Analysis of urban traffic patterns using clustering

Wendy Weijermars

Dissertation committee:

prof. dr. F. Eising	Universiteit Twente, chairman/secretary
prof. dr. ir. E. C. van Berkum	Universiteit Twente, promotor
prof. dr. ir. B. van Arem	Universiteit Twente
prof. dr. ir. M. F. A. M. van Maarseveen	Universiteit Twente
prof. dr. M. C. Bell	University of Leeds
dr. T. Brijs	Universiteit Hasselt
prof. dr. E. Chung	University of Tokyo/EPFL

TRAIL Thesis Series T2007/3, The Netherlands TRAIL Research School

This thesis is the result of a Ph.D. study carried out between 2002 and 2006 at the University of Twente, faculty of Engineering Technology, department of Civil Engineering, Centre for Transport Studies.

TRAIL Research School
P.O. Box 5017
2600 GA Delft, The Netherlands
Telephone: +31 15 2786046
Telefax: + 31 15 2784333
E-mail: info@rsTRAIL.nl

The research is part of the Dutch TRANSUMO (TRansition SUstainable MObility) program.

Cover picture: detail of an artwork designed by Olaf Mooij on a roundabout in Enschede. The picture is taken by Wendy Weijermars

Typeset in LaTeX

Printed by Gildeprint BV, Enschede, The Netherlands.

ISBN 978-90-365-2465-0

Analysis of urban traffic patterns using clustering

PROEFSCHRIFT

ter verkrijging van
de graad van doctor aan de Universiteit Twente,
op gezag van de rector magnificus,
prof. dr. W.H.M. Zijm,
volgens besluit van het College voor Promoties
in het openbaar te verdedigen
op vrijdag 13 april 2007 om 15.00 uur

door

Wilhelmina Adriana Maria Weijermars

geboren op 25 december 1977
te Rijnsburg

Dit proefschrift is goedgekeurd door de promotor:

prof. dr. ir. E. C. van Berkum

Voorwoord

Op sommige momenten en locaties stroomt het verkeer probleemloos door, terwijl het op andere momenten en locaties stagneert. Zo is het ook met onderzoek. Zoals waarschijnlijk iedere promovendus heb ik momenten gehad waarop ik dacht dat ik het nooit af zou ronden. Nu is dan toch het moment gekomen dat ik met het schrijven van dit voorwoord de laatste hand leg aan mijn proefschrift. Dit was niet gelukt zonder de hulp van een aantal mensen en organisaties, die ik hier dan ook voor wil bedanken.

Data is onmisbaar geweest voor dit onderzoek. De volgende bedrijven en instanties wil ik dan ook van harte bedanken voor het leveren van de benodigde data: Vialis, de gemeente Almelo, de regiopolitie Twente en het KNMI. Daarnaast wil ik met name Steven Boerma en Jeroen Konermann van Vialis en Rob Hulleman en Wim Stulen van de gemeente Almelo bedanken voor de goede en prettige samenwerking. Jullie kennis met betrekking tot de data en de lokale situatie heeft me erg geholpen!

Ik wil Martin van Maarseveen bedanken voor de kans die hij mij geboden heeft om na mijn afstuderen bij de vakgroep te komen werken als medewerker onderzoek. In deze tijd hebben Eric van Berkum en ik het promotievoorstel kunnen schrijven dat uiteindelijk heeft geleid tot dit proefschrift. Mijn promotor Eric van Berkum wil ik bedanken voor zijn begeleiding. Eric, met name je wiskundige kennis en je ideeën over verkeersmanagement waren zeer nuttig. Ik ben heel blij dat Tom Thomas uiteindelijk als postdoc binnen Transumo begonnen is. Tom, in het begin was het niet helemaal duidelijk wat jouw taak was en wat we van elkaar moesten verwachten, maar uiteindelijk ben ik heel tevreden over onze samenwerking. De brainstormsessies en je hulp met de analyses waren zeer welkom! Ook alle afstudeerders die ik begeleid heb wil ik bedanken voor de discussies en nieuwe inzichten. Met name Anton en Marcel bedankt, jullie werk is erg nuttig geweest voor mijn promotieonderzoek.

Daarnaast wil ik een aantal andere mensen binnen de UT bedanken voor hun hulp, die opvallend vaak met computerzaken van doen heeft. Bas, bedankt dat je me op gang geholpen hebt met SPSS en Manifold. Kasper, bedankt voor je hulp met Delphi - het is dan uiteindelijk SPSS geworden, maar onze Delphi-exercities hebben me wel nuttige inzichten verschaft- en Omnitrans.

Mijn LaTeX-koningen Mark, Jebbe, Thijs en Andries, bedankt! Zonder jullie was het Word geworden. Axel en Martijn, bedankt dat ik altijd bij jullie binnen kon lopen als mijn computer weer eens kuren vertoonde. Tot slot, Dorette en Maureen bedankt voor de secretariële ondersteuning en de gezelligheid op het secretariaat.

Ik heb een enorm gezellige tijd gehad op de UT! Hiervoor wil ik alle (oud-)collega's van harte bedanken. In het bijzonder wil ik Andries, Anne-Marie, Attila, Bas, Blanca & Freek, Cornelie & Steffen, Daniëlle, Frans & Mascha, Jan-Willem, Jebbe, Judith & Niels, Gio, Kasper, Mako, Mark, Martijn, Pieter van Oel & Bertien, Pieter Roos & Judith, Thijs en Tom bedanken voor de fijne tijd op het werk, op de squashbaan, bij de pubquiz, in de kroeg, in de sportschool en/of op andere locaties.

Frans, na mijn afstudeerbegeleider werd je mijn kamergenoot. Ik had me geen betere kamergenoot kunnen wensen! Bedankt voor de goede gesprekken, de koffie en de lol die we samen hadden en hebben, op en buiten kantoor.

Cornelie en Blanca, ik ben heel blij dat jullie mijn paranimfen willen zijn. Cornelie, we zijn vlak na elkaar begonnen als medewerker onderzoek en zijn vervolgens beiden aan een promotieonderzoek begonnen. Ik ben blij dat je al deze jaren mijn collega geweest bent. Bedankt voor het doorlezen van stukken en de adviezen, je gezelligheid en de memorabele stap-avonden (nee, de toegift is al geweest). Blanca, je kwam twee jaar geleden als tijdelijke medewerker bij de afdeling Water, maar je bent er gelukkig nog steeds. Bedankt voor de Spaanse les en de meidenavonden, je interesse en je spontaniteit.

Van buiten de universiteit wil ik Construktie (Arthur, Bram, Doutsen, Harm, Koen, Marko en Robin) en Cécile bedanken voor hun vriendschap. Papa, mama en Leonie, bedankt voor jullie onvoorwaardelijke liefde en steun en het vertrouwen dat jullie altijd in mij gehad hebben.

Last but certainly not least wil ik Jebbe bedanken. Jebbe, een hele mooie bijkomstigheid van mijn promotieonderzoek is dat ik jou heb leren kennen. Ik had je al bedankt voor je hulp met LaTeX en de fijne tijd op en buiten de UT, maar daarnaast en vooral wil ik je bedanken voor je liefde, je steun, je vertrouwen en je eigenheid. Ik hou van je!

Wendy Weijermars

Enschede, 5 maart 2007

Contents

Voorwoord **v**

1 Introduction **1**
1.1 Background 2
1.2 Research objectives and scope 6
1.3 Scientific and practical relevance 7
1.4 Thesis outline 9

2 Urban traffic data **11**
2.1 Urban traffic information centres 11
2.2 Data 14
2.2.1 Traffic data 14
2.2.2 Data on factors potentially influencing traffic 16
2.3 Data processing 17
2.4 Data interpretation 22
2.5 Summary 23

3 Variations in urban traffic **25**
3.1 Temporal variations 25
3.1.1 Short term variations 26
3.1.2 Variations within a day 26
3.1.3 Variations between days 27
3.1.4 Long term variations 30
3.2 Spatial variations 31
3.2.1 Spatial variations in traffic volumes 31
3.2.2 Spatial-temporal traffic patterns 32
3.2.3 Variations in urban traffic patterns 32
3.2.4 Differences in temporal traffic patterns between locations 33
3.3 Variations in travel behaviour 35
3.4 Discussion 39

4 Analysis of urban traffic patterns **41**
4.1 Design of clustering procedure 42
4.1.1 Pattern representation 42

4.1.2 Clustering procedure 44
4.2 Analysis of temporal traffic patterns 46
4.2.1 Description of resultant clusters 46
4.2.2 Determination of factors on the basis of the clusters . . 48
4.2.3 Variation within the clusters 50
4.3 Analysis of spatial traffic patterns 51
4.3.1 Variations in average daily flow profiles 52
4.3.2 Variations in weekly patterns 53
4.3.3 Variations in seasonal patterns 55
4.3.4 Variations in weather factors 56
4.3.5 Variations in temporal classifications 57
4.4 Traffic patterns on a network level 59
4.5 Summary 61

5 Almelo: Data 63
5.1 ViaContent 63
5.2 Data 64
5.2.1 Traffic data 64
5.2.2 Data on factors potentially influencing traffic 66
5.3 Data processing 67
5.3.1 Pre-processing of individual data records 68
5.3.2 Data validation 69
5.3.3 Evaluation of data control procedure 74
5.4 Available traffic data after processing 77
5.4.1 Data quality 77
5.4.2 Available data 78
5.5 Conclusions 81

6 Almelo: Network and traffic demand 83
6.1 Network structure and major attractions 83
6.2 Main traffic streams 84
6.3 Daily flow profiles 89
6.4 Summary 91

7 Almelo: Traffic patterns 93
7.1 Temporal traffic patterns 93
7.1.1 Definition of a daily flow profile 94
7.1.2 Working day patterns 95
7.1.3 Non-working day patterns 101
7.2 Spatial traffic patterns 103
7.2.1 Variations in average daily flow profiles 103
7.2.2 Variations in weekly patterns 106
7.2.3 Variations in seasonal patterns 109
7.2.4 Variations in weather factors 112
7.2.5 Variations in temporal classifications 114
7.3 Traffic patterns on a network level 116

7.4 Summary 119

8 Applications 123
8.1 Traffic monitoring 124
8.1.1 Direct use of results of cluster analysis 124
8.1.2 Application of results to other cities 127
8.2 Traffic forecasting 127
8.2.1 Method 130
8.2.2 Application and assessment 131
8.2.3 Hybrid model 133
8.3 Traffic management scenarios 134
8.4 Transport modelling 137
8.5 Conclusions 139

9 Evaluation and Discussion 141
9.1 Functioning of the analysis framework 141
9.2 Discussion 142
9.2.1 Clustering algorithm 142
9.2.2 Optimal number of clusters 145
9.2.3 Available traffic data 146
9.2.4 Influence of weather on traffic 148
9.2.5 Congestion 149
9.2.6 Application of results to other cities 149
9.3 Conclusion 152

10 Conclusions and Recommendations 153
10.1 Conclusions 153
10.1.1 Use of data from traffic information centres 153
10.1.2 Analysis of variations in urban traffic volumes 154
10.1.3 Insight into urban traffic patterns 155
10.1.4 Applications 156
10.2 Recommendations for practitioners 158
10.3 Further research 158

Bibliography 160

Notation 172

A Quality control procedure 177

B Daily flow profiles at main arterials 181

C Resulting clusters on network level 187

Summary 189

Samenvatting 193

About the author **197**

TRAIL Thesis Series **199**

Chapter 1

Introduction

As a result of demographic, economic, land use and international developments, mobility is still increasing (MinVenW, 2004). According to the Dutch Ministry of Transport, Public Works and Water Management, mobility is a necessary condition for economic growth and social development (MinVenW, 2004). However, this increase in mobility also has negative side-effects, such as congestion and air pollution. In order to facilitate mobility whilst minimizing its negative side effects, various measures can be employed, for instance the construction of new infrastructure, traffic management measures (e.g. ramp metering, route guidance), land use policy (e.g. compact city) and measures that try to influence travel behaviour (e.g. road pricing). To be able to take adequate measures it is important to have insight into the functioning of the traffic system.

Taylor et al. (1996) describe the traffic analysis process that can be carried out to obtain more insight into the functioning of the traffic system and the underlying phenomena. The process consists of examination of traffic data and models can assist in this process. Issues related to the analysis of traffic systems include accessibility, environment and safety (Taylor et al., 1996). Besides, nowadays, the reliability of travel times is also an important issue (MinVenW, 2004).

In common practice, the traffic analysis process deals with the traffic situation on an average day (Annual Average Daily Traffic: AADT) or an average working day (Annual Average Weekday Traffic: AAWT) and with the Design Hour Volume (DHV). AADT and AAWT are used primarily for network and maintenance planning, and evaluation, whereas DHV – which is mostly described as the n^{th} highest hourly volume – is used for design work (Taylor et al., 1996). However, besides the average traffic situation, also the variability is of crucial importance. Information on the state of the traffic system at different locations and on different moments in time provides insight into the time and locations of bottlenecks and the reliability of travel times. Moreover, it provides insight into the spare capacity on different times and location and can thus be used to analyse the robustness of the road network. Finally, the information can be used to decide when and were dynamic traffic management should be applied.

Until recently, traffic analysis mainly focused on the highway system. This is partly due to the absence of traffic data for the urban network. However, also in urban areas, mobility is still increasing and in cities it is often not possible to extend capacity. Therefore, it is important to take adequate (traffic management) measures to facilitate mobility whilst minimizing its negative side effects and thus to obtain more insight into the functioning of the urban traffic system. Recently, more traffic data is becoming available as a result of the development of urban traffic information centres. This data can be applied to obtain more insight into urban traffic system performance.

The work described in this thesis aims at improving the insight into urban traffic system performance by analysing variations in urban traffic. This chapter discusses the background (Section 1.1) the research objectives and scope (Section 1.2) and its practical and scientific relevance (Section 1.3). The chapter concludes with an outline of the remainder of the thesis.

1.1 Background

The state of the traffic system is influenced by travel demand and traffic supply characteristics. Travel demand is defined by Roess et al. (1998) as the number of vehicles or people that desire to travel past a point during a specified period. The main traffic supply characteristic that influences traffic system performance is capacity. Capacity is defined as the maximum number of vehicles or persons that can reasonably be expected to be served in the given time period (Roess et al., 1998). Also traffic management measures influence traffic system performance. Traffic management in some cases enables a more effective use of the available capacity (direct influence). Moreover, in some cases capacity is increased or decreased or certain trips are stimulated or discouraged, for example by means of road pricing (indirect influence).

Both travel demand and the capacity of a road vary in time and in space and are influenced by external factors. Traffic is a derived demand, caused by the need or desire to employ activities at certain locations (e.g. living, working, shopping, recreation) (Ortuzar and Willumsen, 1994). Most variations in travel demand are due to the distribution of activities over time and space. Additionally, travel demand may vary as a result of changes in modal split, route choice or departure time due to external factors, past experiences or provided information (see fore example Mahmassani, 1997). The capacity of a road obviously depends on the road design and regulations (e.g. maximum speed). Regarding temporal variations, on the urban network, the instantaneous capacity is highly influenced by traffic light cycles: the capacity is zero in case of a red light. Also the weather, road works, accidents and incidents may cause the capacity to vary in time.

The discussed factors and the interaction between them cause the state of the traffic system to vary in time and in space. Regarding temporal variations, different time scales can be distinguished, varying from minute-to-minute to year-to-year variations. The driving forces behind the variations differ by time scale. Short term variations in urban traffic are mainly due to traffic light cycles. Hour-to-hour and day-to-day variations are mainly caused by variations in travel demand, although also variations in capacity (for example due to weather or road works) may play a role. Long term variations in traffic are mainly due to long term demographic, economic and infrastructural developments.

From a traffic analysis point of view, it is interesting to analyse the within and between day variations in the state of the traffic system at different locations and on a network level.

Urban traffic characteristics

Both the travel demand and supply characteristics of urban areas clearly differ from those of highways. Therefore, insight into highway traffic cannot be directly translated to the urban situation. This section briefly discusses the main differences between urban traffic and highway traffic. For an extensive description of urban travel and transportation system characteristics the reader is referred to Meyer and Miller (2001).

A clearly apparent difference between urban traffic and highway traffic is that on the urban road network, multiple traffic modes coexist and interact - for instance pedestrians, bicycles, cars, buses, trucks - whereas highways are mainly used by cars and trucks. This mixture of modes also causes relatively large differences in speed between urban road users. Another characteristic of the urban network is that it contains many intersections. As a result, the traffic situation in urban areas is characterized by many small disturbances, in comparison to highways that in general show less disturbances yet with a higher impact.

Regarding travel demand characteristics, traffic on the urban network is generally more diverse than traffic on highways. First of all, depending on the type of highway, a highway mainly serves medium or long distance traffic. The urban network also serves medium and long distance traffic to and from the highways, yet also a considerable amount of local or short distance traffic. Also the distribution over travel motives is more diverse for urban traffic. Most highways are used for one main travel motive. In general, during working day peak periods, the main travel motives are work and business. Moreover, some highways show peaks on weekend days and during holiday periods caused by leisure traffic, for example to and from the beach. Also most urban roads serve

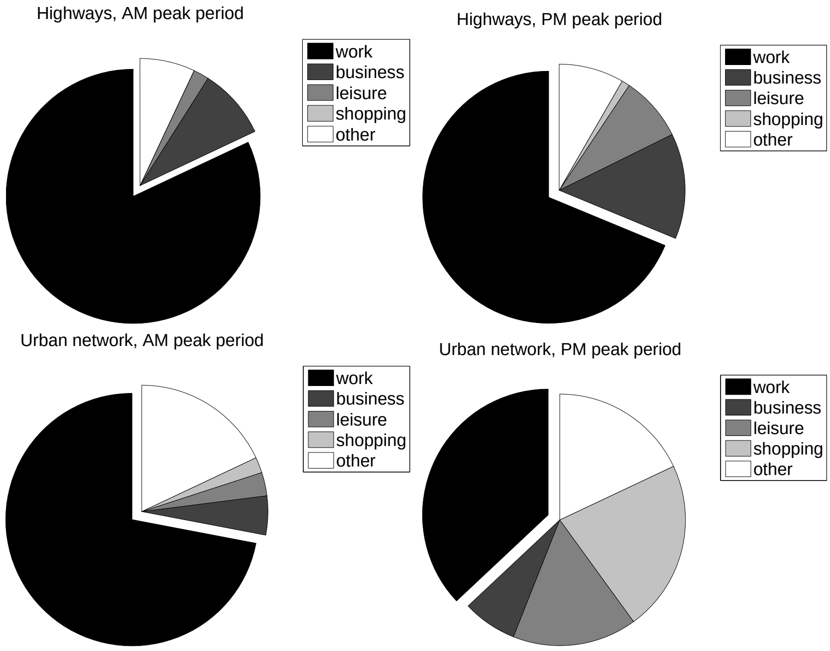


Figure 1.1: Distribution over travel motives. For highways, the distribution over motives is adapted from BGC (1997) which determined the distribution over travel motives on two highways by means of roadside interviews. For urban areas, the distribution over motives is determined on the basis of the National Dutch Travel Survey (OVG) of 1995. All trips departing from medium and large sized cities between 7:00 and 9:00 and between 16:00 and 18:00 were selected.

a considerable amount of work and business related traffic on working days. However, besides commuter traffic also shopping and leisure traffic extensively use the urban network on working days. Figure 1.1 shows that mainly during the P.M. peak period, the share of leisure and shopping traffic is relatively large for the urban network.

Analysis of variations in urban traffic

As mentioned, from a traffic analysis point of view, it is interesting to analyse the within and between day variations in the state of the traffic system at different locations and on a network level. The state of the traffic system is an abstract concept that cannot be measured directly, but that can be described by a number of indicators. Examples of such indicators for the urban traffic system are traffic volume, speed, queue length, delay and travel time. In this thesis, traffic volumes are used to describe the state of the traffic system.

Within and between day variations in urban traffic volumes can be analysed in several ways, using various data sources and different research approaches. First, we briefly discuss the main data sources: travel diary data and traffic volume data. Second, two main research approaches are distinguished: (1) examination of the influence of pre-defined factors on urban traffic and (2) determination and analysis of typical urban traffic patterns using cluster analysis.

Analysis of travel diary data provides insight into (urban) travel demand. Since within and between day variations in urban traffic volumes are mainly due to variations in travel demand, travel diary data can be used for the analysis of these variations in traffic volumes. The advantage of this data source is that it provides insight into underlying travel demand patterns and characteristics of the traffic. Travel diary data is mostly obtained by household travel surveys (Kager, 2005). The main disadvantage of travel surveys is that they are expensive to perform (Kager, 2005). The main surveys in The Netherlands are the 'Onderzoek VerplaatsingsGedrag' (OVG), the 'MobiliteitsOnderzoek Nederland' (MON) and the 'TijdsBestedingsOnderzoek' (TBO). These surveys can be used for the analysis of general travel demand patterns. However, the sample size is too low and the aggregation level of the origin and destination zones is too high to analyse travel demand patterns on a local (city) scale. Besides, travel diary data does not provide information on route choice. Finally, only variations in travel demand are analysed, whereas also variations in supply characteristics (capacity) may cause within and between day variations in the state of the traffic system. In conclusion: travel diary data provides an estimation of variations in traffic volumes in a network, but is not appropriate for the determination of actual variations in traffic volumes.

The advantage of traffic volume data is that it allows analysis of traffic volume patterns on a local level and better represents the actual traffic situation that results from the interplay between travel demand and supply characteristics. A disadvantage of the use of traffic volume data is that it does not provide insight into the underlying travel demand and supply patterns. By combining data from multiple measurement locations some information can be obtained on origins and destinations of the traffic, but the exact distribution over origin and destination zones as well as travel motives remain unknown.

Therefore, in this thesis traffic volume data is exploited to analyse variations in traffic patterns on a local level. Information on general travel demand patterns – obtained by travel surveys – is applied to account for the found variations in traffic volumes.

Most current research on variations in traffic volumes deals with the influence of pre-defined factors like day of the week, holiday periods, season and weather. Some researches (e.g. Keay and Simmonds, 2005) apply regression analysis to determine the influence of different factors. Other researches (e.g. Rakha and Van Aerde, 1995; Stathopoulos and Karlaftis, 2001b) group days on the basis of pre-defined factors (e.g. weekdays) and apply ANOVA-analysis to examine differences between these pre-defined types of days. Alternatively, flow profiles can be grouped on the basis of the data itself. Subsequently, it can be investigated what the characteristics are of the resultant groups. The advantage of this alternative approach is that it groups the data without any (potentially wrong) assumptions. The days within a group show more similar patterns and, possibly new insight can be obtained into factors that influence travel demand or supply.

This thesis adopts this alternative approach: days are grouped on the basis of unsupervised classification or cluster analysis and subsequently it is investigated what factors are responsible for the resulting clusters.

1.2 Research objectives and scope

Research objectives

The main goal of this research is to obtain more insight into urban traffic by analysing within and between day variations in traffic volumes.

The first objective is to design a method for the analysis of temporal and spatial variations in urban traffic volumes using data from urban traffic information centres. First, the data delivered by traffic information centres should be processed to make it appropriate for research. The most important processing task for this research is data validation. In Chapter 2, a data validation procedure is developed. The processed data can subsequently be used for the analysis of variations in traffic volumes. Chapter 3 discusses existing literature on this topic. We propose an alternative approach in Chapter 4. Cluster analysis is applied for the determination of typical urban traffic patterns that can serve as a basis for traffic forecasting, traffic management or traffic modelling scenarios. Besides, basic statistical techniques are adopted to investigate what factors are responsible for these typical traffic patterns. In that way, more insight is obtained into temporal and spatial variations in urban traffic.

The second objective is to apply this method to Almelo, a medium sized city in The Netherlands. It is investigated whether the method is applicable and produces useful and plausible patterns (Chapters 8 and 9).

The third objective is to analyse the patterns for Almelo, resulting in insight into urban traffic patterns (Chapter 7). It is investigated what typical urban traffic patterns can be distinguished, what temporal, circumstantial and spatial factors are on the basis of these patterns and how these patterns can be explained for by variations in travel demand and/or supply.

Scope and limitations

The scope of the research described in this thesis is limited as follows. First, it focuses on variations in the amount of motorized traffic. Variations in public transport use and the number of bicycle trips are not part of this research, although variations in these factors may (partly) explain for variations in the amount of motorized traffic. In this thesis, the term motorized traffic refers to all traffic that uses the main road and is observed by the available detectors, i.e. cars, trucks, buses, motorbikes, mopeds. Furthermore, no distinction is made between different types of motorized traffic.

Second, the research focuses on the urban environment. As explained in the previous section, urban traffic differs from highway traffic in a number of ways as a result of which urban traffic patterns cannot directly be applied to highways, although the proposed analysis framework can be applied to highways as well.

Third, only variations in traffic volumes are analysed. Since travel time data is in general not available for the urban network, the reliability of travel times is not investigated in this thesis. The obtained insight into variations in traffic volumes could however be applied for the analysis of travel time reliability. Also, no information is provided on the time and locations of bottlenecks, yet by linking the traffic volumes to capacity values and data on traffic light cycles, insight can be obtained into traffic system performance (queue lengths, delay etc.).

Finally, the research described in this thesis focuses on within and between day variations in traffic volumes. Short term variations due to traffic light cycles and short term disturbances like the offloading of a truck or a bus stop are not analysed. Moreover, since only one year of traffic data is available, long term variations due to changing land use patterns or infrastructural changes are not taken into account.

1.3 Scientific and practical relevance

Summary of contributions

Next to the insight into urban traffic patterns, three concrete products result from this research:

1. Data control algorithm for the detection of invalid urban traffic data. The data control algorithm that is developed in this research applies a combination of basic maximum and minimum flow thresholds and the principle of conservation of vehicles. Both types of checks were previously applied to highway data, yet in this research they are adjusted for the urban traffic network.

2. Framework for the analysis of urban traffic patterns. The analysis framework designed in this research groups days and locations on the basis of their traffic profiles using Ward's hierarchical clustering. Additionally, it is described which basic statistical methods can be used to determine which factors are responsible for these typical traffic patterns. Finally, it is described how the quality of the resulting classification can be determined.

3. A hybrid model that defines typical traffic patterns that can be used for traffic forecasting, traffic management and transport modelling. The method combines classification on the basis of unsupervised clustering with classification on the basis of weekday and holiday periods and results in better traffic forecasts than forecasts that use historical weekday and holiday period averages and forecasts based on cluster means.

Scientific relevance

The research results in more insight into temporal and spatial variations in urban traffic. The main findings of this research have the following consequences for traffic monitoring, traffic forecasting and traffic modelling and for the further development of urban traffic management.

1. Since daily flow profiles are found to differ between days, average working day volumes do not adequately represent actual traffic volumes on different types of days. With regard to traffic monitoring, it is advisable to remove atypical days (e.g. road works, events) and to make a distinction between different weekdays.

2. Some of the clusters that result from the cluster analyses are caused by location specific factors like road works or footbal matches (events). Cluster analysis is found to be an easy and effective method to automatically detect changes in traffic volumes due to atypical circumstances.

3. Other clusters can be explained by general activity patterns. These findings have implications for the estimation of traffic volumes on roads without detection and can be used for the further development of urban traffic management.

4. As expected, daily flow profiles as well as temporal variations in daily flow profiles vary by location. Differences in distribution of the traffic over travel motives and trip length distribution appear to be on the basis of these spatial traffic patterns. Also these spatial traffic patterns are useful for traffic monitoring and the further development of urban traffic management in general.

Practical relevance

All three contributions are directly of practical relevance. Moreover, the obtained insight can be used in practise to improve urban traffic system performance.

1. The data control algorithm can be applied in other cities as well, although the checks based on the principle of conservation of vehicles have to be adapted to the local detector configuration. The control procedure can also be used for the validation of data that is used for other applications than analysis of traffic patterns.
2. The proposed analysis framework can be used for traffic monitoring. The classification of traffic patterns enables a better estimation of the actual traffic volumes on a certain type of day and a certain location. As a result, more insight is obtained into (potential) times and locations of bottlenecks. Moreover, cluster analysis proved to be an easy and effective way to detect changes in traffic patterns due to road works and other special circumstances. Thus, cluster analysis can also be used for monitoring the influence of road works and events on a network level.
3. Finally, also the hybrid model can be generally applied. This hybrid model can subsequently be used for traffic forecasting, traffic management and transport modelling, in order to provide better traffic information and optimize traffic.

1.4 Thesis outline

Figure 1.2 presents an overview of the structure of this thesis. Chapter 2 deals with urban traffic information centres and the data that is collected at these centres. In addition, we propose a procedure for the validation of urban traffic data to make the data appropriate for the analysis of urban traffic patterns.

Chapter 3 provides an overview of literature on temporal and spatial variations in urban traffic volumes and on underlying variations in travel behaviour. In Chapter 4 we subsequently propose an alternative approach for the analysis of variations in urban traffic. It is described in what way cluster analysis can be used for the determination and examination of urban traffic patterns.

Chapters 5, 6 and 7 deal with the application of the proposed method to the city of Almelo. Chapter 5 describes the available (traffic) data and the processing of this data. Besides, the data validation procedure proposed in Chapter 2 is evaluated. Chapter 6 gives a brief description of the traffic network, the main productions and attractions and the resulting main traffic streams. In Chapter 7, the methods that are proposed in Chapter 4 are applied to Almelo and it is studied what typical traffic patterns can be distinguished and what factors are on the basis of these patterns.

Chapter 8 discusses the potential applications of the obtained insight and resulting clusters, i.e. traffic monitoring, traffic forecasting, traffic management and transport modelling. In Chapter 9, the method is evaluated and discussed. Finally, the main conclusions and recommendations are presented in Chapter 10.

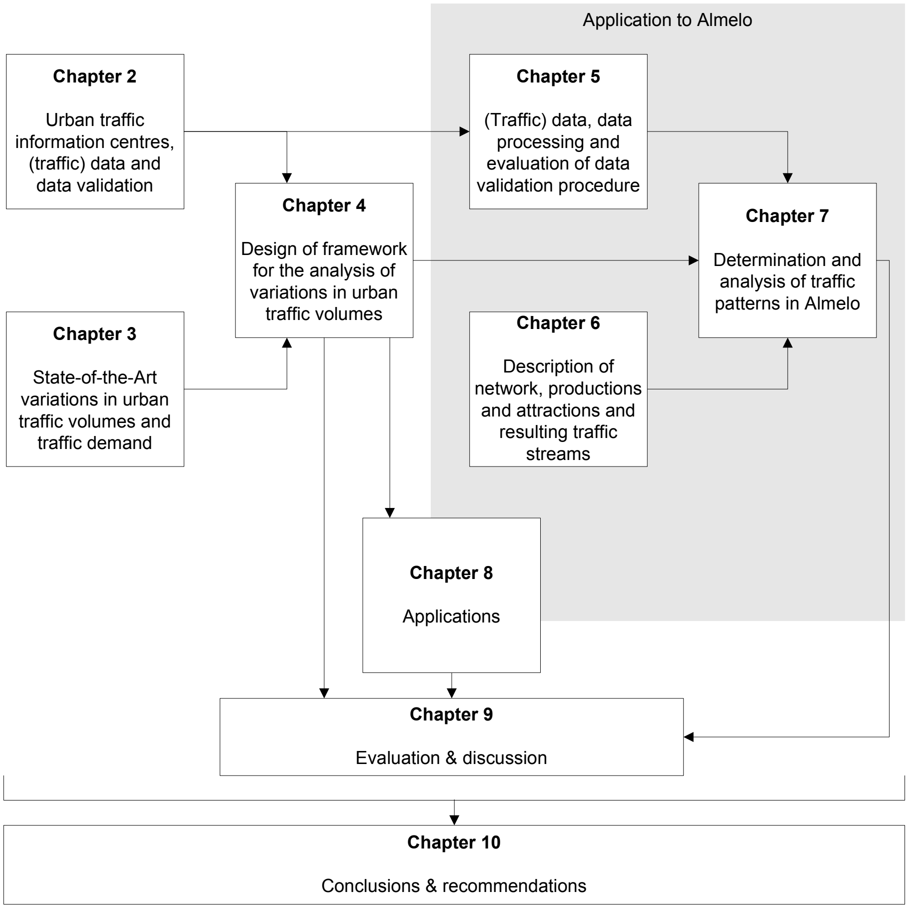


Figure 1.2: Schematic overview of the structure of the thesis.

Chapter 2

Urban traffic data

Highway traffic data has been available for many decades. On the urban network, traffic data is collected by inductive loop detectors at signalized intersections. However, until recently, this traffic data was used only locally for signal control, i.e. the data was not sent to a central database for processing and storage. Nowadays, in several cities initiatives have been taken for the development of urban traffic information centres. In these centres, traffic data is stored and processed in order to provide traffic information. The collected traffic data can also be used for other purposes, one of which is research. This chapter deals with traffic information centres and the use of the data collected at these centres. Section 2.1 provides an overview of the functioning of traffic information centres. The second section discusses the main data sources that can be used for the analysis of urban traffic patterns, the third section discusses what data processing is necessary to be able to use this data for research and the fourth section deals with the interpretation of the data. The chapter ends with a summary.

2.1 Urban traffic information centres

About 15 years ago, the Instrumented city project (Bell et al., 1993; Bell and Gillam, 1994; Bell et al., 1996) was one of the first initiatives for the central collection of urban traffic data. This project dealt with the construction of a database of road traffic data and other relevant data for research, traffic management and traffic information purposes. The last decade, similar facilities have been developed in a number of cities throughout the world, although the focus of these urban traffic information centres is on the provision of traffic information to travellers. Besides, various European projects – for example SCOPE, CAPITALS (PLUS), ENTERPRICE and QUARTET PLUS (see http://cordis.europa.eu/) – deal with the development of urban traffic

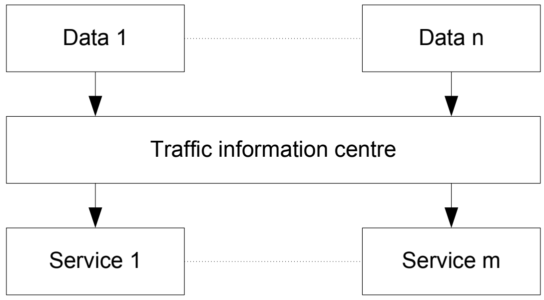


Figure 2.1: Basic principle of a traffic information centre.

information centres. The basic principle of a traffic information centre is shown in Figure 2.1. Collected (traffic) data is sent to the traffic information centre where it is stored and processed in order to be useable for different services. This section discusses the components in Figure 2.1.

