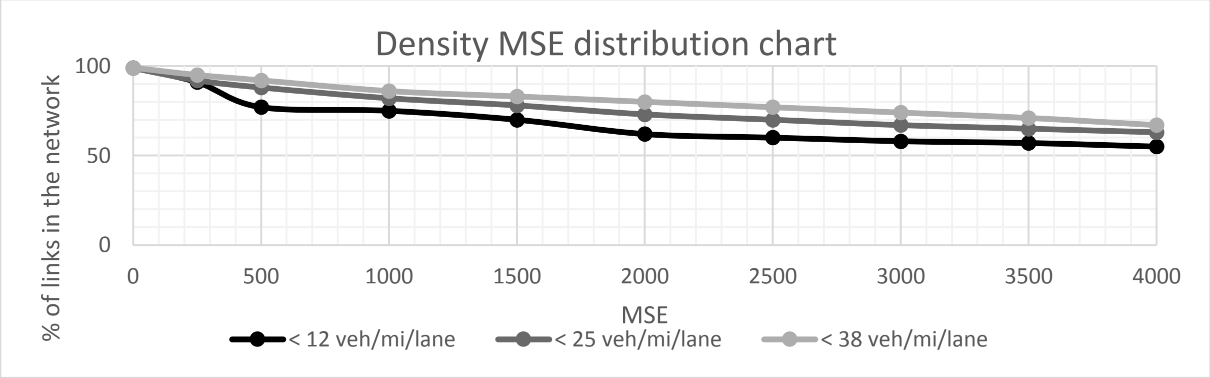

Figure 26: Density MSE distribution chart

14.1.3 Flow MSE distribution

And lastly, the distribution of the flow MSE values are presented.

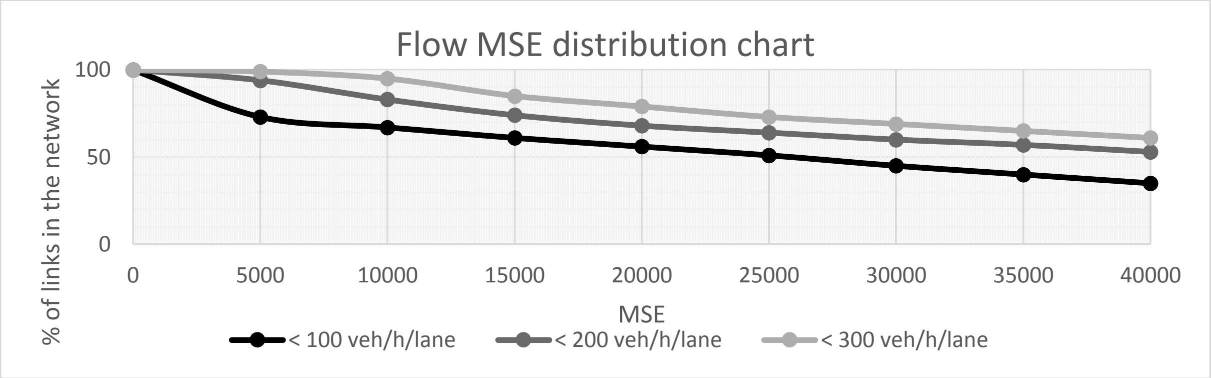

Figure 27: Flow MSE distribution chart

14.2 Global analysis

For comparing both the average MSE and correlation of each variant, the results from the baseline are used as a basis. In the table below the performance differences are presented more clearly. Without a doubt variant 2 performs the best overall with up to 45,3% improvement. Variant 4 and 5 perform better regarding velocity estimations, but show less improvement on other variables. Variant 3 and 6 are comparatively the weakest. Additionally the velocity and density MSE values seem to correspond with a relatively higher reliability interval than the flow MSE value, indicating that the flow estimations are of a relatively lower accuracy.

MSE	**Morning Rush Hour**			**Evening Rush Hour**			**Outside Rush Hour**		
Variant	v	p	q	v	p	q	v	p	q
1	**58,9**	**808**	**24402**	**64,7**	**1387**	**29059**	**75,3**	**1,5**	**919**
2	-11,7%	-45,3%	-23,0%	-6,8%	-44,5%	-24,0%	-8,2%	-13,3%	-8,8%
3	+1,4%	+0,6%	+0,9%	+3,7%	-0,6%	+2,8%	0,0%	0,0%	0,0%
4	-11,4%	+0,7%	-0,7%	-11,9%	+3,0%	-1,4%	-0,1%	0,0%	+0,1%
5	-5,8%	+2,7%	-5,7%	-2,6%	+5,7%	+3,2%	-1,2%	0,0%	-2,0%
6	+0,5%	+6,2%	+1,4%	+0,9%	+1,1%	+6,2%	+0,4%	0,0%	+2,2%

Table 17: Comparison of average MSE in which the results of variant 1 serve as index (=100%).

Before analysing the results in more depth for each of the estimated traffic flow variables, some additional comments are made. Firstly it must be noted that the time it takes to process one full minute of traffic data and to derive the network state estimation for this same minute takes less than one second on the modern laptop used in this research. Therefore an estimation can be provided well within real-time. The most time consuming steps are the determination of the neighbourhood space (correlation) and the solving of the minimization problem (least squares).

Secondly, when looking at the results presented in the previous chapter, in all but the maximum database variant (4), the accuracy of the velocity estimations degrade with each consecutive run. As it does not occur within the maximum database variant, the size of the database used is identified as the cause for this degradation. In more detail, this degradation is caused due to the fact that when the database gets larger, the weighting of the neighbour links converges to an equilibrium solution. This means that most recent data weights relatively less, as opposed to the data in the big traffic database. Therefore the inherently capriciousness traffic flow behaviour is not captured, but smoothened, revealing (unwanted) locally flattened velocity estimations. Two solutions are identified; limiting the database size as in variant #4, or weighting the historical traffic database such that recent data is considered more important. Only the first is researched further.

14.2.1 Morning Rush Hour Analysis

A more detailed look into the morning rush hour results is started by looking at the velocity estimations. In the chart below the average velocity MSE of all four morning rush hour runs within a variant are plotted per time interval of the simulation. Additionally the average number of vehicles in the network is plotted, to give an indication of the average traffic state per minute. The variant numbers refer to the previously defined variants.

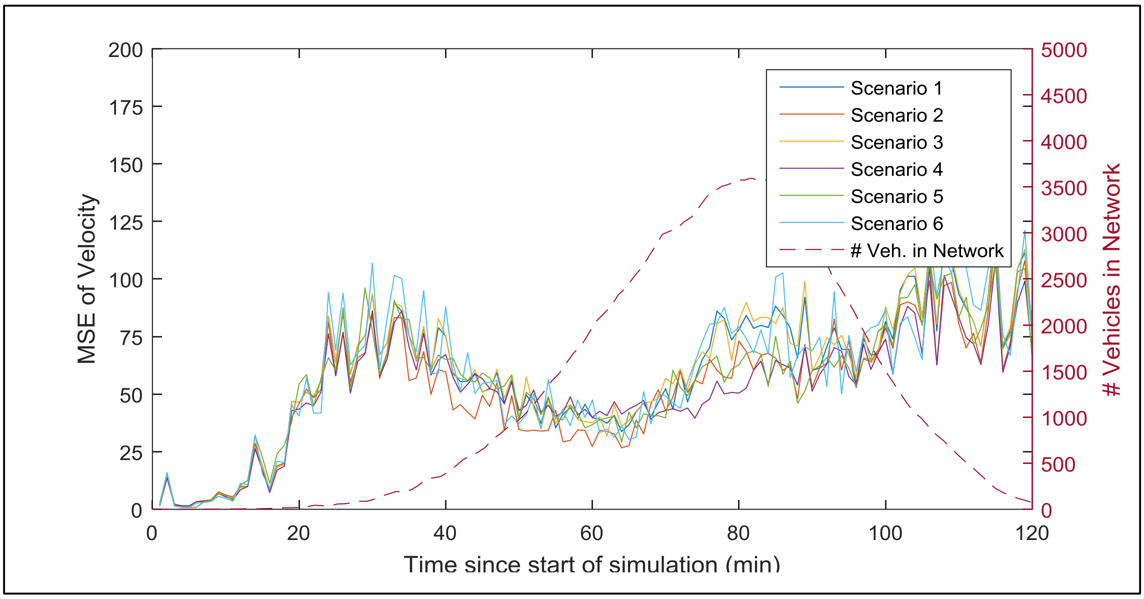

Figure 28: Velocity MSE per variant

As a first observation, it looks like the variants do not vary much between each other. It is only around 80 minutes that the variants give very different results, ranging from a MSE of 50 up to 100. To analyse the chart more thoroughly, the time since the start of the simulation is partitioned into five distinct clusters and discussed more thoroughly as to identify the causes for the MSE during each period.

Time	Description	MSE Range
0-20	[Free-flow] As there is almost no traffic, both the ground truth as the estimated velocities are at v_{max} and therefore the MSE is close to zero. A single car unequipped with GPS which travels slower or over v_{max} is the main cause for estimation errors.	Around 20
20-40	[Bounded] There is some traffic, but even with 5%-10% FCD most of the links have a zero sample rate and thus get estimated near v_{max}, while vehicles can again travel above or lower, generating a higly capricious MSE.	Around 80
40-75	[Traffic Breakdown] Vehicle numbers increase, most links get covered by the sample of floating cars. As on some links congestion starts to develop, vehicle behaviour becomes more uniform, yielding less difference between floating car velocities and the link velocities	Around 40
75-100	[Traffic Recovery] The traffic demand decreases, but congestion is still appearant on some links. For those links the estimations are quite accurate. The recovering links as well as very low visited links (e.g. the highways) are responsible for the increasing MSE. The cause is inherent to the low sampling rate on recovering links. As the velocities estimations on the link itselve are derived from the GPS of only a few vehicles, a badly identified neighbourhood which do not share the same traffic phase (yet) might overrule the estimation from the link itself. As obviously the neighbourhood then yields a lower variance, it gets weighted highly accordingly. It is in this period that differences between variants become appearant.	Around 60
100-120	[Bounded] Again, some traffic is still appearant in the network. Yet a lot of links have zero FC samples and thus get estimated at vmax (whereas in the ground truth they are below or above).	Around 80

Table 18: Detailed causes for varying MSE per identified cluster

For the density estimations, the same type of graph is provided below. Again the average MSE throughout the four runs is plotted against the time since the start of the simulation. Whereas for some runs the average MSE over the whole simulation is below 1000, the peaks within the simulation reveal results up to five times as high.

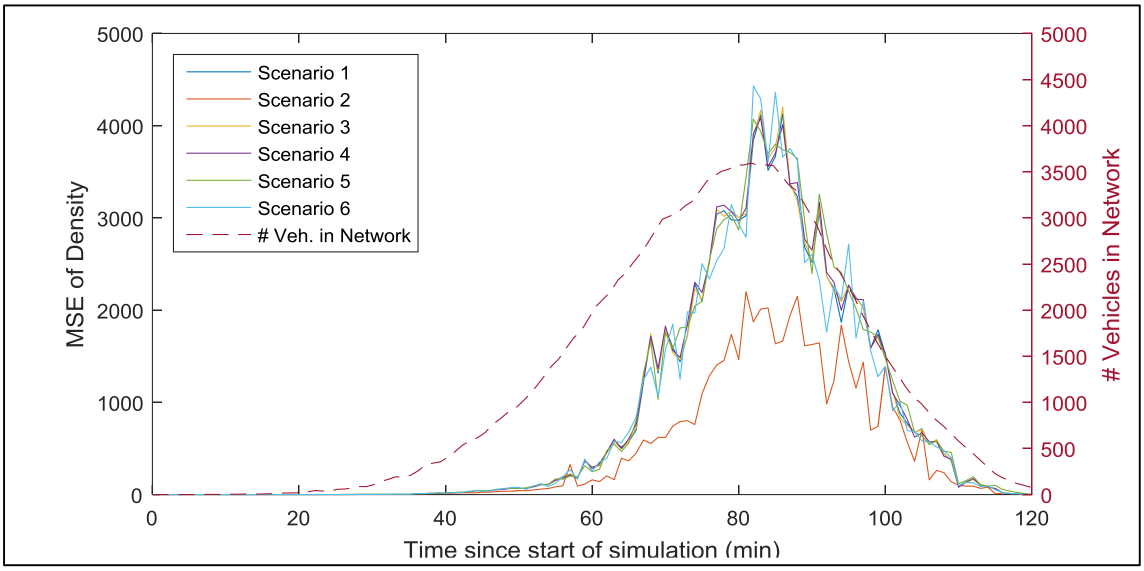

Figure 29: Density MSE per variant

On first sight the density graph moves together with the number of vehicles in the network. This relationship is caused by the fact that when densities increase in the network, the estimations can be wrong on both sides (as in overestimation and underestimation). When densities are near zero, the underestimation errors are negligible. It becomes clear that all but variant 2 yield the same accuracy of results.

Next for the flow estimations, the same procedure is followed. Which shows a surprising result. The flow estimation is the result of the velocity estimation multiplied with the density estimation. But whereas the velocity and density errors are quite small around the t=60 minute mark, the flow error peaks. This is the result of two factors: (1) The flows are highest during the traffic breakdown period and (2) density and velocities are both underestimated at that time, stacking their respective errors into the flow estimation errors (instead of e.g. cancelling each other out).

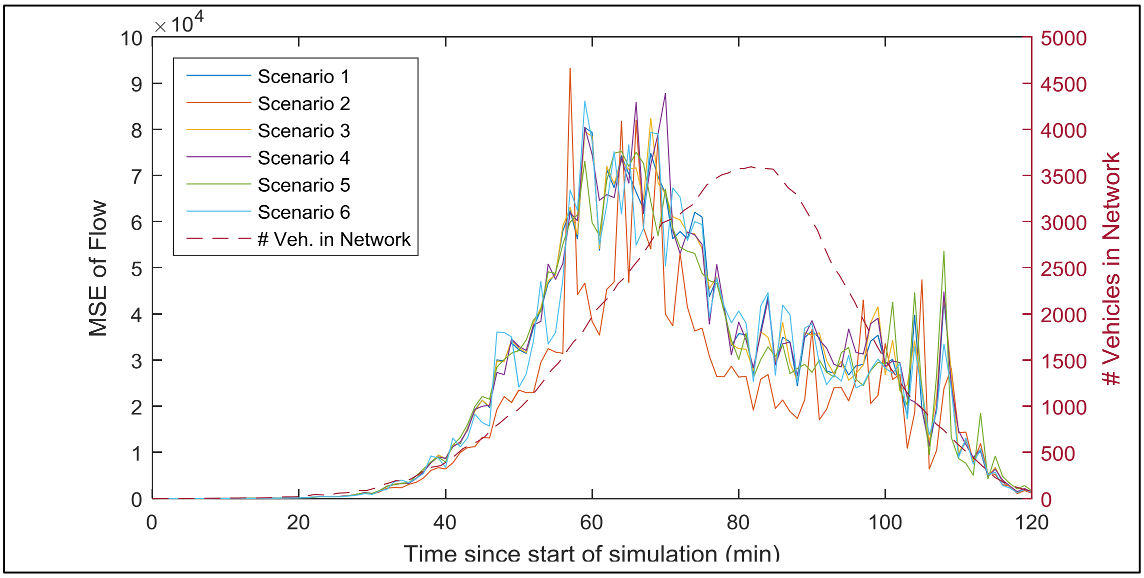

Figure 30: Flow MSE per variant

14.2.2 Evening Rush Hour Analysis

For the evening rush hour, the investigation starts with the MSE of velocity. In the chart below the average velocity MSE of all four evening rush hour runs within a variant are plotted per time interval of the simulation. Additionally the average number of vehicles in the network is plotted, to give an indication of the average traffic state per minute.

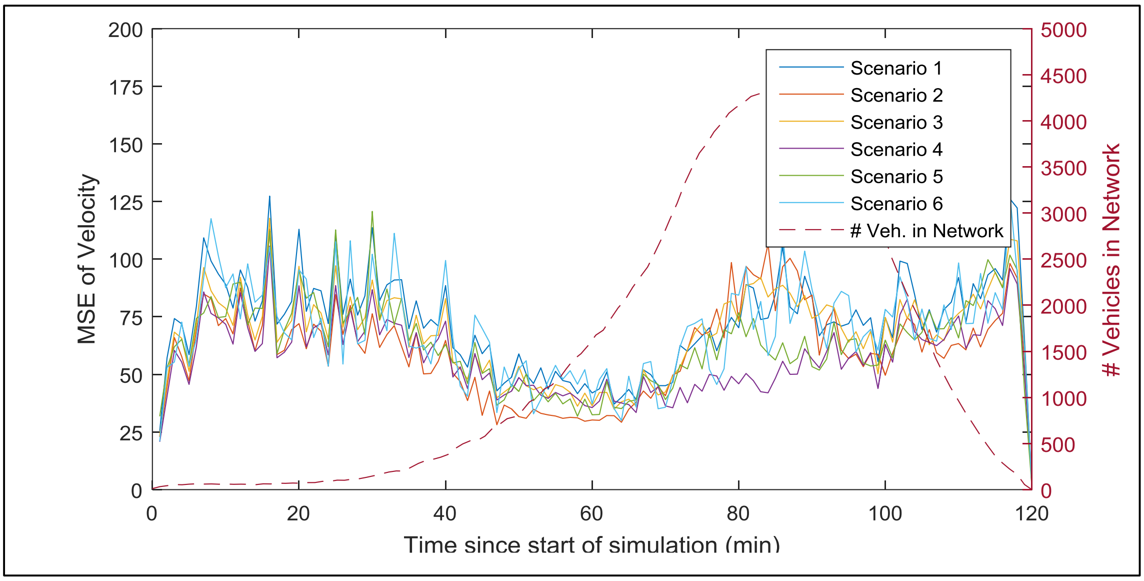

Figure 31: Velocity MSE per variant

The above chart, shows some differences compared with the chart of morning rush hour. Firstly, due to erroneous log data from PARAMICS, the last time interval of traffic data was omitted and therefore yields a MSE of 0. Secondly, the number of vehicles in the network are higher than in morning rush hour. Up to twofold just outside rush hour (both go-and-return trips of secondary activities occur), and from a maximum of around 3500 vehicles to a maximum of 4500 vehicles in the network. Thirdly, the variants again do not vary much between each other, with again around minute 80 a spike as variant 4 yields a MSE of around 40, where variant 6 tops at a MSE of around 110. Again to analyse the chart more thoroughly, the time since the start of the simulation can be partitioned into three clusters using the same descriptions as within the morning rush hour.

Time	Description	MSE Range
0-40	[Bounded] There is quite some traffic, but even with 5%-10% FCD most of the links still have a zero sample rate and thus get estimated near v_{max}, while vehicles can again travel above or lower, generating a highly capricious MSE.	Around 80
40-75	[Traffic Breakdown] Vehicle numbers increase, most links get covered by the sample of floating cars. As on some links congestion starts to develop, vehicle behaviour becomes more uniform, yielding less difference between floating car velocities and the link velocities	Around 40
75-120	[Traffic Recovery] The traffic demand decreases, but congestion is still apparent on some links. For those links the estimations are quite accurate. The recovering links are responsible for the increasing MSE. The cause is again related to the low sampling rate on recovered links. As the velocities estimations on the link itself are derived from the GPS of only a few vehicles, a badly identified neighbourhood which do not share the same free flow traffic phase (yet) might overrule the estimation from the link itself. As obviously the neighbourhood yields a lower variance in results and gets weighted highly accordingly, the measurements on the free flow link itself get overruled. It is in this period that differences between variants become apparent. As the traffic fully recovers, the MSE hints back to the level associated with the outside rush hour cluster.	Around 60

Table 19: Detailed causes for varying MSE per identified cluster

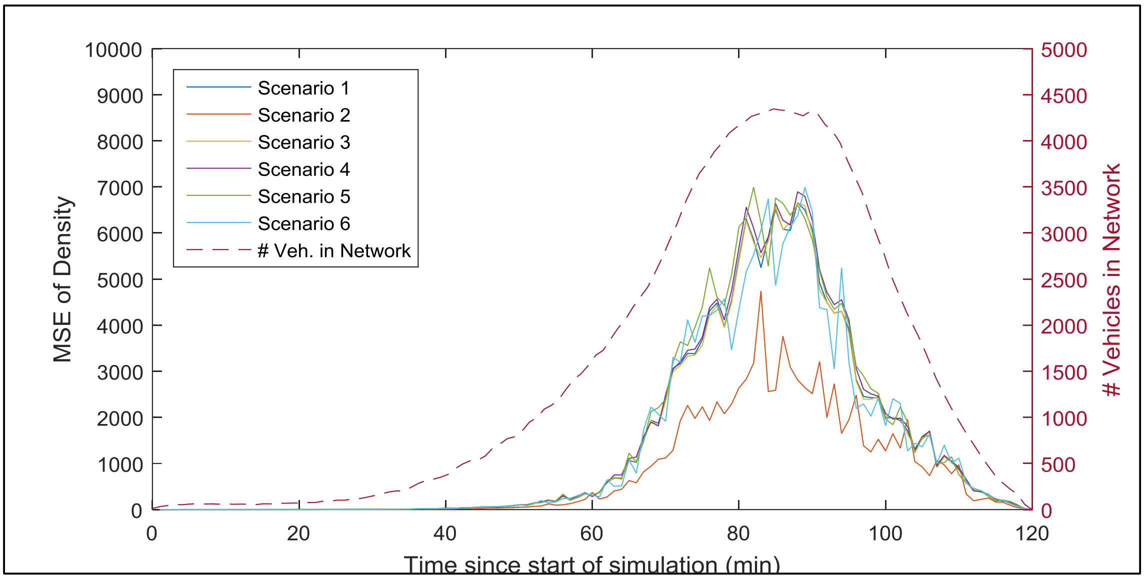

Figure 32: Density MSE per variant

Regarding the density estimations, the same pattern as in the morning rush hour can be found in the evening rush hour. Again the density graph moves together with the number of vehicles in the network and variant 2 shows a lot of improvement over all the other variants. The peaks experienced are the results of not representative initial density estimations due to less to none vehicles passing the ILD at the end of the link. Therefore the density is estimated well below the ground truth density.