Traffic data that are collected include traffic volumes, occupancies, speeds and signal plans. These traffic data are collected by various detection systems. In most cities, inductive loop detectors, infrared detectors and/or radar detectors are implemented (e.g. Bell et al., 1996; Budde, 2002; Richards et al., 2000; Scharrer et al., 2003). Also CCTV cameras (e.g. Bell et al., 1996; Ancidei et al., 2000; Cone et al., 2002; Karl and Trayford, 2000), probe vehicles (e.g. Bae and Lee, 2000; Fellendorf et al., 2000; Ferulano et al., 2000) and cellular phone data (e.g. Karl and Trayford, 2000; Leitsch, 2002) are frequently used data sources. Finally, some cities adapted less common data sources like volunteers that serve as traffic information messengers (Bae and Lee, 2000), Public Transport fleet management or operational control systems (e.g. Henriet and Schmitz, 2000; Hoyer and Herrmann, 2003) and helicopters (FHWA, 2003). Traffic data are often combined with other types of data, such as information on road works, events and incidents (e.g. Bell et al., 1996; Hasberg and Serwill, 2000; Leitsch, 2002), data on the occupancy of parking facilities (e.g. Budde, 2002), weather data (e.g. Bell et al., 1996; Cone et al., 2002; Kellerman and Schmid, 2000), data on emissions (e.g. Bell et al., 1996; Ancidei et al., 2000; Kellerman and Schmid, 2000) and/or noise (e.g. Bell et al., 1996) and calender data (e.g. Kellerman and Schmid, 2000).

In a traffic information centre, collected traffic data is stored and processed. The basic processing tasks of these centres are: (1) combination of data from different sources, (2) data validation and (3) data visualisation. More advanced traffic information centres use traffic models and/or historical traffic data to estimate or forecast travel times or level of service (MIZAR Automazione, 1998; Di Taronto et al., 2000; Karl and Trayford, 2000; Scharrer et al., 2003), for incident detection (MIZAR Automazione, 1998; Kruse et al., 2000; Richards et al., 2000) or to estimate traffic volumes for the entire network (MIZAR Automazione, 1998; Fellendorf et al., 2000; Kellerman and Schmid, 2000).

The main service provided by the urban traffic information centres discussed here is the provision of traffic information to travellers. This is done through different channels like the internet, radio, mobile phone, PDA, on-board navigation system and (dynamic) route information panels. The type of information naturally depends on the collected traffic data and the level of data-processing. Some information centres provide multimodal information and compare travel times of different modes (e.g. Hasberg and Serwill, 2000) and in some cases, route guidance is combined with parking guidance to guide visitors to available (unoccupied) parking facilities via the quickest route (e.g. Budde, 2002). Other services provided by traffic information centres include traffic management (Ferulano et al., 2000), traffic planning (Leitsch, 2002) and research (Bell and Gillam, 1994).

Kirschfink et al. (2000) describe the Mobility and Traffic Information Center (MOTIC) architecture developed in the EU project ENTERPRICE, that includes several tools for intelligent data analysis and decision support. A MOTIC consists of two main functional elements: the traffic and transport information processing component (MOTIC-TIC) and the strategic management component (MOTIC-SMC). Figure 2.2 shows the MOTIC-SMC architecture. The MOTIC-SMC can be used for the generation, simulation and analysis of traffic scenarios. Moreover, since the MOTIC-SMC is an on-line information management platform, the traffic scenarios defined by the user can be adapted to on-line information collected by the MOTIC.

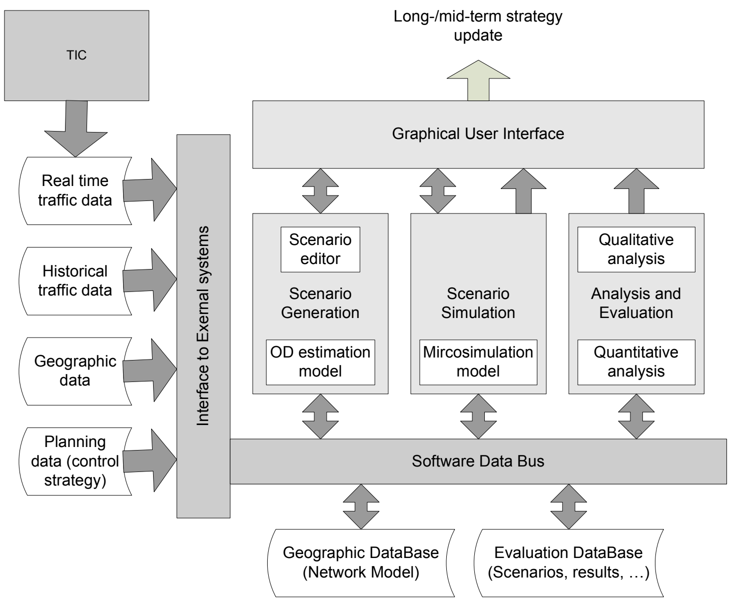


Figure 2.2: MOTIC-SMC software architecture (source: Kirschfink et al (2000)).

2.2 Data

2.2.1 Traffic data

Klein (2001) and Bennett et al. (2005) give an overview of existing traffic data collection systems. They distinguish intrusive and non-intrusive sensors. Intrusive sensors are those that involve placement on top of or in the lane to be monitored, e.g. inductive loop detectors, magnetic sensors, pneumatic tubes and Weight In Motion (WIM) sensors. Non-intrusive sensors do not interfere with traffic either during installation or operation and include infrared sensors, radars and video image detection. Besides these road based sensors, also vehicles or road users can serve as a data source. Examples of such data sources are vehicles that are equipped with a transponder or electronic tag or people that carry a (turned on) cellular phone.

Inductive loop detectors

Inductive loop detectors observe vehicles through the principle of induction. The functioning of inductive loop detectors is explained in Papageorgiou (1991) and Klein (2001) and shown in Figure 2.3. The detector consists of an insulated

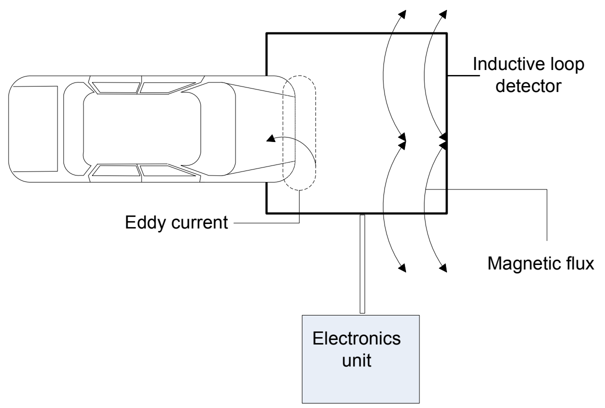


Figure 2.3: Working of inductive loop detector (adapted from Papageorgiou (1991)).

wire buried in a shallow sawcut in the roadway and an electronics unit, located in the controller cabinet. The wire loop is an inductive element in an oscillatory circuit that is energized by the electronics unit. When a vehicle stops on, or passes over the loop, the inductance of the loop decreases as a cause of eddy currents that are induced in a metal vehicle. The decreased inductance increases the oscillation frequency and causes the electronics unit to send a pulse to the controller, indicating the presence or passage of a vehicle. By means of data processing techniques, traffic volumes and occupancy levels can be extracted from these signals.

Both single and dual loop detectors are used. Dual loop detectors are mainly installed on highways and consist of two single loop detectors on a short distance from each other. From the time difference between the signals produced at the first and the second detector, speeds can be estimated. In most Dutch cities, single loop detectors are installed at signalized intersections for actuated signal control. These single loop detectors can only be used for the measurement of flows and occupancies, although algorithms are developed for the estimation of speeds as well (e.g. Wang and Nihan, 2000; Hellinga, 2002).

In general, two types of single loop detectors can be distinguished: short detectors for the detection of vehicles and long detectors for the detection of queues. There is no standard configuration that is implemented in all cities. In most cities, short detectors are located (1) just upstream of the stop line to detect the presence of a vehicle that should get green during that cycle or (2) further upstream to detect (a) queues that exceed a preset maximum queue length or (b) vehicles to anticipate on green. Long detectors are generally located upstream of the first short detector for the estimation of queue lengths with respect to the calculation of green times. Some network optimizing signal control systems use short detectors located downstream of signalized intersections.

Other traffic data sources

Besides single loop detectors, also other detection systems are adopted in urban areas. First of all, pneumatic tubes are often used for short term traffic counts. A pneumatic tube is a hollow rubber tube that detects vehicles by the change in air pressure in the tube. Every vehicle axes that passes the loop is recorded by an air switch. Axle counts can be converted to count, speed and/or classification depending on how the road tube configuration is structured (Bennett et al., 2005). Pneumatic loop detectors are easy to install and remove and are therefore appropriate for short term traffic counts throughout a city.

Non-intrusive sensors that are frequently used include infrared detectors, radars and traffic cameras (see Section 2.1). With regard to infrared sensors, a distinction is made between active and passive sensors (Klein, 2001; Bennett et al., 2005). Active sensors emit a laser beam at the road surface and measure the time for the reflected signal to return to the device, being less when a vehicle is present. Passive sensors measure the infrared energy radiating from the detection zone, which is influenced by the presence of a vehicle. Infrared sensors can be used to record traffic volumes, speeds and classification data (Bennett et al., 2005). Radar (radio detection and ranging) sensors detect vehicles through the transmission of high frequency radio waves. The time delay of the return signals is a measure for the distance of the detected vehicle. Radar technology is capable of recording traffic volumes, speeds and simple classification (Bennett et al., 2005). Traffic cameras provide a picture of the local traffic situation and can for example be used for monitoring the operation of critical intersections and for evaluating signal timing and related functions (Klein, 2001). Besides, by means of video image processing, traffic volumes, lane occupancy and speeds can be estimated. Tracking systems are even able to provide link travel times and OD information by identifying and tracking vehicles as they pass from one camera's field of view to the next. For more information, the reader is referred to Klein (2001).

A relatively new source of urban traffic data are floating car data that are collected by GPS or GSM based systems. These systems track vehicles that are equipped with a GPS system (e.g. taxis, public transport vehicles and probe vehicles) or people that carry cellular phones. By locating vehicles or cellular phones at subsequent moments in time, travel times and information on congestion can be obtained. For more information on this subject see for example Zito et al. (1995), Quiroga and Bullock (1998), Kroes et al. (1999), Zhao (2000) and Huisken (2003). Besides for the estimation of travel times and other traffic system performance measures, GPS and GSM based systems are also applied for the analysis of travel behaviour and trip patterns (e.g. White, 2001; Du and Aultman-Hall, 2007).

2.2.2 Data on factors potentially influencing traffic

As stated in the previous chapter, both variations in traffic demand and variations in road capacity cause variations in the traffic state. Factors that potentially cause variations in traffic demand and/or road capacity are: (1) type and time of day, (2) weather, (3) events, (4) road works and (5) accidents.

Information on the time of the day and the type of day (calendar data) can easily be linked to measured traffic volumes. All traffic volume measurements have a time and date stamp and every date has its characteristics (e.g. day of the week).

The main source of weather data in The Netherlands is the Royal Netherlands Meteorological Institute (KNMI). Historical weather data on a daily basis from ten main weather stations throughout the country are available at http:/www.knmi.nl/. The data include: temperature, cloudiness, hours of sunshine, visibility, humidity, amount and duration of precipitation, wind speed and direction and air pressure. Additionally, data on an hourly basis as well as validated precipitation data (on a daily basis) using various (small) weather stations can be purchased. Also 'weeronline' (http://www.weeronline.nl/) and amateur weather stations throughout the country provide weather data.

Events and road works are known by local governments and are published in local newspapers. However, in many cities, there is no central database in which the times, locations and impacts of events and road works are stored digitally. As a result, data on events and road works cannot be linked automatically to a Geographic Information System (GIS) to visualize them nor can they be linked automatically to a traffic database to estimate impacts of road works and events on traffic volumes. Reefhuis (2005) proposes a design for an information system on road works and events.

Accidents are reported to the police that maintain a database with all reported accidents. It has to be noted that not all accidents are registered. However, in general, accidents that have a high impact on traffic (resulting in for example a temporary road closure) are registered.

2.3 Data processing

The data discussed in the previous section has to be stored and processed. As we mentioned in Section 2.1, the basic processing tasks are the combination of data from different sources, data visualisation and data validation. The way in which data is visualised depends on the use of the data and the objective of the analysis and will therefore not be discussed here. Data storage and the combination of data from different sources are managed by traffic information centres and will only be briefly discussed here. Data validation is discussed in more detail.

Data from various detector stations is sent to a central database in which the data is structured in such a way that useful items can be extracted easily. A database management system (DBMS) can be a useful tool to control the use of a database (Bell et al., 1993). Besides the measured traffic volume, a data record should include a time stamp and a location code. These codes enable the data to be combined with other time and/or location specific information (e.g. calendar data and traffic network data).

With regard to data validity, Turner (2001) states that quality control techniques for archived data should encompass at least:

1. Missing data
2. Suspect or erroneous data: illogical or improbable data values that do not fall within expected ranges or meet established principles or rules
3. Inaccurate data: data values that are systematically inaccurate (but within range of plausible values) because of equipment measurement error

Both erroneous and inaccurate data refer to deviations from true traffic volumes and in the remainder, they will be referred to as invalid data.

Missing and invalid data can be removed from further analysis or be replaced by alternative values (imputation). The best way to deal with missing and invalid data depends on the application. In this research, the data is used to detect and analyse traffic patterns. Replacing missing and invalid data creates a risk that traffic patterns are imported in the data. Therefore missing and invalid data are removed from further analysis. One of the possible applications of the detected and described traffic patterns is imputation of missing and invalid data. Chapter 8 deals with this and other applications. This section deals with the design of a quality control procedure for the detection of erroneous or inaccurate data.

In literature, different types of traffic data quality checks are described. Jacobsen et al. (1990) make a distinction between microscopic and macroscopic tests. Microscopic tests are executed on individual vehicle data, whilst macroscopic tests are executed on aggregated data. Microscopic tests are described by for example Jacobsen et al. (1990), Chen and May (1987) and Coifman and Dhoorjaty (2002). Most macroscopic quality checks that are executed in practice are based on minimum and maximum thresholds that are executed on individual records of traffic volume or occupancy measurements (e.g. Turner et al., 2000; Lomax et al., 2004). More sophisticated tests include checking for implausible combinations of volumes, occupancies and/or speeds (e.g. Jacobsen et al., 1990; Cleghorn et al., 1991; Turner et al., 2000), analysing series of measured traffic volumes together (Chen et al., 2003) and comparing measured traffic volumes with historical data (e.g. Chen and May, 1987; Ishak, 1990; Turner, 2004) or with data from other locations (Kikuchi and Miljkovic, 1999; Wall and Dailey, 2003; Kwon et al., 2004; Vanajakshi and Rilett, 2004; Schoemakers and Van Engelenburg, 2003). Finally, in some traffic information centres discussed in Section 2.1, data is validated by comparing data from a number of sources. Subsequently, data from different sources are combined in order to obtain the best possible dataset (e.g. Bae and Lee, 2000; Henriet and Schmitz, 2000).

The quality control algorithm developed in this research assumes that only aggregated traffic volume data, originating from one data source is available. Therefore, microscopic quality checks (that require individual vehicle data), tests that check for implausible combinations of volumes, occupancies and/or speeds and tests that compare data from a number of sources are not discussed here. Moreover, a disadvantage of the quality check that compares traffic volumes with historical data is that volumes can also deviate from historical values as a result of special circumstances like events or road-works. Since the goal of this research is to analyse traffic patterns, it is undesirable to remove traffic data that deviates from historical values as a result of special circumstances. Therefore, also a check that compares traffic volume measurements to historical data is not applied.

For this research, a data validation procedure is designed that combines basic macroscopic quality checks with checks that compare traffic volumes from multiple locations. All checks are executed on a daily record of traffic volume measurements.

Let us define:

q_{mdt}: measured traffic volume at monitoring detector m on day d for time interval t, and

R_{md}: record of measured traffic volumes at monitoring detector m on day d:

$$R_{md} = (q_{md,1}, ..., q_{mdt}, ..., q_{md,Nt}) \tag{2.1}$$

where Nt is the number of measurement intervals on a day. Nt is determined by the length of the measurement interval.

On all records R a number of quality checks Q are executed that are indexed by i. Each quality check has two possible outcomes, 0 in case the quality check is not passed and 1 in case the check is passed, i.e.:

$$Q_i(R_{md}) = \begin{cases} 0 & \text{if quality check is not passed} \\ 1 & \text{otherwise} \end{cases} \tag{2.2}$$

A record is removed from further analysis in case that one or more of the quality checks are not passed, i.e. when:

$$\prod_i Q_i(R_{md}) = 0 \tag{2.3}$$

The basic quality checks and quality checks based on the principle of conservation of vehicles are discussed in more detail in the remainder of this section. In Chapter 5, the quality control procedure is adjusted for and applied to the traffic data of Almelo. For a more detailed description of the data validation procedure, the reader is referred to Weijermars and Van Berkum (2006a).

Basic quality checks

The basic quality checks are based on minimum and maximum volume thresholds. Regarding the maximum flow threshold, traffic volumes are bounded by the capacity of the measurement location and by the capacity of upstream locations. Naturally, the capacity is not the same for all locations. Moreover, the capacity varies in time as a result of varying conditions (e.g. weather). For reasons of simplicity, one fixed upper limit is used for all locations and all circumstances.

Turner (2001) and Lomax et al. (2004) consider 250 vehicles per 5 minutes (i.e. 3000 vehicles per hour) to be an appropriate upper limit for a link. Since on signalized intersections traffic can only flow during green time, a second, lower threshold is introduced. Measurements above this second threshold are flagged to be suspicious and are further investigated by analysing the daily traffic profile. When the daily flow profile looks abnormal, i.e. when traffic volumes are alternately very high and very low or are very high for consecutive time intervals, a record is assumed to contain erroneous traffic data. When a record does not look abnormal, i.e. when it shows high volumes for some intervals during peak periods, the record is assumed to be valid. The threshold for suspiciously high volumes has to be selected on the basis of an analysis of the available traffic data. The resulting algorithm for the quality check based on the maximum threshold can be represented by:

$$Q_1(R_{md}) = \begin{cases} 0 & \text{if } \exists_t \, (q_{mdt} > 3000 \vee \\ & (T_1 \leq q_{mdt} \leq 3000 \wedge R_{md} \text{ looks abnormal})) \\ 1 & \text{otherwise} \end{cases} \tag{2.4}$$

where T_1 is a threshold for suspiciously high traffic volume measurements.

Three minimum volume thresholds are used. First of all, negative traffic volume measurement are removed from the database:

$$Q_{2a}(R_{md}) = \begin{cases} 0 & \text{if } \exists_t q_{mdt} < 0 \\ 1 & \text{otherwise} \end{cases} \tag{2.5}$$

Secondly, traffic volumes may be zero for one or more measurement intervals on quiet locations and during the evening and night, but traffic counts of zero vehicles for many consecutive time intervals are suspicious. Daily traffic volumes cannot be zero (except in case of road closures, but these data have to be removed as well). Besides, hourly traffic volumes of zero vehicles are suspicious, but might occur. Therefore hourly traffic volumes of zero vehicles are further examined. When present, upstream detectors are used for the verification of zero volume measurements. Because of the time lag between upstream and downstream measurements, low traffic volumes can be measured at upstream detectors in case of zero traffic volumes at a well-functioning monitoring detector. Therefore, upstream hourly volume have to be larger than a certain threshold to report a monitoring detector to be malfunctioning. Also this threshold is selected on the basis of an explorative analysis. In cases where no upstream detectors are available, records with reported hourly traffic volumes of zero vehicles are further investigated by examination of the daily flow profile. When volumes are zero for consecutive hours or alternately zero and very high, a detector is assumed to be malfunctioning. The algorithms that check for zero traffic volumes can be represented by:

$$Q_{2b}(R_{md}) = \begin{cases} 0 & \text{if } \sum_t q_{mdt} = 0 \\ 1 & \text{otherwise} \end{cases} \tag{2.6}$$

$$Q_{2c}(R_{md}) = \begin{cases} 0 & \text{if } \exists_{h \in [8,19]} q_{mdh} = 0 \wedge \\ & (\exists_{u \in Su_m} q_{udh} > T_2 \vee (Nu_m = 0 \wedge R_{md} \text{ looks abnormal})) \\ 1 & \text{otherwise} \end{cases} \tag{2.7}$$

where

$$q_{mdh} = \sum_{j=1}^{\alpha} q_{md,\alpha(h-1)+j}, \tag{2.8}$$

T_2 is a threshold regarding the number of vehicles queueing between two detectors, α is the number of measurement intervals in an hour, q_u are reported traffic volumes at location u upstream of m, Su_m is the set of upstream detectors belonging to m and Nu is the number of upstream detectors.

Quality checks based on the flow conservation law

Traffic volumes are measured at different locations. For two locations between which traffic cannot 'leak away' and new traffic cannot be generated, the principle of conservation of vehicles applies. This implies that the total number of vehicles counted at an upstream detector should be counted at the downstream detector at some future time (Wall and Dailey, 2003). Unfortunately, in the urban transportation network traffic may be generated or leaking away on many locations, like non-monitored intersections and parking lots. Therefore, in general, the principle of conservation of vehicles is difficult to apply for the urban road network. However, there might be situations in which traffic is detected at two (sets of) detectors without traffic leaking away or being created between them.

The quality check based on the principle of conservation compares the amount of vehicles reported during a certain time interval for two or more locations. It is investigated whether the difference in traffic volume is within a certain threshold. Besides invalid data, also changes in the number of vehicles between the detectors cause differences in traffic volume. To minimize this effect, the principle of conservation of vehicles is only applied on hourly and daily traffic volumes. Moreover, differences in traffic volumes are corrected for possible changes in the amount of vehicles between two detectors. The maximum difference resulting from a change in the number of vehicles between two detectors can be calculated using the distance between two detectors and the jam density (in that case it is assumed that the amount of vehicles between the detectors is zero at the start of the measurement interval and equals the maximum amount at the end of the measurement interval).

The general algorithms for the quality control check on the basis of the principle of flow conservation can be represented by:

$$q_a(R_{md}) = \begin{cases} 0 & \text{if } \dfrac{|q_{L_1d} - q_{L_2d}| - T_2}{0.5(q_{L_1d} + q_{L_2d})} > T_3 \\ 1 & \text{otherwise} \end{cases} \tag{2.9}$$

$$q_b(R_{md}) = \begin{cases} 0 & \text{if } \exists_h \dfrac{|q_{L_1dh} - q_{L_2dh}| - T_2}{0.5(q_{L_1dh} + q_{L_2dh})} > T_3 \\ 1 & \text{otherwise} \end{cases} \tag{2.10}$$

where L_1 and L_2 are two (sets of) locations between which the principle of conservation of vehicles is applied and T_3 is a threshold for the percentage difference between the measured traffic volumes.

The maximum allowable percentage difference between two types of detectors is determined by the required data quality and the accuracy of the detectors. The required accuracy depends on the application of the traffic data. For traffic management applications, 10% is a possible accuracy threshold (Turner, 2004). When the detectors are however more accurate, a lower threshold can be adopted. For this study, we determined the threshold on the basis of regular differences in traffic volumes. The expected inaccuracy should however not be larger than 10%.

2.4 Data interpretation

If we assume that the quality checks described in the previous section remove invalid data adequately, the measured traffic volumes are a good estimation of the true traffic volumes. However, even when detectors are functioning adequately, the amount of detected vehicles is not 100% correct. This is due to inaccuracy of loop detectors. The accuracy of a detector is location specific and depends on the installation and tuning of the loop. Deckers (2001) found accuracies between 98% and 100% for single loop detectors in Rotterdam, The Netherlands. The accuracy of detection can be evaluated by additional traffic counts using other detection methods.

As mentioned in Chapter 1, traffic volumes are the result of a combination of traffic demand and traffic supply characteristics. In case that traffic demand is lower than capacity, traffic volumes equal traffic demand, at least when demand does not include latent demand, rerouted trips and future growth. Roess et al. (1998) discuss two basic cases in which volume represent capacity instead of traffic demand:

1. An upstream metering effect; due to signal timing or other capacity limitations traffic does not reach the measurement location without being distorted
2. A queue at the measurement location; the observed volume reflects the downstream discharge instead of the upstream demand

The interpretation of the traffic volumes highly depends on the aggregation level at which the data is analysed. In an urban network, instantaneous traffic volumes are highly influenced by the state of upstream and downstream traffic signals. Section 3.1.1 discusses short term variations in traffic volumes caused by traffic light cycles. The higher the aggregation level, the less the influence of traffic light cycles on traffic volumes. Also in case of upstream or downstream congestion, the aggregation level plays a role. The longer the time period that is analysed, the smaller the probability that traffic volumes are limited by capacity restraints. In case of peak period congestion for example, queues are dissolved during the period after the peak. Assume that the A.M. peak period is from 7:00 - 9:00 and that the queues are dissolved at 9:30. In that case, the peak volume (7:00 - 9:00) does not adequately represent the demand during this period, whereas the traffic volume from 7:00 - 9:30 adequately represents the traffic demand between 7:00 and 9:30.

2.5 Summary

Recently, urban traffic data is becoming available for research and other purposes as a result of the development of urban traffic information centres that collect and process urban traffic data for different services like the provision of traffic information. The data can also be used for the analysis of variations in urban traffic volumes.

In Dutch cities, inductive loop detectors are the major source of traffic data. Besides, pneumatic tubes, infrared detectors, radars, traffic cameras and floating car data may provide additional data. The data is sent to an urban traffic information centre and further processed. For this research, the main processing task is data validation. We proposed a quality control procedure that detects invalid daily records of volume measurements using minimum and maximum volume thresholds and the principle of conservation of vehicles. If we assume that this quality control procedure removes invalid data adequately, the measured traffic volumes are a good estimation of the true traffic volumes. In case that traffic demand is lower than capacity, these traffic volumes represent traffic demand, in case of upstream or downstream capacity restraints they represent capacity.

The traffic data can be combined with calendar data, weather data, data on road works and events and accident data in order to explain variations in measured traffic volumes. Weather data is available from the Royal Netherlands Meteorological Institute (KNMI) and accident data is available from the police. Road works and events are known by local governments, but often there is no central database in which the time, location and impact of events and road works are stored digitally.

Chapter 3

Variations in urban traffic

Urban traffic clearly is not a static phenomenon. The traffic volumes collected by the traffic information centres discussed in the previous chapter vary both in time and in space. This chapter deals with these variations. It provides an overview of existing literature on this topic. When literature dealing with the urban situation is limited, also literature concerning highways is taken into account. On the basis of this overview it is discussed what topics need further research and are addressed in this thesis. The first section deals with temporal variations in urban traffic volumes and the second section with spatial variations. In the third section, variations in traffic volumes are explained by variations in travel behaviour. The chapter concludes with a discussion.

3.1 Temporal variations

Temporal variations in traffic volumes can be analysed at different time scales, ranging from minute-to-minute variations to year-to-year variations. Common time scales are shown in Figure 3.1. In this section, variations on different time scales are described.

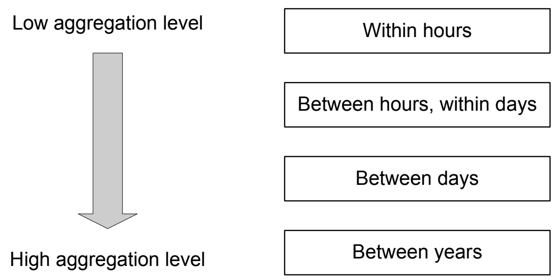


Figure 3.1: Common time scales for analysing temporal variations in traffic volumes.

3.1.1 Short term variations

In urban traffic, minute-to-minute variations are highly influenced by traffic light cycles of both the downstream and upstream intersections. Figure 3.2 illustrates the traffic process at a signalized intersection. Measured traffic

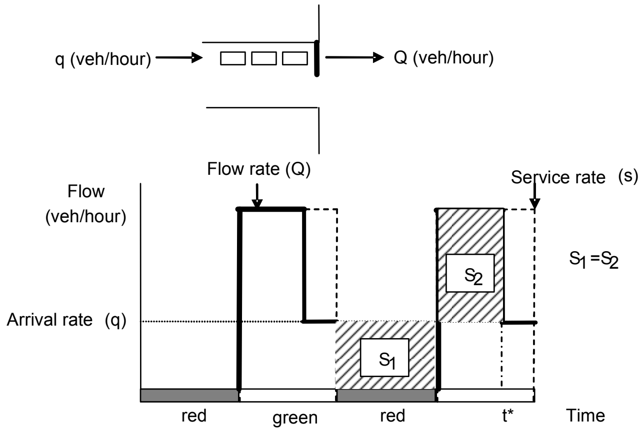


Figure 3.2: Traffic volumes at signalized intersections (based on May (1990) and Taylor et al. (1996)).

volumes (Q) are determined by the service rate (s) and arrival rate (q). The service rate is zero during the red phase and equals the saturation flow rate during the green phase. In the example in Figure 3.2, the arrival rate is assumed to be constant during a traffic light cycle. In practice, the arrival rate may vary as a result of variations in traffic demand or traffic light cycles at upstream intersections. The resultant traffic flow (Q) is zero during the red phase, resulting in a queue at the end of the phase (surface S_1). During the green phase, the queue is dissolved (surface S_2). Until the queue has dissolved, the flow rate (Q) is equal to the service rate. After the queue has dissolved (t^*), the flow rate equals the arrival rate. For more information on the queuing process on signalized intersections the reader is referred to May (1990) and Taylor et al. (1996).

3.1.2 Variations within a day

Various authors deal with the general shape of the daily traffic profile (e.g. Festin, 1996; DLTR, 2001; US DOT, 2001; Chrobok et al., 2004). Most authors distinguish different types of roads as well as different types of vehicles (cars and trucks). All authors make a distinction between working days and weekend days. An average working day show both an A.M. and a P.M. peak period and an off-peak period in between. According to Taylor et al. (1996), in the UK, 8% to 12% are typical values for the peak hour factor, i.e. the ratio between peak hour volume and total daily traffic volume. In general (aggregated over multiple locations) the P.M. peak is higher than the A.M. peak and the P.M. off-peak volume is higher than the A.M. off-peak volume. Lomax et al. (2003) compared delay for different moments in time and found that 47% of total delay occurred during the P.M. peak against 30% for the A.M. peak. Furthermore 22% of the total delay occurred during the afternoon off-peak period against 0% during the morning off-peak period (the other 1% of total delay occurred during the late off-peak period). Transpute (2000) compared delay for different times of the day on highways in The Netherlands and also concluded that the P.M. peak is heavier than the A.M. Peak (respectively 47% and 40% of total delay).

Besides the general shape of the daily flow profile (hour-to-hour variations in traffic volumes), also the shape of the peak periods is important for traffic management. According to Chrobok et al. (2004), the general daily traffic profile of Mondays until Thursdays shows a sharp morning peak and a higher and broader P.M. peak. According to May (1990), flat peaks may indicate that demands exceed capacities and that the peak period is being extended.

Obviously, the exact shape of the daily flow profile differs by location. These differences are discussed in Section 3.2.

3.1.3 Variations between days

Regarding variations between days, a distinction is made between systematic variations and random variations. Systematic variation is defined in this thesis as variation that can be explained by temporal and circumstantial factors like weekdays, seasons, (Public) Holidays, and weather. Random variation is accordingly defined as variation that cannot be explained for.

Systematic variations that are described most often in literature are variations between weekdays, between seasons, and between different weather conditions. Most information on day-of-week and seasonal variations in traffic volumes is found in monitoring reports (Festin, 1996; Wright et al., 1997; DLTR, 2001) and literature concerning the estimation of AADT (Average Annual Daily Traffic) on the basis of short-term traffic counts (Lyly, 1968; Erhunmwunsee, 1991; Schmidt, 1996; Sharma et al., 1996; Hu et al., 1998; Aunet, 2000; US DOT, 2001; Li et al., 2003). These reports describe variations in daily traffic volumes by monthly and day-of-week adjustment factors. Information on variation in traffic volumes due to weather mainly concerns research on the influence of rain on traffic volumes, accidents and driving behaviour.

Seasonal variations are not very strong for urban traffic (Aunet, 2000; Stathopoulos and Karlaftis, 2001b; Keay and Simmonds, 2005). Moreover, seasonal variations appear to differ between Europe and the United States. Researchers in Europe state that traffic volumes are lower in July and August as a result of less commuter traffic during summer holidays (Fox and Clark, n.d.; Lyly, 1968; Schmidt, 1996; Stathopoulos and Karlaftis, 2001b; Chrobok et al., 2004) whilst from research in the USA it is concluded that traffic volumes are slightly higher during summer (Festin, 1996; Aunet, 2000). These differences are probably due to a difference in the amount of holidays, being less for employees in the USA.

Regarding day-of-week variations, traffic volumes are clearly lower on weekend days compared to weekdays. Moreover, various authors conclude that traffic volumes are higher on Thursdays and Fridays, compared to Mondays, Tuesdays and Wednesdays (Schmidt, 1996; US DOT, 2001; DLTR, 2001; Keay and Simmonds, 2005). Stathopoulos and Karlaftis (2001b) tested whether daily traffic volumes differ significantly between the days of the week and found no significant differences between weekdays. However, weekend days differed statistically significant from weekdays and Saturdays differed from Sundays.

Most research into the influence of weather variables on traffic volumes concludes that traffic volumes decrease on rainy days (Tanner, 1952; Codling, 1972; Edwards, 1999; Goodwin, 2002; Keay and Simmonds, 2005; Chung et al., 2005). On the contrary, Hogema (1996) did not find a significant difference in volumes between dry days and days with rain on a Dutch highway. Also Changnon (1996) found no measurable effect of rain on weekday traffic volumes on freeways in the Chicago metropolitan area. Hassan and Barker (1999) investigated the effect of unseasonable weather on traffic volumes and found an increase in traffic in case of higher than expected sunshine hours or temperature and a decrease in traffic in case of lower than expected sunshine hours or temperature or higher than expected rainfall. Finally, most authors report stronger effects on weekend days compared to weekdays (Codling, 1972; Changnon, 1996; Hassan and Barker, 1999; Chung et al., 2005). These results most probably do not apply for Dutch cities, since cycling has a substantial share in modal split in Dutch cities ($> 25\%$ for commuter trips) (CROW, 1996). Hagens (2005) studied the effect of rain on traffic volumes in a Dutch city and found that urban traffic volumes are higher on wet days.

Some authors investigated the effect of Public Holidays or days adjacent to Public Holidays on traffic volumes and report contradictory results. Bexilius and Kengen (1997) found that on highways, days with high traffic volumes mainly occur adjacent to Holidays in the spring. In contrary, Fox and Clark (n.d.) analysed traffic in urban areas and concluded that the volumes on Public Holidays are either much lower than usual or are very similar to volumes on Sundays. These contradictory results appear to be due to different types of roads being investigated. According to Liu et al. (2005) and Liu and Sharma (2006), the effect of Public Holidays and adjacent days on traffic volumes namely depends on the type of road (also see Section 3.2). Moreover, Festin (1996) found that also the day of the week the Holiday is on determines in what way traffic is affected. Chung and Rosalion (2001) examined the effect of school holidays on traffic volumes on an expressway in Melbourne and conclude that the pattern over the holiday period does not appear to be very different from that of the normal traffic flow pattern.

Besides the daily traffic volumes, also the shape of the daily traffic profile may differ between different types of days. Transpute (2000) for example states that total delays on Dutch highways are approximately the same for all weekdays, whilst A.M. peak and P.M. peak delays vary substantially between weekdays. Besides the total daily traffic volume it thus is also interesting to compare the distribution of the traffic over the day for different types of days. Differences in daily traffic flow profiles however have received considerably less attention in literature. Moreover, most of the available literature deals with highways. Edwards (1999) found that the A.M. peak on a highway in Wales is lowest on Fridays. Also BGC (1986) found a different daily flow profile on Friday on all types of Dutch roads. According to them, Mondays to Thursdays show an almost symmetric off-peak period between the A.M. peak and the P.M. peak, whilst on Friday traffic loads increase gradually from 10.00 o'clock to 17.00 o'clock. Rakha and Van Aerde (1995) studied highway traffic volumes and found that the P.M. peak is extended further in the day on Fridays. Moreover, they compared daily traffic profiles and found significant differences between core weekdays (Tuesdays, Wednesdays and Thursdays) and Mondays and between core weekdays and Fridays. Festin (1996) studied traffic volumes on rural and urban highways and states that the Friday P.M. peak is highest of all peaks. Iryo et al. (2005) studied traffic profiles on an urban expressway in Japan and state that the ratio of traffic volume in morning to traffic volume in evening changes depending on seasons. Only Chrobok et al. (2004) dealt with urban daily traffic flow profiles. In the city of Duisburg they investigated which days of the week show comparable daily traffic profiles by means of a matching process. This matching process – that compares the patterns on the basis of an error measure – result in four classes: (1) Monday until Thursday except Holidays or days before Holidays, (2) Friday and days before holidays, (3) Saturday except Holidays, and (4) Sundays and Holidays.

The literature discussed so far compared different types of days that were defined on the basis of daily characteristics. Another way to obtain insight into the variability of daily traffic volumes is to classify days on the basis of their daily flow profiles and to determine what factors are responsible for the resulting groups. In this way, other factors influencing daily traffic profiles might be detected. Only a very limited amount of literature deals with this classification of traffic volumes or travel times. Iryo et al. (2005) classified daily flow profiles on an urban expressway in Osaka (Japan). Days on which an event took place showed dissimilar daily flow profiles. Moreover, weekend days and Public Holidays are classified to different clusters than weekdays. They do however not discuss what other factors are responsible for the classification. Chung (2003) classified A.M. peak and P.M. peak travel times for the Tokyo Metropolitan Expressway and investigates what temporal factors are on the basis of the resulting groups. For the A.M. peak, weekdays are classified to the same group, whilst Saturday and Sunday are classified to two separate clusters. The Sunday cluster also includes Monday Holidays suggesting that it is a Holiday cluster. For the P.M. peak, each day should be treated separately. Moreover, rainfall does not seem to be a factor in the clustering.

Besides obtaining insight into systematic variations in urban traffic it is also important to have information about the amount of random variation, since it provides information on the predictability of traffic volumes. When almost all variation can be explained for by predictable temporal and circumstantial factors, traffic volumes are highly predictable as well. A large influence of unpredictable factors like accidents or a large amount of random variation implies traffic volumes are unpredictable. Keay and Simmonds (2005) designed a linear regression model incorporating trend, day of the week, holidays and weather effects and explained 95% of the variation in daily traffic volumes on two Freeways in Australia. Nowotny et al. (2003) quantified the influence of various classification factors on traffic volumes by means of ANOVA analysis. They found that 75% of the variation in hourly traffic volumes in Vienna could be explained by differences between measurement locations, time of day, different types of days, seasonal variations or weather conditions. The residual 25% of the variation is caused by unregistered factors (e.g. road works) or unpredictable events like accidents. Fox and Clark (n.d.) compared the amount of random variation for different time periods. They state that the variation is higher early in the morning, but remains approximately constant for the rest of the day. Moreover, weekend days show higher variability than weekdays (Rakha and Van Aerde, 1995; Wright et al., 1997; Zhang et al., 2002). Zhang et al. (2002) calculated the correlation coefficients for different days of the week and found the correlation between Tuesdays to be largest. Erhunmwunsee (1991) compared the coefficients of variation for different months and concludes that coefficient of variation is lowest for April and highest for December. Bellamy (1978) found the coefficient of variation to be lowest in April, May, September and October and highest in August. Turochy and Smith (2002) developed a variability index based on multivariate statistical quality control and compared the value of the index for different time periods. They found that the variability on urban freeways in a region of Virginia is greatest for brief periods during the A.M. and P.M. peak periods. Moreover, the variability during the A.M. peak period was highest on Monday and Friday.

3.1.4 Long term variations

As a result of demographic, economic and geographical changes, traffic volumes also vary over years. Festin (1996) studied the trend in traffic volumes between 1970 and 1995 in the United States. Naturally, the amount of traffic increased between 1970 and 1995. Festin (1996) also investigated the trend in monthly and weekly variations in traffic volumes. Between 1985 and 1995, the seasonal variation in traffic was smaller than between 1970 and 1984. Overall though, the patterns appear to remain constant. Also the weekly patterns appear to remain stable over time. Finally, he investigated the trend in the daily flow profile and found that weekday travel is becoming increasingly concentrated between 5:00 and 18:00. The afternoon peak is spreading, but it is spreading faster into the middle of the day than into the evening hours. Locally, long term variations in traffic volumes are mainly determined by land use and infrastructural developments.

3.2 Spatial variations

Analysing spatial variations in urban traffic volumes sec (without taking the temporal component into account) results in the identification of locations with high traffic volumes for a certain time interval. The results of these analyses are location specific and do not provide more insight in urban traffic in general. It is more interesting to analyse spatial variations in combination with temporal variations. As with temporal variations, analyses can be executed at multiple time scales. Moreover, different spatial aggregation levels can be distinguished. Figure **??** shows analyses that can be executed on different combinations of temporal and spatial aggregation levels. These analyses are discussed in this section.

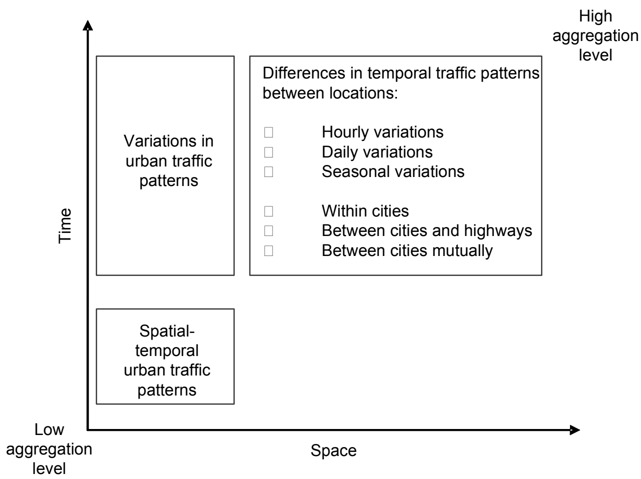


Figure 3.3: Analyses of variations in traffic volumes at different spatial - temporal aggregation levels.

3.2.1 Spatial variations in traffic volumes

Although limited, there is some literature available on spatial variations in traffic volumes. Zhao and Chung (2001) developed a method for the estimation of average annual daily traffic (AADT) for roads on which no traffic counts are carried out. They investigated what factors contribute to AADT on a road. The function of a road was concluded to be the most significant predictor of AADT. This is logical, since the function of a road is determined by the expected traffic volumes. However, the use of function classes in a model fails to capture the underlying causes of varying traffic volumes. Besides function class, also the number of lanes, accessibility to regional employment centres, directness of expressway access and population and employment around a count station influence AADT.

3.2.2 Spatial-temporal traffic patterns

The term spatial-temporal traffic pattern refers to the movement of a traffic stream or the built-up of congestion throughout the network. The movement of a traffic stream through the network is used in traffic forecasting. Multivariate time series models use data from upstream detectors to improve on predictions of downstream locations (e.g. Williams, 2001; Kamarianakis and Prastacos, 2003; Stathopoulos and Karlaftis, 2003). Unfortunately, very little literature discusses spatial-temporal variations in traffic volumes. Only Stathopoulos and Karlaftis (2003) provide some information on spatial-temporal patterns along a signalized arterial in Athens. They found that the general daily traffic pattern was very similar for two detector stations that were close to each other (less than 250 m) and still quite similar even for locations with a larger distance in between. A further analysis of the cross-correlation characteristics of the time series for the various loop locations (also see Stathopoulos and Karlaftis, 2001a) showed strong short-term correlations between upstream and downstream flow measurements. Naturally, the correlation depends on the distance between the points being examined.

Besides the movement of a traffic stream, a spatial-temporal traffic pattern is also used to denote the built-up of congestion in space and time. Evans and Bell (1996) describe a method for the identification of recurrent congestion, using congestion features. A congestion feature is defined as a set of congested links connected together in a group that evolves over time. The method was applied in a region of the city of Leicester and they found 30-40 congestion features for an average weekday. Most features were produced in the morning peak period (8:00-9:30) and in the evening peak period (16:00-18:00). PARTRAS (PAttern Recognition of TRaffic States) is a method to trace back congestions to their supposed cause (Schatz et al., 2003). Congested links are grouped into chains by applying a recursive search for neighboring congestions. Schatz et al. (2003) only provide a brief overview of the method and do not discuss any results.

3.2.3 Variations in urban traffic patterns

For the spatial-temporal congestion patterns discussed in the previous subsection it can be investigated under what circumstances different patterns occur. Analyses are described in Evans and Bell (1996) and Schatz et al. (2003). Evans and Bell (1996) found that peak period congestion features usually develop according to a fixed pattern -i.e. the same links became congested at the same times each day.

Besides, some literature compares traffic patterns for different periods of the day. Stathopoulos and Karlaftis (2003) compared cross-correlations in traffic volumes for different periods of the day and found that cross-correlations varied between different periods. According to them that can be mainly attributed to the difference in the underlying traffic and OD patterns during the different periods of the day. Stathopoulos and Karlaftis (2001a) found that traffic towards the Central Business District of Athens was relatively higher in the morning peak and traffic outward from the central business district was higher in the afternoon. Gram (1996) found comparable patterns for Oslo. Moreover, he states that on weekend days there are only small differences in daily flow patterns between directions.

3.2.4 Differences in temporal traffic patterns between locations

Most literature that discusses differences in temporal traffic patterns between locations deals with the grouping of monitoring locations for the estimation of Average Annual Daily Traffic (AADT) from sample traffic counts (Sharma and Werner, 1981; Sharma et al., 1986; Flaherty, 1993; Faghri and Hua, 1995; Schmidt, 1996; Lingras et al., 2000; Li et al., 2003; Li et al., 2004). Most literature however deals with the grouping of highway or rural monitoring sites. Only Schmidt (1996) deals with the grouping of urban traffic monitoring locations. Since the results of the grouping of rural monitoring sites might also be useful to urban monitoring sites, these studies are described briefly first. Subsequently, the results of the study of Schmidt (1996) are described in more detail.

Sharma and Werner (1981), Sharma et al. (1986) and Lingras et al. (2000) classified monitoring sites in Alberta (Canada) and state that trip purpose and trip length distribution are on the basis of the differences in traffic flow patterns between locations. Faghri and Hua (1995) found that in Delaware (US) road categories could be characterized by type of road use, their attributes being urban or interstate, rural arterial, rural collector, and recreational. In contrary, Flaherty (1993) concludes that the grouping of highways in Arizona is mainly influenced by the location of the highways. BGC (1986) found similar results for the Netherlands. They did not find a correlation between the functional class of a road and the seasonal group a location was classified to. Also in a study by Li et al. (2003) in Florida, functional classification was not identified as a significant factor contributing to seasonal fluctuations in traffic. According to them this is likely because in large urban areas major roads are used by travel for mixed purposes without a single use being dominant. Li et al. (2004) investigated which land use, demographic and socioeconomic characteristics associated with the location of a monitoring site influence seasonal fluctuations in traffic counts and found seasonal movement of seasonal residents and tourists, retired people with high income and retail employment to be significant factors. Summarized, the characteristics that are on the basis of differences between locations appear to differ between study areas.

Most studies grouped sites according to their seasonal variation in traffic volumes only. Hourly and daily variations are not taken into account in most studies. Only Sharma et al. (1986) explicitly include daily and hourly variations in traffic volumes in their classification. They do however not investigate what spatial characteristics influence the variations in traffic volumes. BGC (1986) grouped monitoring sites according to weekday and hourly ratios and found that groups mainly differ from each other with regard to Saturday, Friday and Monday ratio. With regard to the classification on the basis of hourly ratios, groups of locations mainly differed from each other with regard to the A.M. peak on Mondays and Fridays and with regard to the P.M. peak on Tuesdays, Wednesdays and Thursdays.

Liu and Sharma (2006) investigated effects of some holiday weeks on traffic on multiple highways in Alberta, Canada and found that the more recreational traffic a road serves, the more days are significantly affected by holiday weeks[1]. On roads from residential areas towards recreational areas, traffic volumes are relatively high in the beginning of the holiday week, whereas in opposite direction traffic is relatively high on Mondays and Tuesdays after the Holiday. The holiday effects on commuter routes were only significant for Mondays, traffic being lower in holiday weeks compared to surrounding weeks.

Only Schmidt (1996) deals with the classification of urban monitoring locations. He groups locations on the basis of their ratio between the traffic volumes from 16:00 - 18:00 and from 12:00 - 14:00 on a weekday. For West German cities, the following working day flow profile groups are distinguished:

1. $I_{16:00-18:00}/I_{12:00-14:00} \leq 1.40$; streets in the city centre and close to the city centre with mainly business and shopping traffic. These locations show a flat daily traffic profile
2. $1.41 \leq I_{16:00-18:00}/I_{12:00-14:00} \leq 1.80$; partly radial streets to the city centre, partly access roads and roads between major traffic attractors. These locations show an A.M. peak between 7:00 and 9:00 and a high P.M. peak between 16:00 and 18:00
3. $1.81 \leq I_{16:00-18:00}/I_{12:00-14:00} \leq 2.00$; streets on the edge of the city, but also access roads and roads between traffic attractors with a high share of commuter traffic that starts early. These locations show a very high A.M. peak between 7:00 and 8:00, lower traffic in the afternoon and a very high P.M. peak
4. $I_{16:00-18:00}/I_{12:00-14:00} > 2.00$; streets like class 3, yet commuter traffic

[1] A holiday week is defined as the week from the Thursday before a Public Holiday to the Wednesday after a Public Holiday.

is mixed up with occasional traffic or commuter traffic that starts work later. These roads show a very high A.M. peak between 7:00 and 9:00 and a similar high P.M. peak between 16:00 and 18:00 with a maximum between 17:00 and 18:00.

Schmidt (1996) also classifies roads on the basis of the weekly traffic profile. Therefore he uses the Sunday factor ($bso = I_{\text{Sunday}}/I_{\text{Tuesday-Thursday}}$). The following classes are distinguished:

1. $bso \approx 0.5$; streets in city centres of large cities or at the edge of the city with a high share of commuter and business traffic, like urban highways; relatively low traffic on Sunday
2. $bso \approx 0.7$; streets in medium large cities
3. $bso \approx 0.9$; streets on the edge of cities with recreational traffic

3.3 Variations in travel behaviour

As mentioned in the first chapter, variations in traffic volumes are due to variations in traffic demand and variations in supply characteristics. Moreover, within and between day variations in traffic volumes are mainly due to variations in traffic demand, especially when no congestion occurs. This section discusses variations in traffic demand. Variations in traffic demand at a certain location can be explained by variations in travel behaviour. In this context, travel behaviour refers to all four steps of the classical transport model (Ortuzar and Willumsen, 1994), i.e. (1) trip generation, (2) trip distribution, (3) modal choice and (4) route choice, as well as trip timing.

Variations in travel behaviour can be analysed on an individual level or an aggregate level. In case of analysis on an individual level, often a distinction is made between intrapersonal and interpersonal variability (Pas, 1987; Hanson and Huff, 1988). Intrapersonal variability describes the variation in travel behaviour from day-to-day for one person, whereas interpersonal variability focuses on differences between persons. When analysing individual behaviour, single days from different persons are sometimes aggregated into one composite person, yet according to Axhausen et al. (2002) this is an unsatisfactory procedure from a statistical point of view. Moreover, in that case, no distinction can be made between intrapersonal and interpersonal variability (Hanson and Huff, 1988). Ideally, the analysis would be based on longitudinal instead of cross-sectional data. This means activity patterns and travel behaviour of a group of people is monitored for a longer time interval. These longitudinal travel behaviour data have rarely been collected (Schlich and Axhausen, 2003). The emphasis of this thesis is on the analysis of urban travel demand patterns in general, therefore in this case, analysis at an aggregate level, using cross-sectional data is appropriate. Thereby, the main interest is in differences in travel behaviour between different types of days.

The main data sources for travel behaviour research in the Netherlands are the 'Onderzoek VerplaatsingsGedrag' (OVG) and the 'MobiliteitsOnderzoek Nederland' (MON) both national mobility surveys and the 'TijdsBestedingsOnderzoek' (TBO) a national survey on the daily activities. The OVG is held every five years since 1978. Approximately 1% of the population is asked to fill in all their trips in a journey diary on a certain day. The TBO is a survey that questions the activities people employ. The survey is held every five years during two weeks in October. People are asked to write down for every quarter of an hour during one week, what was their main activity during the quarter and whether they were at home, in their municipality or outside their municipality. Both the OVG and the TBO are cross-sectional data sources. In the 1980's additional longitudinal travel behaviour data was collected by means of the 'Longitudinaal VerplaatsingsOnderzoek' (LVO) a National Mobility panel" (Van Wissen and Meurs, 1989). As well the OVG (Hilbers et al., 2004) as the TBO (Batenburg and Knulst, 1993; Harms, 2003), and the LVO (Kitamura and Van der Hoorn, 1987) have been used for the analysis of differences in travel behaviour between different types of days.

Regarding long term variations, between 1975 and 2000 people are travelling more often, for longer distances and for longer time intervals (Harms, 2003). According to Harms, this increase in the number of trips can be explained by spatial, demographic and social-economic developments and by individualization of the society and more intensive use of time.

Hilbers et al. (2004) describe seasonal and weekly variations in the amount of peak and off-peak trips (see Figure 3.4). From the left graph in Figure 3.4 can be seen that the peak travel is relatively low on Monday and Friday and the off-peak travel is relatively low on Monday and high on Friday. Regarding the seasonal variations -after correction for holiday periods- three groups of months can be distinguished (Hilbers et al., 2004): (1) the summer months that show low traffic demand, (2) the end of the year that shows high traffic demand and (3) the other months with an average traffic demand. Finally, holiday periods show lower traffic demand. Harms (2003) distinguished different travel motives when comparing the number of trips between weekdays. From Figure 3.5 can be seen that -as expected- almost all work and educational trips are made on working days, whereas many leisure trips take place on weekend days. These results agree with the results described by Kitamura and Van der Hoorn (1987) as they found that approximately 50% of the social/recreational trips are made on weekend days. Moreover, according to Kitamura and Van der Hoorn (1987) the shopping trip rate is highest on Saturdays. Harms (2003) does not consider shopping trips separately, yet he finds that the amount of personal business trips (that includes shopping trips) is highest on Saturday. With regard to the differences between working days mutually, Kitamura and Van der Hoorn (1987) state that the amount of shopping trips gradually increases in course of the week. This again agrees with findings of Harms (2003) that finds relatively low personal business trip rates on Mondays and relatively high rates on Fridays. Besides, Fridays show relatively many leisure trips, whereas

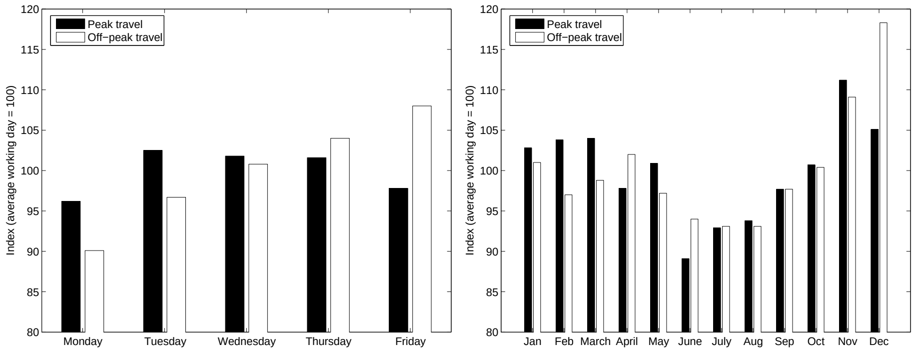


Figure 3.4: Weekly and monthly variations of car trips. Source: OVG (adapted from Hilbers et al. (2004)).

Mondays show relatively few and the amount of work and educational trips are highest on Tuesday (Harms, 2003).

With regard to the shape of the daily travel demand profile, both Harms (2003) and Hilbers et al. (2004) conclude that for car trips, the P.M. peak is higher than the A.M. peak. However, when all trips are considered, the A.M. peak is found to be shorter, yet higher than the P.M. peak (Harms, 2003). According to Harms this difference between car trips and all trips is possibly due to the fact that personal business and leisure trips on working days mainly take place in the afternoon and evening and are often made by car. Fridays show a different traffic demand profile than other working days, the amount of A.M. peak trips being lower and the P.M. peak period starting earlier and extending longer in the evening (Harms, 2003). According to Harms this longer peak period is caused by a mixture of leisure, personal business and work trips. On Monday morning the number of personal business trips is relatively low, whereas on Thursday and Friday the number of personal business trips is relatively high (Harms, 2003). On weekend days, travelling starts later in the morning and the amount of trips gradually increases with a peak around 14.00 or 15.00 o'clock (Harms, 2003).

The variations in trip rates can partly be accounted for by schedules of companies, shops, educational institute and other public services (Batenburg and Knulst, 1993). The variation in the number of shopping/personal business trips can for example be explained by store opening hours (Kitamura and Van der Hoorn, 1987). In the Netherlands, shops are closed on Sundays and Monday mornings, whereas many cities have late shopping nights (until 21.00 o'clock) on Thursdays or Fridays. Moreover, peaks in leisure traffic are caused by trips to and from sports facilities, theatres etc. (Harms, 2003).

In the literature described so far, variations in travel behaviour are due to different activities being carried out at different moments in time. Travel behaviour can also vary in time as a result of adaptations in modal choice,

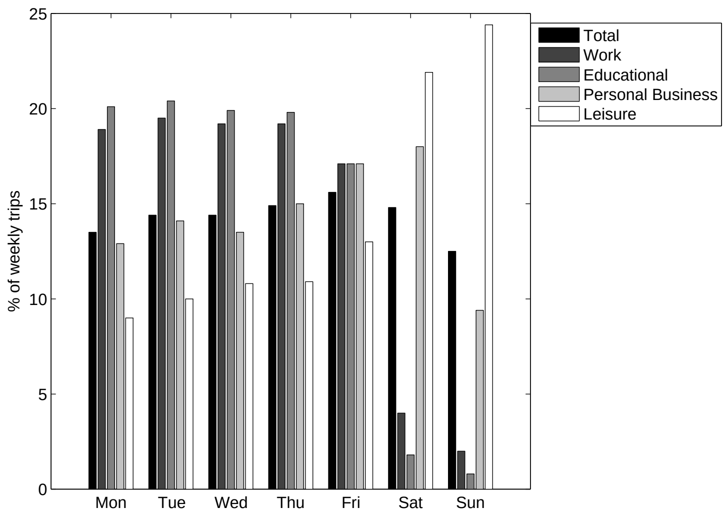


Figure 3.5: % of trips per motive on different weekdays. Source: TBO (adapted from Harms (2003)).

route choice and/or departure time due to external factors, past experiences or provided information (see for example Mahmassani, 1997). On an aggregate level, day-to-day variations in departure time and route choice will in general only result in small variations in daily flow profiles at a certain location. Only in case of special circumstances like road works or an accident, route patterns may also change considerable on an aggregate level. In Dutch cities, day-to-day changes in modal choice are mainly due to weather conditions. Hagens (2005) analysed the effect of rain and season on travel behaviour in Dutch cities using the OVG and found that rain causes a decrease in the total amount of short distance trips and a modal shift from bike to car. The combination of these effects results in an increase in short distance car trips of 12%. Moreover, he states that the seasonal influence is less prominent though still present. As a result of an increase of modal share of car driver and a small decrease in the total number of trips, 3% more car trips in winter than in summer are observed. Also Nankervis (1999) concluded that cycling is lowest in winter and decreases in case of rain. He determined the effect of weather and seasonal fluctuation on bicycle commuting patterns among tertiary students in Melbourne. On base of parked bicycle counts he found that cycling is highest during summer and autumn, declining in winter, and increasing in spring again. Furthermore, he found lower rider number in case of rain, wind and low temperature.

3.4 Discussion

This chapter provides an overview of current research on temporal and spatial variations in (urban) traffic volumes. Temporal variations can be analysed on different time scales, ranging from minute-to-minute variations to year-to-year variations. As we mentioned in Chapter 1, the main focus of this thesis is on within and between day variations. Moreover, spatial variations are most interesting to analyse in combination with temporal variations. Hereby, different combinations of time scale and spatial aggregation level can be distinguished. The focus in this thesis is on differences in temporal variations between different (types of) locations.

Existing literature on within day and between day variations in urban traffic volumes mainly deals with the general shape of a daily traffic flow profile and the influence of the day of the week and the season on total daily traffic volumes. An average working day shows both an A.M. and a P.M. peak period. The P.M. peak period is generally higher than the A.M. peak period and the P.M. off-peak is higher than the A.M. off-peak volume. In the UK, the peak hour volume typically is between 8% and 12% of the total daily traffic volume (Taylor et al., 1996). In Europe, traffic volumes appear to be a little lower in summer than in winter due to holiday periods, although seasonal variations are not very strong for urban traffic. Day of the week variations appear to be stronger, traffic volumes being clearly lower on weekend days, because of the absence of commuter traffic on these days. Moreover, traffic volumes appear to be relatively high on Thursdays and Fridays as a result of relatively many household and leisure trips . Hagens (2005) investigated the effect of rain on urban travel demand and found that traffic volumes are higher on wet days due to a modal shift from bike to car. The effect of Public Holidays and days adjacent to Public Holidays on traffic volumes depends on the type of road (Liu and Sharma, 2006) and the day of the week a Holiday is on (Festin, 1996).

Also the shape of the daily flow profile may differ between different types of days. However, the insight into these differences is limited, especially for urban areas. Only Chrobok et al. (2004) studied variations in urban daily flow profiles between different weekdays and define four groups (1) Monday until Thursday except Holidays or days before Holidays, (2) Fridays and days before holidays, (3) Saturdays except Holidays and (4) Sundays and Holidays. Besides the systematic variation, also the amount or random variation is important, since it provides insight into the predictability of traffic volumes. Keay and Simmonds (2005) could explain 95% of the variation in daily traffic volumes on two highways in Australia, whereas Nowotny et al. (2003) found that 75% of the variation in hourly traffic volumes in Vienna could be explained by known factors (time of day, differences between locations, type of day, seasonal variation and weather).

Most literature on differences in temporal variations between locations discusses differences in seasonal variations between highways or rural roads. Only

Schmidt (1996) deals with urban locations. He grouped roads in German cities according to their daily flow profile and according to their traffic volume on Sunday compared to a weekday. With regard to the shape of the daily flow profile, he found three groups: (1) streets in and close to the city centre with mainly business and shopping traffic show a flat daily flow profile, (2) radials towards the city centre, access roads and roads between major attractors show an A.M. peak and a high P.M. peak, (3) streets on the edge of the city, access roads and roads between attractors with a high share of commuter traffic show very high A.M. and P.M. peaks, and (4) roads like class 3 with occasional traffic or commuter traffic that starts work late that show a similar flow profile to class 3 yet with a longer A.M. peak. With regard to the weekend day factor, Schmidt also found three groups: (1) streets with a high share of commuter and business traffic that show relatively low traffic volumes on Sunday, (2) streets in medium large city with a somewhat higher Sunday factor, and (3) streets on the edge of cities with recreational traffic that show the highest Sunday factor (0.9).

The described variations in traffic volumes can mainly be explained by variations in traffic demand that are due to (1) different activities being carried out at different times and at different locations, and (2) variations in modal split due to weather factors.

From this overview it can be concluded that the insight into variations in daily flow profiles between different types of days is very limited, especially for urban areas. Also, little is known about differences in temporal patterns between different (types of) urban locations. Therefore, the focus on this thesis is on variations in daily flow profiles between different types of days on different types of locations. Moreover, to be able to detect other, unknown, temporal factors that influence the shape of the daily flow profile, we propose an alternative approach. Instead of comparing daily flow profiles between different types of pre-defined days, days are grouped on the basis of their daily flow profiles and it is determined what factors are responsible for the resulting groups. The next chapter proposes a method for the analysis of temporal and spatial variations in traffic volumes by means of classification of days on the basis of their daily flow profiles.

Chapter 4

Analysis of urban traffic patterns

The previous chapter discussed in what way most existing research analyses day-to-day variations in (urban) traffic: pre-defined types of days (e.g. weekdays, seasons) are compared to each other with regard to the total daily traffic volume and/or the shape of the daily flow profile. Another way to obtain more insight into urban traffic is by defining and analysing distinctive (daily) traffic patterns. In that case, days are grouped according to their daily flow profiles by means of unsupervised classification or cluster analysis. Cluster analysis has been applied before in traffic related applications, e.g. in traffic forecasting (Danech-Pajouh and Aron, 1991; Wild, 1997; Chung, 2003), for the detection of abnormal traffic patterns (Venkatanarayana et al., 2006), for the mining of loop detector data (Pribyl and Pribyl, 2005), to account for spatial variability in emissions modelling (Hicks and Niemeier, 2001) and for the estimation of Average Annual Daily Traffic (AADT) on locations with short term traffic monitoring (e.g. Sharma and Werner, 1981; Flaherty, 1993), yet the emphasis in these studies was not on the analysis of traffic patterns.

This chapter deals with the use of cluster analysis for the determination and examination of urban traffic patterns. The first section deals with the design of the clustering procedure. The second section discusses the use of clustering for the analysis of temporal variations in daily traffic flow profiles. The third section deals with cluster analysis in respect to differences in traffic patterns between locations and the fourth section describes in what way cluster analysis can be used for the determination and analysis of urban traffic patterns on a network level. The chapter ends with a summary. The methods proposed in this chapter are applied to the Dutch city of Almelo. The main results are described in Chapter 7. The methods are evaluated in Chapter 9.

4.1 Design of clustering procedure

Clustering is defined by Jain et al. (1999) as the unsupervised classification of patterns (e.g. observations) into groups (clusters). Unsupervised classification means that patterns are grouped on the basis of the data solely, in contrary to supervised classification (discriminant analysis) where patterns are pre-classified (labeled) and the problem is to label a newly encountered, yet unclassified pattern. Before the actual clustering can take place, some choices have to be made concerning the representation of the pattern and the clustering procedure. This section discusses these choices.

4.1.1 Pattern representation

The definition and representation of a traffic pattern depends on the variations that are analysed. This section describes the pattern representation for the analysis of temporal traffic patterns at one location. When analysing spatial variations or temporal variations on network level, the definition and representation of a pattern has to be adjusted. For the analysis of temporal traffic patterns at a link, let us define a cluster as a group of days that show similar daily traffic profiles at link l. A pattern thus is defined as a daily traffic profile at link l.

Cluster analysis requires a pattern to be defined mathematically, by a number of features. In literature, daily flow profiles are defined in multiple ways. Nowotny et al. (2003) used series of traffic counts, whereas Danech-Pajouh and Aron (1991) used the volume profile vector that consisted of half-hourly volumes divided by the total volume. Venkatanarayana et al. (2006) compared three representations of a traffic pattern (that describes the flow profile for a certain time interval) and conclude that a series of traffic counts is better suited for their application (the detection of abnormal daily flow profiles) than the difference in variance and difference in auto covariance. Wild (1997) transformed the time series into a structure of objects. Each object corresponded to one significant point containing information about the beginnings and the maxima of peaks. The algorithm of pattern recognition is however based on the maximum difference in volume only and therefore in the actual clustering procedure, differences in time of peak periods are not used as a distinguishing factor. According to Venkatanarayana et al. (2006), other features that are used for traffic studies include mean or total volume; maximum volume, minimum volume, range of volume; slope and curvature.

Keeping the goal of the clustering procedure in this research in mind, a daily flow profile must be defined in such a way that both the height and the shape of the daily flow profile are described. The simplest definition of a daily traffic profile that takes both the height and the shape of a daily flow profile into account is a series of traffic counts as a function of the time of the day. The number of time intervals depends on the required aggregation level. The optimal aggregation level depends on the amount and frequency of short-term variations. When the aggregation level is too low, differences between days can be due to difference in the number of green periods or other random short term variations in traffic volumes. When the aggregation level is too high on the other hand, differences in time of peak periods or peak volumes might be missed. The optimal aggregation level can be determined by analysing the daily flow profiles on different aggregation levels.

A disadvantage of the definition of a daily flow profile by a series of traffic counts is, that it does not explicitly provide information about the type of differences between traffic patterns. To obtain more insight into the type of differences, a pattern could be defined by multiple characteristics that determine the height and the shape of a daily flow profile, e.g. total daily traffic volume, peak volumes, time of peak periods and ratios between peak volumes and off-peak volumes. These features are not all measured on the same scale and therefore cannot be combined directly into one vector of features. Therefore, in Weijermars and Van Berkum (2005a) a two step procedure was proposed. In the first step, multiple clustering procedures are carried out, each using one of the selected features. In the second step, the results of the clustering procedures are combined and each combination of sub-clusters results in one ultimate cluster.

Both types of definitions were tested using highway data (Weijermars and Van Berkum, 2005b; Weijermars and Van Berkum, 2005a) and from this study it is concluded that the representation by a series of traffic counts results in a better clustering. The second definition resulted in a classification into too many small clusters. Therefore, in this thesis, a daily flow profile is represented by a series of traffic counts. This results in the following definitions for the analysis of temporal traffic patterns at one location:

Pattern (**P**): daily flow profile at link l on day d:

$$\mathbf{P}_{ld} = (q_{ld,1}, ..., q_{ld,t}, ..., q_{ld,Nt}) \tag{4.1}$$

where q is a traffic volume measurement that is indexed by link l, day d and time period t. The number of measurement intervals (Nt) depends on the aggregation level.