The flow estimations errors in evening rush hour show also the same pattern as in the morning rush hour. The errors are higher due to the fact that the link flows around t=60 are higher than in morning rush hour, which again raises the potential error of being wrong. Combine this with the previously mentioned stacking of errors effect and peaks up to 4x the average MSE can be experienced. The link flows are only higher during the traffic breakdown phase and due to a higher number of vehicles in the network at around t=90 minutes, 4500 versus 3600, the flow errors around that time period are actually lower in the evening rush hour than in the morning rush hour respectively.

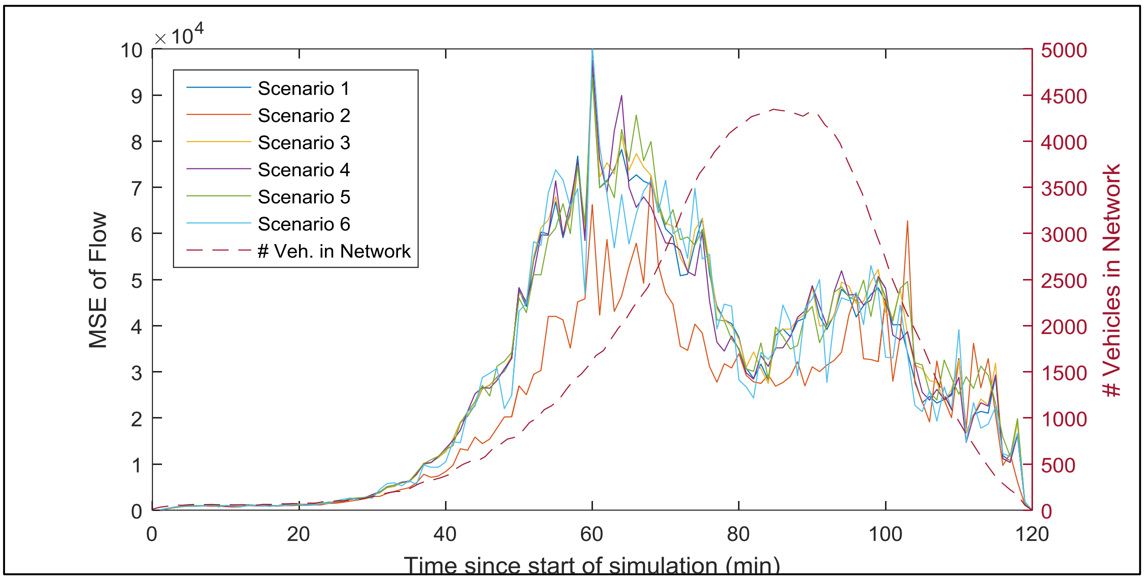

Figure 33: Flow MSE per variant

14.2.3 Outside Rush Hour Analysis

The velocity MSE is plotted for the outside rush hour below. Again separated per variant and with the average number of vehicles in the network included.

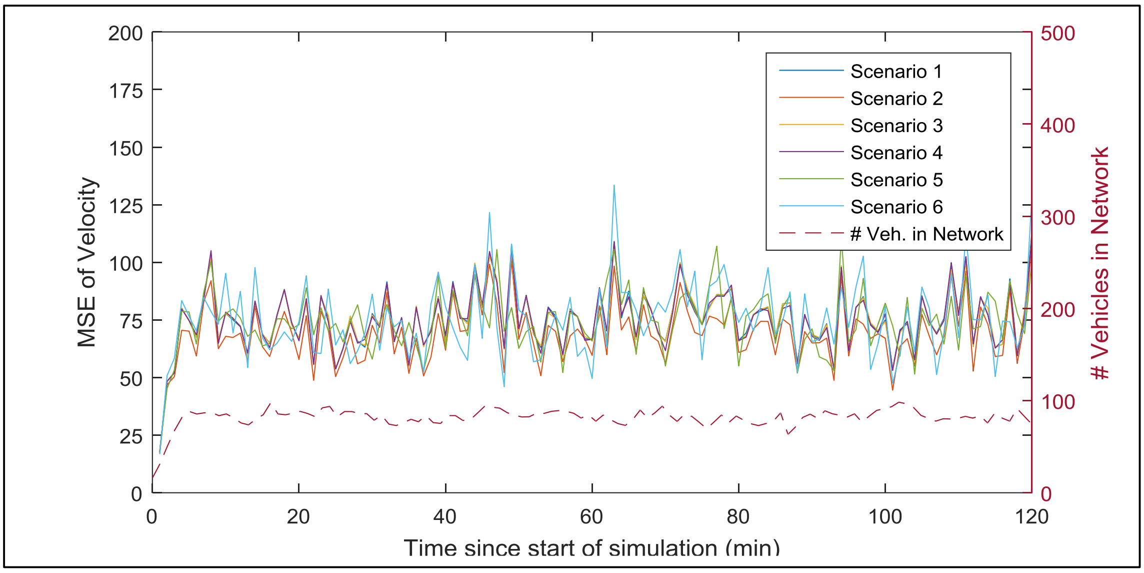

Figure 34: Velocity MSE per variant

The above chart, confirms the earlier identified cluster of outside rush hour. With an average MSE of around 80, the chart cannot be sensibly further partitioned as behaviour is capricious but consistent throughout. For completeness the one cluster is described below.

Time	Description	MSE Range
0-120	[Bounded] There is quite some traffic with 100 vehicles driving around on all 76 links, but even with 5%-10% FCD obviously most of the links have a zero sample rate and thus get estimated near v_{max}, while vehicles can again travel above or lower, generating the highly capricious MSE.	Around 80

Table 20: Detailed causes for varying MSE per identified cluster

Additionally the high MSE during this period is not caused by a single link, but by a group of links. In the graph below, each links' share towards the velocity MSE during the outside rush hour simulations is plotted. Coincidentally the more or less the same graphs can be derived from the rush hour periods. Throughout all simulated periods, the highway links (1 through 8) are responsible for around 60% of the MSE, whereas they only comprise 11% of the network. It's due to both the low traffic intensities as well as the higher v_{max} (70 mi/h vs 30 mi/h) that the PARAMICS distribution and parameters of driving behaviour, generates an already highly capricious ground truth.

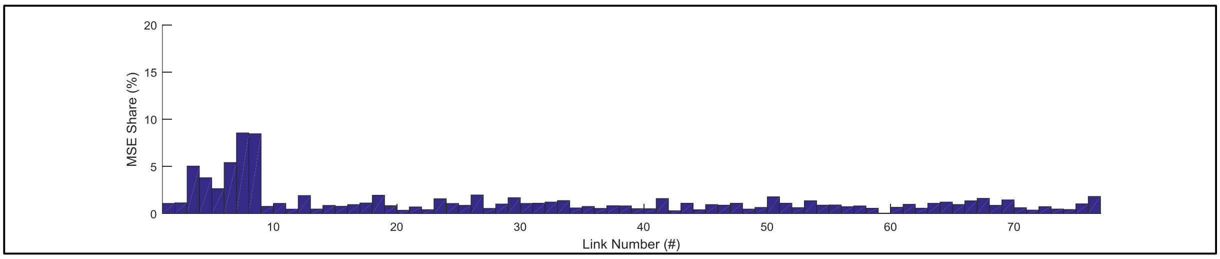

Figure 35: Velocity MSE share towards total MSE of each link in the road network

! Note: As there is little density and therefore almost no flow, the density and flow analysis for the outside rush hour is deemed uninteresting and therefore omitted.

14.3 Detailed analysis

The synthesis part of the results is aimed towards analysing the effects and contents of the different variants more thoroughly. Also included in this synthesis is the estimation biases, estimation correlation with the goal to find out which elements from the variants can be used to improve the baseline variant as to design a final variant. This is achieved by analysing each variant individually.

In general when using the Neighbour Link Method, the velocity and density estimations show high correlations in the rush hour periods (0,85 and 0,75 respectively) with the flow counterpart at around 0,69 throughout most variants. The poor correlation during outside rush hour for density and flow estimations must be put into perspective. As flows and densities are very low, the maximum difference is well below 20 cars. The table below reveals that only the addition of more FCD seems to (positively) effect the correlation of outputs (>0.05).

r	**Morning Rush Hour**			**Evening Rush Hour**			**Outside Rush Hour**		
Variant	v	p	q	v	p	q	v	p	q
1	**0,851**	**0,766**	**0,687**	**0,849**	**0,751**	**0,685**	**0,821**	**0,294**	**0,268**
2	0,000	+0,100	+0,085	-0,005	+0,103	+0,076	+0,003	+0,118	+0,104
3	-0,003	-0,002	-0,006	-0,004	+0,002	-0,012	+0,001	0,000	0,000
4	+0,034	-0,006	+0,004	+0,038	-0,012	+0,003	+0,003	-0,006	-0,005
5	+0,015	+0,006	+0,007	+0,003	-0,015	-0,016	+0,003	+0,005	+0,004
6	-0,039	-0,025	-0,032	-0,026	-0,039	-0,037	0,000	-0,003	-0,011

Table 21: Comparison of correlation in which the results of variant 1 serve as an index.

14.3.1 Variant 1: Baseline

In this variant comparison, the results from the baseline variant are used as comparative basis. Though with an average MSE of 62 for both rush hour periods, this NLM variant already outperforms the LDI method of De Vries (2015) which showed an average MSE of 90 at 5% FCD. A more in-depth look at the estimated versus true velocity scatter plot (in which all estimated link velocities from the evening rush hour are plotted against their ground truth counterparts) reveals a small bias to overestimation. This is considered caused by fact that due to the traffic lights on almost every intersection in the road network, $v_{GT} < v_{max}$. Whereas implementing this different cut-off value for links without samples in the neighbourhood link network might solve this bias, it is difficult to estimate this exact value due to differences in link length and differences in traffic light cycle times. The correlation between velocity estimations and ground truth velocities is 0,849.

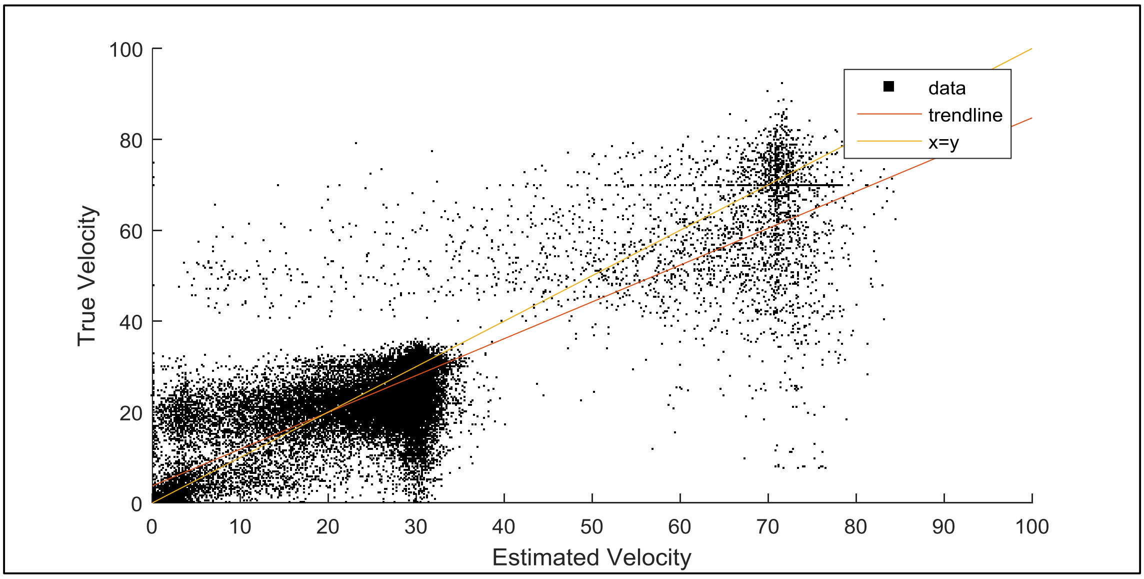

Figure 36: Estimated vs true velocity scatter plot

A further look in the density estimations shows that the biggest deviations are caused by serious underestimation of traffic density or overestimation of traffic density. The density scatter plot of estimated and true densities shows this very clearly. These are however not inherent to any certain link or type of link. They are a result of the way the local penetration rate ⋋ is calculated. A more thorough look reveals two simple improvements which can be made. First of all the penetration rate is equal to zero if none of the floating cars pass the detector, which yields a density of zero, while obviously still floating cars can be driving the link and delivering data. Suggested is to use the global penetration rate instead in this particular case. Secondly the floating car penetration rate can become arbitrarily large if the ratio between floating cars leaving the link and those on the link is small. Yielding unrealistic density estimations of up to 1000 vehicles per mile (one vehicle per 1½ meters). Capping the density estimations at 250 vehicles per mile is suggested for improvement.

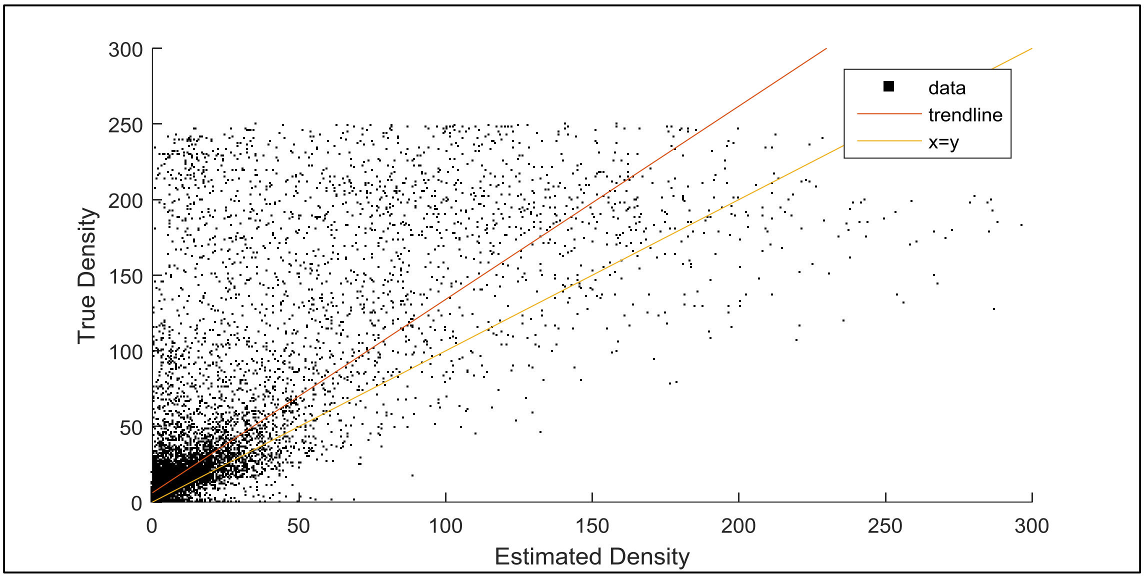

Figure 37: Estimated vs true density scatter Plot (r=0,751)

Regarding the flow estimations, the estimated vs true flow scatter plot shows the stacking of the errors in velocity and density. Flows of up to 5.000 veh/h/lane (outside of x-axis) are (without capping the density) experienced. With a maximum (one-minute peak value) flow of 1.200 veh/h/lane, additional capping of the flow at that threshold is suggested.

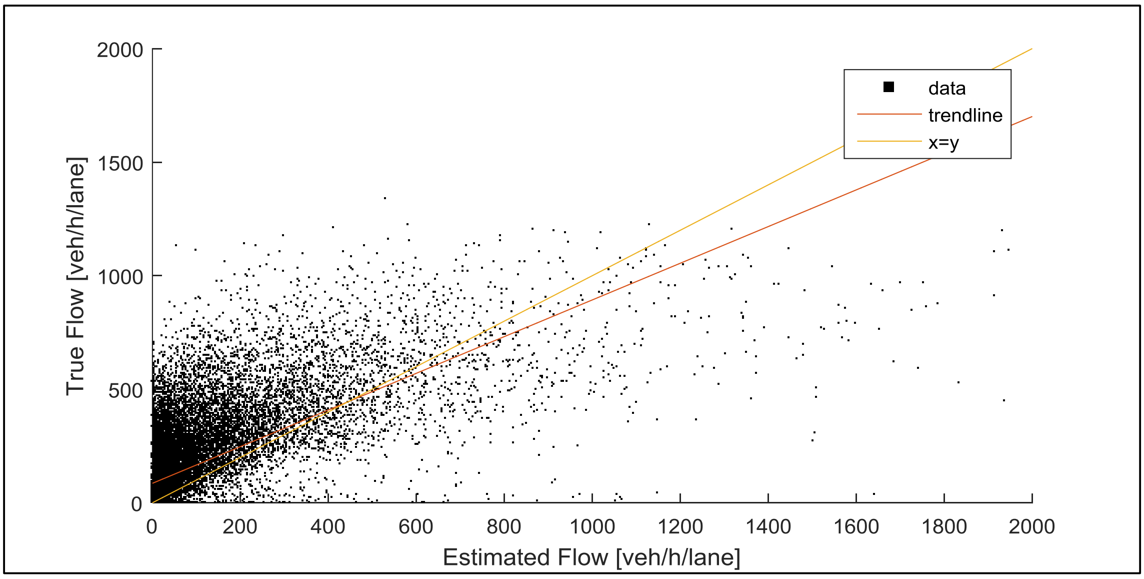

Figure 38: Estimated vs true flow scatter plot

14.3.2 Variant 2: 10% FCD

The results reveal that by doubling the number of cars equipped with GPS to 10% of the total number of vehicles in the network, all estimations improve. This improvement is the result of both an increased sample size (more measurements per link) and an increased link coverage (more links get covered by the sample of FCD). Where for velocity the estimation improves by 6,8% up to 11,7% in the different simulated periods, the density estimations improve vastly by 45,3% and 44,5% in the rush hour periods respectively and therefore comply with the hypothesis. The flow estimations improve by 8,8% outside rush hour to 24,0% inside rush hour. The correlations between estimation and ground truth increase slightly and the biases from the baseline variant still are apparent. While the data size is doubled, the runtime remains well under a second. This variant gives no (additional) hints as how to improve the baseline variant.

14.3.3 Variant 3: Neighbourhood space refreshing variant

The results from this variant are quite interesting. It shows that excessively refreshing the neighbourhood does not necessarily lead to a better estimation as only in one simulation timeframe the density estimation improves slightly by 0,6%. Generally one could consider it even might hamper the accuracy by up to 3,7%. It therefore confirms to the hypothesis that refreshing the neighbourhood space more, does not improve the accuracy. The neighbourhood space itself appears in two steps in the neighbour link method. Primarily it filters the network link list, to four candidates which behave somewhat the same throughout the modelled time periods. Secondarily these candidates are then used to find a weighting which comes closest for all the times in the modelled time period.