Cluster (C): group k of days that show similar patterns on link l.

$$C_{lk} = \{d \in Sd_l \mid \phi(l, d) = k\} \tag{4.2}$$

where Sd_l is the set of days for which valid traffic data is available for all measurement intervals for link l

$$Sd_l = \left\{ d | \prod_{m \in l} \prod_{i} Q_i \left(R_{md} \right) \neq 0 \right\} \tag{4.3}$$

and function ϕ is a function that assigns a Pattern $\mathbf{P}_{ld}$ to a cluster and $k = 1...Nk_l$ and Nk_l is the number of clusters for link l.

For every link l the clustering procedure results in a set of clusters Sc_l:

$$Sc_l = \{Sc_1, ..., Sc_{l,Nk_l}\} \tag{4.4}$$

As mentioned before, these definitions can be adjusted for the analysis of spatial traffic patterns and the analysis on network level.

4.1.2 Clustering procedure

Jain et al. (1999) discuss different clustering techniques. A distinction can be made between hierarchical clustering and partitional clustering. An hierarchical clustering procedure produces a nested series of partitions. Either every pattern starts as a separate cluster and patterns are combined step by step until one cluster remains (agglomerative clustering), or all patterns are initially combined into one cluster and then separated into smaller clusters until all patterns represent individual clusters (divisive clustering). The optimal number of clusters is determined afterwards, using a dendrogram, which visualizes the variation within the clusters for different steps of the clustering procedure.

Unlike hierarchical clustering, partitional clustering leads to only one partition, depending on the number of clusters that is chosen in advance. Chung (2003) distinguishes two types of partitional methods. The first groups the patterns in a single step by either maximizing or minimizing a criterion function (single pass method), whilst the second reallocates patterns from one cluster to another to create better clusters (reallocation method). The k-means algorithm is the most simple and commonly used partitional algorithm.

Both hierarchical and partitional methods are applied in transportation. According to Jain et al. (1999) partitional methods have advantages in applications involving large data sets for which the construction of a dendrogram is computationally prohibitive. According to Kaufman and Rousseeuw (1990) another disadvantage of hierarchical clustering algorithm is that the results of previous steps of the clustering procedure are fixed. In an agglomerative algorithm this means that once two objects are grouped to the same cluster, they cannot be separated anymore, whilst for a divisive algorithm it means that two objects that are split up cannot be reunited anymore. This possibly leads to a non-optimal grouping of the data for a certain amount of clusters. The major disadvantage of partitional algorithms is that the optimal number of clusters has to be chosen in advance. According to Kaufman and Rousseeuw (1990) not all numbers of clusters result in natural classifications. Therefore they state that it is advisable to run the algorithm several times with different numbers of output clusters and to select that number of output clusters for which certain characteristics or graphics look best. Also the so-called silhouette width can be used for the selection of the optimal number of clusters, yet also using this method, the algorithm has to be run for different numbers of clusters. Moreover, according to Jain et al. (1999) a problem with the K-means algorithm is that it is sensitive to the selection of the initial partition and may converge to a local minimum of the criterion function if the initial partition is not properly chosen.

This thesis discusses the clustering of days at multiple locations to obtain insight into urban traffic patterns. The emphasis is thus on the analysis of the resultant traffic patterns and not on the optimal design of the clustering algorithm. Since the optimal number of clusters might differ by location, it is very time consuming to select the optimal number of clusters for all locations in case of a partitional clustering algorithm. Moreover, to be generally applicable, the algorithm should be simple and available in basic statistical software program like SPSS, SAS or STATISTICA. In these software programs, more advanced partitional algorithms are not implemented. Therefore, a hierarchical clustering procedure was adopted for this research.

Within hierarchical clustering, several algorithms can be distinguished. The most commonly used algorithms are (Webb, 2002; Manchester Metropolitan University, 2004):

- Average linkage between groups (UPGMA); the distance between two clusters is defined as the average of the dissimilarities between all pairs of individuals
- Within groups clustering; this is similar to UPMGA, except that clusters are fused so that within cluster variance is minimized. This tends to produce tighter clusters than the UPGMA method
- Complete linkage clustering (furthest neighbour); the distance between two clusters is defined as the maximum of the distances between the patterns in the clusters
- Single linkage clustering (nearest neighbour); the distance between two clusters is defined as the minimum of the distances between the patterns in the clusters
- Ward's method; cluster membership is assessed by calculating the total sum of squared deviations from the mean of a cluster. The criterion for fusion is that it should produce the smallest possible increase in the error sum of squares

From above methods, the single linkage clustering is less appropriate for this application. This method can classify dissimilar patterns in the same cluster, as a consequence of intermediate patterns that reduce the minimum distance between the clusters. All other methods seem to be appropriate for this research. According to Nowotny et al. (2003) the Ward method has been frequently used and has proved to form homogeneous groups of objects. Therefore, also in this thesis Ward's method is applied. The sum of squared deviations for Cluster k is determined by:

$$V_{lk} = \sum_{d \in C_{lk}} \sum_{t=1}^{Nt} \left(q_{ldt} - \bar{q}_{lkt}\right)^2 \tag{4.5}$$

where

$$\bar{q}_{lkt} = \frac{1}{Nd_{lk}} \sum_{d \in C_{lk}} q_{ldt} \tag{4.6}$$

and Nd_{lk} is the number of days in cluster k on link l.

Figure 4.1 shows an example of Ward's clustering procedure. In Chapter 9 the sensitivity of the clustering results for the selected algorithm is discussed.

Sometimes, a pre-classification is executed, using calendar or site-specific data (Wild, 1994; Chung, 2003). The disadvantage of pre-classification is that some of the existing patterns may be disturbed by the pre-classification. On the other hand, pre-classification can help the clustering algorithm to form tighter groups by filtering out large differences. The effect of pre-classification was investigated in an analysis of highway flow patterns (Weijermars and Van Berkum, 2005b; Weijermars and Van Berkum, 2005a) and from this analysis was concluded that pre-classification into working days and weekend days improves the clustering result. Therefore, days are pre-classified into working days and non-working days. An extra index D is attributed to a pattern and cluster, representing the type of day, whereby $D = \{\text{working day, non-working day}\}$.

4.2 Analysis of temporal traffic patterns

This section proposes a framework for the analysis of the clustering results. Section 4.2.1 describes in what way the resulting clusters can be described. Section 4.2.2 deals with the determination of the factors that are on the basis of the resulting clusters and Section 4.2.3 discusses the analysis of the variation within the clusters.

4.2.1 Description of resultant clusters

The resultant clusters can be described by their average daily flow profiles. The average daily flow profile of cluster k, $\mathbf{P}_{lDk}$ is determined by averaging the traffic volumes q over all days in the cluster, i.e.

$$\mathbf{P}_{lDk} = (\bar{q}_{lDk,1}, ..., \bar{q}_{lDk,t}, ..., \bar{q}_{lDk,Nt}) \tag{4.7}$$

The average daily flow profiles of the resultant clusters can be compared to each other by plotting them into one graph. Differences between clusters can

Let us define the following traffic volumes (q_t) on four days d:

d	$t = 1$	$t = 2$	$t = 3$
1	500	550	500
2	400	500	450
3	400	500	500
4	400	450	400

Further, in this example a pattern is defined as a series of three traffic volume measurements on day d:

$$\mathbf{P}_d = (q_{d,1}, q_{d,2}, q_{d,3})$$

These days are grouped by means of a Ward's hierarchical clustering procedure. Every day starts as a separate cluster. In the first step of the clustering procedure, two days are combined into one cluster. Possible clusters (k) are:

$$\begin{array}{ll} C_1^1 = \{\text{day1, day2}\} & C_4^1 = \{\text{day2, day3}\} \\ C_2^1 = \{\text{day1, day3}\} & C_5^1 = \{\text{day2, day4}\} \\ C_3^1 = \{\text{day1, day4}\} & C_6^1 = \{\text{day3, day4}\} \end{array}$$

For all potential clusters, the sum of squared deviations from the mean of the clusters are calculated using

$$V_k = \sum_{d \in C_k} \sum_{t=1}^{3} (q_{dt} - \bar{q}_{kt})^2$$

$$\begin{array}{ll} V_1 = 2500 + 2500 + 625 + 625 + 625 + 625 = 7500 & V_4 = 625 \\ V_2 = 6250 & V_5 = 2500 \\ V_3 = 15000 & V_6 = 6250 \end{array}$$

The fusion of days 2 and 3 results in the smallest possible increase in the sum of squared deviations. Therefore, in the first step of the clustering procedure, days 2 and 3 are combined into a new cluster C_4^1.

For the second step of the clustering procedure, new potential clusters and corresponding sum of squared deviations and increase in sum of squared deviations are determined:

$$\begin{array}{lll} C_1^2 = \{\text{day1, } C_4^1\} & V_1 = 10000 & \Delta V = 10000 \text{ - } 625 = 9375 \\ C_2^2 = \{\text{day1, day4}\} & V_2 = 15000 & \Delta V = 15000 \text{ - } 0 = 15000 \\ C_3^2 = \{C_4^1\text{, day4}\} & V_3 = 6667 & \Delta V = 6667 \text{ - } 625 = 6042 \end{array}$$

The fusion of C_4^1 and day 4 results in the smallest increase in the sum of squared deviations. Therefore, these clusters are combined into a new cluster C_3^2. In the final step of the clustering procedure, day 1 is combined with cluster C_3^2.

Figure 4.1: Example of a Ward's clustering procedure.

be further investigated by comparing total daily traffic volumes, peak volumes, peak times and ratios between peak volumes and off-peak volumes.

Peak times and volumes are determined by means of the moving average. Thereby, the time of the peak period is defined as the hours before and after noon that show the highest traffic volume. Thus, for every time interval $t = a...Nt$ the traffic volume for the past hour (q^*) is determined using:

$$\bar{q}^*_{lDkt} = \bar{q}_{lDkt} + \bar{q}_{lDk,t-1} + ... + \bar{q}_{lDk,t-(a-1)} \tag{4.8}$$

where $a = \frac{Nt}{24}$.

The peak times (t_{AM} and t_{PM}) and peak flows (q_{AM} and q_{PM}) are subsequently determined by:

$$\bar{q}^{\text{AM}}_{lDk} = \max_{t \in \{a,...,\frac{Nt}{2}\}} \bar{q}^*_{lDkt} = \bar{q}^*_{T^{\text{AM}}_{lk}} \tag{4.9}$$

$$\bar{q}^{\text{PM}}_{lDk} = \max_{t \in \{\frac{Nt}{2}+1,...,Nt\}} \bar{q}^*_{lDkt} = \bar{q}^*_{T^{\text{PM}}_{lk}} \tag{4.10}$$

When calculating the ratios between peak and off-peak volumes, fixed peak intervals are used, namely 7:00 - 9:00 as the A.M. peak period and 16:00 - 18:00 as the P.M. peak period.

4.2.2 Determination of factors on the basis of the clusters

From the literature discussed in the previous chapter it was concluded that (1) weekday, (2) season, (3) holiday periods, and (4) weather are potential causes of systematic between day variations in traffic volumes. Descriptive statistics are applied to visualize relations between these potential causes of systematic variations and the resulting clusters and by means of statistical tests it is investigated which characteristics influence the clustering result statistically significant. If the resulting groups cannot be explained by one or more of these factors, the resulting clusters are analysed further to investigate whether other factors are responsible for the classification. Moreover, in case of small clusters with an abnormal daily flow profile it is investigated whether road works, events or accidents are on the basis of the cluster.

The day of the week, season and holiday periods are described by nominal variables. Cross-tabulations are used to investigate whether these variables are on the basis of the resulting clusters. Table 4.1 shows an example of such a cross-tabulation. A cross-tabulation shows the number of days in every cell, whereby each cell represents one combination of values on two variables (in this case cluster and day of the week) (Huizingh, 2002). By comparing the distribution of the days over the weekdays, the seasons and over non-holiday periods and holiday periods for the different clusters, it can be investigated whether these factors influence the classification. In the example,

Table 4.1: Example of a cross-tabulation.

Cluster	Monday	Tuesday	Wednesday	Thursday	Friday	Total
1	25	25	25	0	0	75
2	0	0	0	25	0	25
3	0	0	0	0	25	25
Total	25	25	25	25	25	125

the distribution of the days over the weekdays clearly differs between the clusters. Therefore, in this case, the weekday is stated to be responsible for the resulting classification.

To test whether the influence of these factors is statistically significant, a chi square test is applied. A chi-square test investigates whether the distribution of the days over for example the weekdays differs statistically significant between the clusters.

With regard to weather, in the previous chapter it was stated that in Dutch cities, rain causes a modal shift from bike to car. Further warm and sunny weather possibly causes a modal shift from car to bike and walking. Finally, also extreme weather (e.g. snow, windstorm) probably influences traffic. Warm and sunny weather is highly correlated with the season and is therefore not taken into account as a separate factor. Extreme weather only occasionally occurs in the Netherlands and is therefore only taken into account when explaining for a small cluster with an abnormal daily flow profile. The influence of other weather factors on traffic is expected to be small and other factors are therefore not taken into account.

The influence of rain is examined by comparing the number of dry and the number of wet days for different clusters. A dry day is defined as a day without rain. Since both the intensity (mm/hour) and the duration of the rain might influence modal split, both characteristics are taken into account when defining a wet day. The quantification of these criteria depends on the local weather situation and is a trade off between the amount of wet periods and the robustness of the analysis (Chung et al., 2005). In Chapter 5, the criteria are quantified for the Dutch situation. Finally, besides comparing the proportions of wet and dry days, also parts of the day can be examined separately. Since it is for example expected that especially rain during the A.M. peak period influences modal split, the number of days with and without rain during the A.M. peak period can be compared for different clusters. The available traffic and weather data determine potential aggregation levels at which the analysis can be executed. Moreover, weather data is only collected at a number of stations and rain varies in space as well so one should be careful when using weather data from one location to estimate the amount of rain at another location (Hagens, 2005). In that case, longer periods are more appropriate (Chung et al., 2005).

The number of wet days is relatively small and differs by season. To control for the influence of other variables, matched pair analysis is applied in this research. For a description of matched pair analysis see Andrey et al. (2003). Every wet day is linked to a dry day exactly one or two weeks prior to or after the wet day. For these pairs of days it is analysed whether the distribution differs between the clusters. Again, chi square tests are used to determine whether differences are statistically significant.

4.2.3 Variation within the clusters

When the variation within a cluster is large, the average daily flow profile of the cluster is not representative for all days within the cluster. Plots of the daily flow profiles of all days within a cluster provide a first insight into variations in daily flow profiles. However, in case of large clusters, these plots become less legible. Moreover, they do not provide information about the differences in daily flow profiles within a cluster in relation to the variation within the total dataset. Therefore, the variation within the clusters is further analysed by means of the standard deviation. Per time interval, per cluster, the standard deviation in traffic volume (σ) can be calculated by:

$$\sigma_{lDkt} = \sqrt{\frac{\sum\limits_{d \in C_{lk}} (q_{ldt} - \bar{q}_{lDkt})^2}{Nd_{lDk}}} \tag{4.11}$$

These standard deviations provide information about the thightness of the cluster. To be able to more easily compare clusters, also the total standard deviations per cluster are determined by:

$$\sigma_{lDk} = \sqrt{\frac{\sum\limits_{t=1}^{Nt} \sigma^2_{lDkt}}{Nt}} \tag{4.12}$$

Subsequently, the average standard deviation per link, after classification can be determined by the weighted average over the cluster

$$\sigma^{\text{after}}_{lD} = \sqrt{\frac{\sum\limits_{k \in C_l} Nd_{lDk} * \bar{\sigma}^2_{lDk}}{\sum_{k \in C_{lD}} Nd_{lDk}}} \tag{4.13}$$

This average standard deviation can be compared to the standard deviation before clustering. In that case, the standard deviation is calculated for the entire dataset:

$$\sigma^{\text{before}}_{lD} = \sqrt{\frac{\sum\limits_{d \in S_{lD}} \sum\limits_{t=1}^{nt} (q_{ldt} - \bar{q}_{lDt})^2}{Nd_{lD} * Nt}} \tag{4.14}$$

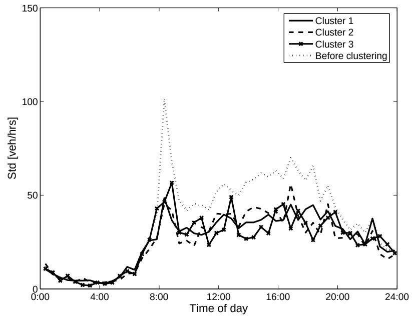


Figure 4.2: Standard deviations before and after clustering as a function of the time of the day.

where

$$\bar{q}_{lDt} = \frac{1}{Nd_{ld}} \sum_{d \in S_{lD}} q_{ldt} \tag{4.15}$$

Finally the ratio (F) between the standard deviations before and after clustering can be calculated:

$$F = \frac{\sigma_{lD}^{\text{before}}}{\sigma_{lD}^{\text{after}}} \tag{4.16}$$

Naturally, the standard deviation before the classification will always be larger than, or equal to the standard deviation after classification. The ratio F provides information about the usefulness of the classification; the larger F, the more useful the classification. The ratio does not provide information about the time of the day differences in standard deviations before and after classification occur. By showing the average standard deviations of the various clusters together with the standard deviation before classification as a function of the time of the day, more insight is obtained into variations in traffic volumes on different times of the day. Figure 4.2 shows an example of such a graph.

4.3 Analysis of spatial traffic patterns

Cluster analysis can also be used for the analysis of spatial variations in traffic patterns by grouping locations that show similar patterns. Subsequently, it can be investigated what (spatial) factors are on the basis of the resultant groups. The most appropriate definition of a traffic pattern depends on the type of temporal variation that is compared for different locations. First, it is interesting to compare locations with regard to their average daily flow profile; this analysis is described in Section 4.3.1. Secondly, from the previous chapter it is concluded that the main characteristics that cause between day variations in traffic volumes are weekday, season and weather. Sections 4.3.2 to 4.3.4 discuss the grouping of locations on the basis of the influence of these characteristics on traffic volumes. Section 4.3.5 describes a method for the classification of locations on the basis of the temporal patterns that resulted from the clustering procedures.

4.3.1 Variations in average daily flow profiles

Cluster analysis has been applied for the grouping of locations that show similar daily flow profiles (BGC, 1986; Hicks and Niemeier, 2001; Niemeier et al., 2002; Pribyl and Pribyl, 2005). However, since their goal of the grouping was another than obtain insight into spatial traffic patterns, none of the authors discussed what spatial factors are on the basis of the classification. Various representations for a daily profile are described in the literature. Hicks and Niemeier (2001) describe a daily flow profile by uncorrelated linear combinations of hourly traffic volumes, determined by a principal component analysis. BGC (1986) and Niemeier et al. (2002) use proportions of daily traffic and Pribyl and Pribyl (2005) used a series of LOS values that are determined by applying Fuzzy linguistic rules to flow and occupancy.

In this research, cluster analysis is applied to discern different types of locations. We are not interested in the distinction between locations with high traffic loads and locations with lower traffic loads, yet in differences in distribution of traffic over the day. Therefore proportions of total daily traffic volume are a more appropriate representation of a daily flow profile than actual traffic volumes. Also for this analysis a distinction is made between working days and non-working days. This results in the following definitions:

Pattern ($\mathbf{P}$): series of proportions of total daily traffic volume (s) for day type D on link l:

$$\mathbf{P}_{lD} = (s_{lD,1}, ..., s_{lD,t}, ..., s_{lD,Nt}) \tag{4.17}$$

where

$$s_{lDt} = \frac{\bar{q}_{lDt}}{\sum_{t=1}^{Nt} \bar{q}_{lDt}} \tag{4.18}$$

Cluster (C): group k of links that show a similar average daily flow profile.

$$C_{Dk} = \{l \mid \gamma(l, D) = k\} \tag{4.19}$$

where function γ is a function that assigns a Pattern $\mathbf{P}_{lD}$ to a cluster and $k = 1...Nk_D$ and Nk_D is the number of clusters for day type D.

The resulting clusters are analysed in a similar way as the temporal clusters discussed in Section 4.1. First of all, the average daily flow profiles of the resultant clusters can be determined by averaging the proportions over the links within the cluster, i.e.:

$$\mathbf{P}_{Dk} = (\bar{s}_{Dk,1}, ..., \bar{s}_{Dk,t}, ..., \bar{s}_{Dk,Nt}) \tag{4.20}$$

where

$$\bar{s}_{Dkt} = \frac{1}{Nl_k} \sum_{l \in C_{Dk}} s_{lDt} \tag{4.21}$$

and Nl_k is the number of links in cluster k.

Secondly, we want to obtain insight into spatial factors that are on the basis of the traffic patterns. Compared with temporal traffic patterns, less literature is available on factors that probably influence the classification. Moreover, only one classification has to be carried out for the entire network instead of a seperate classification for each location. Therefore, instead of determining the influence of potential factors, it is investigated more qualitatively what factors are on the basis of the classification. By means of a Geographical Information System (GIS) tool, it is visualized which locations are classified to each cluster and it is investigated what these locations have in common. Finally, the variation within the clusters is examined in the same way as for temporal traffic patterns, i.e. the average daily flow profiles of all links within one cluster are plotted in one graph and the standard deviations before and after classification are compared to each other and plotted as a function of the time of the day.

4.3.2 Variations in weekly patterns

BGC (1986) grouped locations on the basis of their working day factors that represent differences in daily traffic volumes between different working days. However, in urban areas some locations show relatively high traffic loads on weekend days too. Therefore the distribution of traffic over working days, Saturdays and Sundays is compared as well. Moreover, in Chapter 3 it was demonstrated that weekdays do not only differ with regard to the total daily traffic volume, but also with regard to the shape of the daily flow profile. Thus, both differences in total daily traffic volumes and differences in shapes of daily flow profiles are taken into account. To this end, for different weekdays, the daily flow profile is compared to the average daily flow profile. Since weekend days show dissimilar daily flow profiles compared to working days, weekend days are not taken into account when determining the average daily flow profile. The distribution of the traffic over working days, Saturdays and Sundays is analysed separately.

Thus, the classification of locations on the basis of their weekly traffic patterns consists of two parts. First, the distribution of the traffic over working days, Saturdays and Sundays is compared. Secondly, differences in working day patterns are analysed more extensively by comparing deviations from the average daily flow profile on Mondays to Fridays for different locations. Both classifications are elaborated in this section.

When comparing the distribution of traffic over working days, Saturdays and Sundays, a traffic pattern can be defined by factors that compare the traffic on Saturday and the traffic on Sunday to the traffic on an average working day, i.e.

$$\mathbf{P}_l^{\text{week1}} = \left(\pi_l^{Sat}, \pi_l^{Sun}\right) \tag{4.22}$$

where

$$\pi_l^{\text{Sat}} = \frac{\bar{q}_{l,\text{Sat}}}{\bar{q}_{l,\text{work}}} * 100 \tag{4.23}$$

and

$$\pi_l^{\text{Sun}} = \frac{\bar{q}_{l,\text{Sun}}}{\bar{q}_{l,\text{work}}} * 100 \tag{4.24}$$

It is determined what the differences in Saturday and Sunday factors are between the clusters, which locations are classified to the different clusters and how large the variation is within the clusters. Since in this case a pattern is defined by two factors instead of daily flow profiles, the variation within the clusters is analysed in a different way. For both the Saturday and the Sunday factor, the variation within the different clusters can be shown in a boxplot, that visualises the median, 25^{th} and 75^{th} percentiles, outliers and extreme values of the links within a cluster.

Regarding differences in variations between working days, both differences in height and differences in shape of the daily flow profile are taken into account. The daily flow profile of each working day is compared to the average working day flow profile, by calculating the absolute difference in traffic volume for each time interval. To be able to compare links, these absolute differences are normalized by dividing them by the average traffic volume for that link. This results in the following pattern definition:

Pattern (**P**): series of $5 * Nt$ differences that represent the daily flow profiles on different working days compared to the average working day flow profile:

$$\mathbf{P}_l^{\text{week2}} = (\delta_{l,Mon,1}, ..., \delta_{l,w,t}, ..., \delta_{l,Fri,Nt}) \tag{4.25}$$

where

$$\delta_{lwt} = \frac{\bar{q}_{lwt} - \bar{q}_{lwt}^{\text{work}*}}{\frac{1}{Nt}\sum_{t=1}^{Nt} \bar{q}_{lt}^{\text{work}*}} \tag{4.26}$$

and $w \in$ {Monday, Tuesday, Wednesday, Thursday, Friday}

The traffic volume on an average working day is determined using the average over Monday to Friday to give an equal weight to all weekdays:

$$\bar{q}_{lt}^{\text{work}^*} = \frac{\bar{q}_{l,Mon,t} + \bar{q}_{l,Tue,t} + \bar{q}_{l,Wed,t} + \bar{q}_{l,Thu,t} + \bar{q}_{l,Fri,t}}{5} \tag{4.27}$$

The resulting traffic patterns are analysed in the same way as described in Section 4.3.1. Per weekday, per cluster the average difference between the weekday flow profile and the average daily flow profile can be shown as a function of the time of the day. Moreover, using a Geographical Information System (GIS) it can be visualized what links are classified to each cluster. Finally, the variation within the clusters is analysed by plotting the difference profiles for all links within each cluster and by comparing the standard deviation before and after classification.

4.3.3 Variations in seasonal patterns

Grouping locations according to their seasonal factors has been subject of research in literature on AADT (Average Annual Daily Traffic) estimation on the basis of short term traffic counts (e.g. Bellamy, 1978; Sharma and Werner, 1981; Flaherty, 1993; Li et al., 2003). Permanent traffic counters are grouped on the basis of their monthly factors, i.e. for each month the average daily traffic volume is compared to the AADT. Since it is concluded from the previous chapter that also the shape of the daily flow profile might differ between seasons, also in this case both the height and the shape of the daily flow profile are taken into account. Analogue to the analysis of differences in weekly variations, the working day flow profile of each month is compared to the average working day flow profile. This results in the following pattern definition:

Pattern (**P**): series of $12 * Nt$ differences that represent the daily flow profiles on different months compared to the average working day flow profile:

$$\mathbf{P}_l^{\text{season}} = (\delta_{l,jan,1}, ..., \delta_{l,f,t}, ..., \delta_{l,dec,Nt}) \tag{4.28}$$

where

$$\delta_{lft} = \frac{\bar{q}_{lft} - \bar{q}_{lft}^{\text{work}**}}{\frac{1}{Nt}\sum_{t=1}^{Nt} \bar{q}_{lt}^{\text{work}**}} \tag{4.29}$$

and $f \in$ {jan, feb, march, april, may, june, july, aug, sept, oct, nov, dec}

Again, the traffic volume on an average working day is calculated by averaging the traffic volumes of different months:

$$\bar{q}_{lt}^{\text{work}^{**}} = \frac{\sum_{f=jan}^{dec} \bar{q}_{l,f,t}}{12} \tag{4.30}$$

The resulting traffic patterns are analysed in the same way as the working day patterns described in the previous section.

4.3.4 Variations in weather factors

As noted before, rain is probably the most important weather factor affecting urban traffic volumes. Moreover, its influence may differ by location. It is expected that locations with a relatively large proportion of short distance traffic are more influenced by rain than locations with a relatively small share of short distance traffic. Therefore, locations are also grouped according to the influence of rain on traffic volumes. Traffic volumes are compared for dry and for wet periods using matched pair analysis (see Section 4.2). For the definition of wet and dry periods, the same considerations as discussed in Section 4.2 play a role. According to Hagens (2005), the influence of rain on the modal split is larger in summer than in winter, so when enough data is available, a distinction could be made according to season or month of the year. A pattern thus is defined as a series of ratios of traffic volumes during wet periods and corresponding dry periods. The number of ratios depends on the selected aggregation level and whether or not a distinction is made between different seasons or months. Suppose traffic volumes between wet and dry periods are compared for every measurement interval t and that a distinction is made between different months f. In that case a traffic pattern ($\mathbf{P}^{\text{weather}}$) is defined as:

$$\mathbf{P}_l^{\text{weather}} = (\pi_{l,jan,1}, ..., \pi_{l,f,t}, ..., \pi_{l,dec,Nt}) \tag{4.31}$$

where

$$\pi_{lft} = \frac{\bar{q}_{lft}^{\text{wet}}}{\bar{q}_{lft}^{\text{dry}}} * 100 \tag{4.32}$$

$\bar{q}_{lft}^{\text{wet}}$ and $\bar{q}_{lft}^{\text{dry}}$ are the traffic volumes on wet and corresponding dry t's.

The resulting patterns are analysed in the same way as the weekday vs weekend day patterns described in Section 4.3.2.

4.3.5 Variations in temporal classifications

Spatial variations in temporal traffic patterns can also be analysed by comparing clustering results for locations. First, it can be analysed to what extent the resulting clusters at different locations correspond. In that case, the sets of days in the resulting clusters are compared for different locations. Thus:

Pattern (**P**): series of Nd_l clusters to which dates are assigned to:

$$\mathbf{P}_l^c = (\phi_{l,1}, ..., \phi_{l,d}, ..., \phi_{l,Nd_l}) \tag{4.33}$$

These patterns have to be adjusted in order to be appropriate as input for a clustering procedure. The features are not measured on a regular interval scale but on a binary scale. For example, when day 1 is classified to cluster 1 at location A and to cluster 3 at location B, the distance between the clusters is not equal to 2, yet the day is classified to different clusters. Therefore, distances between patterns have to be calculated in a different way. The distance (Δ) between two classifications is defined as the proportion of days that are classified to different clusters:

$$\Delta\left(Sc_1, Sc_2\right) = \frac{1}{Nd_{l1 \cap l2}} \sum_{d \in Sd_{l1 \cap l2}} \Delta_{d,l1,l2} \tag{4.34}$$

where

$$\Delta_{d,l1,l2} = \begin{cases} 1 & \text{when } \phi(l_1, d) \neq \phi(l_2, d) \\ 0 & \text{otherwise} \end{cases} \tag{4.35}$$

and $Sd_{l1 \cap l2}$ is the set of days for which traffic data is collected at both link 1 and link 2 and $Nd_{l1 \cap l2}$ is the number of days for which traffic data is collected both at link 1 and link 2.

Distances are calculated for all combinations of locations. Beforehand, clusters have to be reordered in such a way that $\mid C_{l1,k} \cap C_{l2,k} \mid$ is maximized for all k. Moreover, in case of a dissimilar number of clusters at two locations, the clusters at the location with the highest number of clusters have to be combined into new clusters in such a way that $\mid C_{l1,k} \cap C_{l2,k} \mid$ is maximized for all k. The calculation of the distance between two locations is demonstrated in Figure 4.3. The distances between all pairs of locations are recorded in a matrix that is used as input for a clustering procedure. Ward's hierarchical clustering is subsequently used for the grouping of locations. It can be determined what types of days can be distinguished at different types of locations and what spatial factors are responsible for the cluster a location is classified to. It has to be noted that although the resulting groups of days are similar for locations within a cluster, the typical daily flow profile might differ. Thus, the set of days in a cluster is similar, whilst the daily flow profile is not necessarily.

Suppose traffic data on 50 days is collected at locations 1 and 2. Further, suppose that the clustering procedure resulted in $Sc_1 = \{C_{1,1}, C_{1,2}, C_{1,3}\}$ and $Sc_2 = \{C_{2,1}, C_{2,2}\}$ and that:

$$\begin{array}{lcl@{\qquad}lcl} C_{1,1} & = & \{1,...,20\} & C_{2,1} & = & \{36,...,50\} \\ C_{1,2} & = & \{21,...,40\} & C_{2,2} & = & \{1,...,35\} \\ C_{1,3} & = & \{41,...,50\} & & & \end{array}$$

Clearly, $Nk_1 > Nk_2$, thus at location 1, 2 clusters have to be combined into one larger cluster so that $\mid C_{1,1} \cap C_{2,1} \mid$ and $\mid C_{1,2} \cap C_{2,2} \mid$ are maximized. In this case, the combination of $C_{1,1}$ and $C_{1,2}$ yields the best solution. Moreover C_2 has to be reordered in order to maximize $\mid C_{1,1} \cap C_{2,1} \mid$ and $\mid C_{1,2} \cap C_{2,2} \mid$. Thus, for the calculation of the distance between locations 1 and 2, the following clusters are defined:

$$\begin{array}{lcl@{\qquad}lcl} C'_{1,1} & = & \{1,...,40\} & C'_{2,1} & = & \{1,...,35\} \\ C'_{1,2} & = & \{41,...,50\} & C'_{2,2} & = & \{36,...,50\} \end{array}$$

Days 36 to 40 are classified to different clusters at the two locations so

$$\Delta\left(Sc_1, Sc_2\right) = \frac{5}{50} = 0.10$$

Figure 4.3: Example of the calculation of the distance in classification between two locations.

The method described here compares the locations on the basis of all resulting clusters. It does not provide insight into the occurrence of specific traffic patterns at different types of locations. However, it might also be interesting to obtain more insight into the occurrence of specific patterns. Therefore, it is also useful to investigate at what locations specific patterns occur. The specific patterns that are analysed follow from the analysis of the results of the temporal classification.

4.4 Traffic patterns on a network level

Cluster analyses for all links in the network may result in many different patterns. In order to obtain more insight into recurrent traffic patterns in an entire city, it is useful to determine traffic patterns on a network level. Two approaches can be distinguished for the determination of these network patterns. First, the results of the temporal and spatial analysis can be combined. Secondly, a new definition of a traffic pattern can be formulated that describes a pattern on network level. These approaches are discussed in this section.

Section 4.3.5 described the classification of locations on the basis of the temporal patterns they show. The classification results in different types of temporal traffic patterns. Subsequently, a date can be classified to the different temporal traffic patterns on the basis of its characteristics. Possible combinations of temporal traffic patterns determine the patterns on a network level. Figure 4.4 demonstrates this approach.

Patterns on a network level can also be determined straight away, by executing a cluster analysis on network level. In that case, a pattern consists of a series of daily flow profiles on multiple links on one day. When all links in a network are included in the analysis, this approach would lead to a pattern representation that consists of many features. This could be problematic for the clustering algorithm. Moreover, traffic flow patterns at different links along one arterial are expected to be correlated. In that case the number of links along one arterial determines the weight of that arterial in the clustering result. Therefore, instead of defining a pattern by daily flow profiles on all links, it is preferable to define a pattern by daily flow profiles on some key locations that are selected on the basis of their function in the network. An analysis of the major OD-relations served by different links in the network helps in selecting these key locations. This results in the following definitions:

Pattern ($\mathbf{P}$): daily flow profiles at a set of selected links Sl^* on day d :

$$\mathbf{P}_d = \left(q_{1,1,1}, ..., q_{ldt}, ..., q_{Nl_{Sl^*},d,Nt}\right) \tag{4.36}$$

where $l \in \{Sl^*\}$, Sl^* is a set of links that are selected as key locations and Nl_{Sl^*} is the number of links that is selected as key location.