The first observation that must be made regarding the filtering, is that the operation of finding the highest correlating links, does not guarantee that any *appropriate* weighting can be found. Consider for example a link with 5 velocity measurements: (30, 29, 30, 27, 28) and two possibly correlated links with their respective measurements: (31, 27, 24, 21, 17) and (29, 30, 29, 31, 30). In this example one link is filtered. The correlation coefficients of the two links are +0,71 and -0,55 respectively, which means the first link is selected. However with the best weighting conceivable for link 1 at w=1,17 a squared error (SE) of 122 is generated, while for link 2 with a handpicked weighting of w=1,00 the SE is 23, five times smaller. Obviously there is no weighting for the 1st correlated link thinkable which will come even near a SE of 23. A different approach might provide better results e.g. Spatial Correlation (Esaway, 2012) or the correlation in fluxes of the traffic variables, instead of the values of the traffic variables themselves, however these are not further investigated in this research.

The second observation regards the actual effect of refreshing the neighbourhood space during the run. In this variant for all times the most correlated links are found and weighted accordingly. This means that instead of looking at the previous run(s) as a whole, local capricious variances in traffic variables effect the neighbourhood space. This effect becomes stronger when later on in the run a new (better) localized neighbourhood space is found and then used to minimize the MSE to find the best weighting solution for all times up to this time. When minimizing the MSE, it must take the time into account at which another neighbourhood space would be more correlated. Due to the bespoken effect in the first observation that another neighbourhood does not necessarily implies a better or worse solution, it can go either way. The graph presented on the next page shows the effects of in-run neighbourhood space refreshing on the neighbourhood space of link x=10 in evening rush hour. Whereas the neighbouring space commences at $N_{10} = [7, 12, 18, 71]$, it cycles through a lot of links [7, 12, 18, 39, 40, 41, 43, 50, 51, 57, 71] especially in the second half of the run. It is noted that the links in the centre of the network, show less extreme symptoms.

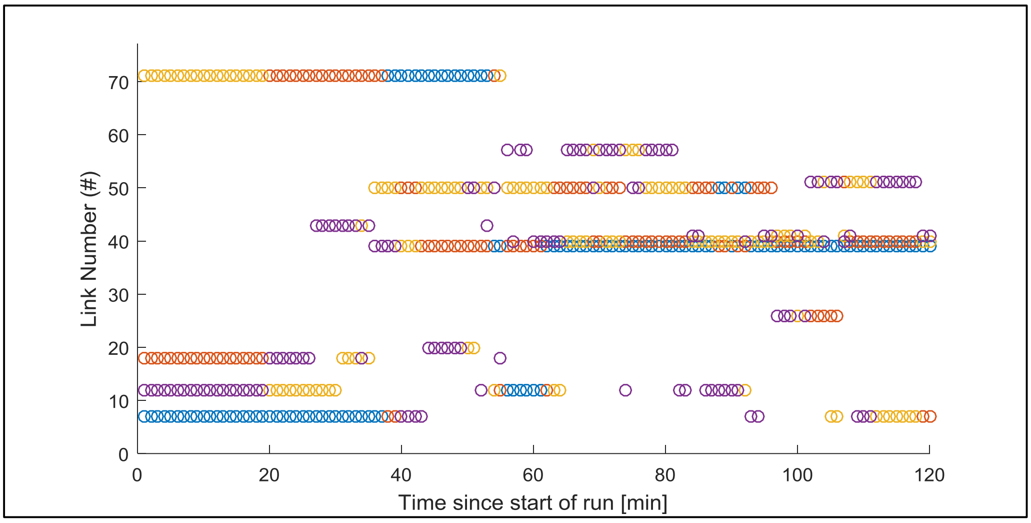

Figure 39: Changing of the neighbourhood space of x=10 during M=2 at evening rush hour.

Additionally the runtime in this variant increases approximately twofold due to the refreshing of the neighbourhood spaces once per minute, yielding a runtime of just under 2 seconds per minute of data considered. The bias of the results as well as the correlation to the ground truth does not change. Due to the capricious results showing no clear improvement with double calculation times, leaving the refreshment rate at the default of once per run is suggested.

14.3.4 Variant 4: Maximum Database Size Variant

Setting a maximum size to the traffic database used, has a clear effect on the performance of velocity estimations. As previously mentioned, they degrade with each consecutive run in all other variants and limiting the database size effectively from 5 to 2 runs, solves this degradation issue. The cause for the degradation is identified to be related to the inter-between day differences, simulated by the random seed in PARAMICS. Whereas using a larger database means that due to the smoothening a less capricious velocity estimations are outputted, using a smaller database current traffic data and slight deviations from the historical traffic situation are weighted more and therefore are captured more accurately just as expected from the hypothesis, which is hereby confirmed. The result is velocity estimations which improve by up to 11,9% in rush hour periods. However a smaller database size comes also with a disadvantage regarding the density. As there is already only a small fraction of data available during the real congested periods (i.e. 25 to 40 minutes of the full 120), a smaller database means there is even less data on the actual congestion phases (where densities are high). Whereas the density estimations worsen with 0,7% and 3,0% in the rush hour periods, the positive effects regarding velocity outweighs the negative effects regarding density such that the flow estimations also improve by 0,7% and 1,4%. Minor performance changes occur in the outside rush hour period as higher densities and flows are non-existent. The correlation with the ground truth improves big for the velocity estimations but remains somewhat the same for every other traffic flow variable. Additionally the runtime decreases slightly compared to the baseline variant. As the effects on velocity are overwhelming and also effect the flow estimations positively, suggested is to improve the baseline variant by defining a maximum database size of 2.

14.3.5 Variant 5: Reversed ordering

In the reverse ordering variant the sensitivity towards the run(s) stored in the historical database is assessed. The results show that less suitable or less correlated runs with the run of investigation has a negative impact of up to 3% on the MSE of velocity, 6% on the MSE of density and 3% on the MSE of flow. In this research, the variance between runs (mimicking the day-to-day-variance in traffic) can therefore be responsible for different outcomes confirming the hypothesis. A profit of e.g. 2% in MSE therefore does not necessarily mean that the variant outperforms, but can just as well be a result of variance within day-to-day-traffic. The results of variant 3 and 6 might therefore not be a result of the changes in the method proposed, but a result of the sensitiveness due to a lesser selection of traffic data used as database seed. Devising up a method which selects the most relatable partition of data to aid improvement in a current run might help here, but is not researched further.

14.3.6 Variant 6: Bagging

This last variant defined two bins for the traffic database. Using the time at which the lowest average network velocity was measured, the database was partitioned at t=90 minutes. The results show that using this method, the accuracy of all traffic flow variables and their correlation with the ground truth worsen. This is related to (a combination) of the following reasons; firstly the database is split (meaning less data is used), which might have the same effect on density and flow as in variant 4. Secondly the local minimum of a link velocity is likely not to coincide with the global network velocity minimum. Whereas t=90 minutes coincides with the moment number of vehicles in the network is the largest, the graph below reveals that individual links might already be recovering at this time, or even just starting to get congested. In this graph the ground truth of the 5th run from the morning rush hour is shown in which for every link their velocity is plotted.

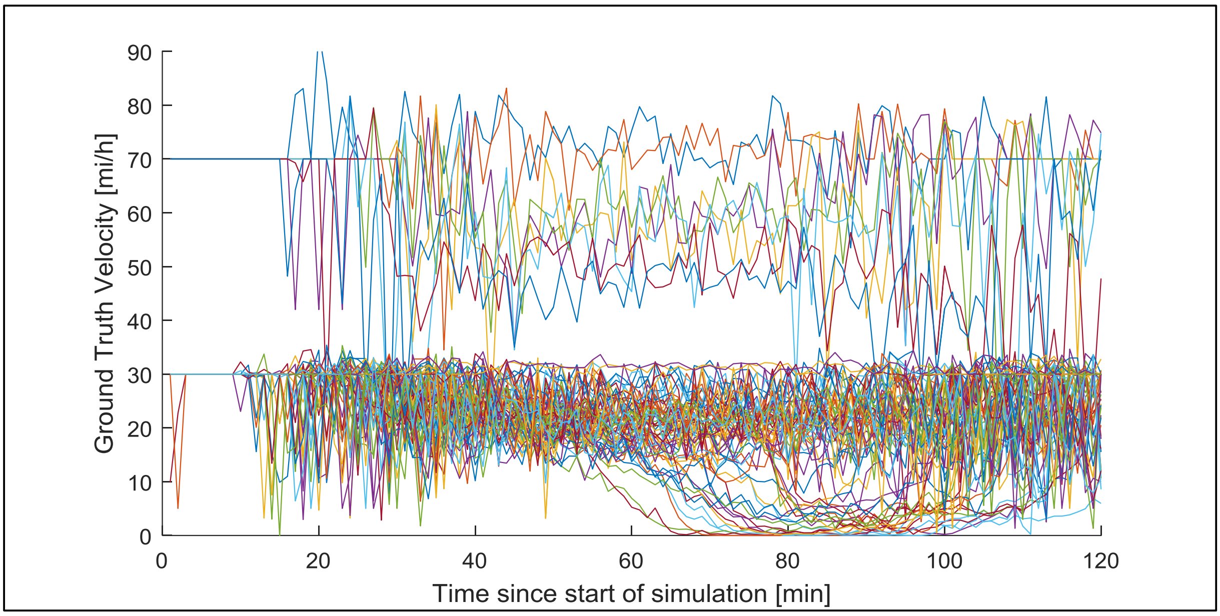

Figure 40: Ground truth velocities of all links in run M=5 of morning rush hour.

Some more trial-and-error runs were executed as to find out whether a split time lower or higher than 90 minutes would improve the traffic flow variables. This search yielded no results in which the estimation accuracy would increase. While in this variant the complexity is raised due to the partitioning, the runtime remains well below the one second mark per minute of data. Suggested is to not use additional bagging in the traffic state estimation part of this research as it provides no additional benefits. Therefore the hypothesis that bagging would improve accuracy is disproved. However in the traffic state prediction part of this research, bagging might seriously be required as the interest shifts from solely the ‘most current’ estimation to a future traffic state estimation which might require bagging to separate the traffic breakdown state with the traffic recovery state. Devising up a method to dynamically determine the partition time for each individual link might prove useful and will be further researched in the traffic state prediction part of this research.

14.4 Synthesized final variant

Combining both the analysis and synthesis of all variants, allows to design a final variant of the NLM in which both the knowledge and experience from all previously discussed variants are implemented. Using the baseline variant as basis, the improvements that make up this final variant are listed next.

14.4.1 The improvements

The list below reveals the improvements which were derived from the result analysis and synthesis and implemented in this final variant. No modifications to the framework are made as all changes can be implemented using fairly simple if-loops.

Capping of Density:	Every traffic estimation which yields an unrealistic density of over 250 veh/mi/lane, is capped at 250 veh/mi/lane.
Capping of Flow:	Every traffic estimation which yields an unrealistic one minute flow of over 1200 veh/h/lane, is capped to that value.
Hotfix ⋋ :	Whenever from the subset of floating cars, none pass the detector, ⋋ was set to 0, yielding 0 density. The fix assures that in this case the global FCD penetration rate is used $\overline{\leftthreetimes}\ (t) \approx 20$ in case of 5% FCD.
Capping database:	The database size is maximized to 2 runs, by dropping the less recent run whenever this limit is reached.

14.4.2 Results

The performance results regarding the estimations outputted by this final variant are:

	Morning Rush Hour			Evening Rush Hour			Outside Rush Hour		
	v	p	q	v	p	q	v	p	q
MSE of Run #2	56,3	172	17286	59,2	424	20800	72,9	1,7	1156
MSE of Run #3	49,5	249	16418	56,6	467	22783	73,9	1,8	1206
MSE of Run #4	52,8	262	18536	55,6	362	19815	74,8	1,8	1281
MSE of Run #5	50,3	306	18363	56,6	424	22879	79,1	1,7	1157
$\overline{\mathbf{MSE}}$	52,2	247	17651	57,0	419	21569	75,2	1,7	1200
r	0,885	0,933	0,774	0,886	0,931	0,756	0,824	0,181	0,181

Table 22: MSE per run, mean MSE and correlation for the final variant

Additionally, as from the global analysis became clear that the highway links can be responsible for up to 60% of the velocity MSE, whereas they only comprise 11% of the network, the same result table is presented in which the highway links are excluded. These results are presented in the table below. Excluding these highway links has only a small effect on flow and density MSE.

	Morning Rush Hour			Evening Rush Hour			Outside Rush Hour		
	v	p	q	v	p	q	v	p	q
MSE of Run #2	33,7	167	15951	35,8	392	19028	45,2	1,6	1032
MSE of Run #3	29,7	237	14923	34,9	443	20717	46,0	1,6	1083
MSE of Run #4	31,9	245	17058	34,7	325	17878	45,1	1,7	1116
MSE of Run #5	30,8	291	16827	34,7	377	21134	47,0	1,6	1053
$\overline{\mathbf{MSE}}$	31,5	235	16197	35,0	385	19689	45,8	1,6	1071
r	0,912	0,934	0,781	0,913	0,933	0,758	0,837	0,185	0,184

Table 23: MSE per run, mean MSE and correlation for the final variant ***excluding highway links***

14.4.3 Evaluation

Comparing these results with the baseline variant show serious improvements over both MSE and correlation. The slight degradation within the outside rush hour is the result of the previously discussed smaller database size. A more detailed look at the differences is provided next.

MSE	**Morning Rush Hour**			**Evening Rush Hour**			**Outside Rush Hour**		
Variant	v	p	q	v	p	q	v	p	q
Baseline	**58,9**	**808**	**24402**	**64,7**	**1387**	**29059**	**75,3**	**1,5**	**919**
Final	-11,4%	-69,4%	-27,7%	-11,9%	-69,8%	-25,8%	-0,1%	+13,3%	+30,6%
Final (Excl HW)	-46,5%	-70,9%	-83,4%	-45,9%	-72,2%	-83,1%	-39,2%	+6,7%	+16,5%

Table 24 Comparison of average MSE in which the results of variant 1 serve as an index.

r	**Morning Rush Hour**			**Evening Rush Hour**			**Outside Rush Hour**		
Variant	v	p	q	v	p	q	v	p	q
Baseline	**0,851**	**0,766**	**0,687**	**0,849**	**0,751**	**0,685**	**0,821**	**0,294**	**0,268**
Final	+0,034	+0,167	+0,087	+0,037	+0,180	+0,071	+0,003	-0,113	-0,087
Final (Excl HW)	+0,061	+0,168	+0,094	+0,064	+0,182	+0,073	+0,016	-0,109	-0,084

Table 25: Comparison of average correlation in which the results of variant 1 serve as an index.

By design the improvement of the velocity estimations is equal to the improvement achieved at variant #4. Exclusion of highway links in the MSE calculation yields a very visible difference, throughout the whole simulated period. For density and flow estimations, the same graph is plotted which shows the serious MSE improvement achieved in this final scenario as result of the implemented changes. The density and flow MSE comparison graphs referring to the morning rush hour can be found on the next page.

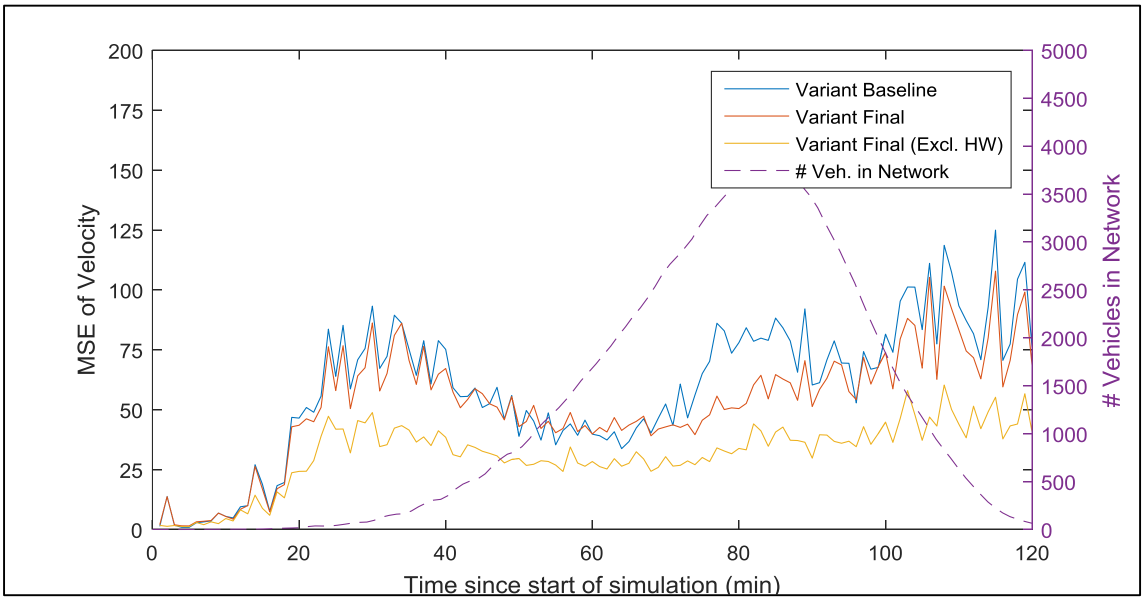

Figure 41: Velocity MSE compared with the baseline and exclusion of highways variants in morning rush hour

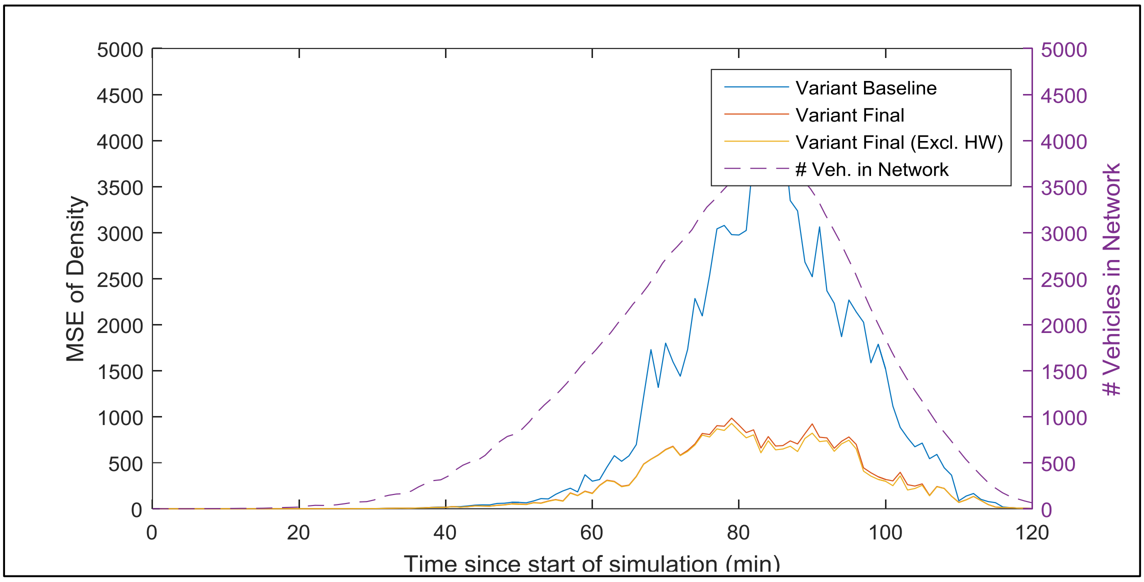

Figure 42: Density MSE compared with the baseline and exclusion of highways variants in morning rush hour

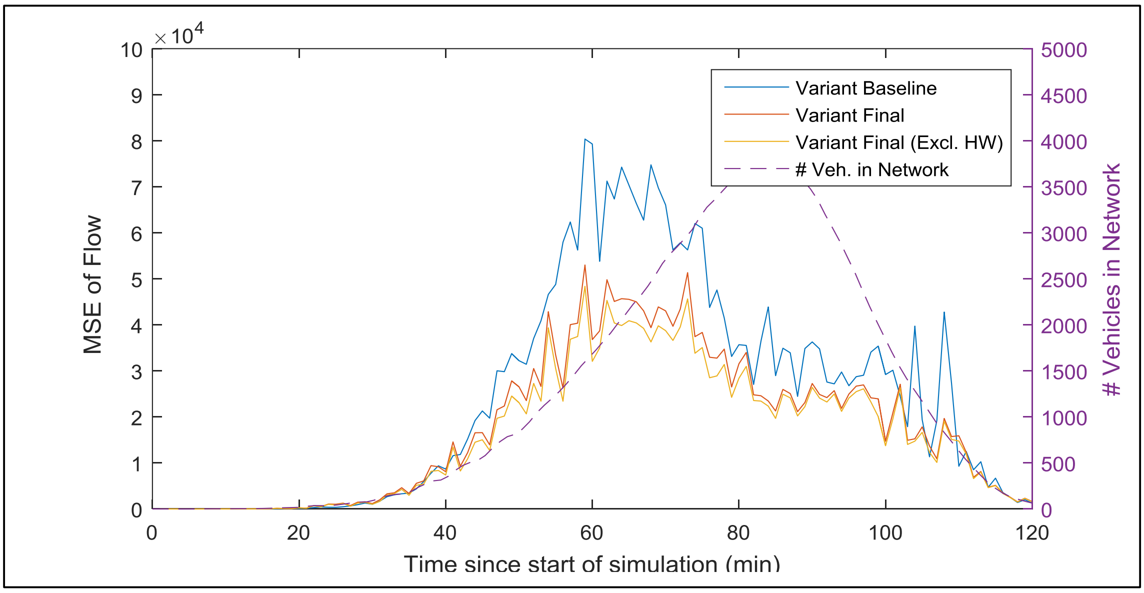

Figure 43: Flow MSE compared with the baseline and exclusion of highways variants in morning rush hour

From both figures it can be concluded that the settings chosen for this final variant are quite successful at both overall improvement as well as in peek reduction. All identified possible improvements are therefore confirmed to improve the results with the exception regarding outside rush hour. The above presented graphs show the average results of multiple runs and can be combined with the calculated reliability intervals to give an indication of how accurate the estimations given by the NLM are. When excluding the uncommonly traversed highway links, the velocity estimation accuracy on urban links is around 60% lower throughout every simulated period. It is however hard to transfer this global picture to a single link. To that extent the situation of one links x=13 and x=40 during run M=5 at morning rush hour is presented on the next page. This centralized link x=40 is selected as it is one of the first to show congestion and cycles through all the traffic states (free flow, breakdown, congestion and recovery). Link x=13 is one of the last links to show congestion. These links can therefore be considered to be both representative links in the network.