Suppose the classification of locations results in two types of locations:

Type 1: This type of location shows three clusters:

- Cluster 1: contains Mondays, Tuesdays and Wednesdays and shows a relatively peaked daily flow profile at all locations
- Cluster 2: contains Thursdays and shows relatively high traffic volumes in the evening at all locations
- Cluster 3: contains Fridays and shows a relatively flat daily flow profile at all locations

Type 2: This type of locations shows four clusters:

- Cluster 1: contains road works and shows low traffic volumes on some locations and high traffic volumes on other locations
- Cluster 2: contains Thursdays and shows relatively high traffic volumes in the evening at all locations
- Cluster 3: contains days in Spring (except from Thursdays) and shows high traffic volumes at all locations
- Cluster 4: contains the remaining days

On a network level, this results in 7 patterns:

1. Mondays, Tuesdays and Wednesdays on which road works occurred. These days show a relatively peaked daily flow profile at type 1 locations, relatively low traffic volumes at some type 2 locations and relatively high traffic volumes at other type 2 locations
2. Mondays, Tuesdays and Wednesdays in spring. These days show a relatively peaked daily flow profile at type 1 locations and relatively high traffic volumes at type 2 locations
3. Mondays, Tuesdays and Wednesdays in other seasons without road works. These days show relatively peaked daily flow profiles at type 1 locations and an average daily flow profile at type 2 locations
4. Thursdays; they show relatively high traffic volumes in the evening at all locations
5. Fridays on which road works occurred. These days show a relatively flat daily flow profile at type 1 locations, relatively low traffic volumes at some type 2 locations and relatively high traffic volumes at other type 2 locations
6. Fridays in spring. These days show a relatively flat daily flow profile at type 1 locations and relatively high traffic volumes at type 2 locations
7. Fridays in other seasons without road works. These days show relatively flat daily flow profiles at type 1 locations and an average daily flow profile at type 2 locations

Figure 4.4: Example of the determination of patterns at network level.

Cluster (C): group k of days that show similar patterns on set of links Sl^*.

$$C_{Dk} = \{d \in Sd_{Sl^*D} \mid \kappa(d) = k\} \tag{4.37}$$

where Sd_{Sl^*D} is the set of days for which valid traffic data is available for all time intervals on all links in Sl^* and function κ is a function that assigns a Pattern $\mathbf{P}_d$ to a cluster

4.5 Summary

This chapter proposes methods for the analysis of urban traffic patterns by means of cluster analysis. Temporal traffic patterns can be analysed by grouping days that show similar daily flow profiles. Spatial traffic patterns can be analysed by grouping links that show similar temporal variations. Links are grouped with regard to their average daily flow profile, their weekly and seasonal variations, the influence of weather on their traffic volumes, and with regard to the results of their temporal clusterings. Finally, two approaches for the analysis of urban traffic patterns on a network level are proposed. The first combines the results of the temporal and spatial analyses, whilst the second defines a pattern by series of daily flow profiles at multiple locations.

A Ward's hierarchical clustering algorithm is proposed for all classifications. For all proposed analyses, appropriate pattern definitions are formulated and it is discussed in what way the results of the cluster analyses can be analysed to obtain more insight into urban traffic patterns. The average traffic patterns of the resulting clusters provide information on typical urban traffic patterns. Moreover, the characteristics of the days or links that are classified to a certain cluster provide information on the temporal and spatial factors that are on the basis of temporal or spatial variations in urban traffic volumes. Finally, the variation within the clusters is examined to analyse to what extent the cluster represents all days or links within the cluster.

Chapter 5

Almelo: Data

The traffic data for this research is provided by the ViaContent system in Almelo. Almelo is a medium sized city (approximately 70.000 inhabitants) in the east of The Netherlands. The next chapter discusses the traffic network and major attractions of Almelo and the resulting traffic streams. This chapter deals with the available data. The first section describes the traffic information system ViaContent. The second section discusses the data that is available and the third section deals with the data processing. The fourth section provides an overview of the available traffic data after processing and the chapter ends with conclusions.

5.1 ViaContent

ViaContent (http://www.viacontent.nl/) is a traffic information system. Figure 5.1 illustrates the functioning of ViaContent. Single loop detectors at signalized intersections are the main source of information for the urban network. When available, also traffic data from other sources (e.g. Floating Car Data, data on the occupancy of parking facilities or data on the location of Public Transport vehicles) can be stored and processed by ViaContent. The data are processed using Visum online, a simulation model that models and predicts the current and future situation. Visum online determines the most likely traffic situation on the basis of an approximate (historical) OD-matrix and real-time traffic data using the Path Flow Estimator (Bell et al., 1997). This process is called data completion and it generates estimated traffic volumes for locations for which no measured traffic data is available. The working of VISUM online is discussed in more detail in Fellendorf et al. (2000).

At the time of the research, only data collected by single loop detectors was available. Moreover, only raw traffic data are used for this research. This means that volumes estimated by traffic models are not used.

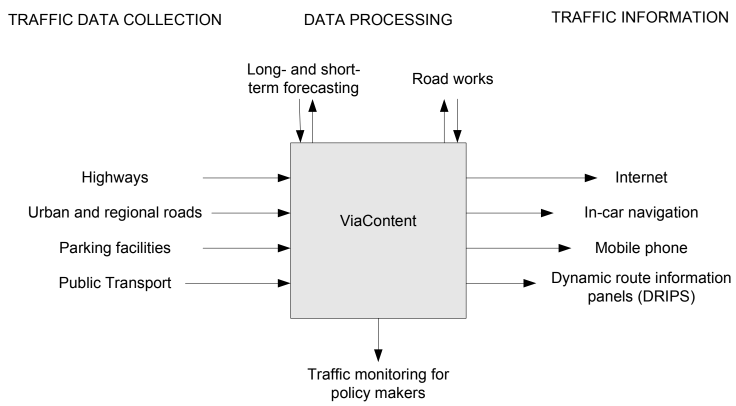


Figure 5.1: ViaContent system (adapted from www.viacontent.nl).

5.2 Data

This section discusses the data that is available in Almelo. The structure is comparable to the structure of Section 2.2. First, the available traffic data in Almelo is discussed and subsequently it is discussed which data is available on the factors that potentially influence traffic volumes.

5.2.1 Traffic data

As was mentioned in Section 5.1, all available traffic data is produced by inductive loop detectors at signalized intersections. Figure 5.2 shows at which intersections traffic data is collected. The data used for this research is collected in the period September 2004 - September 2005.

The number of vehicles counted by the detectors is aggregated into 5 minute, 15 minute or 30 minute traffic volume measurements, depending on the type of controller. The data is provided in ASCII files. Every record contains one traffic volume measurement for a certain detector for a certain time period on a certain day. A record contains the following variables: date, intersection, detector, type of detector, start time of measurement interval, length of measurement interval, state of the detector (reflecting the result of the microscopic control procedure), and measured traffic volume.

In general, each lane contains one or more detectors. In Almelo, four types of detectors are distinguished. Type 1 detectors are short detectors that are selected as monitoring detector in ViaContent. In most cases, the monitoring detectors are located just upstream of the stop line. Type 4 detectors are short detectors that are not used as monitoring detector by ViaContent but only for traffic control purposes (anticipate on green and/or detection of queues that exceed maximum length). Type 2 and type 3 detectors are long detectors

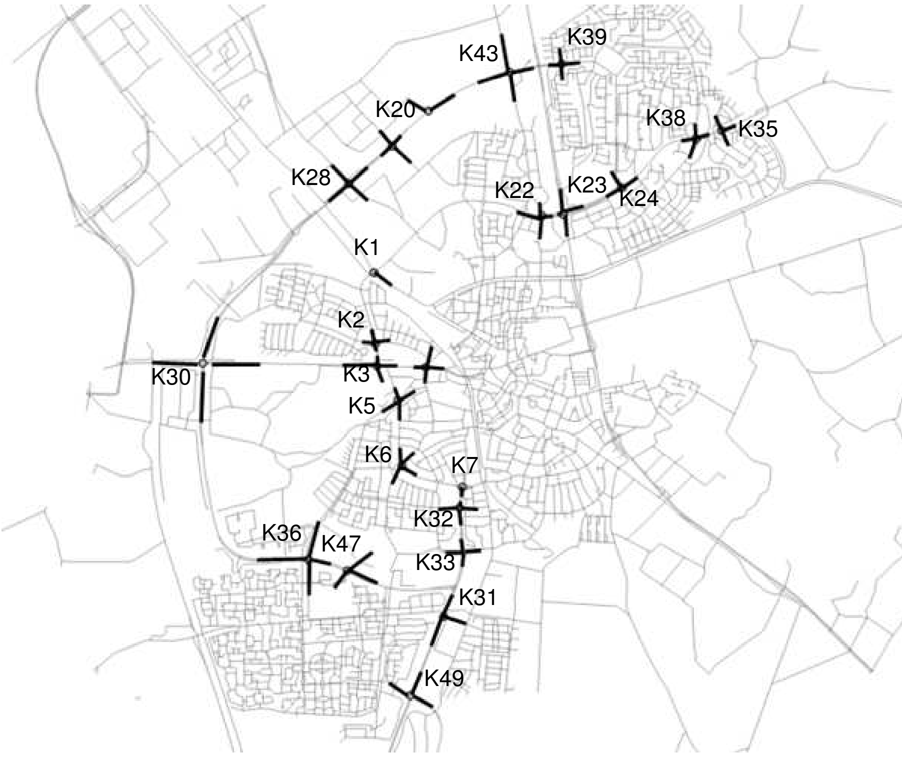


Figure 5.2: Intersections at which traffic data is collected.

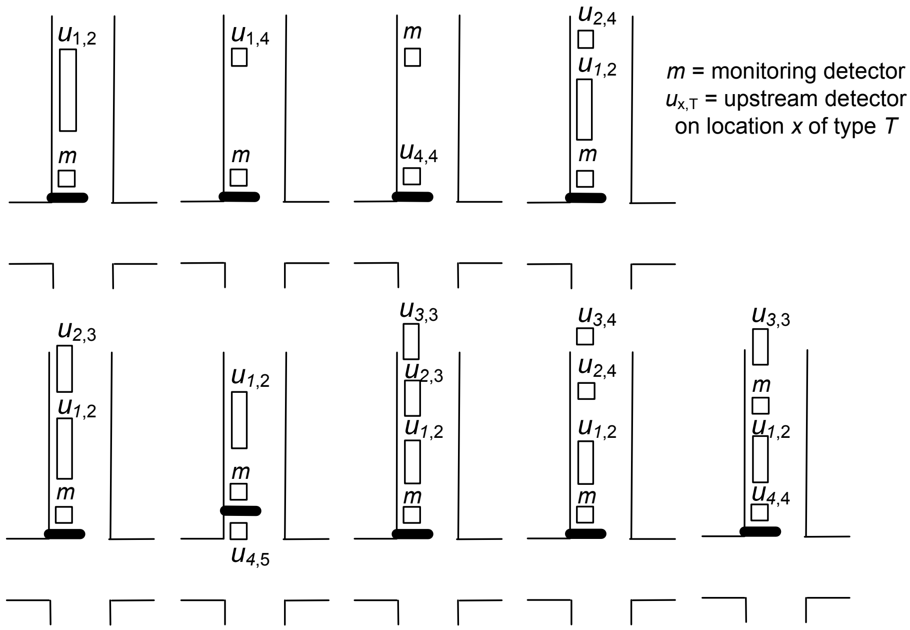


Figure 5.3: Detector configurations that occur in Almelo.

that are located upstream of the first short detector and that are used for the detection of queues. The exact configuration differs by intersection and by lane. The detector configurations that occur in Almelo are shown in Figure 5.3. Lanes that form one link do not always show the same detector configuration. Moreover, in some cases, one detector covers multiple lanes. Since the detectors that are selected as monitoring detectors (type 1 detectors) are expected to produce the most reliable and accurate traffic data, only these detectors are used for the analysis of urban traffic patterns. Other types of detectors are only used in the data validation procedure, for the quality check that applies the principle of conservation of vehicles.

5.2.2 Data on factors potentially influencing traffic

As stated in Section 2.2, factors that potentially cause variations in measured traffic volumes are: (1) type and time of day, (2) weather, (3) events, (4) road works and (5) accidents. Here it is shortly described what data is available on these factors in Almelo. As was already mentioned in Chapter 2, that information on the time and type of day (calendar data) can easily be linked to measured traffic volumes.

Regarding weather data, the main weather station nearest to Almelo is Twente airport which is located approximately 20 kilometres from Almelo. Via the

KNMI, weather data from this station is available on a daily level. Moreover, from http://www.weeronline.nl/ data on a six hour aggregation level from this weather station can be downloaded. It has to be noted that the weather at the weather station might differ from the weather in Almelo. As a result, a period that is classified as wet can in reality be dry and the other way around. The probability of a wrong classification depends on the definition of a wet period and the aggregation level at which the data is analysed. Hagens (2005) discusses potential misclassifications in more detail. The requirements for a period or day to be wet are adapted from Hagens (2005), a period is labeled wet in case of a rain duration of > 2 hours with an average intensity of > 0.5 mm/hour. Table 5.1 provides an estimation of the probability of different types of misclassifications.

Table 5.1: Probabilities of misclassifications of wet and dry days. These values result from a comparison between the weather situations at two weather stations that are located 43 km apart.

	Classification	
Actual situation	wet	dry
wet		1.0 %
dry	1.8%	

Data on events and road works is provided by the municipality of Almelo. The data consists of a list of upcoming events and road works that are published on the internet and in a local newspaper. However, there is no central database in which the time, locations and impacts of events and road works are stored. Therefore, this data cannot be automatically linked to the traffic data. Accident data is provided by the Police Department of Twente. The data consist of an Excel file with the date, time, location and severity of accident (only material damage, injury, or casualty). Unfortunately, the location of the accident cannot be automatically linked to the locations of the detectors.

5.3 Data processing

The traffic data provided by the ViaContent system is processed in order to be useable for the analysis of traffic patterns. First of all, some records show deficiencies and are pre-processed. This pre-processing is discussed in Section 5.3.1. The pre-processed data are restructured in such a way that the new records contain all traffic volume measurements for one detector on one day. These records of traffic volume measurements are input for the data validation procedure that is discussed in Section 5.3.2. This data validation procedure is evaluated in Section 5.3.3. Finally, the valid records of traffic volume data are aggregated to link level and the traffic data is linked to weather data and data on the type of day using the date as a key variable.

5.3.1 Pre-processing of individual data records

Some records show a deficient start time or a deficient period length. A deficient start time implies that a measurement period is shifted in time; traffic is for example collected from 12:01 to 12:06 instead of 12:00 to 12:05. A deficient period length implies that traffic is not detected for 5, 15 or 30 minutes but for another time period. In general these deficient period lengths and time stamps are due to restarting of a controller after it was down. A controller is for example restarted at 14:07:25 and traffic data is aggregated into 5 minute flows. The first time stamp is in that case 14:07:25 and the first period is 155 seconds.

Only very little data appears to show deficient period length (< 0.01% of the records) or deficient time stamps (2.3% of the records). However, in some cases deficient data is structural, i.e. some of the intersections show high percentages of deficient records. Table 5.2 shows the intersections for which 1% or more of records shows deficient start times or lengths of measurement period. Intersections 5, 39 and 43 show very large percentages of data with a deficient start time. Further analysis shows that start times of measurement intervals vary in time at these intersections. These shifted measurement periods are probably due to synchronization of the clocks of controllers of an older type. Since the records with deficient start times are unevenly distributed over the intersections, relatively a lot of information would be lost when these records would be removed from further analysis. It would mean that traffic data of three out of twenty-two intersections could not be analysed. Therefore, records with deficient start times are processed and included in further analysis. Time shifts are removed by interpolating the available traffic counts (see Figure 5.4). Most records with deficient period lengths occur at intersection K4. On the basis of further analysis is concluded that most of these records show a period length of 2 hours (82.5%) or a negative length (16.9%), the latter representing errors in detection. Since there are relatively few records with a deficient period length, processing the data separately does not lead to a clearly better data set. Therefore it is not worthwhile to process these data separately and the data are removed from further analysis.

Table 5.2: Intersections with deficient start times and/or period lengths.

Intersection	Deficient start time	Deficient period length
K4	6%	6%
K5	95%	0%
K39	100%	0%
K43	100%	0%

Suppose that the following traffic data is collected:

12:05 - 12:20	30 vehicles
12:20 - 12:35	45 vehicles

If the traffic within a time period is assumed to be evenly distributed in time, time stamps can be assigned to the traffic counts. Next, the amount of traffic between 12:15 and 12:30 can be calculated by means of interpolating the available counts:

period	Time stamp	number of vehicles
12:05 - 12:20	12:12:30	30 vehicles
12:15 - 12:30	**12:22:30**	**40 vehicles**
12:20 - 12:35	12:27:30	45 vehicles

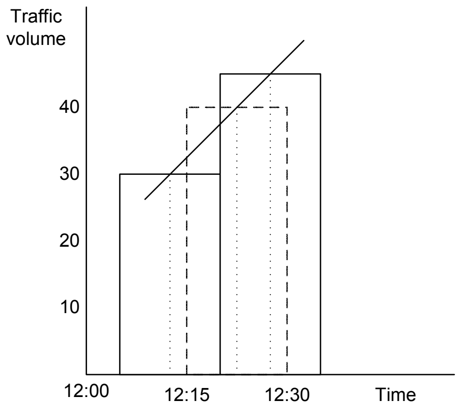


Figure 5.4: Processing of records with a deficient start time.

5.3.2 Data validation

Chapter 2 discussed the design of a data control procedure for urban traffic data. In this section, this data control procedure is adjusted to the specific situation of Almelo. As mentioned in Chapter 2, the data control procedure is executed on daily records of traffic volume measurements. A record is removed when it does not pass one or more of the quality checks (see equations 2.2 and 2.3).

The first quality check deals with the removal of records that show missing data. The second quality check checks for failures detected by the detector itself. The third quality check deals with the maximum volume threshold and the fourth deals with the minimum volume thresholds. The fifth and sixth quality checks apply the principle of conservation of vehicles, respectively between monitoring detectors and upstream detectors and between two sets of monitoring detectors. Appendix A provides an overview of the data control procedure for Almelo.

This section discusses some of the quality checks in more detail.

Microscopic quality checks

Loop detector stations themselves execute some quality checks. These microscopic quality checks result in a value for a variable called state. Four state reports are distinguished: (1) state = 0 when a detector is functioning adequately; (2) state = 2 when the current in the loop is interrupted which means that the loop is broken; (3) state = 4 when the occupancy of a detector is 100% for a certain time interval and (4) state = 8 when the occupancy is 0% for a certain time interval. In case of occupancy of 100%, the length of the time interval is fixed and set on 10 minutes for most links and 30 minutes for very congested links. In case of zero occupancy, the length of the time period is based on the occupancy during the previous period. This data on the state of the detector is only available for traffic data in 2005. Moreover, only controllers of a new type are capable of reporting state.

A record (containing all traffic volume measurements for one detector for one day) is removed from further analysis when one of the microscopic quality checks is not passed (i.e. when state $\neq 0$) for one or more of the time intervals.

Basic macroscopic quality checks

As discussed in Chapter 2, the basic macroscopic quality checks use minimum and maximum volume thresholds to detect erroneous data records. A daily record of traffic volume measurements is removed when (1) one or more of the measurements exceed a maximum threshold, (2) the record contains one or more negative traffic volumes, (3) the total daily traffic volume equals zero or (4) traffic volume equals zero for one or more hours during day time and (a) upstream detectors measure traffic volumes higher than a certain threshold or (b) the daily flow profile shows abnormalities. These quality checks contained some parameters that had to be calibrated to the specific situation.

Regarding the maximum volume threshold, a threshold for suspiciously high volumes had to be chosen. This threshold should not be too low, because in that case, too many data records have to be analysed manually. On the other hand it should not be too high, since erroneous records may remain undetected in that case. The value was set on 1000 vehicles per hour and after an explorative analysis it was concluded that this threshold is reasonable.

Regarding the quality check concerning zero hourly traffic volumes, a threshold had to be chosen for the maximum number of vehicles detected by upstream detectors in case of zero hourly traffic volumes at a monitoring detector. This threshold equals the maximum number of vehicles queuing between two detectors and could be determined on the basis of the distance between the detectors and the jam density. Since the distance between detectors varies by

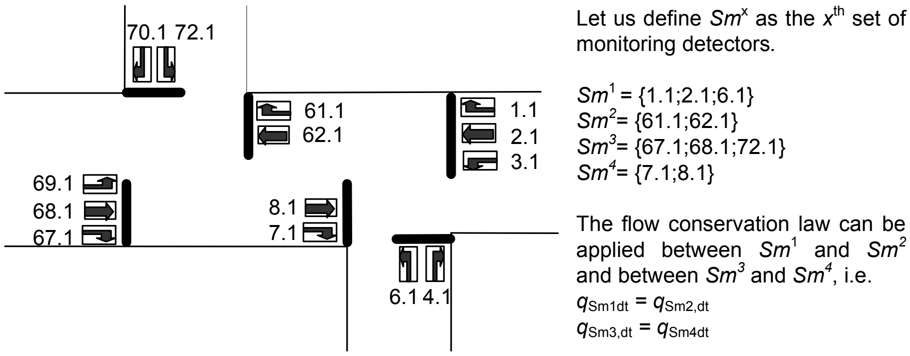


Figure 5.5: Principle of conservation of vehicles between two sets of monitoring detectors.

location, a separate threshold should be calculated for each pair of detectors. For reasons of simplicity, one threshold is adapted for all locations. Also this threshold is selected on the basis of an explorative analysis and is set on 20 vehicles.

Quality checks based on the principle of conservation of vehicles

As mentioned in Chapter 2, in general it is difficult to apply the principle of conservation of vehicles in urban areas, since traffic is generated and/or leaking away at unsignalized intersections, parking areas, etc. In Almelo, there are some pairs of signalized intersections between which there are no unsignalized intersections. However, between these intersections other facilities like a gasoline station or small companies are located. Therefore, it is decided that the principle of conservation of vehicles is not applied between two intersections. However, there are two intersections that consist of two parts. On these intersections, the principle of conservation of vehicles can be applied between two sets of monitoring detectors. The principle of flow conservation between two sets of monitoring detectors is shown in Figure 5.5.

The principle of conservation of vehicles can also be applied between monitoring detectors and upstream detectors. As discussed in Section 5.2, there is no standard detector configuration for all locations. Figure 5.6 shows the general principle of conservation of vehicles between monitoring and upstream detectors.

The general formula for the quality check based on the principle of conservation of vehicles contained two thresholds; one for the maximum allowable number of vehicles queuing between two detectors and one for the maximum allowable percentage difference in traffic volume measurements between two detectors. A threshold for the maximum number of vehicles queuing between two detectors was already determined for the quality check based on the minimum volume

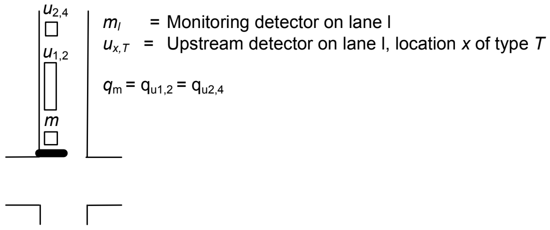


Figure 5.6: Principle of conservation of vehicles between monitoring and upstream detectors.

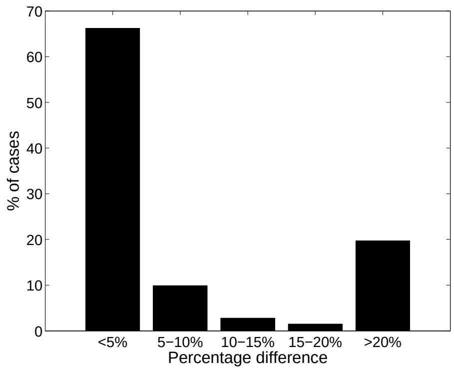


Figure 5.7: Percentage difference in daily traffic volumes between two sets of monitoring detectors.

thresholds and was set on 20 vehicles.

As already stated in Chapter 2, for this research, the maximum acceptable percentage difference is determined on the basis of regular differences in volumes between detectors. The expected inaccuracy should however not be larger than 10%. Different thresholds are determined for both quality checks (i.e. principle of conservation of vehicles between two sets of monitoring detectors and principle of conservation of vehicles between monitoring detectors and upstream detectors).

When traffic volume measurements are compared for two sets of monitoring detectors, in most cases, the difference is smaller than 5% (see Figure 5.7). When the threshold is set on 10%, only a small amount of additional data would pass the check. Therefore, for the quality check based on the principle of flow conservation between two sets of monitoring locations, 5% is selected as the threshold for the maximum allowable percentage difference.

When traffic volume measurements are compared between monitoring detectors and upstream detectors, the differences in traffic counts vary between different types of detectors (see Figure 5.8). From the relatively large differences in reported volumes between monitoring detectors and type 2 detectors, it can be

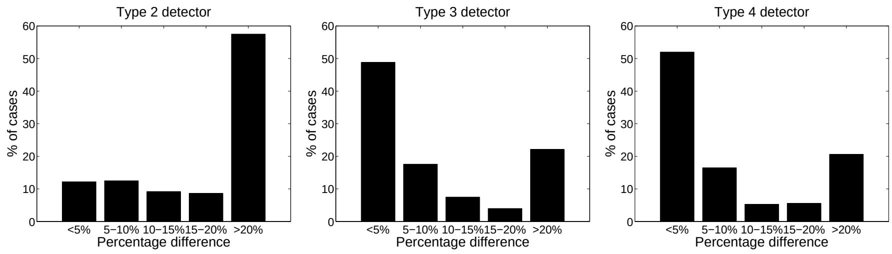


Figure 5.8: Percentage difference in daily traffic volumes between monitoring detectors and different types of upstream detectors.

concluded that type 2 detectors produce less accurate counts than other types of upstream detectors. Moreover, short detectors (monitoring detectors and type 4 detectors) produce more accurate counts than long detectors. This was expected as the main task of long detectors is the detection of queues instead of vehicles. Finally, in general type 4 detectors will produce less accurate counts than short detectors that are selected for traffic monitoring. Thus, in case of a difference between a monitoring detector and one upstream detector it is most likely that the upstream detector is producing inaccurate counts.

These findings have some implications for the quality check. First of all, volumes measured at monitoring detectors are the reference value, i.e. the difference in relation to the volume measured at the monitoring location is determined instead of the difference in relation to the average volume. Moreover, in order to use this quality check, the monitoring detector has to meet the following requirements:

- Two or more upstream detectors are available
- At least one of the upstream detectors is of type 3 or type 4
- All upstream detectors report either higher or lower traffic volumes

In Chapter 2, 10% is mentioned to be a common maximum allowable inaccuracy. Therefore, the threshold T_3 is set to 10%[1] for the quality check that compares traffic volumes reported by monitoring detectors with volumes reported by upstream detectors.

These thresholds and the resulting quality checks can be found as Q_{5a} and Q_{5b} in Appendix A, that gives an overview of the final quality control procedure that is executed on the traffic data from Almelo.

[1]When it holds that the probability of overestimation equals the probability of underestimation, a difference of 10% implies that on average the inaccuracy of both detectors is about 10%. In some cases, both detectors will underestimate or overestimate the true traffic volume and the true inaccuracy of the monitoring detector is larger than 10%. In other cases, one of the detectors underestimates the true traffic volume whereas the other detector overestimates the true count and then the true inaccuracy is smaller than 10%.

5.3.3 Evaluation of data control procedure

Ideally, the control procedure would detect all measurements that do not represent actual traffic volumes. However, invalid data might be undetected by the quality checks. On the other hand, data might be flagged to be invalid, whilst in reality the data is of sufficient quality. Since it is unknown whether traffic volume measurements are valid or not, it is not possible to determine the frequency of occurrence of both types of errors. Yet, some analyses were executed to get an impression of the functioning of the data validation procedure. This section discusses these analyses. First, the individual quality checks are evaluated to investigate whether records that did not pass the quality checks were rightfully removed. Secondly, the results of the different quality checks are compared to each other and finally, by means of visual inspection of the remaining data it is investigated to what extent records unrightfully pass the data validation procedure. The analyses are executed on part of the total dataset, namely data from the period September 2004 - May 2005.

Individual quality checks

By a visual inspection of the removed records, it is investigated whether the data records that failed the maximum flow threshold were rightfully removed. All records that contained volume measurements exceeding the maximum threshold of 3000 veh/h showed abnormal daily flow profiles, i.e. volumes were alternately very high and very low or were very high for a lot of consecutive time intervals. All these records thus were rightfully removed. Moreover, only about one third of the cases that showed suspiciously high traffic volume measurements appeared to contain erroneous data. Therefore, it is recommended to further investigate suspiciously high traffic volumes instead of immediately removing them to limit the amount of unrightfully removed data.

Regarding the zero daily and hourly traffic volumes, in some cases, vehicles were detected during the night, whilst no vehicles were detected during the day. Therefore, the check for zero daily traffic volumes is executed on a record that contains the traffic volume measurement of the day period (8:00 - 19:00). Moreover, hourly traffic volume measurements of zero vehicles are often due to the absence of traffic (11% of the flagged records appeared to contain erroneous data), for example on Saturday and Sunday mornings. Therefore, it is important to further analyse these cases before removing the data.

Quality checks based on the principle of conservation of vehicles can only be applied on about one third of the records. Moreover, for the quality checks that compare traffic volume measurements of monitoring detectors with volume measurements from upstream detectors, it is not always clear whether the outcome is reliable. In many cases, traffic volume measurements are not consistent between upstream detectors mutually. First of all, for about three quarters of the cases that failed the quality check, also upstream detectors mutually show inconsistencies in traffic volume measurements (difference in volume is >10%). These cases are analysed further. In about 2% of these cases, the difference between the upstream detectors mutually was around 10%, whereas the differences between the monitoring detector and the upstream detectors were much larger. These cases are assumed to be rightfully removed from further analysis. At other locations, the difference between the upstream detectors mutually is in the same order of magnitude or larger than the differences between the monitoring detector and the upstream detectors About two third of these cases occurred on K30 or K49. At these locations, field observations were carried out [2]. From these field observations it is concluded that the monitoring detectors function adequately (the differences between manual counts and counts reported by ViaContent being between 1% and 4%). Therefore, data from these locations is not removed from further analysis (at least in case that all basic quality checks are passed). Data from locations were no field observations were carried out is removed from further analysis, yet possibly wrongfully.

Moreover, in about 6% of the cases that the principle of flow conservation is applied, one of the upstream detectors reported higher volumes, whilst the other upstream detector reported lower volumes than the monitoring detector. These cases are not removed from further analysis (see the requirements in the previous section). It is however unclear whether the monitoring detector is functioning adequately in these cases. Most of these cases occurred on K2, K22 and K35. K2 is reconstructed in the meanwhile, as a result of which field observations were not possible. On K22 and K35, field observations were carried out. On K22, the difference between the manual count and the count reported by the monitoring detector was only 1%. The cases are thus rightfully included in the final database. However, on K35, the difference is larger (21%). Therefore, data from this location is removed from further analysis. The data from locations at which no field observations were carried out are included in the database, yet possibly unrightfully.

Finally, sensitivity analyses are executed on the thresholds regarding the maximum number of vehicles queuing between two detectors and the maximum percentage difference in volume. Regarding the threshold on the maximum number of vehicles queuing between two detectors, the percentage of records that failed this quality check increased from 22% to 27% when the threshold was set to 10 vehicles and decreased to 20% when the threshold was set to 30 vehicles. Regarding the threshold on the maximum percentage difference between two sets of monitoring detectors, the percentage of failures decreased from 40% to 34% when the difference was set to 10%. Regarding the maximum percentage difference between monitoring detectors and upstream detectors, the percentage of failures increased from 21% to 23% when the difference was

[2]Other locations were less suited for field observations since the mentioned problem only occurred on a limited amount of days. In that case, the probability that the problem occurs on the day field observations were carried out is too small.

Table 5.3: Cross tabulation results microscopic quality checks by macroscopic quality checks.

	Passed micro checks	Failed micro checks
Passed macro checks	21090 records (91%)	114 records (< 1%)
Failed macro checks	1752 records (8%)	257 records (1%)

set to 9% and decreased to 19% when the difference was set to 11%.

Comparison of results different quality checks

The macroscopic quality checks were compared to the microscopic quality checks executed by the detector station itself. Table 5.3 shows a comparison of the results of these quality checks.

The microscopic quality checks did not report all invalid data. In 8% of the cases, invalid data was reported by the macroscopic quality checks, whilst the detector station itself did not report invalid data. On the other hand, in less than 1% invalid data was detected by the detector station itself whilst it was not detected by the macroscopic quality checks. Besides, these cases mainly concerned records for which the detector was reported to be malfunctioning for one 5 minute interval or for a couple of intervals during the night or evening. The macroscopic quality checks thus are a useful addition to the quality checks executed by the detector station itself.

Besides, the results of the basic quality checks are compared to the quality checks based on the principle of conservation of vehicles. In almost all cases where traffic volumes were inconsistent between two sets of monitoring detectors, one of the detectors did not pass the basic quality checks. In these cases, only traffic measurements of the malfunctioning detector are removed from further analysis. On the other hand, most cases that did not pass the quality check based on the principle of conservation of vehicles between monitoring detectors and upstream detectors, were not reported to contain erroneous data by the basic macroscopic quality checks. The quality checks based on the principle of conservation of vehicles thus are a useful addition for the detection of inaccurate data. However, as mentioned in the previous section, some of the data might be unrightfully removed by the quality check based on the principle of conservation of vehicles between monitoring detectors and upstream detectors.

Visual inspection of remaining data

As mentioned in the introduction of this section, it is unknown whether actual traffic data is valid or not. However, data records that show implausible daily flow profiles (e.g. alternately high and low traffic volumes) can be assumed to contain invalid data. Thus, the amount of records that passed all quality checks and show an implausible daily flow profile provides an estimation of the percentage of records that unrightfully passed the data validation procedure. Note that this only concerns records that show implausible flow profiles. Also records that show plausible flow profiles could contain invalid data. However, in this case it is not possible to detect these records.