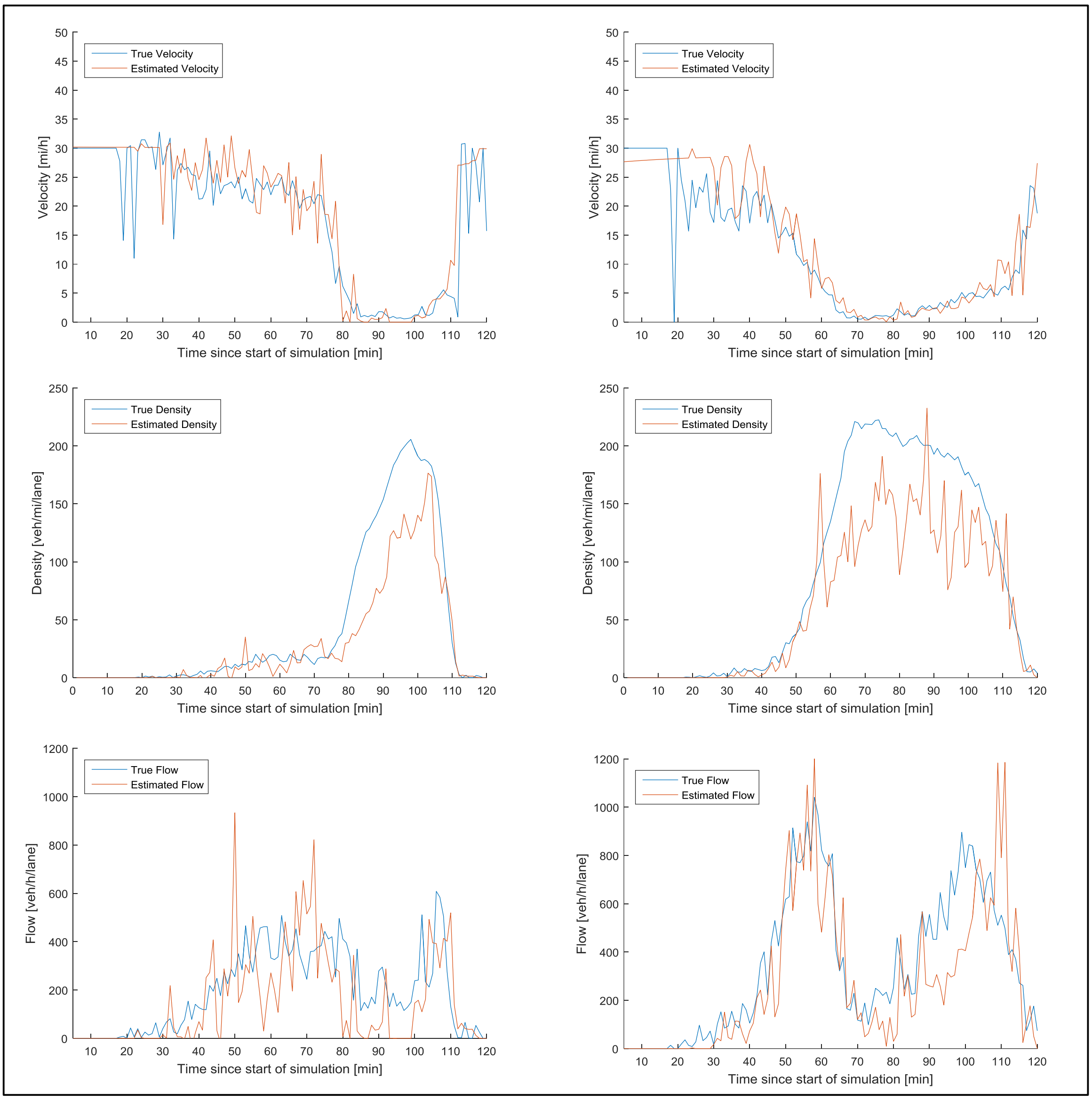

Figure 44: Velocity, density and flow comparison chart for link x=13 and x=40, run M=5, during morning rush hour

PART III

URBAN TRAFFIC STATE PREDICTION

15 NLM prediction framework

In this third part of this research, the previously used NLM for traffic state estimation is edited to output traffic state predictions. An important notice is that in this prediction framework the lessons learned from the traffic state estimation part of this research are already incorporated. NLM for prediction is based on the idea that by using patterns from historical traffic data, the current situation of neighbouring links can be extrapolated to serve as indicators for the future traffic state on a link. To assess the performance of the predictive abilities the same traffic data as used in the previous part of this research is used. The global structure of the NLM, repeated in the figure below, remains unmodified. The challenge in the prediction part of this research is therefore to derive an as accurately as possible real-time state prediction for a time in the future, beyond the most recent available traffic data.

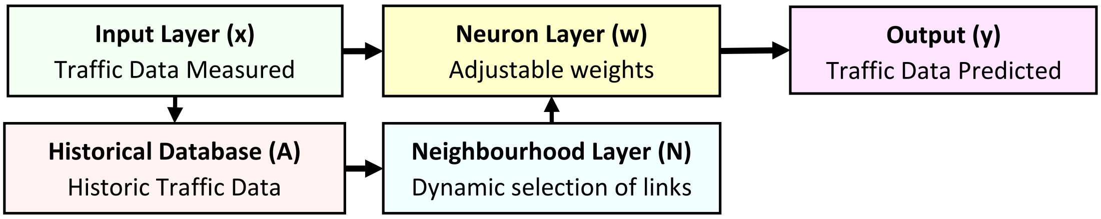

Figure 45: Global structure of the NLC method for traffic state prediction

15.1 Framework overview revisited

The goal of the NLM is again twofold. It must find the appropriate neighbourhood to be used as indication for the future traffic situation on each link and it must derive again a robust, complete and accurate picture of the urban traffic state on all links in the network for this time in the future. The difference between the time of most recent traffic data (noted: t_{max}) and the time of the prediction (noted: t_f) is branded as the prediction horizon and is noted as Δt_p. More specifically, if t_{max} would be the time of the most current data, the traffic state prediction is for time $t_{max} + \Delta t_p$. In the figure below these notations are visualised for clarity.

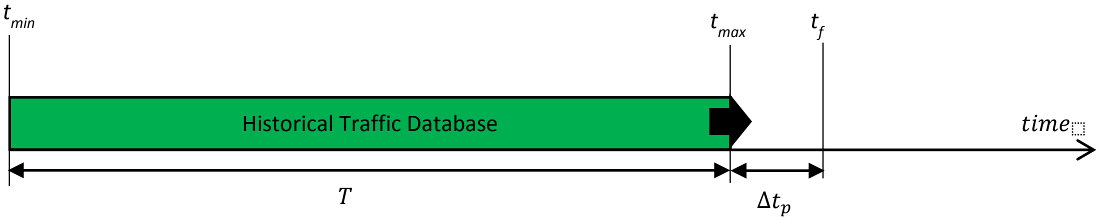

Figure 46: Visualized overview of the time notations

For completeness the legend of the symbols used:

t_{min} =	Time of first entry in database	[h]
t_{max} =	Time of last entry in database	[h]
$T = t_{max} - t_{min}$ =	Historical Traffic Database size	[h]
Δt_p =	Prediction horizon	[h]
$t_f = t_{max} + \Delta t_p =$	Time of traffic state prediction from data of time t_{max}	[h]

! *Note: In the final scenario of the traffic state estimation part, the database size (T) was capped at 2 runs, which resulted into a shifting* t_{min} *as the 'less recent' traffic data was dropped.*

To adjust the NLM from traffic state estimation to traffic state prediction, suggested is to remove the forced fusion of traffic data from the neighbourhood links and the link itself, because solely the link itself might not be a good indicator of the future traffic situation on the link. The exclusion of the link itself from the link neighbourhood is relaxed in the case it actually is a good indicator. This means that the steps 5 and 6 from the NLM framework for traffic estimation are omitted, creating the following step by step structure for the revised NLM framework for traffic state prediction. This chapter serves again as the guide-book, in which all subsequent steps taken to arrive at the output (y), are explained in more detail. The (mathematical) content and adjustments made in each step are explained in the next paragraphs.

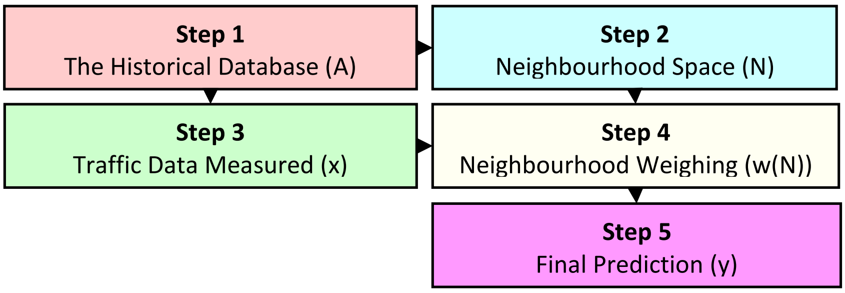

Figure 47: Step by step structure of the NLM framework for traffic state prediction

15.1.1 Step 1: The revised historical traffic database

No changes are made to the way the traffic data is stored.

15.1.2 Step 2: The revised neighbourhood space

The neighbourhood space of a link in the network estimation part of this research consisted out of links which could be used as indicators for a current (time t) traffic state on the link in question. The new definition of the neighbourhood space for traffic state prediction, takes the temporal difference of links into account. Therefore the revised neighbourhood space of a link consists of links which can be used as indicators for the future traffic state (at time t_f) on the link in question.

The first change applied to the neighbourhood space for a link $x \in X$ at a time t, is that there is no restriction applied on the inclusion of link x, as the traffic state on the link x at time t can be an indicator for the traffic state on link x at a time: $t + \Delta t_p$. Where the neighbourhood space size remains at 4 links, the mathematical description of the neighbourhood space becomes:

$$N_x(\mathrm{t}) = \{x_1, x_2, x_3, x_4\} \subset X$$

Again the links are selected by calculating for each link, the correlation between the historical database records for the link itself and every other link and selecting the four links yielding the highest Pearson's correlation coefficient. However, a slight modification is applied to implement the intended temporal difference Δt_p between the link measurements at a time t and the link prediction on the link in question at time $t + \Delta t_p$. In simple words, the correlation between a link x and all other links is determined by comparing the traffic data from times $t_{min} + \Delta t_p$ up to t_{max} on link x, to the traffic data of all links from all times t_{min} to $t_{max} - \Delta t_p$.The condition for choosing the 4 links that make up the neighbourhood space becomes mathematically:

$$corr\left(A_x(t), A_j\left(t - \Delta t_p\right)\right) \geq corr\left(A_x(t), A_k\left(t - \Delta t_p\right)\right) \quad | \, \forall \mathrm{j} \in N_x(\mathrm{t}) \;\; , \;\; \forall \mathrm{k} \in X \backslash N_x(\mathrm{t})$$

In which:

$A_x(t)$ = The historical database of e.g. velocities on link x for times: $t \in (t_{min} + \Delta t_p\ , t_{max})$

$A_x(t - \Delta t_p)$ = The historical database of e.g. velocities on link x for times: $t \in (t_{min}\ , t_{max} - \Delta t_p)$

15.1.3 Step 3: Considering newly arrived data revisited

The consideration of new data remains unmodified. That is for all links in the network, $\bar{v}(x, t)$: an indication of space mean speed and $\bar{p}(x, t)$: an indication of average density per lane, become available. The data from t_{max}, is in the next steps is however not used to derive the traffic state estimation at time t_{max} but to derive the traffic state at the time in the future: t_f.

15.1.4 Step 4: The revised neighbourhood space weighting

With the data from the previous steps, the goal of this neighbourhood space algorithm is to find for each link (x) a linear weighting of the links in the neighbourhood space $w_v(N_x(t))$ such that the linear combination of weights and traffic data for all times: $t \in (t_{min}\ , t_{max} - \Delta t_p)$ is equal to the indicator of velocity for all times $t \in (t_{min} + \Delta t_p\ , t_{max})$.

The optimization problem for finding these weights requires a small adjustment in the notation of the least squares problems for both velocity and density:

$$\min_{w_v(N_x(t))} \left|\left|A_{N_{v,x}}(t - \Delta t_p) * w_v(N_x(t)) - \bar{v}(x,t)\right|\right|^2 \qquad \forall t \in (t_{min} + \Delta t_p\ , t_{max}) \qquad [19a]$$

$$\min_{w_p(N_x(t))} \left|\left|A_{N_{p,x}}(t - \Delta t_p) * w_p(N_x(t)) - \bar{v}(x,t)\right|\right|^2 \qquad \forall t \in (t_{min} + \Delta t_p\ , t_{max}) \qquad [19b]$$

In which:

$A_{N_{v,x}}(t - \Delta t_p)$ = The historical database of velocities on neighbourhood links N_x. [mi/h]

$A_{N_{p,x}}(t - \Delta t_p)$ = The historical database of densities on neighbourhood links N_x. [veh/mi/lane]

15.1.5 Step 5: Revised final prediction

As there is no traffic data available for the time of prediction t_f, the per se fusing of neighbourhood traffic data and traffic data from the link itself is omitted. However, the algorithm might include its own link in its neighbourhood space now and weighting it accordingly. The final prediction of velocity and density at a time t_f in the future is then calculated using simple multiplication of the weights and the most current traffic data at time t_{max} using the formulas below:

$$\hat{v}(x, t_f) = \sum_{j=1}^{|N_x|} \bar{v}(N_{x,j}(t_{max})) * w_v(N_x(t)) \qquad [20a]$$

$$\hat{p}(x, t_f) = \sum_{j=1}^{|N_x|} \bar{p}(N_{x,j}(t_{max})) * w_p(N_x(t)) \qquad [20b]$$

$$\hat{q}(x, t_f) = \bar{v}(x, t_f) * \bar{p}(x, t_f) \qquad [20c]$$

16 NLM prediction variants

The next step in this research is again to apply the neighbour link framework on the whole set of case study data. However this time it serves as to assess the state prediction performance of the neighbour link method throughout the different modelled and simulated periods. Again variants are set up to allow for comparison, assessment of performance, sensitivity, analysis and synthesis. The changing basis for the two primary variants is a different prediction horizon of 5 and 15 minutes. Additionally two dynamic clustering variants are tested in which a more intelligent approach towards traffic database bagging are proposed.

16.1 Variant #1: Baseline

In the baseline variant the prediction horizon is set to 5 minutes, whereas all settings from the final variant regarding state estimation remain unchanged. Again the database is initially filled with the traffic data from run #1 of the respective period. Subsequently the simulated runs are iteratively added to the databases and for each of these runs (2,3,4,5) the minute to minute predictions are evaluated on accuracy. Whenever the database size exceeds 2 runs, the oldest run is removed. For completeness; 5% of FCD is used and the neighbourhood space is refreshed at the end of each run in this variant.

Overview Characteristics Scenario #1:

FCD coverage:	5%
Neighbourhood Refresh:	Every run (M)
Database Size:	Maximized
Run order:	[1,2,3,4,5]
Prediction horizon:	5 minutes
Binning:	None

16.2 Variant #2: 15 minute prediction horizon

The second variant changes the setting regarding the prediction horizon from 5 to 15 minutes as opposed to the baseline variant. Again the minute to minute predictions are evaluated on accuracy. The goal of this variant is to analyse the sensitivity of the performance towards an increasing prediction horizon.

Overview Characteristics Scenario #2:

FCD coverage:	5%
Neighbourhood Refresh:	Every run (M)
Database Size:	Maximized
Run order:	[1,2,3,4,5]
Prediction horizon:	**15 minutes**
Binning:	None

16.3 Variant #3 & #4: Dynamic clustering

The third and fourth variants are more elaborate and computational approach to the used bagging variant in the traffic state estimation part of this research. The reason for formulating a more intelligent approach is due the weak results of the naïve, rudimentary method used. Suggested is to adopt two aspects of the local traffic state prediction framework formulated by Antoniou et al. (2013) into the neighbourhood prediction framework of the previous chapter. These two aspects are the dynamic clustering (DC) of the measurements in the historical database, and the measurement classification (MC) of the most recent measurement used to derive a final state prediction. In the figure on the next page, these two aspects are added to the design a dynamic clustering version of the NLM framework in which the partition time of each link can be different.

Overview Characteristics Scenario #3:

FCD coverage:	5%
Neighbourhood Refresh:	Every run (M)
Database Size:	Maximized
Run order:	[1,2,3,4,5]
Prediction horizon:	**5 minutes**
Binning:	**Dynamic**

Overview Characteristics Scenario #4:

FCD coverage:	5%
Neighbourhood Refresh:	Every run (M)
Database Size:	Maximized
Run order:	[1,2,3,4,5]
Prediction horizon:	**15 minutes**
Binning:	**Dynamic**

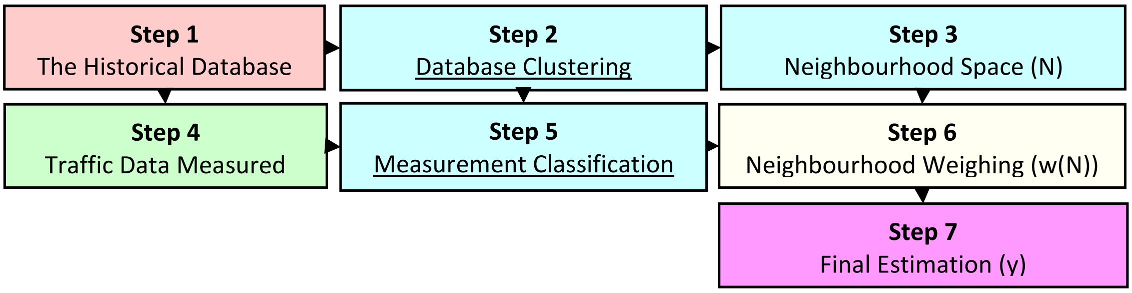

Figure 48: Modified structure of the NLM framework for dynamic clustering

The contents of all steps (1 through 7) will be described more thoroughly next. Again a distinction is made for a prediction horizon of 5 minutes in variant #3 and a prediction horizon of 15 minutes in variant #4 to allow fair comparison with variant #1 and #2 respectively.

16.3.1 Step 1: The historical database

The entries in the historical database as a whole remain unmodified.

16.3.2 Step 2: Database clustering (DC)

For the traffic estimation bagging, the transition point between decreasing and increasing of the average velocity on all links was used. Trivially the time interval in which each individual link shows the minimum velocity within a simulation time period probably does not coincide with the calculated global transition point. Therefore a more robust approach is suggested to determine for each link the transition point between decreasing and increasing of the link velocity more accurately.

Proposed is to firstly use a moving average of 5 measurements. As each individual link's average velocity shows inherent capricious behaviour. This moving average calculated for each link from the stored velocities in the historical database. The time interval for which the moving average is minimum, is then selected as the transition point upon which the traffic database of that link is partitioned in two clusters. The first cluster captures the traffic state breakdown (if at all present), the second cluster the traffic state recovery. The neighbourhood space is then determined for each of the separate clusters. Mathematically the following minimization problem is solved for each link x:

$$\min_{i} \left| \sum_{i=t-2}^{i=t+2} v(x,i) \right| \quad where\ t \in (t_{min}, t_{max}) \qquad [21]$$

Consider the solution found is time t $= \acute{t}(x)$. Thereafter, the database of both velocity and density is split at its found transition point into: $A_x(t_{min}, \acute{t})$ and $A_x(\acute{t}+1, t_{max})$. For simplicity the first part is considered the first cluster (A_x^1), and the second part is considered the second cluster (A_x^2) of measurements. As for each link the solution is different, each link holds onto a different partitioning of the main traffic database.

16.3.3 Step 3: Neighbourhood space

The determination of the neighbourhood space remains unedited, but is performed on both database clusters separately. Therefore two neighbourhood spaces are created per link, N_x^1(t) and N_x^2(t) respectively. With each (possibly) a different selection of neighbour links.

16.3.4 Step 4: Traffic data measured

No changes are made to the most recent traffic data that is received.

16.3.5 Step 5: Measurement classification (MC)

Due to the simulated day-to-day variations between runs, as new measurements become available, another algorithm is needed to find out to which regime or cluster the measurement belongs. Suggested is to use the naïve method of k-nearest neighbourhood (a different neighbourhood in this case) developed by Yakowitz (1987). A more complex but intuitively more logical approach of identifying the cluster by velocity trend is not adopted because of the practical difficulty that comes with implementing this. During trial and error runs, the most difficulty was experienced due to the inherently capricious raw velocity measurements, making finding the correct trend a hassle already. The k-nearest neighbourhood method has proven to be a reliable non-parametric regression method for short-term traffic flow forecasting, and therefore fits well within this step of measurement classification. The goal is to find within the historical database the measurement with the least distance to the most recent measurement. This is done by comparing the most current array of traffic data ($v(t_{max})$) with all previously logged data (($v(t)$) where $t \in \{t_{min}\,, t_{max} - 1\}$ and finding the one with the least distance in between them. That is, searched is for some time t_i where the sum of squared deviations is minimal:

$$\min_{t_i} | \sum_{x=1}^{|X|} (\bar{v}(x, t_i) - \bar{v}(x, t_{max}))^2 \quad | \, for \; \forall \; t \in \{t_{min}\,, t_{max} - 1\} \qquad [22]$$

The found time t_i corresponds for each link to a possibly different cluster (as for each link the clustering operation of step 2 is repeated). Bluntly selecting for each link the cluster in which this found time t_i lies (either cluster 1 or 2) is however not sufficient. The reason for this is that the prediction horizon might cross the gap between the clusters, as the prediction is not for time t_{max}, but for the time in the future $t_f = t_{max} + \Delta t_p$. The correct cluster would be the cluster in which the time $t_i + \Delta t_p$ can be found. Therefore for each link the cluster corresponding to the time $t_i + \Delta t_p$ is identified upon which the corresponding neighbourhood spaces are loaded.

16.3.6 Step 6: Neighbourhood weighting

The neighbourhood weighting step is not modified other than that only loaded is the correct part of the neighbourhood space cluster and respective database cluster. Solving the least squares optimization problem yields the weighting arrays: w_v and w_p again.

16.3.7 Step 7: Final prediction

The final prediction becomes (where c represents the selected cluster):

$$\hat{v}(x, t_f) = \sum_{j=1}^{|N_x^c|} \bar{v}(N_{x,j}^c(t_{max})) * w_v(N_{x,j}^c(t)) \qquad [23a]$$

$$\hat{p}(x, t_f) = \sum_{j=1}^{|N_x^c|} \bar{p}(|N_{x,j}^c(t_{max})) * w_p(N_{x,j}^c(t)) \qquad [23b]$$

$$\hat{q}(x, t_f) = \bar{v}(x, t_f) * \bar{p}(x, t_f) \qquad [23c]$$

17 NLM prediction results

In this chapter the raw results of each of the traffic state prediction variants are displayed. The results are again presented separately for each simulation interval and for each run of that respective interval. Additionally the mean MSE ($\overline{\text{MSE}}$) and correlation of estimation versus ground truth are given. A more thorough look at the individual results, comparison between results, discussion and synthesis is saved for the next chapter.

17.1 The baseline prediction variant: 5 minutes

	Morning Rush Hour			Evening Rush Hour			Outside Rush Hour		
	v	p	q	v	p	q	v	p	q
MSE of Run #2	55,9	368	52532	64,8	745	69525	75,8	1,7	1716
MSE of Run #3	60,6	399	53464	71,9	1044	70897	77,7	1,8	2191
MSE of Run #4	61,8	452	56234	69,7	640	74228	80,8	2,0	2045
MSE of Run #5	59,7	475	57547	69,3	782	78180	83,2	1,7	1331
$\overline{\text{MSE}}$	59,5	424	54944	68,9	803	73208	79,4	1,8	1821
r	0,877	0,870	0,677	0,870	0,851	0,606	0,817	0,040	0,031

Table 26: MSE per run, mean MSE and correlation for baseline variant

17.2 Prediction variant: 15 minutes

	Morning Rush Hour			Evening Rush Hour			Outside Rush Hour		
	v	p	q	v	p	q	v	p	q
MSE of Run #2	64,3	762	104033	71,1	1714	137614	76,1	1,9	2099
MSE of Run #3	70,5	964	106023	78,2	1554	131532	77,6	4,6	1971
MSE of Run #4	68,8	931	102209	78,3	1747	146046	81,8	1,9	2458
MSE of Run #5	63,9	665	99704	74,4	1818	121621	84,5	1,7	1700
$\overline{\text{MSE}}$	66,9	831	102992	75,5	1708	134203	80,0	2,5	2057
r	0,864	0,756	0,475	0,854	0,676	0,369	0,815	0,021	0,029

Table 27: MSE per run, mean MSE and correlation for variant 2

17.3 Dynamic Clustering: 5 minutes

	Morning Rush Hour			Evening Rush Hour			Outside Rush Hour		
	v	p	q	v	p	q	v	p	q
MSE of Run #2	83,2	655	42874	73,7	1161	47290	78,4	11	1906
MSE of Run #3	65,0	704	36132	83,4	1168	54505	80,4	321	204166
MSE of Run #4	66,7	711	55521	70,5	943	51874	83,5	11	1835
MSE of Run #5	64,2	707	33827	72,1	1090	51333	85,9	2	1327
$\overline{\text{MSE}}$	69,8	694	42089	74,9	1091	51251	82,1	86	52309
r	0,811	0,789	0,499	0,838	0,795	0,487	0,794	0,003	0,000

Table 28: MSE per run, mean MSE and correlation for variant 3

17.4 Dynamic Clustering: 15 minutes

	Morning Rush Hour			Evening Rush Hour			Outside Rush Hour		
	v	p	q	v	p	q	v	p	q
MSE of Run #2	91,7	1523	119342	96,2	2794	80874	88,3	3,9	2272
MSE of Run #3	88,2	1223	68910	130,9	2054	87768	88,4	5,4	2534
MSE of Run #4	84,1	1591	104157	98,9	1965	72434	97,3	2,4	1432
MSE of Run #5	90,2	1332	73643	108,5	2015	91190	102,3	7,3	2983
$\overline{\text{MSE}}$	88,6	1417	91513	108,6	2207	83067	94,1	4,8	2305
r	0,776	0,555	0,262	0,517	0,578	0,277	0,755	0,017	0,015

Table 29: MSE per run, mean MSE and correlation for variant 4

18 NLM prediction synthesis

In this chapter, the results from the traffic state prediction variants are analysed further. First a general comparison between variants and the final estimation variant is presented. Next the results during the different simulated periods are assessed in more detail. Finally an in-depth variant analysis is performed, on three representative links in the network.

18.1 Global analysis

For comparing both the average MSE and correlation of each variant, two comparison tables are presented. Firstly a table in which the results of the baseline variant are compared to the results of all other variants. Secondly the results of the 2nd variant are compared to the results of variant 4, because they both share a 15 minute prediction horizon instead of 5 minutes. Both tables are presented below. It becomes directly clear that whereas the dynamic clustering was intended to improve the performance of the NLM for traffic state prediction, it fails at that for both velocity and density. However for flow predictions during rush hour, it shows serious improvement of 23,4% and 30,0% for both rush hour periods for a 5 minute prediction horizon and 11,1% and 38,1% for both rush hour periods with a 15 minute prediction horizon. The reason for this improvement is related to an average flow prediction improvement over the whole simulation period. While performance on velocity and density decreases, it is the effect of bagging that removes the slightly present overestimation bias of density and underestimation bias of velocity (as discussed in the traffic estimation part). As density and flow errors stack due to the multiplicative property of flow ($err(q) \approx err(p) + err(v)$), they are in the bagging scenarios more unbiased, and thus have a higher likelihood of cancelling each other out, especially when the flow predictions are potentially larger (at medium density and medium velocity and thus at high flow) and thus yield a more accurate flow prediction.

MSE	Morning Rush Hour			Evening Rush Hour			Outside Rush Hour		
Variant	v	p	q	v	p	q	v	p	q
1	**59,5**	**424**	**54944**	**68,9**	**803**	**73208**	**79,4**	**1,8**	**1821**
2	-12,4%	-96,0%	-87,4%	-9,6%	-112,7%	-83,3%	-0,8%	-38,9%	-13,0%
3	-17,3%	-63,7%	+23,4%	-8,7%	-35,9%	+30,0%	-3,4%	-4677,8%	-2772,5%
4	-48,9%	-234,2%	-66,6%	-57,6%	-174,8%	-13,5%	-18,5%	-166,7%	-26,6%

Table 30: Comparison of average MSE in which the results of variant 1 serve as the index (=100).

MSE	Morning Rush Hour			Evening Rush Hour			Outside Rush Hour		
Variant	v	p	q	v	p	q	v	p	q
2	**66,9**	**831**	**102992**	**75,5**	**1708**	**134203**	**80**	**2,5**	**2057**
4	-32,4%	-70,5%	+11,1%	-43,8%	-29,2%	+38,1%	-17,6%	-92,0%	-12,1%

Table 31: Comparison of average MSE in which the results of variant 2 serve as the index (=100).

Additionally the results in the outside rush hour period show huge differences throughout the variants. From -0,8% worsening regarding velocity predictions in variant #2 to -4678% worsened density predictions in variant #3. The cause for this excessive difference is related to the inherent challenges within the dynamic clustering variant and will be discussed in more detail later on. Again the time it takes to process one full minute of traffic data and to derive the network state prediction for this same minute takes less than one second on the modern laptop used in this research. The most time consuming steps are again the determination of the neighbourhood space (correlation) and the solving of the minimization problem (least squares) in the baseline variant. For the dynamic clustering method one additional intensive operations is performed which is the measurement clustering of the most recent traffic data.

18.1.1 Morning Rush Hour Analysis

The morning rush hour period is analysed first. In the chart below the average velocity MSE of all four morning rush hour runs within a variant are plotted per minute of the simulation. The variant numbers refer to the previously defined variants as; variant 1: the baseline prediction variant, variant 2: is the increased prediction horizon variant, variant 3: the dynamic clustering variant of the 1st variant and variant 4: the dynamic clustering version of the 2nd variant.

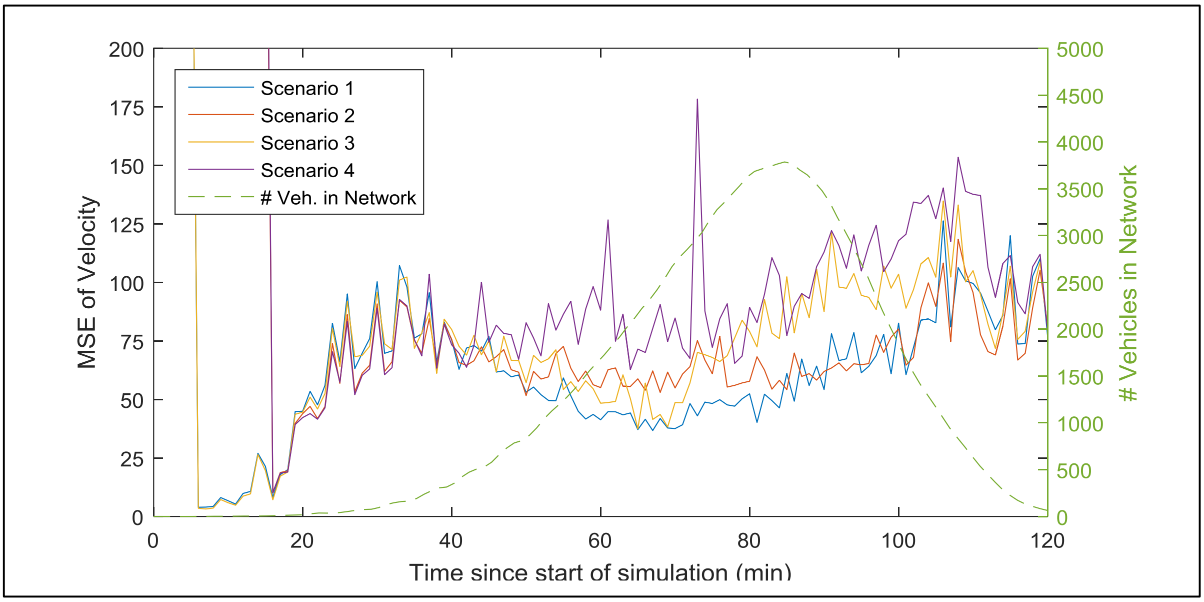

Figure 49: Velocity MSE per variant during morning rush hour

At first observation, both pairs of variants (1 & 3, 2 & 4) seem to move more or less hand-in-hand throughout the simulations. There are only a few moments visible in which the velocity prediction of the dynamic clustering counterpart is better than the basic version. The peeks experienced at t=5 minutes and t=15 minutes should be ignored as they are inherent of the method used (the data used at t=1 minutes yields no prediction earlier than t=6 or t=16). The pattern of the graph complies with the pattern of the graph in the estimation part of this research, as partitioning by characteristics of the graph yields the same partitions. The most difficulty is again experienced during the traffic state recovery phase, especially for when the prediction horizon is set at 15 minutes. The same can be concluded from the density MSE graph below.

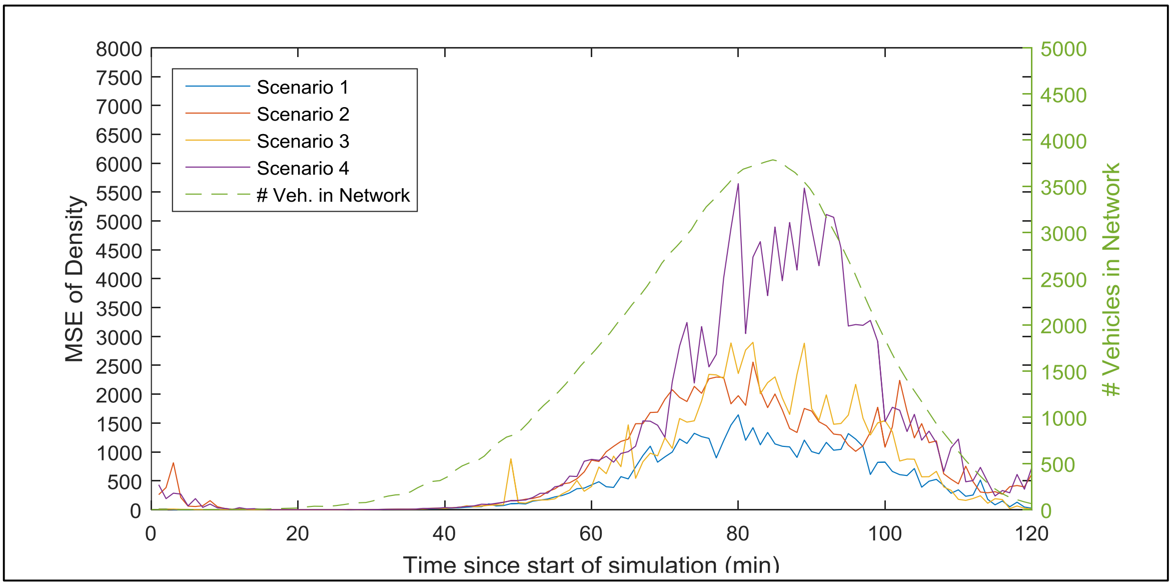

Figure 50: Density MSE per variant during morning rush hour

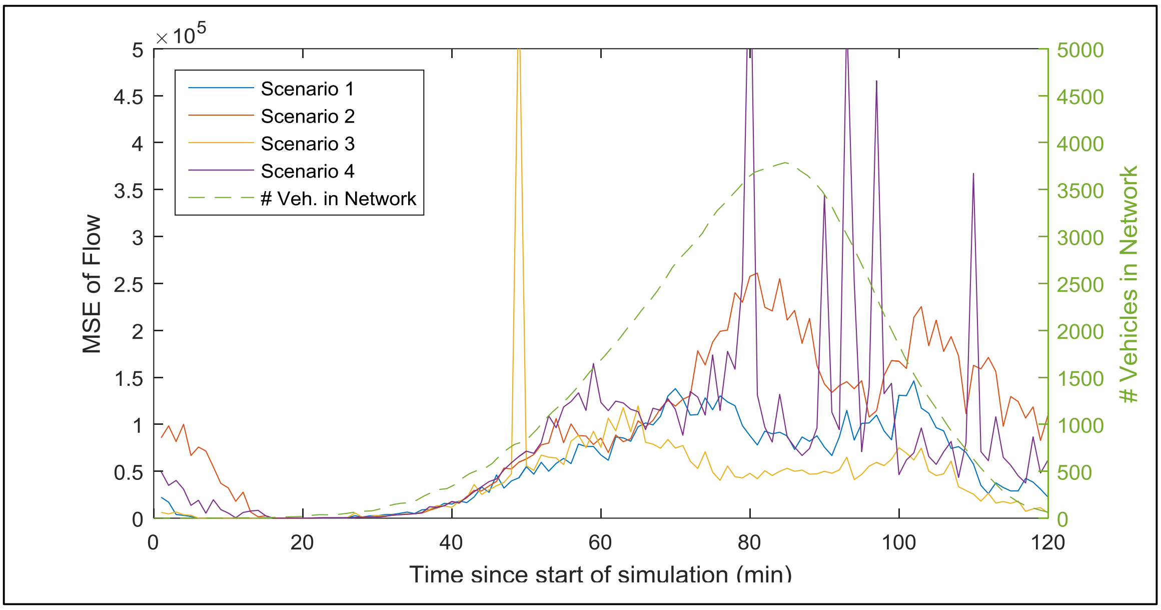

Figure 51: Flow MSE per variant during morning rush hour

The flow chart above reveals the time period for which the dynamic variants provide relatively more accurate predictions. From t=60 to t=120 both dynamic clustering variant's mostly outperform their respective counterparts. However, some obvious spikes are visible around the t=48, t=80, t=90, t=93, t=96 and t=110 marks which are coming from the dynamic clustering variants. The cause is further investigated in the individual variant assessments.

18.1.2 Evening rush hour analysis

The MSE plots of the evening rush hour period show mostly the same patterns as within the morning rush hour. Again the peaks until t=5 and t=15 should be disregarded, as well as the drop near the end due to the data cut-off at the last interval. Again the fourth variant performs worst overall while the first and third variant show less difference. The traffic state recovery phase from t=70 to t=120 is again identified as the most challenging phase.