Since a visual inspection of all 31.750 records would be too laborious, For all monitoring detectors, 10% of the records were randomly selected for visual inspection. Only 5 out of 3125 investigated records show suspicious daily flow profiles. Four of these records show traffic volumes of zero vehicles for 45 or 55 minutes, whereas the fifth shows very high traffic volumes for some 30 minute periods in the night. These abnormal flow profiles are probably due to a malfunctioning detector. The traffic volumes of zero vehicles all occurred at one detector. From a visual inspection of all data collected at this detector is concluded that this detector was probably malfunctioning on about one third of the days (26 records). The percentage of records that show implausible daily flow profiles and passed data validation can however be concluded to be very small.

5.4 Available traffic data after processing

The percentage of the data that failed one of the quality checks is a measure for the quality of the traffic data. This section discusses the quality of the traffic data and describes the data that is available after processing. Note that this only concerns data produced by monitoring (type 1) detectors.

5.4.1 Data quality

Missing data is in most cases due to a communication failure between the detector station[3] and the central computer. Since there is one detector station per intersection, data is either present of missing for an entire intersection. In some cases, data is missing for part of the day whereas in other cases data is missing for an entire day or for successive days. The amount of missing data is high. Traffic data is collected at 226 monitoring detectors from September 2004 to September 2005. When data would be collected on all days at all locations, this would lead to 395 days * 226 locations = 89.270 records, whereas in reality, only 52.740 records are available. This means that entire records of daily volume measurements are missing in 41% of the cases. Moreover, for 19% of these 52.740 records, traffic data is missing for one or more measurement intervals. It has to be noted that part of the missing data is due to reconstructions (intersection K2 is reconstructed into a roundabout without detection) and temporary road closures. Moreover, as a large part of

[3]A detector station collects and processes that data from a number of detectors and sends the data to the central computer.

the missing data is due to start-up problems and a relocation of the server, the amount of missing data will probably be much smaller in the future.

The amount of invalid data is smaller than the amount of missing data. Detector stations executed microscopic quality checks on about two third of the records. In about 2% of these cases a monitoring detector failed one of these tests. Also the basic macroscopic quality checks are passed by most of the records. Only 1% of the records fail the maximum threshold test and only 2% failed the minimum threshold tests. The principle of conservation of vehicles could be applied in about one third of the cases. The principle of conservation of vehicles between two sets of monitoring detectors is not met in 38% out of 791 cases. As mentioned in section 5.3.3.2 in most of these cases (96%), one of the detectors did not pass the basic macroscopic quality checks. The flow conservation law between monitoring detectors and upstream detectors is not met in 16% out of 11.461 cases.

5.4.2 Available data

After removal of the incomplete and invalid data records, 38.927 data records are left. For the analysis of urban traffic patterns, data is aggregated to link level, i.e. data is aggregated over all lanes entering an intersection from one direction. A record of volume measurements for link l is only included in the analysis in case that valid traffic data is available for all detectors that are part of the link. Table 5.4 shows for each link the number of days for which valid traffic data is collected during all day.

Besides the amount of data per link, also the distribution of the data over weekdays and seasons is important. A skew distribution may distort average traffic volumes and existent traffic patterns. From Figure 5.9 can be seen that the distribution of the available traffic data is uneven over the seasons. In general, less traffic data is collected during summer and autumn, yet also the differences between links mutually are relatively large, considering the distance between the minimum and maximum percentage per season. Chapter 9 discusses the consequences of this uneven distribution for the outcome of the analysis. The distribution over the days of the week is more even.

Table 5.4: Available traffic data per link NB = Northbound; EB = Eastbound; SB = Southbound; WB = Westbound. Intersections K20 and K47 consist of five approaches.

	Working day (N of days)					Saturday (N of days)					Sunday (N of days)				
	NB	EB	SB	WB		NB	EB	SB	WB		NB	EB	SB	WB	
K1				122					26		1			26	
K2	100	75	83	96		19	14	15	19		27	21	23	26	
K3	103	104	102	104		21	20	21	21		24	24	24	24	
K4	100	101	91	99		12	13	10	13		21	22	20	22	
K5	186	186	186	101		32	32	32	18		39	39	39	22	
K6	147	147	147	147		32	32	31	32		35	35	35	35	
K7	171					36					35				
K20	138	150	154	145	145	29	12	32	31	29	30	8	34	31	28
K22	118	118	118	119		23	23	24	23		26	25	25	25	
K23	84	106	103	104		5	19	18	20		1	22	22	19	
K24		65	64	64			13	14	13			14	13	15	
K28	133	132	129	108		31	32	30	30		34	34	33	30	
K30	122	123	122	122		29	29	29	29		31	31	31	31	
K31	70		84	56		13		19	17		12		27	28	
K32	159	130	161	146		30	29	31	29		32	29	33	30	
K33		18	6	17			3	1	4			3	1	4	
K35	152		152	137		30		30	27		30		31	25	
K36	117	68	118	117		22	22	22	22		30	31	31	31	
K38	67	70	62	69		13	13	11	13		13	12	11	13	
K39	200	182	79	200		34	32	13	34		2741	36	10	41	
K43	200	200	200	200		33	33	33	33		42	42	42	42	
K47	114		59	82	115	18	10	23	23	23	18	21	23	23	23
K49		133	132	133			24	24	24			26	26	26	

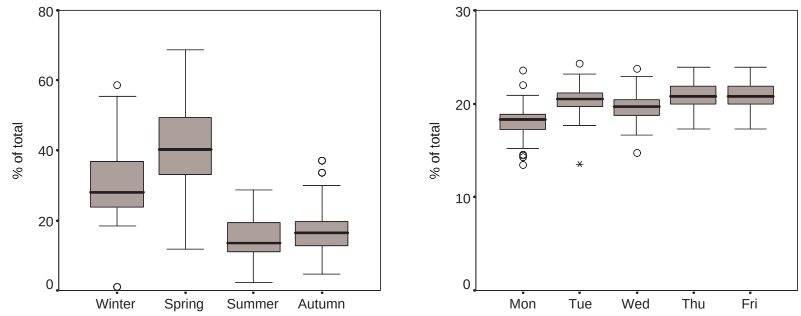


Figure 5.9: Distribution of the data over the seasons (left) and weekdays (right). Only links for which traffic data are available for more than 50 working days and working day data are selected. The winter is defined as the months December, January and February etc.

Finally, also the accuracy of the loop detectors is relevant. Since the accuracy of a detector depends on the installation and the tuning of a loop, the exact accuracy differs by location. Considering the amount of detectors, it would be too labour-intensive to investigate the accuracy of all detectors. However, the executed field observations provide an impression of the general level of accuracy. The field observations were executed during the P.M. peak on a Tuesday in July (12-07 2005) outside the school holiday period under sunny and warm weather conditions. Traffic volumes were counted for 12 successive 5-minute periods, resulting in a total measurement period of one hour. The clocks of the field observed have been synchronized with the atomic clock to be able to compare the manual counts with the ViaContent counts. Table 5.5 shows a comparison between the total hourly traffic volumes. The difference between traffic volume measurements reported by ViaContent and number of vehicles counted by manual counts is very small for most locations. These traffic volume measurements of ViaContent are thus very accurate. These results agree with the accuracy Deckers (2001) found for the loop detectors in Rotterdam. Vialis executed field observations on five other locations and found somewhat lower accuracy levels (inaccuracy of approximately 8%). At one of the locations ViaContent reported higher traffic volumes as a result of lateral influence of freight traffic on adjacent links. At the other locations traffic volumes were underestimated by ViaContent due to a less sensitive tuning of the detector and short following distances as a result of which two successive vehicles are detected as one. Moreover, these field observations took only 20 minute and the traffic volumes were lower (50-75 vehicles) as a result of which relatively small absolute differences lead to large percentage differences. As mentioned before, because the accuracy depends on the installation and tuning of the detector, field observations should be carried out at each individual detector to obtain insight into the accuracy of all individual detectors.

Table 5.5: Comparison between manual counts and Viacontent data.

Location	Manual [veh/hour]	ViaContent [veh/hour]	Difference [%]
1	236	238	1%
2	356	249	-2%
3	261	251	-4%
4	289	282	-2%
5	722	710	-2%
6	149	148	-1%

5.5 Conclusions

The data for the analysis of urban traffic patterns is provided by the traffic information system ViaContent. Traffic volumes are collected by inductive loop detectors at 23 signalized intersections throughout the city of Almelo. Weather data is available from the weather station at Twente airport which is located approximately 20 kilometres from Almelo and can be obtained from the KNMI or from http://www.weeronline.nl/. Data on road works and events is provided by the municipality of Almelo and accident data is provided by the police department of Twente. Unfortunately, the locations of road works, events and accidents cannot be linked directly to the locations of the loop detectors, because they are not stored in a central database.

After pre-processing, the traffic data is restructured in such a way that each record contains the measured traffic volumes for one detector for one day. These records are input for the quality control procedure that consists of six checks: (1) missing data, (2) microscopic tests computed by the detector station itself, (3) maximum flow threshold, (4) minimum flow thresholds (5) conservation of vehicles between a monitoring detector and upstream detectors, and (6) conservation of vehicles between two sets of monitoring detectors. Traffic data is used for the calibration of the thresholds for Almelo.

The quality control procedure is evaluated and we conclude that it functions adequately. Since it is unknown whether traffic volume measurements are truly valid or not, it is not possible to determine what percentages of the records (1) unrightfully passed the data validation procedure and (2) were unrightfully removed from further analysis. The frequency of occurrence of the first type of error is investigated by means of a visual inspection of part of the records that passed data validation. From this inspection it is concluded that the percentage of records that pass data validation and show implausible daily flow profiles is very small. However, it has to be noted that we have no information on the percentage of records that showed a plausible daily flow profile yet contained invalid data and unrightfully passed the data validation procedure. Since the quality checks based on the principle of conservation of vehicles – that could only be applied on about one third of the records – detect more invalid data than the basic quality checks, it is expected that some of the records unrightfully passed the data validation procedure. With regard to the second type of error it is concluded that all records that are removed by the basic macroscopic quality checks (3 and 4) show implausible daily flow profiles and thus are rightfully removed. However, from field observations is concluded that some of the records that were removed by the 5^{th} check (principle of conservation of vehicles between monitoring detector and upstream detectors) were unrightfully removed. Therefore, it is recommended to use this check with caution. In case differences are large between upstream detectors mutually as well, field observations may be carried out to check whether the volumes reported by the monitoring detector are truly inaccurate.

After validation, the traffic data is aggregated to link level. Despite a significant amount of missing data, for most links valid traffic data is available for a considerable number of days (on average 118 working days). Unfortunately, the traffic data is unevenly distributed over the seasons. The distribution varies by link, yet in general less traffic data is collected during summer and autumn months. The distribution of the data over the days of the week is quite even. The data seems very accurate, at least for the locations at which field observations were carried out. At these six locations, the difference between manual counts and reported traffic volumes by ViaContent was less than 5%. However, since the level of accuracy depends on the installation and tuning of the loop detectors, accuracy levels at other locations might be different.

Chapter 6

Almelo: Network and traffic demand

This chapter describes the traffic situation in Almelo. The first section deals with the network structure and the major attractions. The second section discusses the main traffic streams. It is investigated which main roads show the highest traffic loads and what trip patterns can explain these traffic loads. The third section describes the average daily flow profiles for working days, Saturdays and Sundays. The chapter concludes with a summary.

6.1 Network structure and major attractions

Figure 6.1 shows the main arterials and major attractions in the city of Almelo. The main roads roughly form two semicircles that are completed by the Van Rechteren Limpurgsingel. The outer ring is formed by the Weezebeeksingel, the Schuilenburgsingel and the Bleskolksingel. The ring is not complete yet, but will be completed in the future by the Nijreesingel that will connect the Weezebeeksingel to the Bornsestraat. The inner ring is formed by the Violierstraat, Schoolstraat, Nachtegaalstraat, Egbert ten Catelaan, Aalderinkssingel and Kolthofsingel. Besides these ring roads, there are some other main roads: The Henriette Roland Holstlaan connects Almelo to the highway (A35); the Wierdensestraat is the main road towards the city centre and the Sluitersveldsingel forms the main route for the traffic to the residential areas in the east of Almelo and for traffic in the direction of Ootmarssum and Denekamp. Besides local traffic and traffic to and from Almelo, some of the roads also serve a considerable amount of through traffic. The Henriette-Roland Holstlaan and Weezebeeksingel are part of the N35 that connects cities like Nijverdal and Zwolle to the A35. The Schuilenburgsingel, Bleskolksingel and a section of the Van Rechteren Limpurgsingel are part of the

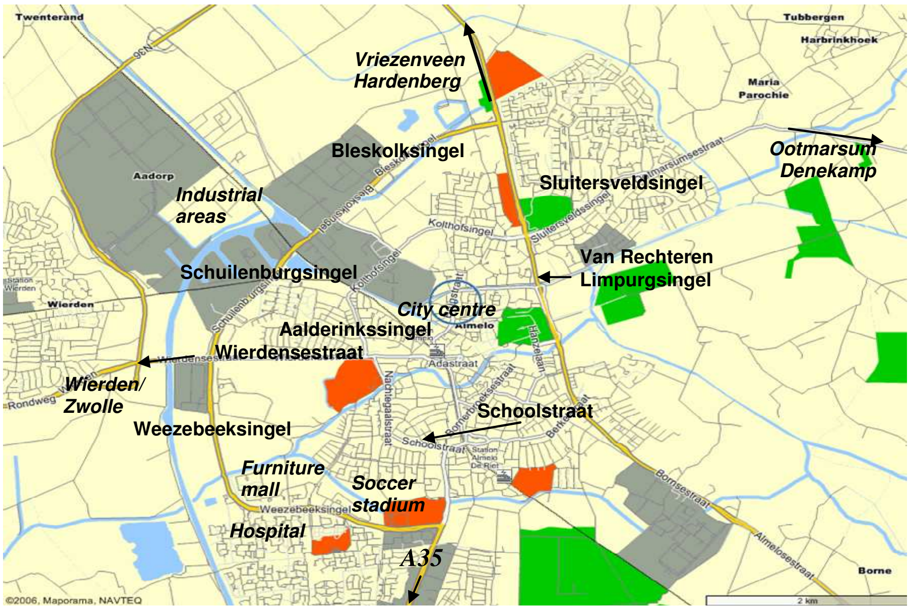


Figure 6.1: Main roads and attractions in Almelo (Source: http://www.maporama.com/share/).

N349 that is used by traffic between the A35 (connected by the Weezebeeksingel and the Henriette Roland Holstlaan) and Zwolle/Nijverdal on one hand and Vriezenveen/Hardenberg on the other hand. Finally, the Bornsestraat is part of the N743 that connects Almelo to Borne and Hengelo. Figure 6.2 shows the position of Almelo in the region as well as the main regional roads.

6.2 Main traffic streams

By aggregating traffic volumes over all lanes entering an intersection from one direction, average daily traffic volumes are determined for all links entering intersections. When all turns have separate lanes, average daily traffic volumes are also calculated for links exiting an intersection. Figures 6.3 and 6.4 show the daily traffic volumes on an average working day, an average Saturday and an average Sunday. The thickness of the line represents the amount of traffic. For clearer presentation, only traffic loads of the links entering the intersection are depicted. All lines thus represent traffic loads for traffic in one direction.

Traffic loads appear to be highest on the Henriette Roland Holstlaan and the Weezebeeksingel. These high traffic loads are mainly caused by through traffic between the highway (A35) on the one hand and the N35 and N349 on the other hand (also see Figure 6.5). Also the Sluitersveldsingel, the

Figure 6.2: Position of Almelo in the region (Source: http://www.routenet.nl/map.asp#).

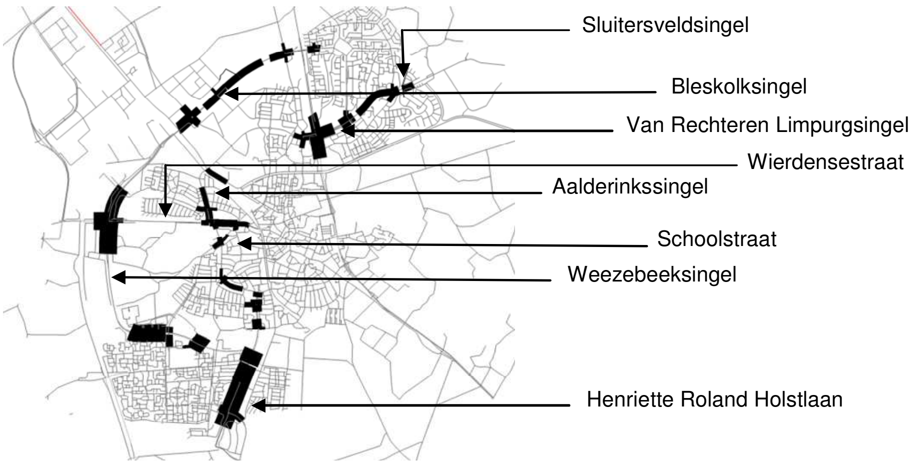


Figure 6.3: Traffic volumes on an average working day.

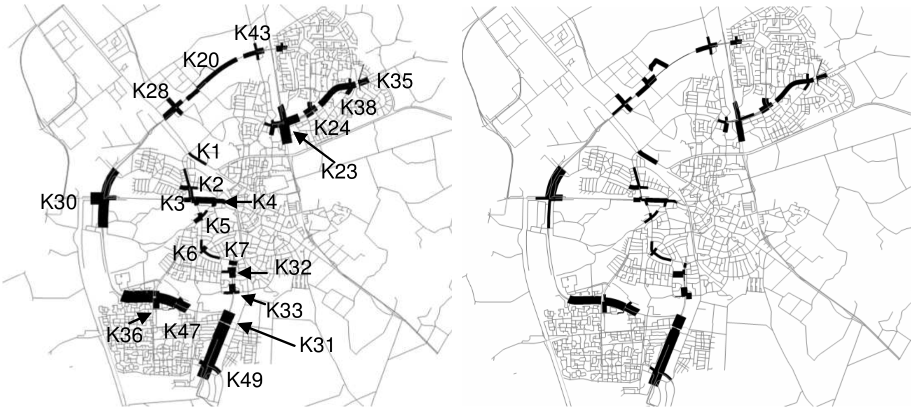
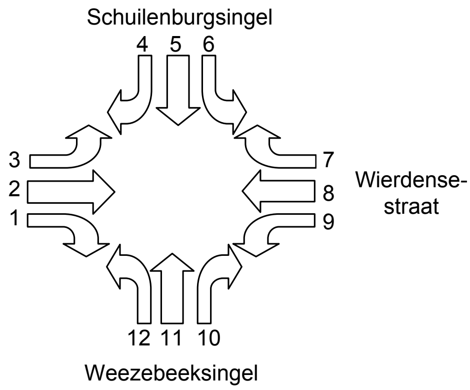


Figure 6.4: Traffic volumes on an average Saturday (left) and Sunday (right).



	Work	Sat	Sun
1	8946	73%	57%
2	3639	80%	54%
3	4202	74%	56%
4	4633	71%	52%
5	6398	69%	45%
6	559	74%	36%
7	566	84%	38%
8	3336	80%	54%
9	700	75%	55%
10	470	63%	44%
11	6632	67%	43%
12	7921	76%	59%

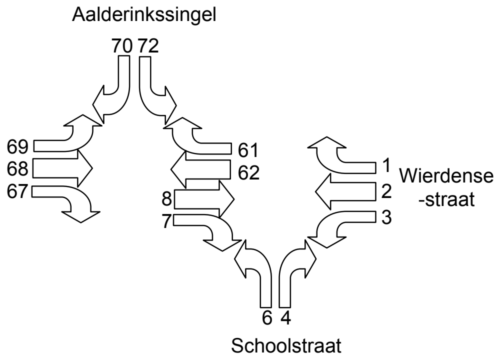
Figure 6.5: Traffic volumes at K30. Sat and Sun refer to percentages of working day traffic volume on an average Saturday and Sunday.

Table 6.1: Traffic volumes on main arterials. All weekday averages are based on at least 58 days. Saturday and Sunday values show percentages of weekday traffic volume. In case averages are based on less than 20 days, the number of days is shown in brackets.

Arterial	location	Weekday [veh/day]	Saturday	Sunday
H. Roland Holstlaan	K49 from K31	15905	71%	57%
	K32 from K7	9767	81%	63%
Weezebeeksingel	K30 to K36	16041	71%	52%
Wierdensestraat	K30 to K3	4668	78%	51%
	K3 to K4	3151	78% (18)	44%
Schoolstraat	K5 from K6	3310	77%	60%
	K3 from K5	4351	82% (18)	59%
Aalderinkssingel	K3 to K2	5160	78% (18)	56%
	K2 from K1	3137	67% (15)	48%
Sluitersveldsingel	K23 to K24	10214	84% (14)	62% (14)
	K38 from K24	6450	85% (13)	67% (12)
Schuilenburgsingel	K30 to K28	11395	70%	48%
/ Bleskolksingel	K28 from K30	9472	70%	52%
	K28 to K20	6115	74%	57%
	K43 to K44	6784	77%	58%

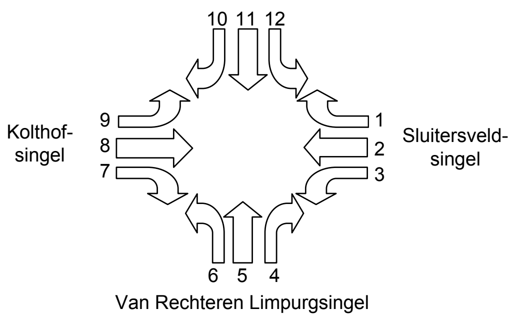
Van Rechteren Limpurgsingel and the Schuilenburgsingel/ Bleskolksingel show higher traffic loads than the roads near the city centre. The other roads show lower traffic loads but are also interesting to analyse, since they probably show other traffic flow patterns because they serve other traffic (more local and less regional traffic). These roads are the Aalderinkssingel, Schoolstraat and the Wierdensestraat.

Table 6.1 shows average traffic volumes on some key locations along the main roads. Traffic volumes at other locations along the roads as well as traffic volumes in opposite direction are comparable to the volumes shown in the table. From this table it can be concluded that the traffic loads vary along some of the main routes. At the Henriette Roland Holstlaan, traffic volumes are clearly higher between the highway and Weezebeeksingel than between the Weezebeeksingel and the city centre. This demonstrates the function of the first part of the H. Roland Holstlaan as a regional route. On the Wierdensestraat, traffic volumes are clearly lower between K3 and K4. Besides, on the Aalderinkssingel and the Schoolstraat, traffic volumes are highest around K3. From this, and from Figure 6.6 it is concluded that most traffic on these roads travels between the Schoolstraat (residential areas) and Aalderinkssingel (residential and industrial areas) and between the Schoolstraat/ Aalderinkssingel and the part of the Wierdensestraat between K30 and K3 that connects the residential areas and industrial areas to the regional routes. It has to be noted that no data is available for traffic



det	work	Sat	Sun
1			
2	2209	80%	43%
3	209	77%	48%
4	252	75%	49%
6	4099	82%	60%
7	4002	81%	59%
8	2899	78%	43%
61	3803	82%	57%
62	3112	82%	48%
67	1002	79%	59%
68	2342	81%	44%
69	1358	68%	55%
70	1521	72%	58%
72	3553	80%	56%

Figure 6.6: Traffic streams at K3.



	Work	Sat	Sun
1	2035	77%	62%
2	3607	83%	59%
3	4649	84%	65%
4	5081	87%	66%
5	5011	73%	59%
6	1715	67%	54%
7	1707	67%	51%
8	3216	82%	56%
9	1062	90%	58%
10	1073	88%	54%
11	4934	71%	57%
12	1918	77%	61%

Figure 6.7: Traffic streams at K23.

entering K7. Probably, traffic loads are relatively high at this location as well, caused by traffic leaving the residential areas in the direction of the highway. At the Sluitersveldsingel, traffic volumes decrease along the route, indicating that the route is mainly used by residential traffic between the Van Rechteren Limpurgsingel and Kolthofsingel on the one hand and residential areas along the Sluitersveldsingel on the other hand. Figure 6.7 shows that the major part of the traffic from the Sluitersveldsingel turns left on the Van Rechteren Limpurgsingel. Also through traffic on the Van Rechteren Limpurgsingel is high. At the Schuilenburgsingel/ Bleskolksingel, traffic volumes vary considerably along the route. Traffic volumes decrease between K30 and K28 and at K28, whilst at K43 traffic volumes are a little higher than at K28 and K20. This indicates that relatively a lot of traffic travels between K30 and the industrial areas between K30 and K28 and between the industrial areas and the Van Rechteren Limpurgsingel.

From Table 6.1 and Figure 6.4 can also be seen that traffic volumes are clearly lower on Saturdays and Sundays. As expected, the difference in traffic loads between working days and Saturdays is smallest for routes between residential areas and the city centre. Moreover, roads that mainly serve traffic to and from residential areas (e.g. the Sluitersveldsingel) show relatively high traffic volumes on both weekend days.

6.3 Daily flow profiles

The shape of a daily flow profile provides information about the within day variation in travel demand. In this section, daily traffic profiles are determined for an average weekday (excluding Holidays), an average Saturday and an average Sunday. Here, only the general daily flow profiles on network level are discussed. In Appendix B daily flow profiles are compared for different routes.

The determination of one general daily flow profile for the entire network in Almelo can be carried out in two ways:

1. On the basis of the total amount of traffic in the entire network (sum up traffic loads of all links and determine the daily flow profile for these loads)
2. On the basis of the traffic flow profiles of individual links (determine the average daily traffic profile for each link and calculate the mean)

Since only for two days traffic data was available for all intersections, the second option is chosen in this case. The method consists of the following steps:

$$s_{tdI} = \frac{q_{tdI}}{\sum_t q_{tdl}} \tag{6.1}$$

$$\bar{s}_{tI} = \frac{\sum_{d=1}^{Nd} q_{tdl}}{Nd} \tag{6.2}$$

$$\bar{s}_t = \frac{\sum_{l=1}^{Nl} q_{tl}}{Nl} \tag{6.3}$$

The average daily flow profiles are determined on the basis of 15-minute traffic volumes and are shown in Figure 6.8. Peak times and peak volumes are calculated using moving averages. A distinction is made between two hour peak periods, peak hours and peak intervals (15-minute interval). The peak times and relative peak volumes are shown in Tables 6.2 and 6.3. The average daily flow profiles agree with daily flow profiles found in literature (e.g. Chrobok et al., 2004). An average working day shows a daily flow profile with both an A.M. peak and a somewhat higher and broader P.M. peak. The P.M. peak volume is clearly higher than the A.M. peak volume and the P.M. off-peak volume is clearly higher than the A.M. off-peak volume. Moreover, the P.M. peak period is broader than the A.M. peak period. These peak factors are

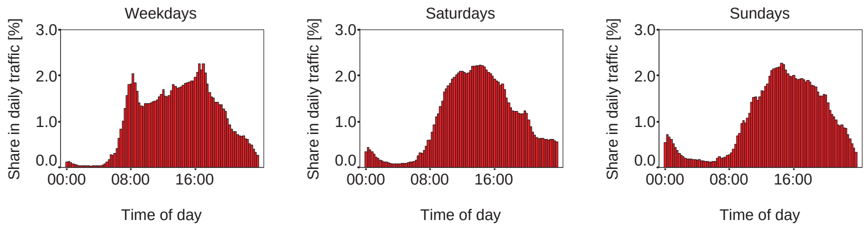


Figure 6.8: Average daily flow profile on working days, Saturdays and Sundays.

Table 6.2: Peak times.

	Working day		Saturday	Sunday
	A.M. peak	P.M. peak		
Two hour	7:30 - 9:30	15:30 - 17:30	13:15 - 15:15	13:15 - 15:15
Hour	7:45 - 8:45	16:15 - 17:15	13:45 - 14:45	14:00 - 15:00
15-minute interval	8:15 - 8:30	16:30 - 16:45	14:15 - 14:30	14:30 - 14:45

Table 6.3: Relative peak volumes.

	Working day		Saturday	Sunday
	A.M. peak	P.M. peak		
Two hour	13.5%	16.5%	17.6%	17.4%
Hour	7.5%	8.7%	8.9%	8.9%
15-minute interval	2.0%	2.3%	2.2%	2.3%

within the range of 8% to 12% mentioned by Taylor et al. (1996). Weekend days show bell-shaped traffic profiles with a peak period in the beginning of the afternoon.

From Appendix B it is concluded that the shape of the daily flow profile varies by location. For some arterials, the main direction of travel clearly differs between both peak periods. Traffic towards the city centre as well as traffic from residential areas shows a relatively high A.M. peak whereas traffic from the city centre and traffic to residential areas show a relatively high P.M. peak. These traffic patterns agree with existing literature that is described in Chapter 3 (Stathopoulos and Karlaftis, 2001b; Gram, 1996). Other routes serve multiple types of traffic (i.e. traffic with different travel motives) and yield a daily flow profile that shows an A.M. peak and a P.M. peak in both directions.

For most arterials it is possible to derive one general daily flow profile that is representative for all locations in one direction. However, on some routes the daily flow profile varies between links as a result of different traffic streams on different parts of the arterial. At the Weezebeeksingel for example, the daily flow profiles are highly influenced by a hospital, a furniture mall and a residential area halfway the arterial. Furthermore, along the Schuilenburgsingel the main direction of travel changes along the route as a consequence of industrial areas that attract traffic both from the Wierdensestraat/ Weezebeeksingel and the Van Rechteren Limpurtsingel.

6.4 Summary

In this chapter the traffic situation in Almelo is described. The main road network of Almelo roughly consists of two rings and some tangential roads towards the city centre, residential areas and the highway. The first part of the H. Roland Holstlaan – that connects Almelo to the highway – and the Weezebeeksingel – that is part of the outer ring and part of the regional route N35 towards Wierden/Zwolle – show the highest traffic loads. Also the other part of the outer ring – that mainly serves traffic to and from industrial areas and through traffic – and the Sluitersveldsingel – that mainly serves traffic to and from residential areas in the east of Almelo – show relatively high traffic loads. The roads that form the inner circle and the roads towards the city centre show the lowest traffic loads.

In general, Saturday traffic volumes are 70% to 80% of the weekday traffic volumes and Sunday traffic volumes are 50% to 60% of the weekday volume. As expected, the difference between weekday volume and weekend day volume is relatively large for roads that mainly serve through traffic and for roads to and from industrial areas and relatively small for roads to and from residential areas and roads to and from the city centre.

The general daily flow profiles agree with daily flow profiles found in literature (e.g. Chrobok et al., 2004; Taylor et al., 1996). An average working day shows a daily flow profile with an A.M. peak and a somewhat higher and broader P.M. peak. Weekend days show bell-shaped traffic profiles with a peak period in the beginning of the afternoon. The shape of the daily flow profile varies by location. For arterials that serve traffic to and from the city centre and to and from residential areas, the main direction of travel clearly differs between both peak periods. Moreover, on some arterials, the shape of the daily flow profile changes along the route, as a consequence of major attractions that attract traffic from different directions.

Chapter 7

Almelo: Traffic patterns

In Chapter 4 it was explained how cluster analysis can be applied for the determination and analysis of temporal and spatial traffic patterns. In this chapter, cluster analysis is applied to the traffic data of Almelo. The first section discusses temporal variations in daily flow profiles, the second section deals with spatial variations and in the third section traffic patterns are analysed on a network level. The chapter ends with a summary of the main results.

7.1 Temporal traffic patterns

This section deals with temporal traffic patterns. Daily flow profiles are grouped using a Ward hierarchical clustering procedure. Both working days and non-working days are grouped for all links for which enough traffic data was collected (more than 50 days). The individual clustering procedures are not discussed in detail in this thesis. A more extensive description of the results for all links along the main arterials can be found in Weijermars and Van Berkum (2006b).

This chapter discusses the main results. First, it is investigated what definition of a daily flow profile is most appropriate for describing the general shape of the flow profile. Subsequently, the main results of the cluster analyses are described briefly by addressing the topics presented in Section 4.2. Section 7.1.2 discusses the results for working days, whereas Section 7.1.3 discusses the results for non-working days.

7.1.1 Definition of a daily flow profile

In Chapter 4 it was explained that a daily flow profile is represented by a series of traffic counts. It was stated that the number of measurement intervals depends on the selected aggregation level. In this section it is investigated what aggregation level is most appropriate for the description of the general shape of the daily flow profile. For that purpose, the daily flow profiles of random days are drawn using 5-minute, 15 minute, 30 minute, and hourly traffic volumes. The daily flow profile that does not show too many short term fluctuations, yet describes the general shape of the flow profile adequately is selected as a representation of a traffic pattern. Figure 7.1 shows an example of the daily flow profiles on a random day for one of the locations. This example is representative for the other locations.

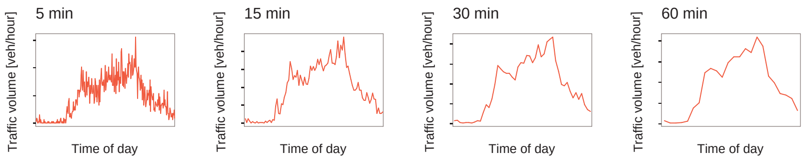


Figure 7.1: Example of daily flow profiles on different aggregation levels.

From figure 7.1 can be seen that a daily flow profile defined by 5-minute traffic volumes shows many short term variations. When this definition of a daily flow profile is used, days with a similar general daily flow profile can could be classified to different clusters as a result of differences in short term variations. Therefore, 5-minute traffic volumes are not appropriate for the definition of a daily flow profile. Also the daily flow profile that is determined by 15 minute traffic volumes shows considerable short term variations. In case of 30 minute traffic volumes, the short term variations are almost filtered out, whilst the general shape of the daily flow profile is still described adequately. A daily flow profile on the basis of hourly traffic volumes does not show any short term variations. However, from Figure 7.1 can be seen that the flow profile shows a lower A.M. peak than the flow profile at other aggregation levels. Thus, the general shape of the flow profile is less adequately described. Therefore, this definition of a daily flow profile is less appropriate as well. From an examination of the daily flow profiles on all locations and a explorative analysis of the clustering results for different aggregation levels, it is concluded that 30 minute traffic volumes are most appropriate for describing the general shape of a daily flow profile. Therefore, in this research, a traffic pattern on link l and day d is defined as a series of 48 traffic volumes as a function of the time of the day, i.e.

$$\mathbf{P}_{ld} = (q_{ld,1}, ..., q_{ldt}, ..., q_{ld,48}) \tag{7.1}$$

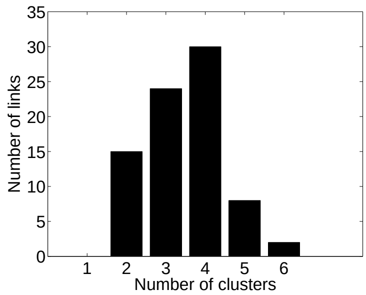


Figure 7.2: Histogram numbers of clusters.

7.1.2 Working day patterns

Cluster analyses are executed for 79 links. The number of clusters as well as the resultant traffic patterns and the factors that are on the basis of these traffic patterns, differ by location. Figure 7.2 shows a histogram with the frequency of different numbers of clusters. In most cases the clustering results in two, three or four clusters. In general, weekday, holiday periods and road works appear to be the main decisive factors.

In some cases, a cluster is mainly determined by one factor. Often, a holiday cluster exists that shows low traffic volumes throughout the day compared to the other clusters, yet the difference is largest during the A.M. peak period. Also seasonal clusters mainly differ from each other with regard to the height of the daily flow profile, although these clusters only occasionally occur. At a couple of locations, summer days show relatively low traffic volumes. Clusters that are characterized by a weekday occur more often and mainly show differences in the shape of the daily flow profile. Friday clusters show less peaked daily traffic profiles than other clusters. Often, the off-peak flow is relatively high for a Friday cluster, whilst especially the A.M. peak is relatively low. Mondays and to a lesser extent Tuesdays often are classified to a cluster that shows a relatively peaked daily flow profile and a relatively low daily traffic volume compared to other clusters. Thursdays are sometimes classified to a separate cluster that shows relatively high traffic volumes in the evening. Only one of the locations shows a Wednesday cluster. Figure 7.3 shows an example of a classification that is based on the day of the week and holiday periods.

The clusters discussed thus far agree well with the literature discussed in Chapter 3. Most variations in daily flow profiles can be explained by variations in activity patterns. The relatively high traffic volumes on Thursday evening on some of the locations can be explained by late shopping night in the city centre of Almelo. The relatively flat flow profile on Fridays is caused by the fact that relatively many employees go home early on Friday or have a day off, as a result of which the amount of shopping, personal business and leisure traffic is relatively high and the amount of commuter traffic is relatively low on Fridays.

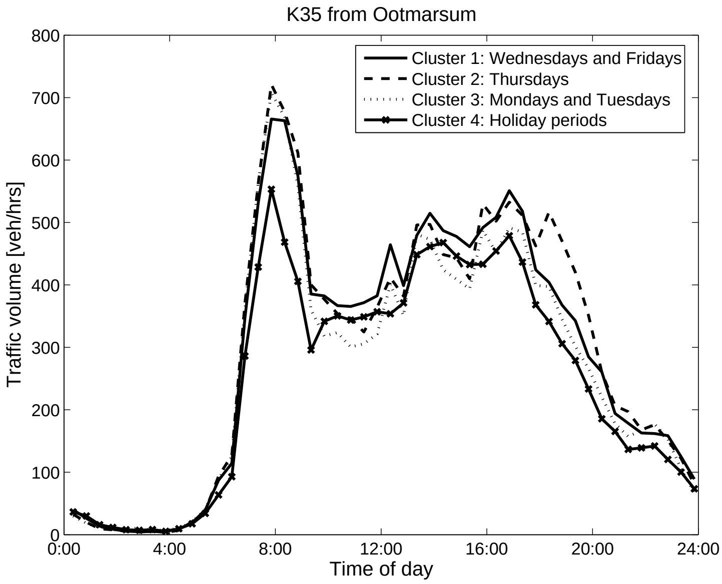


Figure 7.3: Example of a classification on the basis of weekday and holiday periods.

The relatively low A.M. off-peak flow on Mondays can be accounted for by the fact that most shops are closed on Monday. The relatively low traffic volumes on days within holiday periods and the relatively flat daily flow profile on these days is due to relatively little commuter and educational traffic on these days. Also low traffic volumes on summer days may be due to less commuter traffic, yet also a modal shift from car to bike might play a role.

Also road works appear to play an important role. At almost one third of the locations, road works are on the basis of one or more of the clusters. In some cases, road works result in a very dissimilar flow profiles (see left graph of Figure 7.4) whilst in other cases, traffic is only slightly different as a result of the re-routing of traffic due to road works elsewhere in the network (see right graph of Figure 7.4). Since road works continue for multiple months in some cases, re-routing effects of road works are easily mixed up with seasonal effects, so one should be careful when analysing seasonal variations.

At some locations, events result in a separate cluster. Football-matches of Heracles[1] result in clusters that show peaks in traffic in the evening. At a few locations, one of the clusters shows a somewhat shifted daily flow profile (see Figure 7.5). This cluster contains the first three working days after the adjustment of the clock to winter time. It seems that the clock of the controller is adjusted to winter time a couple of days too late. This case illustrates the use of cluster analysis for the detection of invalid loop detector data. Besides,

[1] Heracles is a local football club that played in the First Division at the time of the data collection.

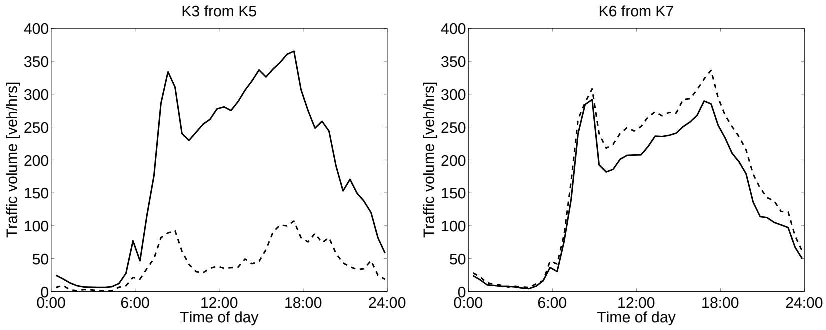


Figure 7.4: Effects of road works on two locations. The dashed lines represent the average daily flow profiles of the road works clusters and the solid lines represent the average flow profiles of the remaining working days.

at one of the locations, the cluster analysis resulted in two clusters that could not be explained for by weekday, season, holiday periods, road works or events. Probably, invalid data is on the basis of this clustering as well.

In some cases, a combination of factors is responsible for a cluster. On one hand, a certain type of day (e.g. a certain weekday) can be distributed over multiple clusters on the basis of another characteristic (e.g. season or road works). At some locations, Fridays are for example distributed over multiple clusters on the basis of the season. In other cases, one of the clusters contains Tuesdays, Wednesdays and Thursdays in spring. On the other hand, multiple types of days can be combined into one cluster, Mondays and Tuesdays are for example sometimes grouped with winter and/or autumn days in a cluster that shows a relatively low daily traffic volume and a relatively peaked flow profile. Days within holiday periods are sometimes classified with Fridays in a cluster that shows a low A.M. peak flow compared to other clusters. At other locations, holiday periods are combined with winter, autumn and/or summer days in a cluster that shows relatively low traffic volumes throughout all day, yet mainly during the A.M. peak.

The resulting clusters are hardly ever totally homogeneous and complete. Let us define a function τ that assigns a day to a certain type. A cluster is labeled complete when for all days in the cluster it holds that:

If $\tau(d_1) = \tau(d_2)$ then $\phi(d_1) = \phi(d_2)$

In most cases, a cluster contains most days of a certain type, yet not all days.

A cluster is labeled homogeneous when for all days in the cluster it holds that:

If $\phi(d_1) = \phi(d_2)$ then $\tau(d_1) = \tau(d_2)$

In most cases, a cluster contains relatively many days of a certain type, yet not exclusively.

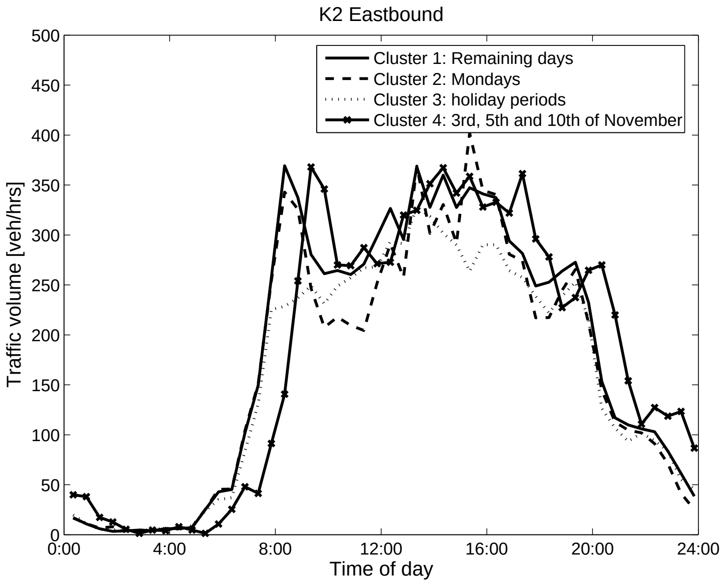


Figure 7.5: Example of a cluster that is based on invalid traffic data. Cluster 4 shows a shifted daily flow profile.

The extent to which clusters are homogeneous and complete varies by location. In general, holiday clusters are quite complete, yet not homogeneous. This means that (almost) all days within holiday periods are classified to one cluster that also contains days outside holiday periods. Moreover, clusters with Thursdays, with Mondays and with Fridays are more complete than clusters with Tuesdays and with Wednesdays. Besides, Friday clusters generally are quite homogeneous. In contrary, clusters with all or most Mondays also contain many other days. Wednesdays often are distributed over multiple clusters. As mentioned before, seasonal clusters are less common than weekday clusters. In most cases that clusters differ with regard to the distribution of the days over seasons, the season is not the only factor on the basis of the cluster, i.e. clusters are not homogeneous. Figures 7.6 and 7.7 show examples of a relatively complete and homogeneous classification and a relatively incomplete and mixed classification.

Rain does not seem to be a decisive factor for the clustering result at most locations. The percentages of wet and dry days differed between the clusters at only 10% of the analysed locations. The conditions for applying a Chi-square test were not met in any of the cases[2], so no significant influence of the weather can be determined. However, for eight locations, one of the clusters appeared to contain relatively many wet days in comparison to the other clusters. At these locations,traffic volumes are higher for clusters with relatively many wet

[2]This is due to a limited amount of pairs of wet and dry days (see Chapter 9).

	Weekday			Total
	Mon - Wed	Thu	Fri	
Cluster 1	70	2	2	74
Cluster 2	0	25	0	25
Cluster 3	0	0	23	23
Total	70	27	25	122

The second cluster solely consists of Thursdays and most Thursdays are classified to cluster 2. Cluster 3 solely contains Fridays and most Fridays are classified to cluster 3. Finally, cluster 1 mainly contains Mondays to Wednesdays and all Mondays, Tuesdays and Wednesdays are classified to this cluster.

Figure 7.6: Example of a rather homogeneous and complete classification.

days.

As mentioned in Chapter 4, a Ward's hierarchical clustering procedure consists of a number of steps in which clusters with the most similar daily flow profiles are combined. By analyzing the resulting classifications for different steps of the clustering procedure, more insight may be obtained into the influence of different factors. Therefore, for eleven locations, we analysed the results for all steps of the clustering procedure. It is investigated which characteristics are on the basis of the clusters, what the differences in daily flow profiles are between the clusters and which clusters are combined during the different steps. Also these results differed by location, yet here some general conclusions are drawn.

In general, Thursdays, Fridays and days within holiday periods show more distinct daily flow profiles than Mondays, Tuesdays and Wednesdays. Mondays, Tuesdays and/or Wednesdays are often combined in the first or second step of the clustering procedure, whereas Thursdays, Fridays and holiday periods continue to be separate clusters for multiple steps of the clustering procedure. Also summer days and spring days are often classified to separate clusters at the first steps of the clustering procedure. In subsequent steps of the clustering procedure, holiday clusters are sometimes combined to Friday clusters and in other cases combined with spring and summer clusters. Moreover, Thursdays are sometimes combined with Fridays, spring days or Tuesdays and Wednesdays.

With regard to differences in daily flow profiles it is concluded Wednesdays often show a relatively flat daily flow profile compared with Mondays and Tuesdays. Days within holiday periods show the flattest daily flow profile. Moreover, summer days show a higher off-peak flow than spring days. Finally, from a combination of a cluster with Tuesdays and Thursdays in summer and autumn with a cluster with Tuesdays and Thursdays in spring and winter is concluded

	Holiday period		Total
	No	Yes	
Cluster 1	52	0	52
Cluster 2	27	0	27
Cluster 3	6	10	16
Cluster 4	13	0	13
Total	98	10	108

	Season (outside holiday periods)				Total
	Dec - Feb	March - May	June - Aug	Sept - Nov	
Cluster 1	16	25	1	10	52
Cluster 2	16	0	8	3	27
Cluster 3	6 (1)	3 (3)	5 (2)	2 (0)	16
Cluster 4	3	6	1	3	13
Total	29	53	5	21	108

	Weekday (outside holiday periods)					Total
	Mon	Tue	Wed	Thu	Fri	
Cluster 1	8	13	15	16	0	52
Cluster 2	8	7	5	3	4	27
Cluster 3	3 (2)	3 (1)	3 (1)	3 (0)	4 (2)	16
Cluster 4	0	0	0	0	13	13
Total	19	23	23	22	21	108

Cluster 3 contains all holidays, yet not exclusively holidays. On the other hand, the fourth cluster contains solely Fridays, yet not all Fridays are classified to this cluster. Cluster 2 contains relatively many Mondays and Tuesdays, yet less than half of all Mondays and Tuesdays are classified to cluster 2. Cluster 1 contains most Thursdays and relatively many Tuesdays and Wednesdays, yet not all of these days and also some Mondays.

Mondays, Tuesdays, Wednesdays and Thursdays are divided over clusters 1 and 2 on the basis of the season. From a further analysis of the data is concluded that cluster 1 contains relatively many days in Mach, April, May and December, whereas cluster 2 mainly contains days in January, February, March and September.

Figure 7.7: Example of a rather homogeneous and complete classification.

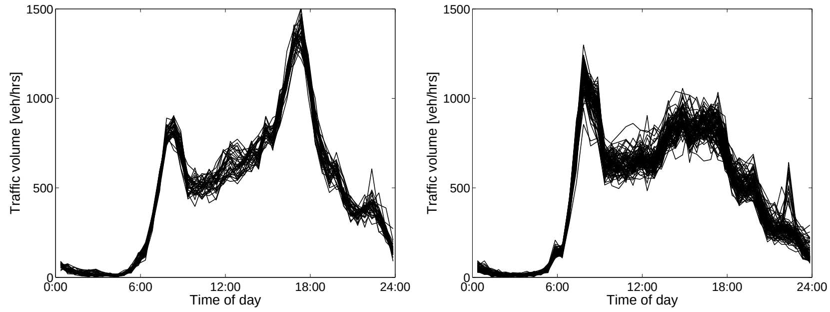


Figure 7.8: Examples of clusters with relatively similar (left) and relatively dissimilar (right) daily flow profiles.

that the spring and winter cluster shows a more peaked flow profile.

The variation within and between the clusters varies by location. Some clusters show substantial variations between the days within the cluster, whilst the days in other clusters show highly similar daily flow profiles. Figure 7.8 shows examples of a cluster with relatively large variations between the days within the clusters and a cluster with days that show very similar daily flow profiles. The ratio F that compares the standard deviation before classification to the standard deviation after classification is between 1.08 and 1.90. The average value for the ratio over all links is 1.28. The ratio is highest for locations with abnormal daily flow profiles due to road works. For these locations, the clustering classifies days with road works that cause high standard deviations to a separate cluster. The ratio is lowest for locations that show relatively low traffic volumes. On these locations, the amount of systematic variations is relatively small whereas the amount of random variation is relatively large. Since these locations do not show typical traffic patterns, cluster analysis is not useful. For many locations, the standard deviation decreases substantially during one or both peak periods. Figure 7.9 shows an example of such a location. For locations with a Thursday evening or Heracles cluster, the standard deviation mainly decreases during the evening period, since for these locations the clustering distinguishes days with and days without high traffic volumes in the evening. Finally, in general, Friday and holiday clusters show higher standard deviations than Monday-Tuesday clusters and mixed clusters show higher standard deviations than homogeneous clusters.

7.1.3 Non-working day patterns

For non-working days, the resulting traffic patterns are similar for most locations. In general, the optimal number of clusters is two. One of the clusters mainly contains Saturdays, whilst the other cluster mainly contains Sundays. The Saturday cluster shows higher traffic volumes than the Sunday cluster

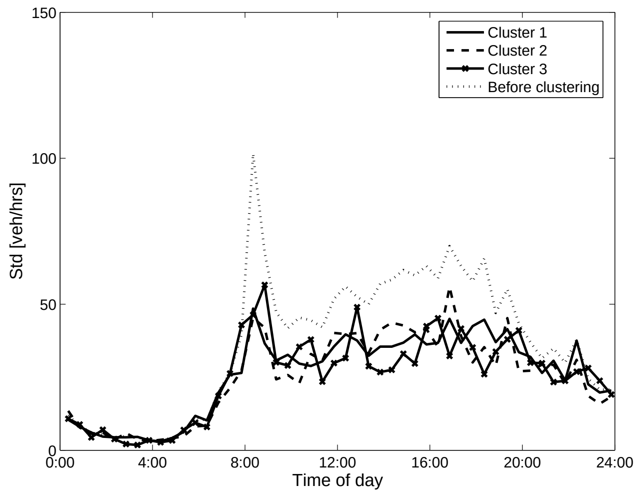


Figure 7.9: Location for which the standard deviation after clustering mainly decreases during the A.M. peak period.

(see left graph of Figure 7.10). Most special days (days between Boxing day and New Years day, Good Friday, day after Ascension day) are classified to the Saturday cluster, whereas New Year's day, Pentecost, Christmas day and Boxing day are classified to the Sunday cluster. The other Public Holidays are classified to different clusters at different locations.

In general non-working day clusters are more complete and homogenous than working-day clusters. At all locations, none or only a few Saturdays are classified to the Sunday cluster and the other way around. Saturdays and Sundays show more distinctive daily flow profiles than different types of working days. This also can be seen from the ratios between the standard deviations

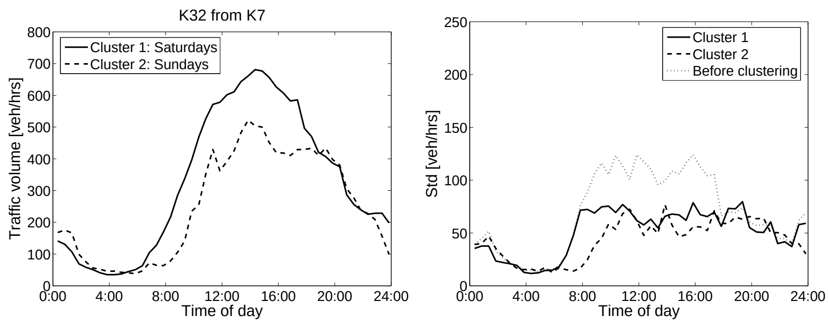


Figure 7.10: Example of a classification of non-working days. The left graph shows the resulting clusters and the right graph shows the standard deviation before and after classification as a function of the time of the day.

before and after classification. In general, these ratios are higher for non-working day clusters (between 1.30 and 2.17 with an average value of 1.63). On the other hand, for most non-working day clusters, the variation within the clusters is quite large as well, i.e. the days within the clusters show quite dissimilar daily flow profiles. From plots of the standard deviations before and after clustering as a function of the time of the day can be seen that the standard deviation decreases during the entire day (see the left graph of Figure 7.10 for an example).

7.2 Spatial traffic patterns

7.2.1 Variations in average daily flow profiles

The average working day flow profiles of all links entering intersections for which enough traffic data is available (more than 50 days) are grouped by means of a Ward's hierarchical clustering procedure. As explained in Chapter 4, a pattern is defined by a series of proportions of total daily traffic volume for a certain day type:

$$\mathbf{P}_{lD} = (s_{lD,1}, ..., s_{lD,t}, ..., s_{lD,48}) \tag{7.2}$$

In this thesis, only the classification of working days is discussed, Saturdays and Sundays can be clustered and analysed in a similar way. The classification of working days results in four clusters that are shown in Figure 7.11. The distribution of the links over the clusters is shown in Figure 7.12.

Cluster 1 contains links that are used by several types of traffic (several travel motives). During the A.M. peak period commuter traffic leaves the residential areas or uses for example the Henriette Roland Holstlaan to reach work. In the afternoon, the same links are used by household, shopping, leisure and educational traffic. Traffic on the Schuilenburgsingel and Aalderinkssingel probably is traffic returning home from industrial areas along the routes.

Cluster 2 does not show a real A.M. peak. These links are mainly used by traffic returning home from work or other activities during the afternoon and P.M. peak period, whereas they are not intensively used by commuter traffic going to work during the A.M. peak. Cluster 3 in contrary, shows a relatively high A.M. peak and no real P.M. peak. This cluster partly consists of links with traffic in opposite directions of the links in cluster 2. These results agree with the finding of Gram (1996) and Stathopoulos and Karlaftis (2001b) that traffic towards the Central Business District (CBD) shows a peak in the morning and traffic from the CBD shows a peak in the afternoon.

Cluster 4 shows a very high P.M. peak and no A.M. peak and contains some side streets of the Schuilenburgsingel at K20. These side streets show very low traffic loads and connect some industrial areas to the Schuilenburgsingel.

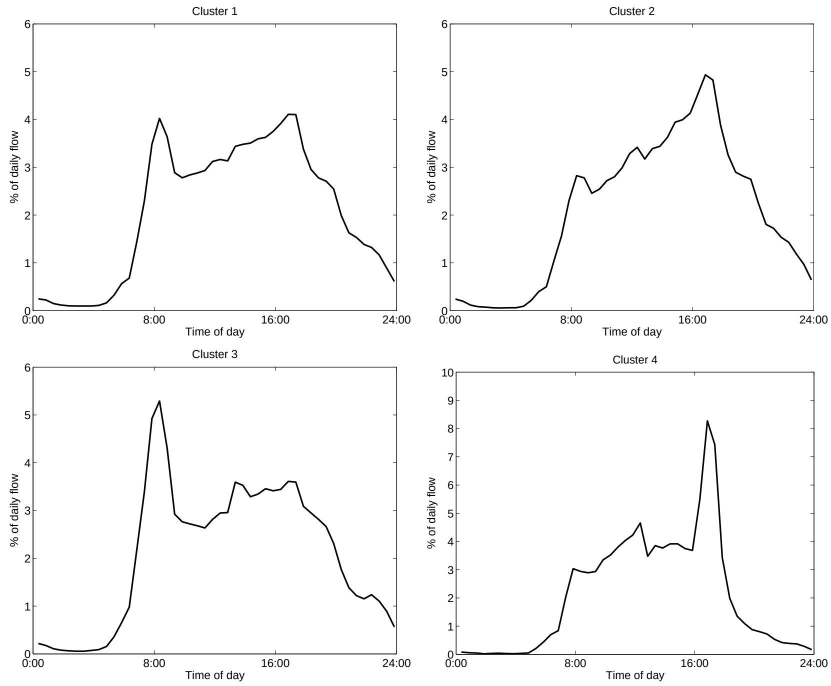


Figure 7.11: Daily flow profiles of resultant average working day profile clusters.

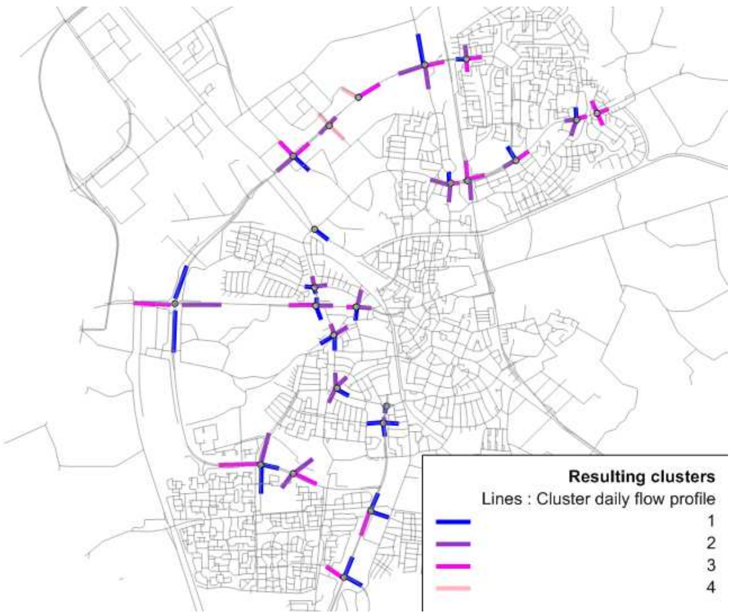


Figure 7.12: Distribution of the links over the average flow profile clusters.

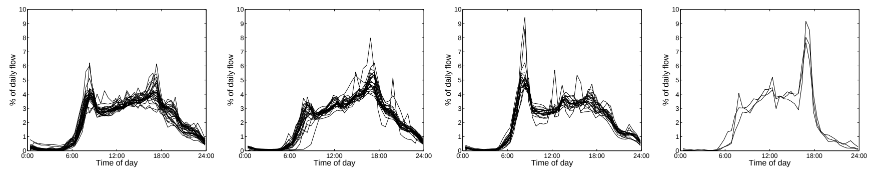


Figure 7.13: Average daily flow profiles of links within the clusters.

Figure 7.13 shows the average daily profiles of all links within the different clusters. From this figure it can be seen that the variation within the clusters is substantial. The total variation within the clusters is comparable for all clusters, yet the distribution of the variation over the day differs between the clusters. Cluster 3 for example shows a relatively high variation during the A.M. peak period (see Figure 7.14). This is caused by the fact that links in cluster 3 are characterized by the absence of a P.M. peak period. All links that show no real P.M. peak period are classified to cluster 3, regardless of their A.M. peak period. The fourth cluster contains three links with low traffic volumes that show a very high and short P.M. peak period. The ratio between the standard deviation before classification and the average standard deviation after classification equals 1.47 and the difference in variation before and after clustering mainly occurs during the peak periods (see Figure 7.14).

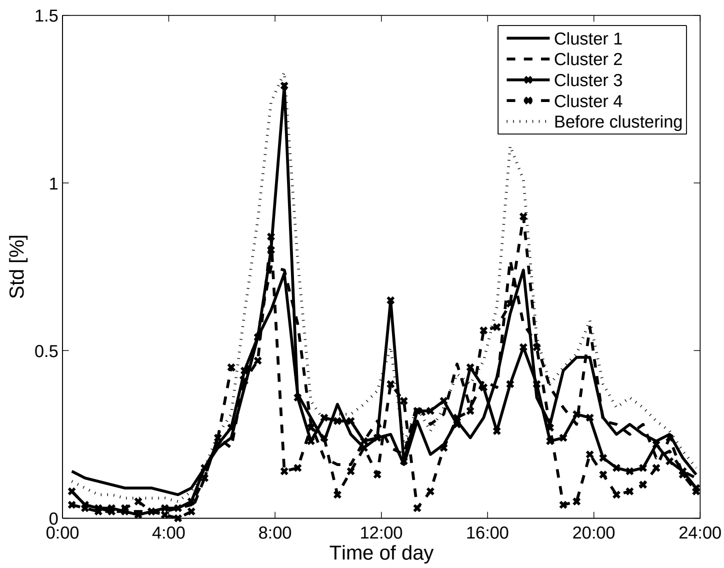


Figure 7.14: Standard deviation before and after clustering of average daily flow profiles as a function of the time of the day.

7.2.2 Variations in weekly patterns

As discussed in Chapter 4, locations are clustered both regarding the distribution of the traffic over weekdays, Saturdays and Sundays and regarding the distribution of the traffic over the different working days.

With regard to the distribution of the traffic over working days, Saturdays and Sundays, only links for which traffic data is collected on at least ten working days, ten Saturdays and ten Sundays are selected. The clustering results in three clusters. In Figure 7.15 it can be seen that cluster 3 shows very low weekend day indices. This cluster contains only three links with very low traffic loads. The first and second clusters mainly differ from each other with regard to the Saturday index, which is higher for the second cluster. From Figure 7.17 can be seen that locations that are classified to the second cluster in general connect residential areas to the city centre, other shopping areas or locations where leisure activities take place. Locations that are classified to the first cluster mainly serve commuter travel.

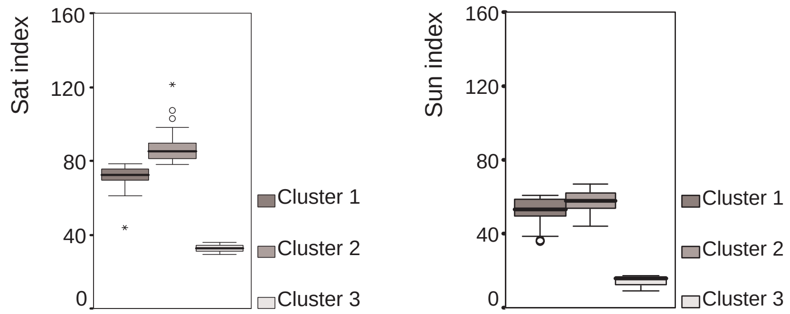


Figure 7.15: Boxplots of Saturday and Sunday indices (average working day = 100).

Regarding the distribution of the traffic over the different working days, only links for which for every weekday traffic data is collected on ten or more days are selected. The classification results in ten clusters, of which six consist of three links or less that in most cases show (very) low traffic volumes. These clusters are not discussed here. Figure 7.16 shows the mean difference from the average daily flow profile for different weekdays for the four larger clusters. Figure 7.18 shows which links are classified to each cluster.

From figure 7.16 it can be seen that in general, traffic volumes are relatively low during the off-peak period on Monday and the peak periods on Friday, whereas traffic volumes are relatively high during the off-peak period on Friday and on Thursday morning and evening. These results agree with the results from the cluster analysis discussed in Section 7.1. The peak on Thursday morning can be explained by the weekly market in the city centre. Clusters 1 and 4 clearly differ from the clusters 2 and 3, as they show relatively large differences from

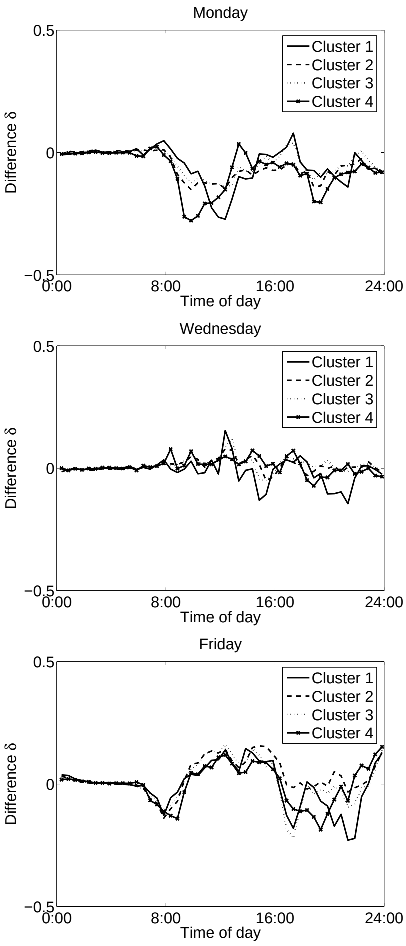


Figure 7.16: Distances to average daily flow profile for different weekdays.

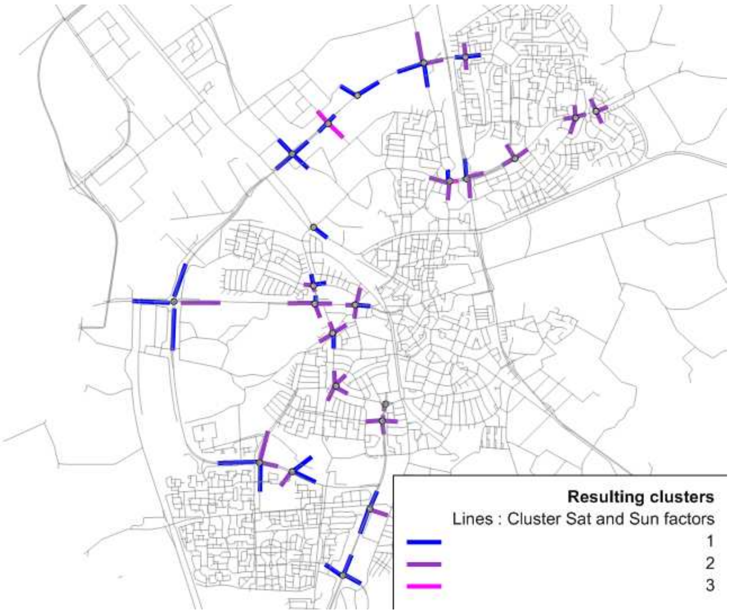


Figure 7.17: Distribution of the links over the clusters based on Saturday and Sunday factors.

the average daily flow profile, i.e. weekly variations are strongest for these clusters. Clusters 1 and 4 mainly differ from each other with regard to the time of the deviations from the average daily flow profile. Cluster 4 mainly contains links that are used by traffic towards the city centre and shows a relatively early peak on Thursday evening compared to cluster 1 that mainly contains links that are used by traffic originating from the city centre. Clusters 2 and clusters 3 both show relatively little variations between the weekdays and are quite similar to each other. The main difference between these clusters is the relatively low P.M. peak flow on Friday and the relatively high P.M. peak flow on Monday and Tuesday for cluster 3. Also the distribution of the links over these clusters can be explained: Links that are classified to cluster 3 are mainly used by commuter traffic towards home, whereas links that are classified to cluster 2 are mainly used by commuter traffic towards work. Thus, links in cluster 3 are more influenced by less commuter traffic during the Friday P.M. peak.

The variation is clearly smaller after classification; the ratio F equals 1.59. Moreover, the variation is similar for all clusters. From Figure 7.19 can be seen that the variation mainly decreases during the P.M. peak periods.

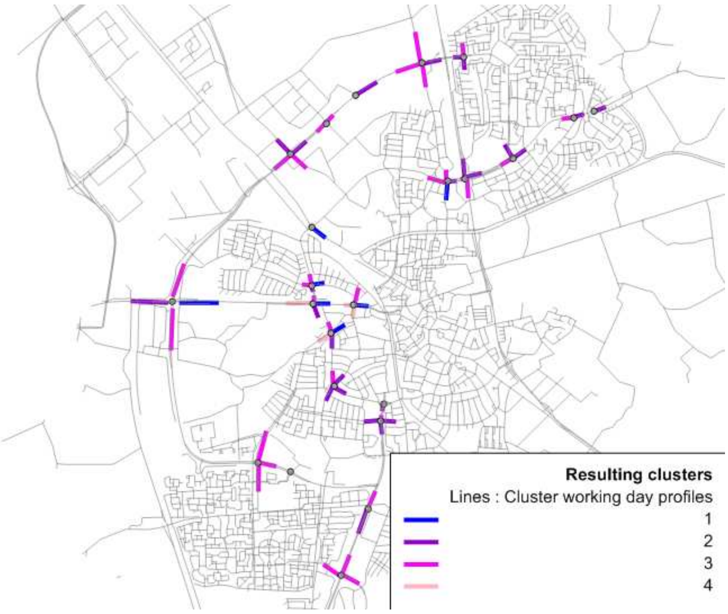


Figure 7.18: Distribution of the links over the weekly flow profile clusters.