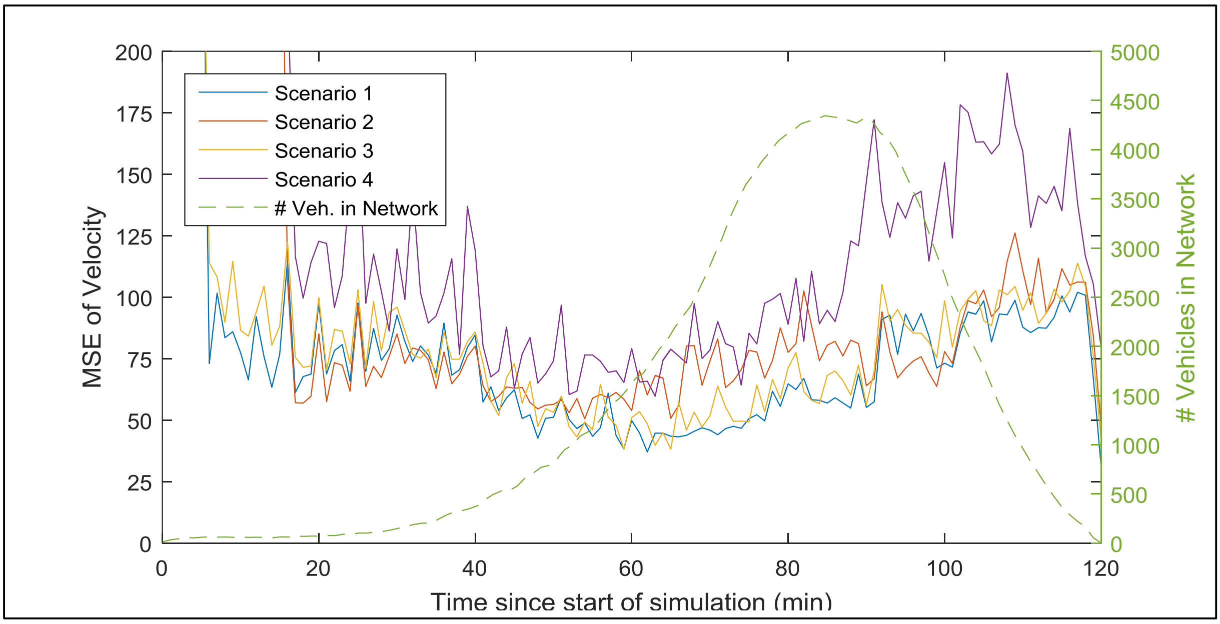

Figure 52: Velocity MSE per variant during evening rush hour

The density chart presented below shows a difference compared to the morning rush hour period. Whereas the density MSE between variants did not differ much from t=15 to t=70 and from t=100 to t=120 in morning rush hour. This chart shows that the density estimation performance of variant #2 is challenged during the last few minutes of the simulations. The cause for this is again further investigated in the individual variant assessments.

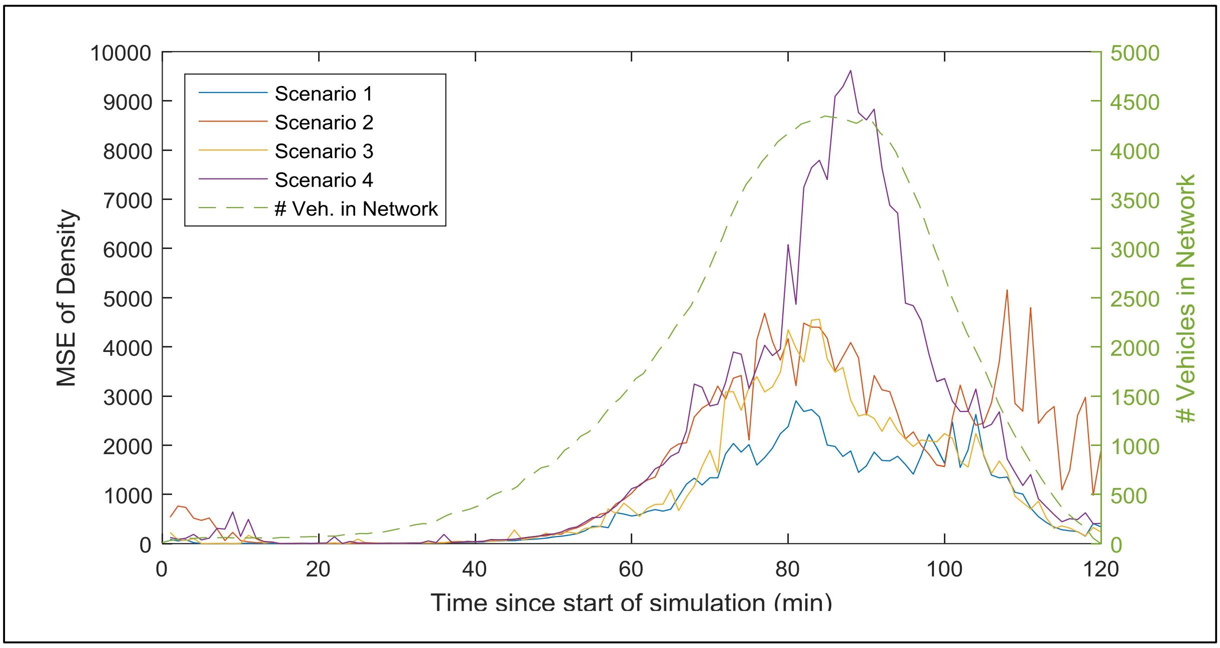

Figure 53: Density MSE per variant during evening rush hour

Ignoring the peaks before t=15, the flow chart below shows behaviour which was also seen in the morning rush hour. Again the dynamic variants provide (much) better predictions from t=60 to t=120, as they mostly outperform their respective counterparts. Also some spikes are visible, at t=22 and t=44.

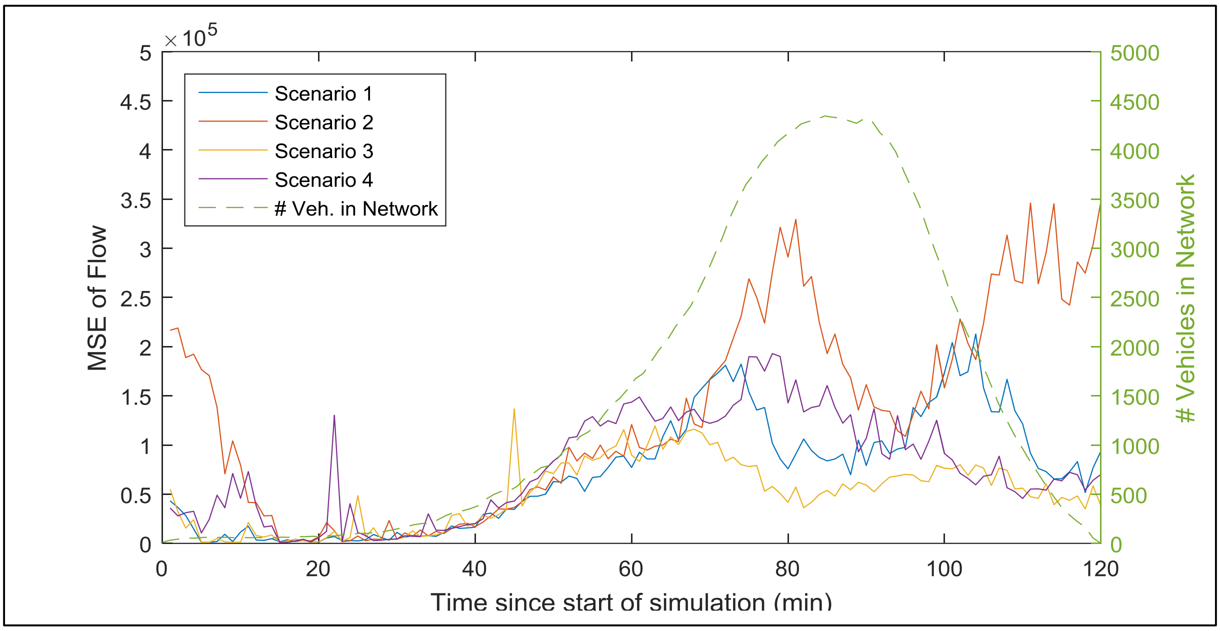

Figure 54: Flow MSE per variant during evening rush hour

18.1.3 Outside rush hour analysis

The same plots for the outside rush hour period, yield no new findings other than that the heights of the peaks experienced in density and flow MSE seem to dim as the time within the run increases. For completeness these plots are presented below.

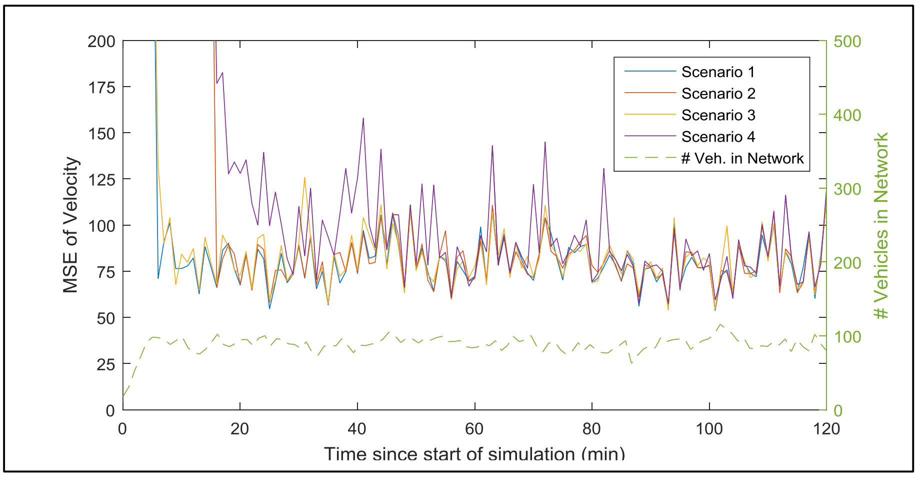

Figure 55: Mean velocity MSE per variant during outside rush hour

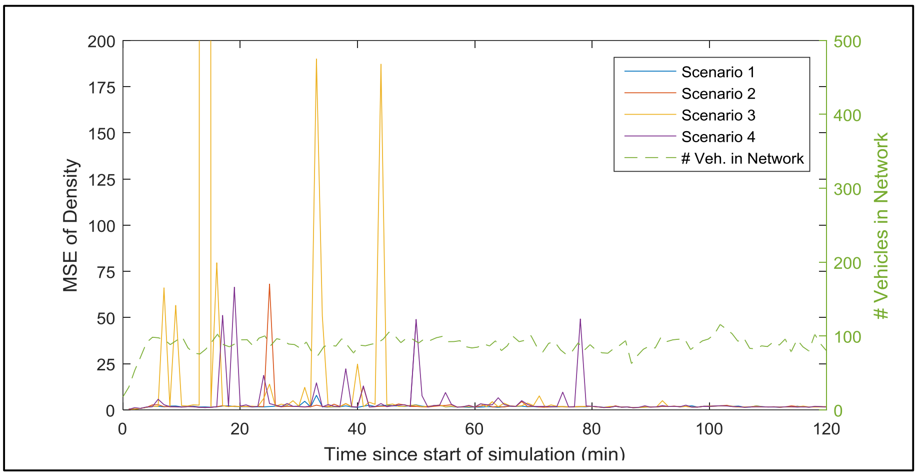

Figure 56: Mean Density MSE per variant during outside rush hour

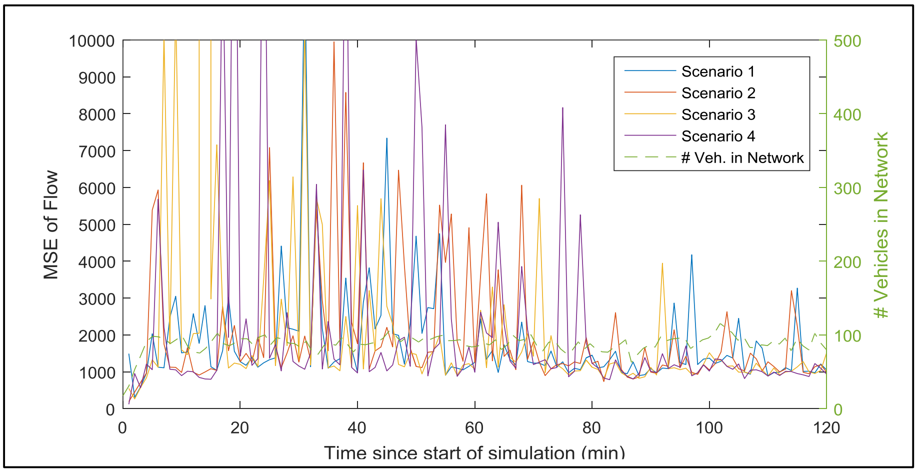

Figure 57: Mean flow MSE per variant during outside rush hour

18.2 Detailed analysis

The evaluation of the predictive ability of the NLM starts with looking at the correlation between prediction and ground truth. While for short term prediction the correlation coefficients do not differ much from the previously obtained state estimation results, the correlations of the 15 minute prediction variants are showing a less strong relationship. Using a general statistic threshold of 0,5 (indicating that 25% of the variance in the truth is predicted by the NLM), it is concluded that the flow prediction performance is very low in variant 2, 3 and 4. Additionally in the outside rush hour period, the density prediction performance is low as well. While the results from the outside rush hour period can be safely disregarded due to the fact that there is almost no flow and no density and thus not much to predict, the relatively bad flow predictions are unsatisfactory.

r	**Morning Rush Hour**			**Evening Rush Hour**			**Outside Rush Hour**		
Variant	v	p	q	v	p	q	v	p	q
1	**0,877**	**0,870**	**0,677**	**0,870**	**0,851**	**0,606**	**0,817**	**0,040**	**0,031**
2	0,864	0,756	0,475	0,854	0,676	0,369	0,815	0,021	0,029
3	0,811	0,789	0,499	0,838	0,795	0,487	0,794	0,003	0,000
4	0,776	0,555	0,262	0,517	0,578	0,277	0,755	0,017	0,015

Table 32: Correlation results per traffic flow variable

The questions that remain are; which reasons are there for this bad performance and are there ways to improve the prediction quality? To answer these questions, each individual variant result is investigated in more detail for a link in the network. It is chosen to present the full run data (M=5) for a selection of three different links, because of internal differences in prediction performance within the urban network. For this evaluation a link without congestion (x=73), a link with a limited duration of congestion (x=13) and a link which is the first to get congested (x=40) are presented. Their locations are designated in the figure below. The evaluation is continued after presenting the prediction dashboards for each of these links.

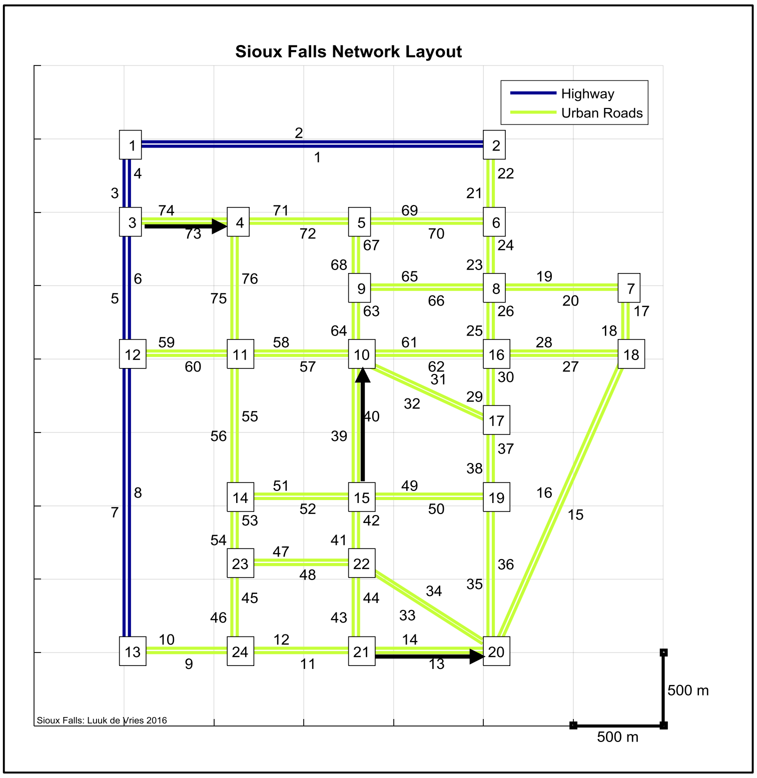

Figure 58: Network layout with the selected links highlighted.

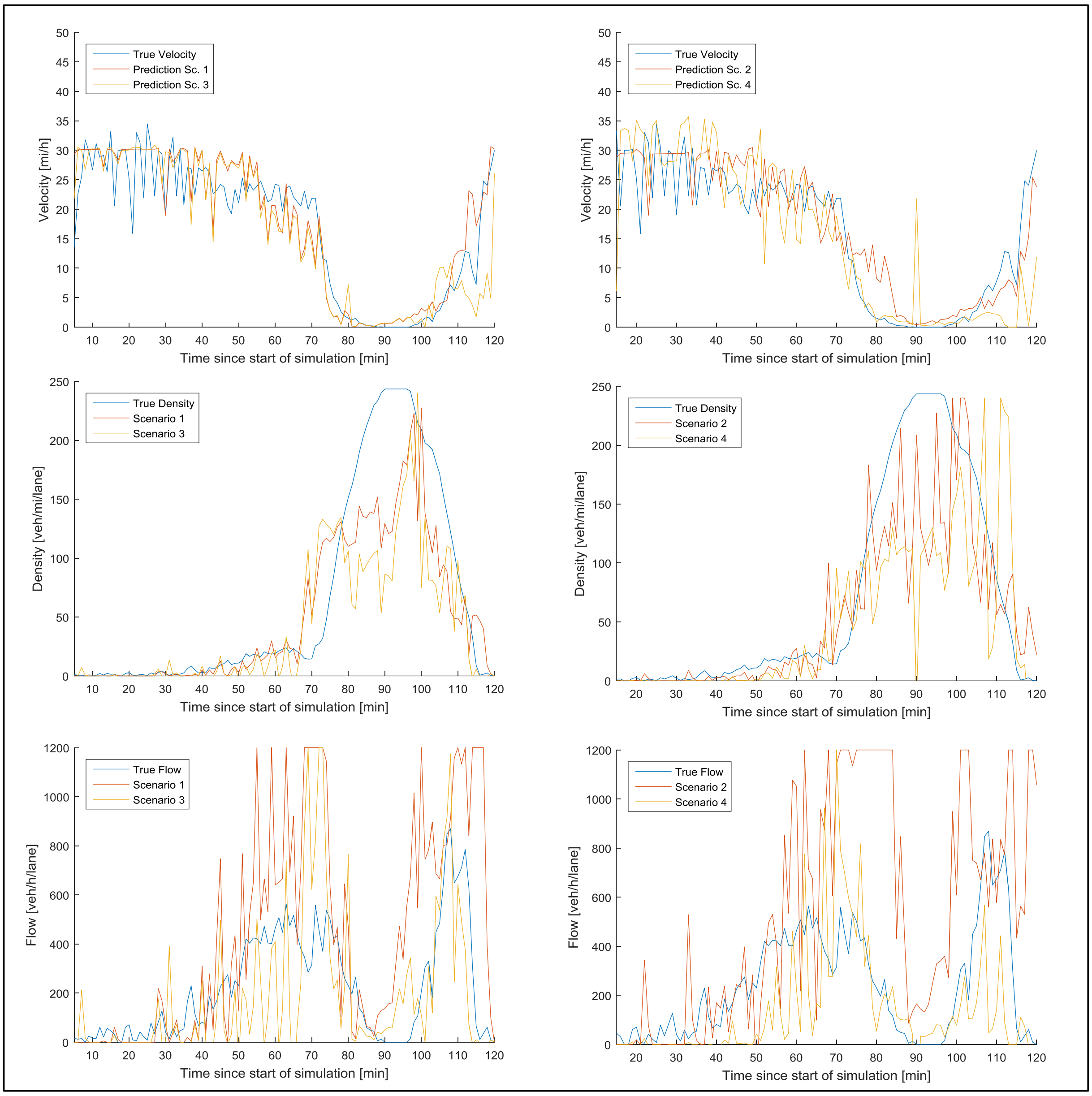

Figure 59: Velocity, density and flow comparison charts for link x=13, run M=5, during evening rush hour

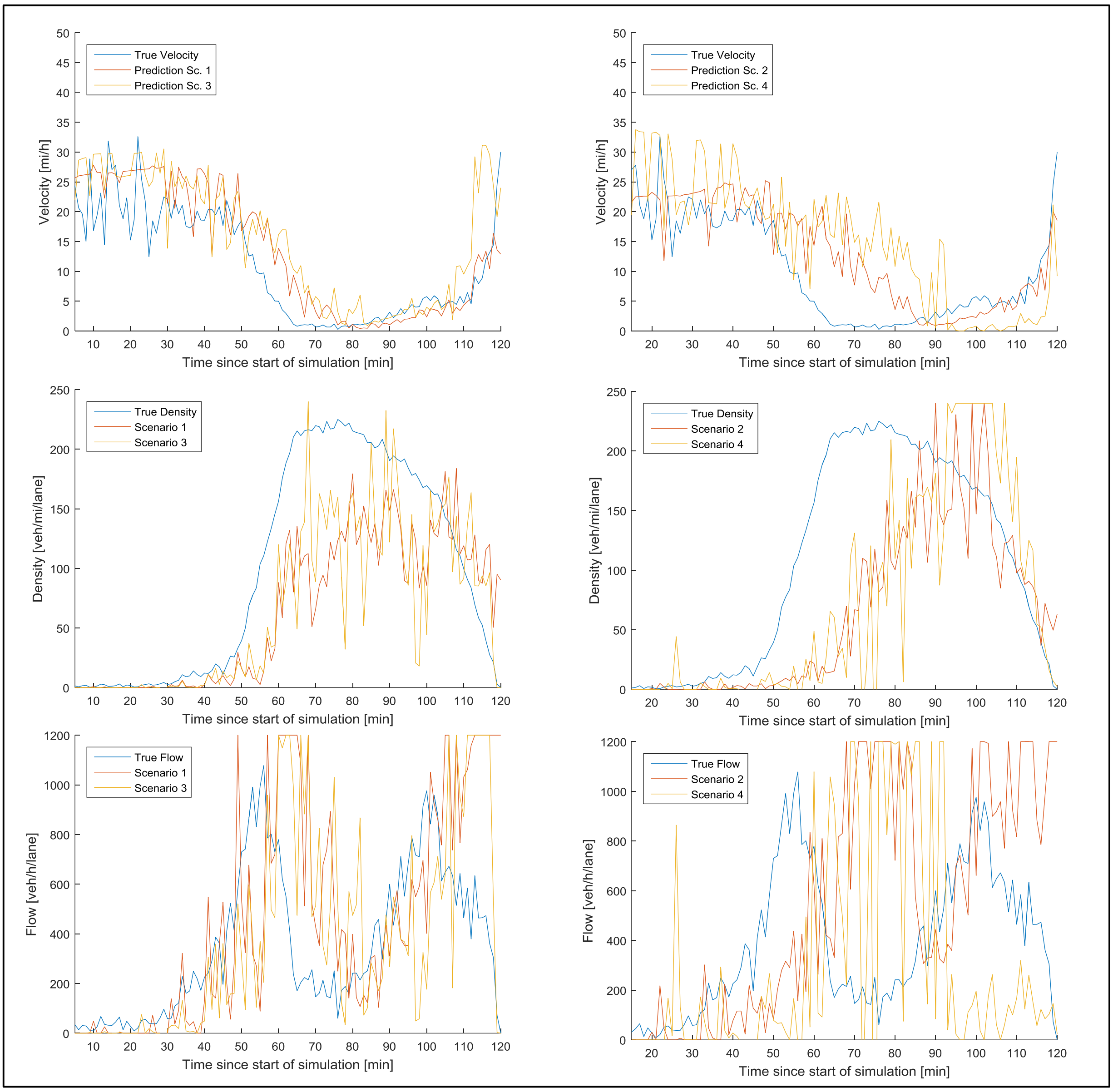

Figure 60: Velocity, density and flow comparison charts for link x=40, run M=5, during evening rush hour

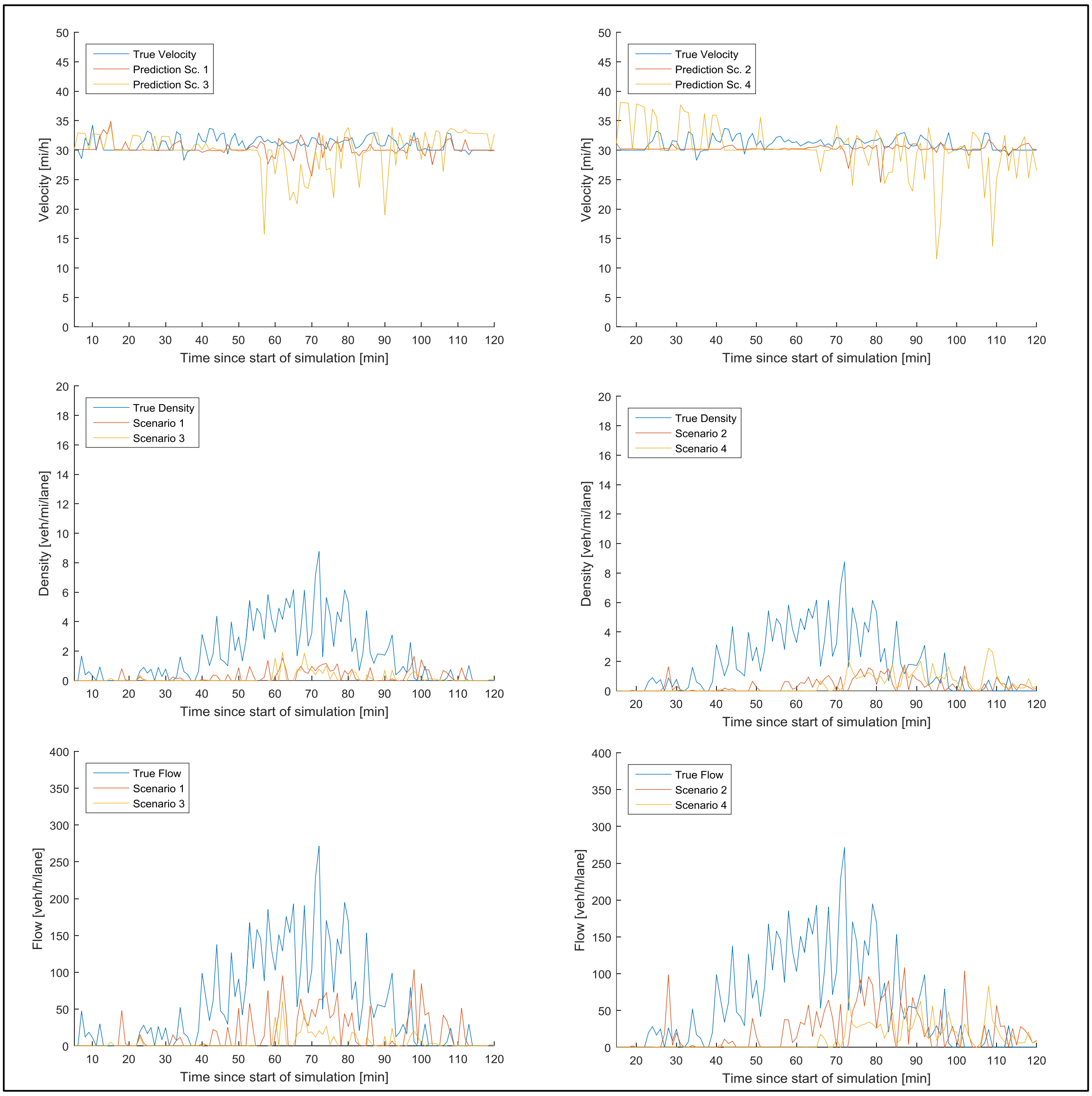

Figure 61: Velocity, density and flow comparison charts for link x=73, run M=5, during evening rush hour

18.2.1 Variant #1: Baseline

The predictive ability of the baseline variant for the velocity estimation of link x=13 is considered to be spot on. With a prediction horizon of 5 minutes, the up and down movements of predicted velocities and true velocities coincide throughout the full run. The reason for this success can be found by looking at the network map. Link x=13 is on the outskirts of the network. Whereas the congestion starts to form at t=60 in the centre of the network, it spreads towards x=13 approximately 15 minutes later. Additionally it is not one of the first links to recover fully as well. This means that it is theoretically possible to find link neighbours which predict the traffic state breakdown and traffic state recovery 5 minutes ahead of time. The velocity result for link x=13 complies with a MSE of 32,2 which is 54% better than average. The density prediction reveals a pattern in which a 5 minute delay can be observed between prediction and true density, which is caused by inclusion of the highly correlating links which ‘predict’ the situation 10 minutes into advance. The density result for link x=13 complies with a MSE of 966, 128% worse than average. The same pattern becomes visible in the flow prediction, where big errors in density and velocity estimations get amplified. The flow performance with a MSE of 77000, is 33% worse than average.

For link x=40, the same line of reasoning can be applied, although some interesting different results are observed. Again the velocity prediction follows the true velocity closely, but for the traffic breakdown phase it can be considered 5 minutes too late. The traffic recovery phase is predicted spot on. Again the link location is the reason for this behaviour. Link x=40’s centre location, makes it one of the first links to experience congestion and one of the last to recover from it. This means that the NLM cannot find a neighbourhood space in which links experience congestion earlier than link x=40, instead it finds the neighbourhood links that experience recovery earlier than link x=40. The velocity result for link x=40 complies with a MSE of 45,5 which is 24% better than average. For density the same applies, though due to the poor selection of links which predict the traffic state breakdown, a bias to underestimation of the link density is observed. The density result complies with a MSE 400% worse than average. The resulting flow predictions show the same 5 minute shift and more capriciousness due to the stacking and cancellation of prediction errors. The flow result complies with a MSE 350% worse than average.

For link x=73, the first variant interpolates the received traffic data in a way that a more or less stable velocity prediction is outputted at the v_{max} of 30 mi/h. The velocity prediction complies with a MSE of 25,8 which is 57% better than average. Some minor fluctuations can be seen due to small changes in the neighbouring links. The density prediction is clearly underestimating the true density, which is caused due to the insecurity of the initial density estimations as a result of the use of the global/local FCD penetration rate $(\leftthreetimes \text{ and } \overline{\leftthreetimes})$ on both the link itself as on neighbouring links. As a result the density prediction is relatively high up to a relative factor of 3, but in absolute terms negligible (<0.01% of MSE(p)). The same applies to the flow prediction (<0.1% of MSE(q)).

18.2.2 Variant #2: Increased prediction horizon

For a prediction horizon of 15 minutes, the velocity prediction of link x=13 is again quite close to the true velocity. Only in the traffic breakdown phase, the prediction slightly lags behind (5 to 10 minutes), which is related to the links in the neighbourhood not being a perfect *15*-minute indicator. This same reason applies as to why the density estimation is somewhat off. Again due to stacking of errors the amplitude of the flow graph is exaggerated. For link x=40, the previous effect of the inability to find a predictive neighbourhood for the traffic breakdown phase is amplified. The velocity prediction seems to be 15 minutes behind, until the traffic recovers, at which the prediction is more accurate. The same applies to the true density and as a result the flow is again all over the place. For link x=73, a larger prediction horizon does not yield a visible difference between the first and this second variant.

18.2.3 Variant #3: Dynamic clustering: 5 minutes

As previously mentioned, in this run, the dynamic clustering method finds the database split time to be at t=92 for link x=13. The neighbourhood space up until t=92 therefore contains the links which correlate most during the traffic breakdown phase, while after t=92 correlation is sought with the links which predict the traffic recovery phase. The measurement classification works well in this example, as the same classification is outputted until t=75. At around t=76 a misclassification is performed, yielding a prediction spike at t=81. This same misclassification occurs between t=100 and t=120 multiple times yielding falsely predicted traffic breakdown behaviour, where recovery is true. The same applies to the density predictions. Consequently the worse performance on velocity and density especially on the last 60 minutes of the run, yields a (slightly) worse flow prediction.

On link x=40, the database split occurs at t=73. This splitting does not change the fact that there are no neighbour links available which predict the traffic state breakdown, yielding the same 5 minute offset experienced as with the baseline variant. A surprising negative offset of 5 minutes is experienced during the traffic recovery phase in this variant. This strange behaviour is the result of the size of the congestive state in this second phase. While the congestion is worst at t=73, the true traffic state up to t=110 minutes can be considered to be congested. As measurements can be mistakenly wrongly classified as well, the algorithm finds the link itself to be most correlated to its future state, yielding a delay equal to the prediction horizon. The result is a slightly less accurate density prediction and a relative less accurate flow prediction as well.

For the link x=73, the database split time happens at t=52. In this example is becomes clear from the true velocity that splitting the run at that time does not yield two very different partitions. Logically the NLM has difficulty categorizing measurements in the ‘correct’ cluster, yielding slightly more capricious predictions for all traffic flow variables. Consequently it performs slightly worse than the first variant throughout the prediction run for this variant.

18.2.4 Variant #4: Dynamic clustering: 15 minutes

In the 15 minute prediction variants, the effects of misclassification of data becomes more visible. For link x=13, the newly arrived traffic data at t=75 is for example classified as belonging to the traffic recovery phase, yielding a high velocity and low density at t=90. It is found to be otherwise as the measurement from the next minute is again *correctly* identified. Whereas the density prediction is not much different from the density prediction without dynamic clustering, in the flow predictions obvious improvement can be observed up until about midway.

For link x=40, the velocity prediction performance with dynamic clustering enabled is even worse. Whereas the true velocity reveals congestion up from approximately t=55, it is at time t=85 that congestion is predicted (even though it is already there!). A combination of lack of predictive links for the traffic state breakdown, as well errors in the measurement classification process are identified as the cause. A selection of links showing very little correlation is initially made, and thereafter the measurements received get identified as belonging to the 2nd cluster of recovery, yielding high velocities as a prediction. It is not until the measurements received get identified as belonging to the first cluster, that the velocity prediction comes closer to the truth. The density prediction reflects this idea and as a result the flow prediction is considered *unsatisfactory,* due to a very low correlation with the truth as well as a relatively high flow MSE.

For link x=73, no big difference can be observed with respect to the variant 3. Performance is likewise worse than without any bagging.

19 Conclusion, discussion and further research

19.1 Conclusion

The objective of this research was to design a performing traffic state estimation and traffic state prediction method and framework, which by utilizing both floating car- and inductive loop detector data, delivered real-time link- velocities, -densities and -flows within an urban traffic environment. The focus of this research was to design this framework built around three pillars. Firstly it must be based on the idea that patterns in historical traffic data can be used to allow the current traffic state of links to serve as indicators for the current and future traffic state on neighbouring links. Secondly it is classifiable as a non-parametric method, assuring no parameters need to be found and calibrated. Lastly the real-time run ability must not be forfeited. The research question of this research was formulated:

> *"How can a NLM framework best be designed as to deliver both a traffic state estimation and state prediction of all relevant traffic flow variables within an urban network, and how well does it perform in delivering an accurate estimation and prediction for different traffic conditions within the urban network?"*

As a starting point for development of the NLM framework, the frameworks used by Morita (2011) and Esaway (2012) were completely modified and expanded. The best performing NLM estimation variant goes through 7 steps. For completeness the framework developed is presented below. The summarized content of each step is:

1) Defining a historical traffic database with the aggregated traffic flow variables;
2) Finding for each link the 4 links which correlate the most w.r.t. the traffic data in the database;
3) Adding the most recently arrived traffic data to the database;
4) Finding the best weighting of the neighbourhood links using linear regression;
5) Applying the weighting for each link to the most recent traffic data of the neighbouring links;
6) Fuse the neighbourhood estimation and own link estimation based on reliability (variance);
7) Formulate and calculate a final estimation of all traffic flow variables;

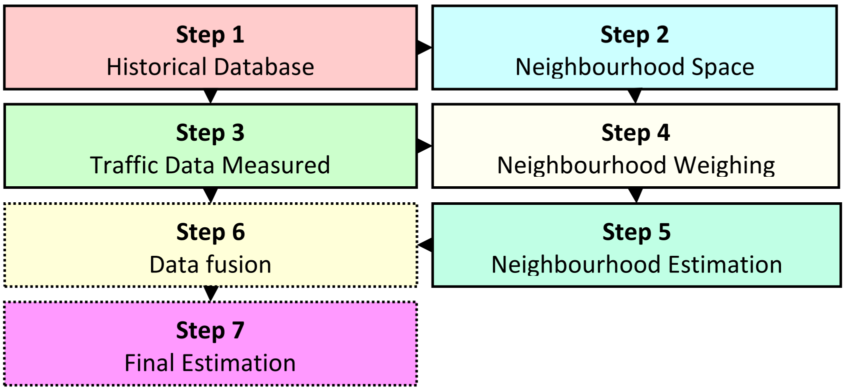

Figure 62: Structural overview of the designed NLM estimation/prediction framework.

The best performing NLM prediction variant uses a slightly modified framework in which steps 6 and 7 are omitted (the neighbourhood prediction is the final prediction). Internally the biggest changes are related to that the correlation, weighting and output of NLM at a time t is a prediction of the traffic flow variables not at the same time t, but at a time t_f in the future. The performance of the best performing estimation and prediction variants of the NLM tested upon the *Enriched Sioux Falls Scenario* from Chakirov (2014) are presented next.

The NLM framework for estimation shows at 5% FCD the average result of a velocity MSE of 52 and 57 for morning and evening rush hour respectively and an average velocity MSE of 75 for the outside rush hour period. This means that on average throughout the morning rush hour, 50% of the links can be estimated within 3 mi/h, 65% within 6 mi/h and 80% within 9 mi/h. Outside rush hour a MSE of 75 yields percentages of 45%, 60% and 75% respectively. The average MSE for density is 247 and 419 in each respective rush hour period and 2 for the outside rush hour period. These values comply 80% of the links showing deviations less than 12 veh/mi/lane, 89% less than 25 veh/mi/lane and 92% less than 38 veh/mi/lane. For the outside rush hour these percentages are 100%, 100% and 100% respectively. The average MSE for flow is 17651 and 21569 for the rush hour periods and 1200 for the outside rush hour period. This means that throughout the morning rush hour, on average, 58% of the links can be estimated within 100 veh/h/lane, 70% within 200 veh/h/lane and 82% within 300 veh/h/lane. Outside rush hour a MSE of 1200 yields percentages of 93%, 100% and 100% respectively. The correlation of estimations and ground truth for all traffic flow variables during the rush hour periods shows values higher than 0,75. Only the correlation values of flow and density during the outside rush hour period show lower values due to the very low sampling rate and therefore already capricious ground truth.

The NLM framework for prediction shows at 5% FCD and a prediction horizon of 5 minutes again promising results. The average velocity MSE increases to 59 and 69 for the rush hour periods and to 79 for the outside rush hour period. This means that on average throughout the morning rush hour, 49% of the links can be predicted within 3 mi/h, 64% within 6 mi/h and 80% within 9 mi/h. Outside rush hour a MSE of 79 yields percentages of 43%, 58% and 72% respectively. The average MSE for density is 424 and 803 in each respective rush hour period and 2 for the outside rush hour period. These values comply 75% of the links showing deviations less than 12 veh/mi/lane, 83% less than 25 veh/mi/lane and 87% less than 38 veh/mi/lane. For the outside rush hour these percentages are 99%, 100% and 100% respectively. The average MSE for flow is 55000 and 73000 for the rush hour periods and 1800 the outside rush hour period. This means that throughout the morning rush hour, on average, 20% of the links can be estimated within 100 veh/h/lane, 41% within 200 veh/h/lane and 49% within 300 veh/h/lane. Outside rush hour a MSE of 1200 yields percentages of 88%, 98% and 100% respectively. The correlation of the predictions with the ground truth underline the relatively weak flow estimations as the r drops from 0,774 and 0,756 to 0,677 and 0,606 respectively in each rush hour period.

The NLM framework for prediction at a prediction horizon of 15 minutes, shows less promising results in terms of both density and flow predictions. The average velocity MSE increases even further to 67 and 75 for the rush hour periods and to 80 for the outside rush hour period. This means that on average throughout the morning rush hour, on average 47% of the links can be predicted within 3 mi/h, 63% within 6 mi/h and 78% within 9 mi/h. Outside rush hour a MSE of 80 yields percentages of 44%, 59% and 72% respectively. The average MSE for density doubles to 831 and 1708 in each respective rush hour period and is 3 for the outside rush hour period. The morning rush hour values comply with 75% of the links showing deviations less than 12 veh/mi/lane, 85% less than 25 veh/mi/lane and 90% less than 38 veh/mi/lane. For the outside rush hour these percentages remain all at 100% respectively. The average MSE for flow is 102.000 and 134.000 for the rush hour periods and 2.000 for the outside rush hour period. This means that throughout the morning rush hour, on average, 10% of the links can be estimated within 100 veh/h/lane, 15% within 200 veh/h/lane and 20% within 300 veh/h/lane. Outside rush hour the MSE of 2000 yields percentages of 85%, 96% and 100% respectively. The correlation of the predictions with the ground truth reflects this serious degradation as well. Whereas the velocity predictions still show high correlation throughout all three simulated periods (>0.81), the density correlation drops to 0,76 and 0,68 within the rush hour periods and the flow correlation drops even below the 0,5 threshold, at 0,48 and 0,37 for each rush hour period respectively. Indicating that both flow and density predictions must be labelled to be (very) weak.

19.2 Discussion

The possible biggest discussion point in this research regards the way the case study was set up. The choice for modelling the *Enriched Sioux Falls Scenario* within PARAMICS meant that some practical modifications were required. Firstly the network size was reduced by 25%, the run length reduced from 24h to separate slots of 2 hours, an OD-table trip reduction due to multiplication of a factor of 0,12 and node junction adjustments, as some nodes are equipped with a fixed time traffic signals, independent of flows on branches. Additionally PARAMICS might not fully capture the real urban network dynamics; due to exclusion of other modalities, every link consisting of 2 or more lanes, trips released by ratio throughout the modelled period and the homogeneity of the vehicle mix. However, by running each simulation with 5 different seeds, actual vehicles are simulated to show (route) choice behaviour and variance in departure times, mimicking real world behaviour. Lastly it is assumed that a 5% penetration rate seems plausible and feasible in near future application. Though the assumption that no sample bias would be present in this urban network, might not hold in real-world conditions.

Consequently the NLM framework can be considered to be optimized throughout this research for this specific scenario. E.g. the exact maximal flow and maximum density values used to cap off unrealistic estimations and predictions, might not be available in a real world scenario. Though argued is that these parameters can be substituted for theoretical estimations of these, keeping the non-parametrized nature of NLM alive.

To be able to derive a density and flow estimation/prediction, the FCD and ILDD were compared within a minute interval to calculate the local floating car penetration rate and thereby enabling to calculate the density estimation from the FCD. The global floating car penetration rate was used for links unequipped with ILDs, which did not yield a more erroneous indication of the density as compared with other links. Though other ways on how to derive a local density from FCD and ILD remained unsearched. Trivially the flow was determined by multiplying velocity and density, yielding all three traffic flow variables for each link in the network. However due to inhomogeneity's and differences in link length of the segments, the representativeness of a traffic variable for a localized segment might be affected.

The performance assessment in this research has been executed by means of expressing the deviations in result between estimation/prediction and the ground truth. The quantity of measurement is the mean square error (MSE) and Pearson's' correlation coefficient (r). While these values allow fair comparison, they do not reveal any details about in-run-differences; e.g. the MSE of density is obviously prone to be higher at actual higher densities in the network. Therefore the average MSE of both density and flow might not be fully representative of the average accuracy achieved. Additionally the presented reliability intervals of an estimation in which an average percentage of links is estimated within a predefined absolute difference of the ground truth is case specific. Whereas sometimes the mean absolute percentage error (MAPE) is adopted (e.g. Van Lint, 2015), this was infeasible in this specific case due to zero flows and densities present.

19.3 Other findings

In this research the most difficult period to be estimated and predicted by the NLM is during the traffic recovery phase of a link. The traffic demand during this recovery period decreases, while on neighbouring links it is likely that congestion is still present. It is for a recovering link, that inclusion of data from link neighbours can influence the result hugely. The reason for this is that the drop in vehicle numbers on the link in question, coincides with a naturally increasing variance of the traffic data received from this link. Therefore when weighting the traffic data from both the neighbourhood space and the link itself, it is the data from the neighbourhood space which most likely fully dominates the result. This implies that during this period it is key that the links which show exact the same behaviour are included in the neighbourhood space. Due to between-day-variabilities in the traffic, it is however not necessarily true that if two links show exact the same minute-to-minute behaviour yesterday, that they will do so today. Subsequently a higher than average MSE is the result. The second most difficult period is found to be the outside rush hour periods, related to the use of FCD in general. In this research, approximately 50 to 150 vehicles were present in the network during the outside rush hour periods. Yielding ground truth intensities of only up to 10 vehicles per link per minute. With a 5% FCD coverage rate, most links in the network therefore were not sampled. As with few vehicles on each link the heterogeneous traffic dynamics of vehicles plays an important role, the ground truth is highly capricious. Assuming that with no samples, the travelled velocity is v_{max} and both density and flow are zero, the ingredients are created for lesser estimation and prediction accuracy.

Interestingly in this research, additional bagging of results aiming to separate the periods of traffic breakdown and traffic recovery prove to be quite difficult. Both statically and dynamically partitioning historical traffic data and assigning real-time traffic data with a nearest neighbourhood algorithm proved to yield no better results than without any bagging. Lastly limiting the database size – used to store the traffic data – to two full runs has shown to yield the best results, besides the additional if-then statements included to capture unrealistic high output values.

Regarding the expansion of the NLM to include density and flow outputs, combining both FCD and ILDD was found to be difficult as well, due to the spatial differences in measurements. Point speeds measured from ILD located at the link's top showed a serious underestimation bias towards the mean speed on a whole link. The same applies for flow and density measurements. Arguably more representative positioning of inductive loops on highway links might therefore improve the performance of NLM hugely, especially as in this research vehicle counts on the highway were low.

The determination of the neighbourhood space is key within the NLM framework. During this research clues were found that using the adopted approach of linear correlation in the raw traffic data might not be the most rewarding. Firstly because using Pearson's correlation coefficient for linear correlation, does not necessarily guarantee finding a neighbourhood space in which a relatively good weighting for the traffic data can be found. Secondly, difficulty arises due the fact that the neighbourhood space needs actual neighbours to function properly. For the estimation part the weighting scheme can compensate for a slight mismatch in the timing of the traffic phases on neighbouring links (e.g. a neighbouring link might be fully congested some minutes earlier or another some minutes later, therefore together with the traffic data from the link itself, still describing more or less the correct situation) because in the used scenario there are ample links that experience congestion and free-flow and the traffic data from the own link is included. For the prediction part being the first link to experience a traffic breakdown or first to experience recovery, no neighbour links or even the link itself, can be used to predict this occurrence and thus create a time lag equal to the prediction horizon for these links. Density and flow predictions are as a result inaccurate.

19.4 Research implications

This research shows that the designed NLM framework can yield very reasonable traffic state estimation results in a modelled and simulated environment. Due to the fact it is simple in essence and algorithmically not very complex, NLM can be easily transferred to a real world scenario. Additionally other traffic data sources can be effortlessly implemented in the process, improving the estimation accuracy even more. For the traffic state prediction the results show a more clouded image, as specifically the flow accuracy (as a result of stacking of errors of both velocity and density predictions) is deemed low. NLM in its' current state might therefore not be sufficiently intelligent to be utilized in the field of ITS.

The advantages of utilizing the state estimation part of NLM in the ITS field of study are however already ample; (1) it can deliver a complete and robust overview of the urban traffic network, showing what's going on anywhere in the city in the form of real-time travel times, velocities, densities, flows for every link in the network; (2) NLM enables the controller to communicate more effectively to the users by allowing for real-time dynamic routing and load balancing via e.g. VMS (Treiber, 2012); (3) it enables the users of the traffic network to make dynamic real-time route choices via e.g. traffic-dependent navigation devices and (4) it provides tactical information on (real-time) fixes possible in the city, e.g. where to improve traffic flows by tweaking traffic lights or whether to impose variable speed limits.

19.5 Further research

This research is concluded by recommending areas of possible further research. It is found that frameworks that are based on FCD inherently struggle within the traffic state phase of recovery and the traffic state of free-flow, due to declining (or low) sampling rates of floating cars. As arguably the network controller is more interested in finding congestion (with any cause), determining the exact free flow velocities experienced in the death of night it is of little interest. Though improvements of the NLM framework that specifically target the traffic state recovery phase are a viable area for further research. Additionally the fusing of FCD and ILDD could play a bigger role here. Whereas in this research only vehicle counts of ILDD were used due to velocity and flow measurements found to be not representative of the velocities and flows on the whole link, further research on how to make these measurements usable for data fusion might prove to be of huge help.

The biggest area for improvement within the NLM framework itself, is found during the neighbour space determination step. By no means is finding correlation a guarantee for finding a proper weighting later on. Referred is to the previously mentioned example in which three links of data are compared: (30, 29, 30, 27, 28) and two possibly correlated links with their respective measurements: (31, 27, 24, 21, 17) and (29, 30, 29, 31, 30) with then a r of 0,71 and -0,96 respectively. Further research on finding a better method for finding neighbouring links can start by for example using the correlation of not the traffic variables itself but the absolute values of fluxes (derivative), which in this example yields better results. Another thought is to identify the neighbourhood space by replacing the correlation step by the unrestricted version of the linear regression problem. The link neighbours can then selected from the links which get weighed above a set threshold. Further research along these lines and the use of spatial correlation or even autocorrelation might lead to better selections of neighbourhood spaces.

In this thesis, the performance of NLM's traffic state estimation and prediction has been assessed, based on a ground truth derived simplified version of an urban network. In this urban network, the dynamic of on-street-parking was captured. However, other traits that typically describe an urban environment were not included in the used traffic model (e.g. user-interaction, mode-interaction, a heterogeneous vehicle mix and dynamic traffic lights). Further work is needed to improve and test the current implementation into more complex and realistic urban traffic networks.

The ground truth used in this research is fictitious. Assumed is that the OD-table is the same for every working day and no outside influences are incorporated. However, real world circumstances can effect driving behaviour or the OD-table significantly. E.g. rain, events, temperature, day of the week all might describe different bags with each a distinct traffic pattern. Additional research on how to improve the current bagging implementation is very much needed to cope with these changing real-world circumstances, let alone that its current implementation, shows no promising results yet.

Additionally within this thesis three distinct periods were used to assess the performance of NLM. Within these regular situations NLM is able to output an accurate traffic state- estimation and (to some extent) also a –prediction. Typically the traffic controller is also interested in how traffic will behave on irregular situations, due to e.g. road-works, accidents, events or after intervening. It is clear that the current version of NLM is unable to yield the appropriate answers for these irregular activities. Incorporation of an actual traffic model, which is able to yield these answers, in the form of e.g. a plug-in for NLM is identified as another area for future research.

On a final note, it is by no means implied that the search for improvements on NLM is exhausted. Especially as both the traffic state estimation framework and the traffic state prediction framework can be easily modified, redeveloped and edited. More data sources can be incorporated, the currently applied techniques can be refined or even replaced by different approaches.

■

20 References

Aron, M. & Danech-Pajouh, M. (1991). *Athena: a method for short-term inter-urban motorway traffic forecasting*. Recherche Transports Securite, pages 11–6, 1991.

Bell, M. C., Bennett, L. D. and Evans, R. G. (1992). *"The 'Instrumented City' Project: towards an Integrated Transport Database"*. Proceeding of the First Meeting of the EURO Working Group on Urban Traffic and Transportation.

Bar-Gera, H. (2016) Transportation Network Test Problems. Url: http://www.bgu.ac.il/~bargera/tntp/.

Cayford, R. and Johnson, T. (2003). "Operational Parameters Affecting the Use of Anonymous Cell Phone Tracking for Generating Traffic Information," Transportation Research Board Annual Meeting, Washington DC, 2003.

Cheu, R, Xie C. and Lee, D. (2002). *Probe Vehicle Population and Sample Size for Arterial Speed Estimation.* Computer-Aided Civil and Infrastructure Engineering 17 (2002) 53–60.

Chakirov, A. & Fourie, P.J. (2014). *Enriched Sioux Falls Scenario with Dynamic And Disaggregate Demand.* Working paper, Future Cities Laboratory, Singapore - ETH Centre (SEC), Singapore.

Daganzo, C. F. (1994): The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory, Transportation Research Part B: Methodological, Vol. 28, No. 4, pp. 269–287, 1994.

Daganzo, C. F. (1995): Requiem for second-order fluid approximations of traffic flow, Transportation Research. Part B: Methodological, Vol. 29, No. 4, pp. 277–286, 1995.

Dai, X., Ferman, M.A., Roesser, R.P. (2003). A simulation evaluation of a real-time traffic information system using probe vehicles. Intelligent Transportation Systems, 2003. Proceedings. 2003 IEEE. DOI: 10.1109/ITSC.2003.1251999

De Vries, L.O. (2015). Network State Estimation. *Evaluation of both a variance-based and learning-database interpolation technique for an urban network, using the Ground Truth derived from a microscopic traffic model.* Master Thesis.

Deng, W., Zhou, X. (2011). Freeway Traffic State Estimation and Uncertainty Quantification based on Heterogeneous Data Sources: Stochastic Three-Detector. Transportation Research Part B 00 (2011)000–000.

Esawey, M. and Sayed, T. (2012). *A framework for neighbour links travel time estimation in an urban network*. Transportation Planning and Technology, 35:3, 281-301, DOI: 10.1080/03081060.2012.671028

Ezell, S. (2010). Intelligent Transportation Systems. The Information Technology & Innovation Foundation, January 2010.

Gajewski, B.J. and Rilett, L.R. (2003). Estimating link travel time correlation: an application of Bayesian smoothing splines. In: Proceedings of the 82nd annual meeting of the transportation research board, January, Washington, DC.

Georgiou, T., Abbadi, A., Yan, X. and Georg, J. (2015). Mining Complaints for Traffic-Jam Estimation: A Social Sensor Application.

Gosh, B. and Smith, P. (2015). *Customisation of Automatic Incident Detection Algorithms for Signalised Urban Arterials*.

Habtemichael, F.G., Cetin, M. (2015). *Short-term traffic flow rate forecasting based on identifying similar traffic patterns*. Transport. Res. Part C (2015). http://dx.doi.org/10.1016/j.trc.2015.08.017

Herrera J. C. (2010). "Evaluation of Traffic Data Obtained via GPS-Enabled Mobile Phones: The Mobile Century Field Experiment," Transportation Research Part C, Vol. 18, No. 4, 2010, pp. 568-583.

INRIX® Traffic (2016). INRIX Traffic. Url: http://inrix.com.

Josefsson, M., Patriksson, M. (2007). Sensitivity analysis of separable traffic equilibria, with application to bilevel optimization in network design. Transportation Research Part B 41, 4–31.

Kalman, R.E., Bucy, R.S. (1961). New results in linear filtering and prediction theory. Transactions of the ASME Series D 83, 95–108.

Laval, J. A. and Leclercq, L. (2013). The Hamilton–Jacobi partial differential equation and the three representations of traffic flow, Transportation Research Part B: Methodological, Vol. 52, pp. 17–30, 2013.

LeBlanc, L.J. (1975). An Algorithm for the Discrete Network Design Problem, *Transportation Science*, **9** (3) 183–99.

Li, J., Xie, L. and Lai, X. (2013). *Route Reconstruction from Floating Car Data with Low SamplingRate Based on Feature Matching*. Research Center of Intelligent Transportation System, Sun Yat-Sen University, China.

Lighthill, M. J. and Whitham, G. B. (1955). On kinematic waves. II. A theory of traffic flow on long crowded roads, Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences, Vol. 229, No. 1178, pp. 317–345, 1955.

Mahmassani, H. S., Chang, G. L. (1987). On Boundedly Rational User Equilibrium in Transportation Systems. Transportation Science, 21(2):89-99.

Morita, T. Yano, J. Kagawa, K. (2009) "*Interpolation System of Traffic Condition by Estimation/Learning Agents*" in Proceedings of 12th International Conference on Practice in Multi-Agent Systems, Nagoya,pp. 487-499.

Morita, T. Yano, J. Kagawa, K. (2010) "*Multiagent Based Interpolation System for Traffic Condition by Estimation/Learning*" Proc. of the 9th International Conference on Autonomous Agents and Multiagent Systems, Toronto, pp. 1697-1704.

Morita, T. (2011). *High Performance Spatial Interpolation System for Traffic Conditions by Floating Car Data*. SEI Technical Review, number 72, april 2011

Morlok, E.K., Schofer, J.L., Pierskalla, W.P., Marsten, R.E., Agarwal, S.K., Stoner, J.W., Edwards, J.L., LeBlanc, L.J. and Spacek, D.T. (1973). Development and Application of a Highway Network

Design Model, Volumes 1 and 2. Final Report: FHWA Contract Number DOT-PH-11. Northwestern University.

Nantes, A., Ngoduyb, D., Bhaskara, A., Miskaa, M. and Chunga, E. (2015). *Real-time traffic state estimation in urban corridors from heterogeneous data*. Transportation Research Part C: Emerging Technologies.

Newell, G. F. (1993): A simplified theory of kinematic waves in highway traffic, part I: General theory, Transportation Research Part B: Methodological, Vol. 27, No. 4, pp. 281–287, 1993.

Park, J. (2011). "Real time vehicle speed prediction using a Neural Network Traffic Model," *Neural Networks (IJCNN), The 2011 International Joint Conference on*, San Jose, CA, 2011, pp. 2991-2996. doi: 10.1109/IJCNN.2011.6033614

Ristic, B., Arulampalam, S. and Gordon, N. (2004). *Beyond the Kalman filter*. Artech House, 2004, ISBN: 1-58053-631.

Sen, A., Thakuriah, P., Zhu, X., and Karr, A. (1997). *Frequency of probe reports and variance of travel time estimates*. Journal of Transportation Engineering, ASCE, 123 (4), 290297.

Snelder, M. and Calvert, S. (2015). *Real-time reistijdvoorspellingen voor routekeuze- en vertrektijdstipadvies - een toepassing in de Praktijkproef Amsterdam*. TNO. Bijdrage aan het Colloquium Vervoersplanologisch Speurwerk 19 en 20 november 2015, Antwerpen.

Srinivasan, K. & Jovanis, P. (1996), Determination of the number of probe vehicles required for reliable travel time measurement in an urban network, Transportation Research Record 1537, TRB, Washington, D.C., 15–22.

Smith, B. L., Williams, B. M. and Oswald, R. K. (2002). Comparison of Parametric and Nonparametric Models for Traffic Flow Forecasting. Transportation Research Part C, Vol. 10, 2002, pp. 303–321.

Tao, S., Manolopoulos, V., Rodriguez, S., Rusu, A. (2012). *Real-Time Urban Traffic State Estimation with A-GPS Mobile Phones as Probes*. Journal of Transportation Technologies, 2012, 2, 22-31

Tampère, M.J., Corthout, R., Cattrysse, D., Immers, L. (2011). *A generic class of first order node models for dynamic macroscopic simulation of traffic flows*. Transportation Research Part B 45 (2011) 289–309.

Treiber, M., Kesting, A. (2012). Traffic Flow Dynamics. Data, Models and Simulation. ISBN 978-3-642-32459-8.

Van Hinsbergen, C., Van Van Lint, J. and Sanders, F. (2007). Short-Term Traffic Prediction Models. Proc., 14th World Congress on Intelligent Transport Systems: ITS for a Better Life, Beijing: Research Institute of Highway, Chinese Ministry of Communications, 2007.

Van Lint, J., Hoogendoorn, S. and Hegyi, A. (2008). *Dual EKF State and Parameter Estimation in Multi-Class First-Order Traffic Flow Models*. 17th IFAC World Congress, Vol. 17, 2008.

Van Lint, J. and Van Hinsbergen, C. (2011). *Short-Term Traffic and Travel Time Prediction Models*. Transportation Research Number E-C168 November 2012.

Van Lint, J. (2015). *Traffic state estimation basics*. Traffic simulation & Computing, Civil Engineering & Geosciences, TU Delft. Mathematical Approaches for Traffic Flow Management. Lectures 2015.

Yakowitz, S. (1987). Nearest-Neighbour Methods for Time Series Analysis. Journal of Time Series Analysis 8, 10-26. http://dx.doi.org/10.1111/j.1467-9892.1987.tb00435.x

Yperman, I. (2007). *The Link Transmission Model for Dynamic Network Loading*. Doctorate Thesis.

Wan, A. and Merwe, R. van der (2000). *The unscented Kalman filter for nonlinear estimation*. IEEE Symposium on Adaptive Systems for Signal Processing, Communication and Control, 2000, pp. 153–158.

Wang, Y., Papageorgiou, M. (2005). *Real-time freeway traffic state estimation based on extended Kalman filter: a general approach*. Transportation Research. Part B 39, 141–167.

Wang, Y., Papageorgiou, M., Messmer, A. (2008). *Real-time freeway traffic state estimation based on extended Kalman filter: Adaptive capabilities and real data testing*. Transportation Research Part A 42 (2008) 1340–1358

Wang, Y., Papageorgiou, Tzimitsi, A., Coppola, P., M., Messmer, A. (2011). *Real-Time Freeway Network Traffic Surveillance: Large-Scale Field-Testing Results in Southern Italy*. DOI: 10.1109/TITS.2011.2107901

Watson, S.M., Clark, S.D., Redfern, E. and Tight, M.R. (1992). *Setar Modelling of Traffic Count Data*. Working Paper. Institute of Transport Studies, University of Leeds, Leeds, UK.