7.2.3 Variations in seasonal patterns

In Chapter 4 it was explained that differences in seasonal variations between links can be analysed by comparing monthly factors for different locations. However, the data available in Almelo is unevenly distributed over the year. There are only a few locations for which traffic data is available for all months and even for these locations, the amount of data is limited for some of the months. Therefore we executed the analysis on a seasonal level, i.e. seasonal flow profiles instead of monthly flow profiles are compared to the average daily flow profile. Thereby, the winter is defined as the months December, January and February etc. Holiday periods were excluded from the analysis. Besides, only links for which 10 or more days were available for each season were selected.

Thus, a traffic pattern was defined by 48*4 features that represent the difference in daily flow profile between each season and the average daily flow profile, i.e.

$$\mathbf{P}_l^{\text{season}'} = (\delta_{l,winter,1}, ..., \delta_{lgt}, ..., \delta_{l,autumn,48}) \tag{7.3}$$

Where g is an index for season. The δ's are calculated in a similar way as the monthly δ's in Chapter 4.

The 41 links that met the requirements were grouped into five clusters of which two only contained one link. These clusters are not discussed here. The differences from the daily flow profiles of the other clusters in relation to the average daily flow profile are shown in figure 7.20. Figure 7.21 shows which

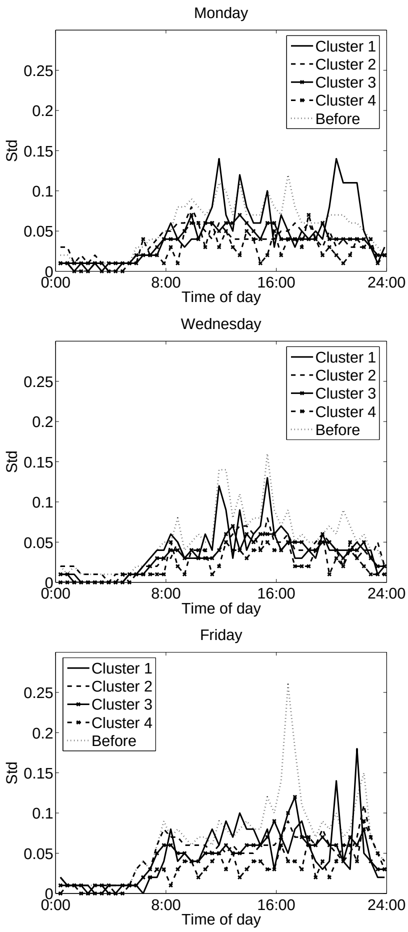


Figure 7.19: Standard deviations on different weekdays as a function of the time of the day.

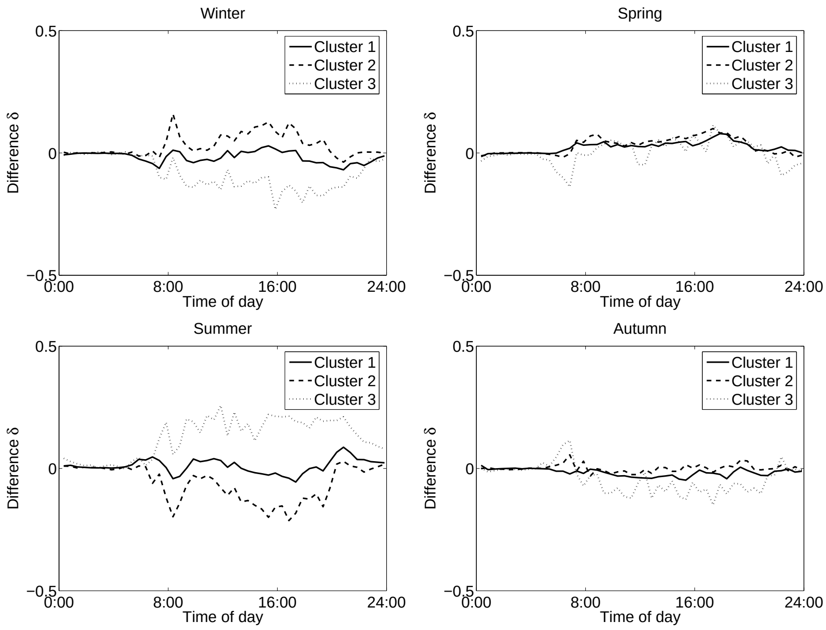


Figure 7.20: Distances to average daily flow profile for different seasons.

links are classified to each cluster.

Seasonal differences are somewhat smaller than weekday differences. Moreover, traffic volumes are either higher or lower for the entire day, thus daily flow profiles of different seasons mainly differ from each other with regard to the height of the flow profile in contrary to weekdays that mainly differed from each other with regard to the shape of the profile. However, seasonal effects are not constant over the day. For cluster 2 for example the decrease in traffic in summer and the increase in traffic in winter are relatively large during the peak periods. Thus, summer days show a relatively flat daily flow profile compared to winter days.

Most links are classified to cluster 1 that shows the least seasonal variations. Cluster 2 contains 12 locations and shows higher than average volumes in winter and relatively low volumes in summer. The locations are mainly located near residential areas with relatively much short distance travel, so possibly in summer the modal share of the car is smaller than in winter. The third cluster shows opposite seasonal effects as there is relatively much traffic in summer as a result of re-routing effects due to road works.

The variation is clearly smaller after classification, although the ratio between the standard deviations before and after classification is a little smaller than for weekday patterns (1.47). From Figure 7.22 can be seen that the largest difference in variation before and after classification occurs in summer,

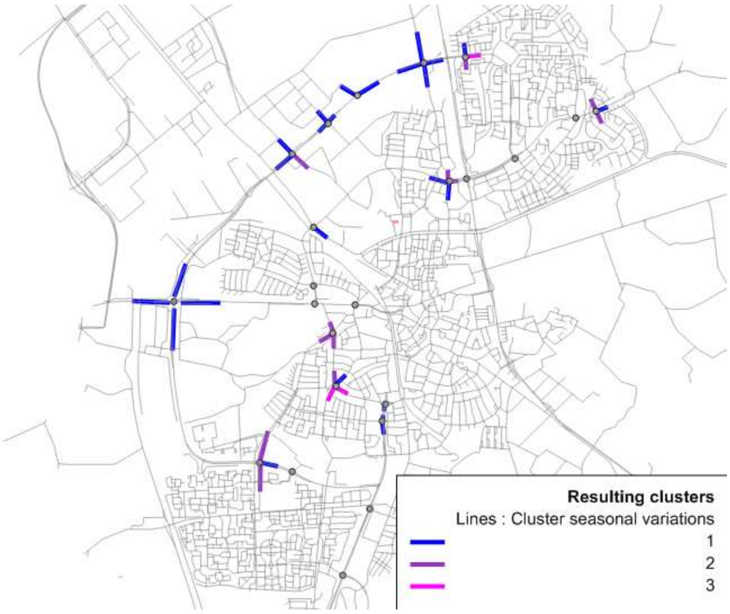


Figure 7.21: Distribution of the links over the seasonal flow profile clusters.

especially during the P.M. peak period.

7.2.4 Variations in weather factors

As was mentioned in Chapter 5, for this research weather data from Twente airport on a six hour aggregation level is available. Moreover, it was decided that a period is labeled wet when the duration of rain is more than two hours and the intensity is more than 0.5 mm/hour. As a consequence of this classification, 75% of the 6-hour periods is labeled dry, whereas about 5% is labeled wet. For most links, the amount of wet periods appeared to be very small. Therefore, analyses are only executed on a daily level. Moreover, due to the limited amount of data, no distinction is made between different seasons.

A pattern thus is defined as:

$$\mathbf{P}_l^{\text{weather}} = (\pi_l) \tag{7.4}$$

where

$$\pi_l = \frac{\bar{q}_l^{\text{wet}}}{\bar{q}_l^{\text{dry}}} * 100 \tag{7.5}$$

The classification results in four clusters for which the average ratios between wet and dry days are shown in Table 7.1. By means of a selected link

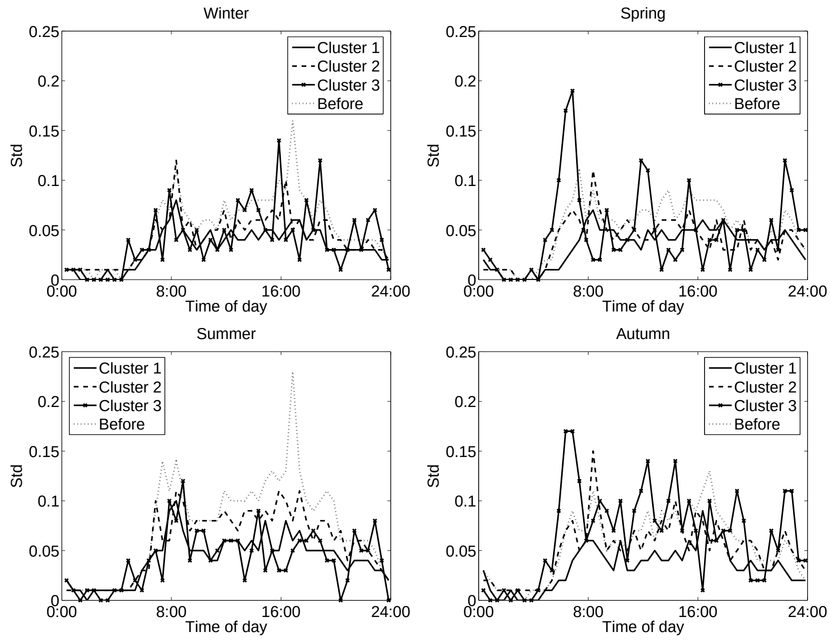


Figure 7.22: Standard deviations for different seasons as a function of the time of the day.

analysis in the traffic simulation model Omnitrans (http://www.omnitrans-international.com/), the percentages of short distance trips are determined for the different clusters. These percentages are also shown in Table 7.1. These percentages are for a P.M. peak period only, but they give an impression of the difference in the amount of short distance traffic between the clusters.

Cluster 2 shows the highest index and thus shows relatively high traffic volumes on rainy days. This cluster mainly contains locations that connect residential areas to main roads. The average percentage of short distance traffic is highest for this cluster. Clusters 1 and 3 show somewhat lower indices and mainly contain links on main arterials that serve relatively much long distance traffic.

Table 7.1: Average ratios between wet and dry days and percentages of short distance trips for different clusters.

	Nl	$\bar{\pi}$	% trips $\leq$ 7 km
Cluster 1	36	103	45%
Cluster 2	8	109	67%
Cluster 3	29	100	35%
Cluster 4	2	91	77%

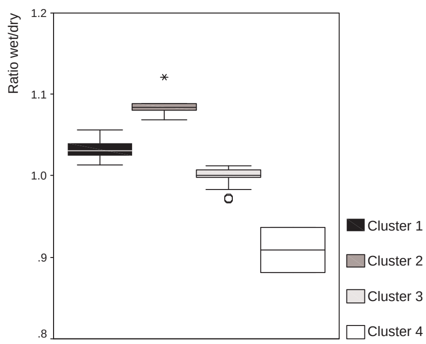


Figure 7.23: Boxplot of $\bar{\pi}$'s.

Furthermore, cluster 1 shows a little higher ratio than cluster 3 and mainly consists of links on the Sluitersveldsingel and the Schoolstraat, whereas cluster 1 mainly contains links on the Schuilenburgsingel, Wierdensestraat and Henriette Roland Holstlaan. The low index for the links in cluster 4 cannot be explained for. From a further analysis of the data it is concluded that for some pairs of wet and dry days, the traffic volume is much higher on the dry day. Possibly special circumstances are subject to these differences.

Figure 7.23 shows the variation within the clusters and from the figure can be seen that the variation within the clusters is relatively small compared to the variation between the clusters.

7.2.5 Variations in temporal classifications

In Chapter 4 we proposed a method for the classification of locations on the basis of their temporal clusters. For all pairs of locations the proportion of the days that are classified to different clusters was determined. These proportions were recorded in a matrix that was used as input for the clustering procedure. The clustering resulted in many clusters that only consisted of a few links. In case of fewer clusters, many days are classified to different clusters at different locations. In case of 4 clusters for example, on average around 30% of the days are classified to different clusters at two links within one cluster.

From a further analysis of the distance matrix is concluded that the distances between some pairs of locations are relatively small whereas different factors are on the basis of the classifications. On the other hand, between some pairs of locations the distances between the classifications are relatively large whereas the same factors are on the basis of the classification. The first problem mainly occurs when a classification into many clusters is compared to a classification into two clusters. Figure 7.24 illustrates this problem. The second problem mainly occurs in case of many clusters at both locations. Since the clusters

Classifications are compared for the following locations:

A: Thursdays (cluster 2) and remaining days (cluster 1)
B: Mon and Fri (cluster 1), holidays (3), road works (4), remaining days (2)

A cross-tabulation is drawn that shows the number of days for each combination of clusters:

	B, C1	B, C2	B, C3	B, C4	Total
A, C1	18	20	14	3	55
A, C2	0	6	3	1	10
Total	18	26	17	4	65

To be able to compare the classifications, clusters at location B are combined into new clusters in such a way that the set of days that is classified to the same cluster is maximized. Clusters 1, 2 and 3 are combined into a new cluster that is compared to cluster 1 at location A. The road works cluster of location B thus is compared to the Thursday cluster of location A. 6+3+3=12 out of 65 days = 18% of the days are classified to different clusters. The distance is relatively small compared to other distances whereas different factors are responsible for the classifications.

Figure 7.24: Example of a small distance between locations in case of dissimilar classifications.

are rarely complete and homogeneous and classifications are based on different datasets (data is collected on different groups of days at different locations), distances can be relatively large whereas the same factors are on the basis of the classifications. An example of such a situation is shown in Figure 7.25.

Although the absence of clear clusters can partly be explained by the problems mentioned above, clear groups of locations with regard to the groups of days that result from the temporal classification appear to be absent in Almelo anyway. Classifications between pairs of links differ from each other in multiple ways, yet there are no clear groups of links that show classifications similar to each other and dissimilar from links in other groups. As discussed in the first section of this chapter, at most locations, weekday, holiday periods, road works or a combination of these factors are on the basis of the classification. At many locations, the clustering results in a combination of clear, complete and homogeneous clusters and less complete and heterogeneous clusters. Often, locations differ from each other with regard to the incomplete and mixed clusters, e.g. at one location a certain cluster contains 80% of the winter days and at another location 70%. Moreover, in some cases a certain type of day is classified to dissimilar clusters, e.g. at one location Wednesdays are classified to the Friday cluster whereas at another location they are classified to the Monday and Tuesday cluster. Finally, in some cases a typical cluster exists at one location and does not exist at another location, e.g. one of the locations shows a holiday cluster whereas the other location does not.

Classifications are compared for the following locations:

A: Fri (cluster 3), many holidays + all autumn (2), remaining (1)
B: Fri (cluster 1), all holidays + many summer (2), remaining (3)

A cross-tabulation is drawn that shows the number of days for each combination of clusters:

	B, C1	B, C2	B, C3	Total
A, C1	0	4	24	28
A, C2	1	18	0	19
A, C3	12	2	0	14
Total	13	24	24	61

Although the classifications are based on the same factors, (4+1+2)/61 = 12% of the days is classified to different clusters.

Figure 7.25: Example of a large distance between locations in case of similar classifications.

Although no clear groups of links can be distinguished with regard to the total outcome of the cluster analysis, there are some groups of locations that show the same specific traffic pattern. Figure 7.26 shows locations with a Heracles cluster (this cluster contains days on which Heracles played a home match and shows a peak in the evening) and locations with a Thursday cluster (this clusters contains Thursdays and shows a peak in the evening).

7.3 Traffic patterns on a network level

In Chapter 4, we described two approaches for the analysis of traffic patterns on a network level. The first approach uses the results of the classification of links on the basis of the temporal patterns. Since no clear groups of links could be detected for Almelo, this approach is not elaborated here. The second approach classifies days on the basis of a series of daily flow profiles from multiple locations and is applied in Almelo.

Seven locations are selected as key locations, one on each arterial for which traffic data is collected. These locations were selected on the basis of the amount of valid traffic data. On the Schuilenburgsingel/ Bleskolksingel, two locations are selected since at this arterial, different main traffic streams were detected at different parts of the arterial (see Chapter 6). The key locations are shown in Figure 7.27

The clustering results in four clusters. Figure 7.28 shows the average daily flow profiles of these clusters on the key locations. Appendix C gives a more extensive description of the clusters. The first cluster contains relatively many

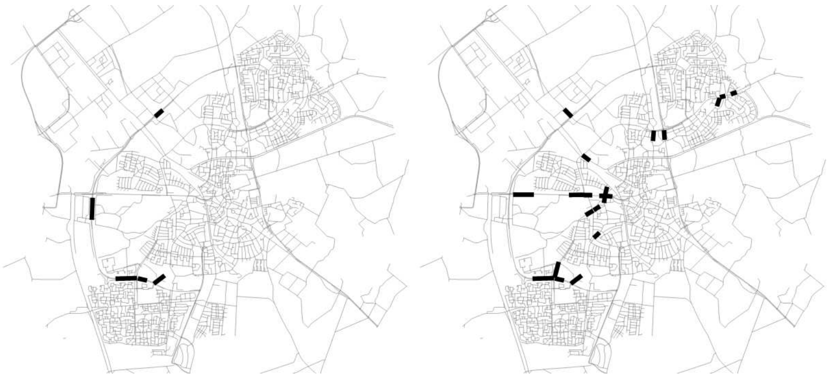


Figure 7.26: Links that show a Heracles cluster (left) and links that show a Thursday (right) cluster.

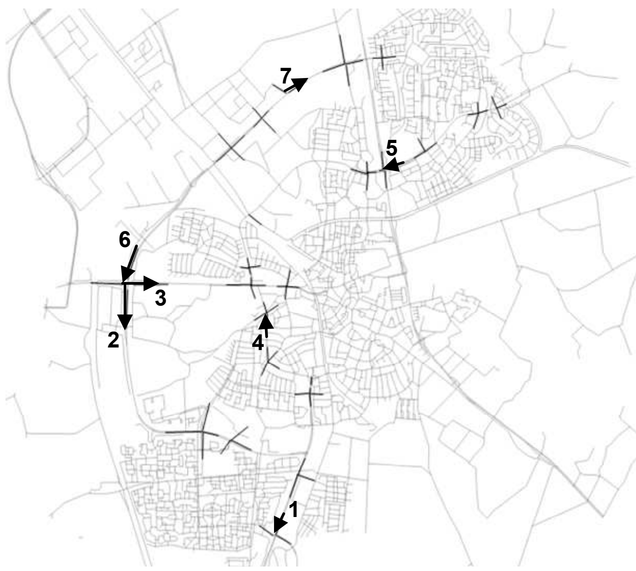


Figure 7.27: Links that are selected for the cluster analysis on network level.

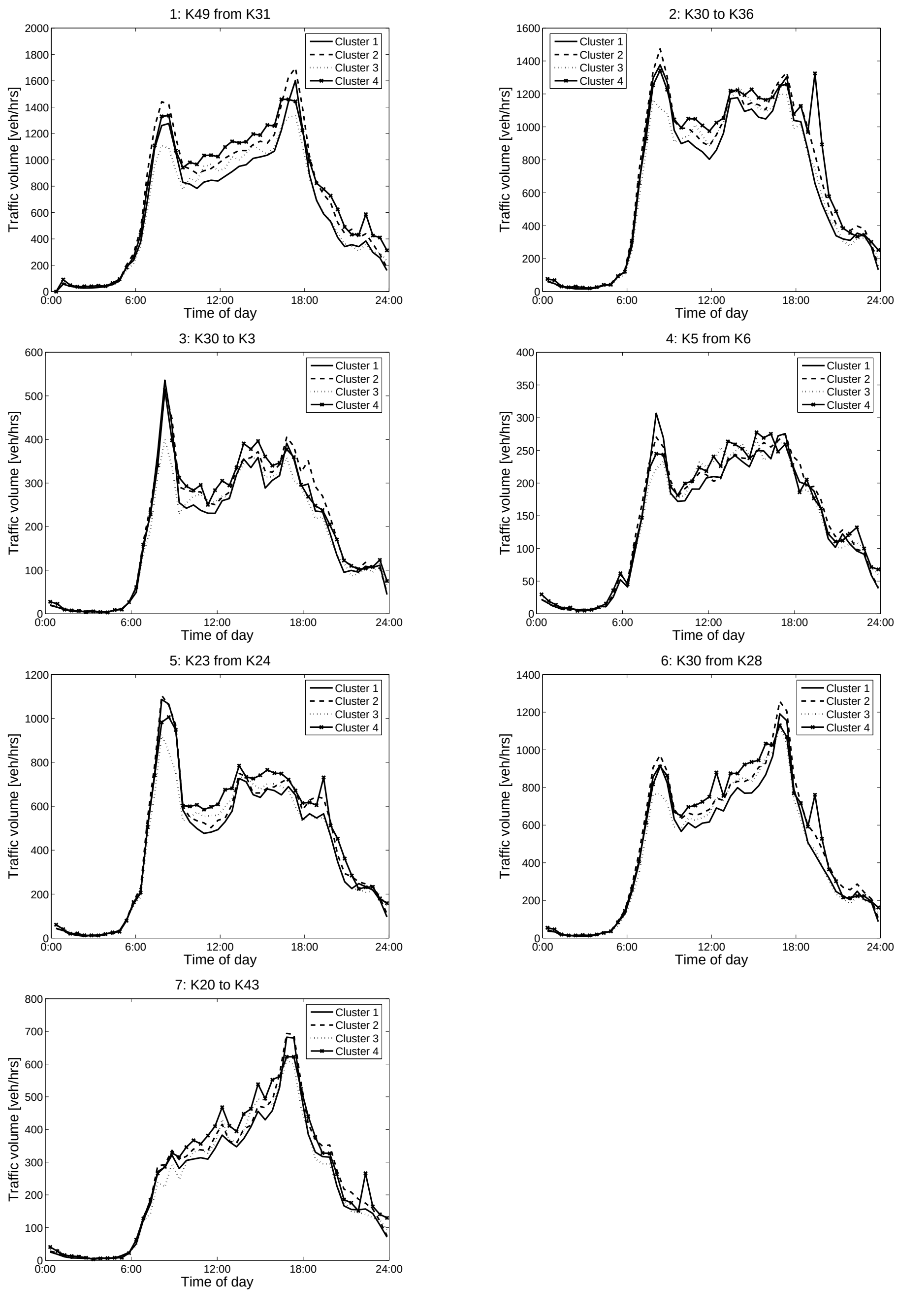


Figure 7.28: Average daily flow profiles of resulting clusters.

winter days and relatively many Mondays, Tuesdays and Wednesdays and shows a relatively peaked daily flow profile with a relatively low off-peak flow at all locations (91 - 95% of the average off-peak flow). Cluster 2 shows the highest traffic volumes and contains relatively many Thursdays and relatively many spring days. The third cluster consists of Fridays and days within holiday periods and shows a relatively flat daily flow profile with relatively low peak flows (the A.M. peak flow is 85 - 90% of the average A.M. peak flow and the P.M. peak flow is 90 - 98% of the average). The fourth cluster contains days on which Heracles played a home match and shows a peak in the evening at the Weezebeeksingel and the Schuilenburgsingel.

Surprisingly, the clustering does not lead to a clear Thursday cluster that only consists of Thursdays and shows a peak in the evening due to late night shopping. The absence of a Thursday cluster is probably due to the fact that the key locations with high traffic loads (1, 2, 5, 6, 7) do not show a real Thursday cluster. Only location 3 shows a real Thursday cluster (see Figure 7.26), yet the influence of this location on the clustering result is small since the traffic loads are relatively low. The selection of key locations is determinative for the clustering result and policy makers can define different sets of key locations dependent on the OD-relations and/or locations they want to analyse. Moreover, they can prioritize links by manipulating traffic loads.

The average standard deviation is clearly smaller after classification. The ratio between the standard deviation before and after classification is 1.34. Moreover, the standard deviation is smallest for cluster 1 and largest for the Friday and holiday cluster. The ratios for the different locations are similar to the ratios that were obtained from the cluster analyses at the individual locations. Figure 7.29 shows the standard deviations before and after classification as a function of the time of the day. From these plots it can be seen that for most locations, the standard deviations of most clusters are clearly lower than before classification for one or both peak periods. However, some clusters show high standard deviations for one peak period. For location 4, the standard deviation of one or more of the clusters is higher than the standard deviation before clustering during the entire day, . This results in a relatively low ratio (1.10). This high variation is partly due to the fact that road works were carried out on some of the days. These days show dissimilar daily flow profiles but are divided over clusters

7.4 Summary

This chapter discussed the application of the method proposed in Chapter 4 for the analysis of traffic patterns in Almelo. A daily flow profile represented by half hourly traffic volumes is selected as the most appropriate definition of a traffic pattern. Cluster analyses are executed both for working days and non-working days. For non-working days, results are similar for all locations. The

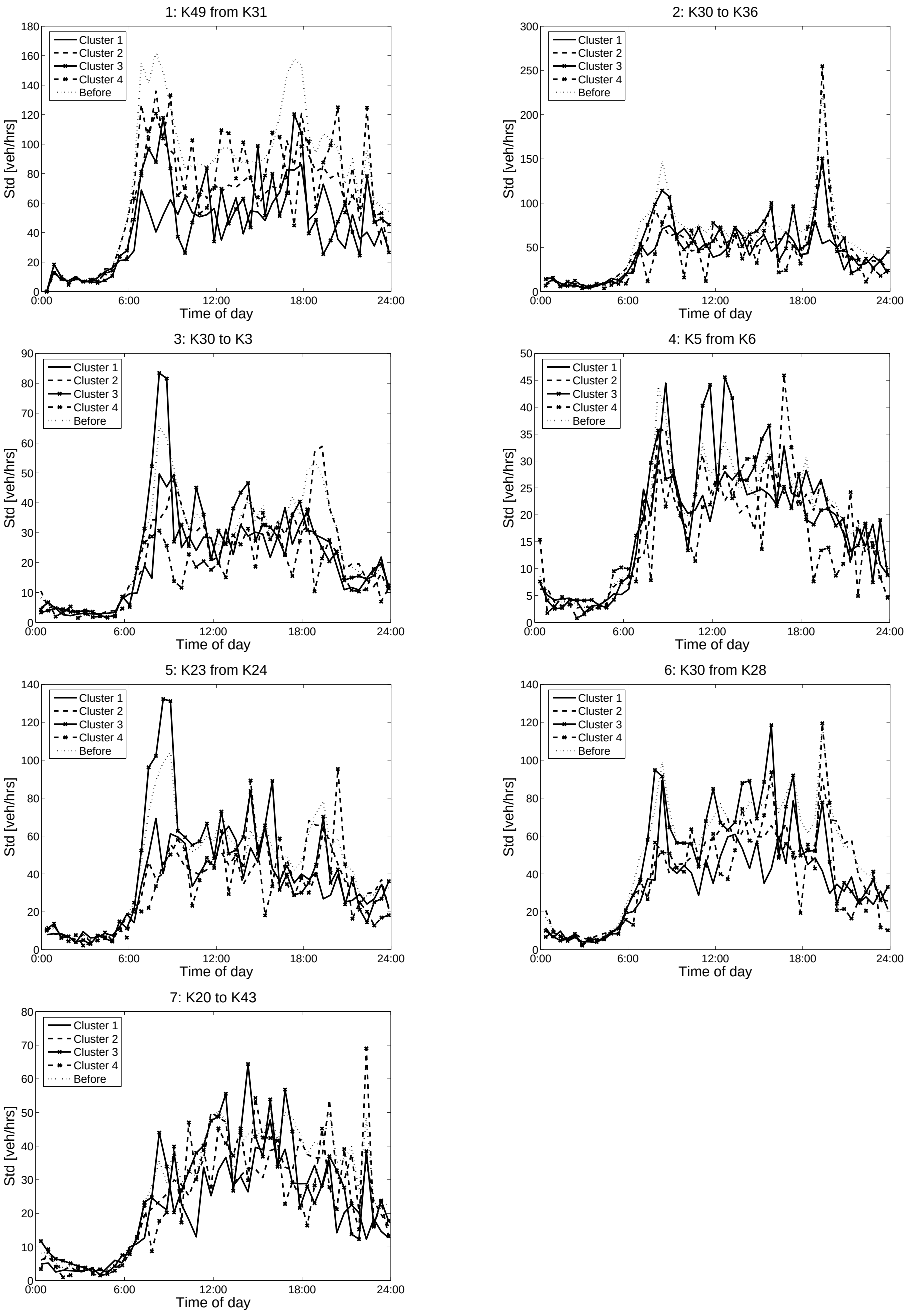


Figure 7.29: Average daily flow profiles of resulting clusters.

clustering results in a Saturday cluster and a Sunday cluster and the Saturday cluster shows higher traffic volumes throughout all day.

For working days, the resulting clusters differ by location. In general, the clustering result in two to four clusters that are determined by weekday, holiday periods or road works. Weekday and holiday clusters can be explained for by variations in activity patterns.

- Friday clusters show a relatively flat daily flow profile with a high off-peak flow and low peak flows, as a result of relatively little commuter traffic and much leisure and shopping traffic.
- Thursday clusters show a peak in the evening due to late night shopping traffic to and from the city center.
- Clusters with relatively many Mondays (and to a lesser extent Tuesdays) show a relatively peaked daily flow profile with a low A.M. off-peak flow due to the fact that shops are closed on Monday morning.
- Holiday clusters show relatively low traffic volumes throughout all day and a relatively flat daily flow profile as a result of less educational and commuter traffic.

Road works are on the basis of a cluster at almost one third of the locations. In some cases, road works result in very dissimilar daily flow profiles, whilst in other cases traffic volumes are only slightly higher as a result of re-routing due to road works elsewhere in the network. Also other location specific factors like events and invalid data play a role. Rain plays a role only at about 10% of the locations and also season is not a main determining factor in the clustering. In some cases, a combination of season and weekday are on the basis of a cluster; at some locations Mondays and Tuesdays are for example grouped together wit winter days and at other locations spring days are grouped together with Thursdays. The resulting working day clusters are hardly ever totally homogeneous and complete. In most cases, a cluster contains relatively many days of a certain type, yet not exclusively and not all.

Links are clustered on the basis of various type of temporal patterns. A classification on the basis of the average working day flow profile resulted in three common types of daily flow profiles (and a small cluster). One of the clusters shows both an A.M. peak and a P.M. peak and contains links that contain a mixture of travel motives. The other clusters show only an A.M. peak or a P.M. peak and are mainly used by commuter traffic, respectively to and from work. Also the clusters that result from the classification on the basis of weekly patterns can be explained by differences in travel motives. The classification on the basis of Saturday and Sunday factors resulted in two main clusters that mainly differed from each other with regard to the Saturday factor. The cluster with the highest factor contains links that are mainly used by traffic between residential areas and shopping or leisure areas. The classification on the basis of flow profiles on different weekdays resulted in four main clusters, two of which showed relatively large variations between weekdays compared to the other two. The links in the clusters with relatively large variations are mainly used by traffic to and from the city centre, whereas the links in the clusters with relatively little variations contain relatively much commuter traffic to and from work. Seasonal patterns and the clustering on the basis of the influence of rain can be explained by differences in the amount of short distance traffic. Locations with relatively much short distance traffic show relatively low traffic volumes in summer and relatively high traffic volumes on rainy days as a result of a modal shift between bike and car. Locations with little short distance traffic in general show little seasonal or rain related variations.

The clustering of locations on the basis of their temporal traffic patterns did not lead to clear groups of links. Classifications between pairs of links differ from each other in multiple ways, yet there are not clear groups of links that show classifications that are similar to each other and dissimilar from links in other groups. However, there are some groups of links that show specific patterns, like a Thursday cluster or a Heracles cluster.

On a network level, a clustering on the basis of daily flow profiles at seven key locations resulted in four clusters: (1) a cluster with relatively many Mondays, Tuesdays and Wednesdays and relatively many winter days that shows a relatively peaked flow profile with a low off-peak flow at al locations, (2) a cluster that contains relatively many Thursdays and spring days that shows relatively high traffic volumes throughout all day at all locations, (3) a cluster with a relatively flat daily flow profile and low peak volumes at all locations that consists Fridays and days within holiday periods and (4) a cluster that shows a peak in the evening at the Weezebeeksingel and Schuilenburgsingel and contains days on which Heracles played a (home) match.

All classifications result in a clear decrease in standard deviation. Moreover, in many cases, the standard deviation decreases substantially during one or both peak periods.

Chapter 8

Applications

In literature, some applications of cluster analyses have been discussed. First of all, cluster analysis has been applied in traffic forecasting (Chung, 2003; Danech-Pajouh and Aron, 1991; Wild, 1994). Days or periods of a day are grouped by means of cluster analysis and subsequently detected traffic profiles are matched with the most similar profiles in a cluster (pattern recognition/ pattern matching), or regression models are fit to forecast traffic volumes. In a similar way, clusters can also be used for the imputation of missing or invalid traffic data. Literature on AADT estimation on the basis of short term traffic counts (e.g. Bellamy, 1978; Sharma and Werner, 1981; Li et al., 2003) exploits spatial patterns to estimate traffic volumes on locations without continuous monitoring. Venkatanarayana et al. (2006) use cluster analysis for the detection of 'abnormal' traffic patterns. They applied a clustering algorithm to group traffic profiles that represent a period of the day and define patterns that are classified to a small cluster as 'abnormal'. Pribyl and Pribyl (2005) apply cluster analysis for the mining of loop detector data. They group detectors that show similar daily traffic patterns to limit the amount of data that has to be stored. Likewise, days can be grouped to limit the amount of historical data. Cluster analysis can also be used to account for variability in the data. In that case, cluster averages are used instead of averages over all locations or all days. Hicks and Niemeier (2001) and Niemeier et al. (2002) for example, model emissions and account for spatial variability by clustering locations. Similarly, temporal traffic patterns account for variability in time and besides for modelling emissions, the resultant patterns can also be used for transport modelling. Finally, cluster analysis has also been applied for traffic management. Smith et al. (2001) and Wang et al. (2005) for example, use cluster analysis for the development of signal timing plans. Finally, the insight into urban traffic patterns can be used for traffic monitoring.

This chapter discusses potential applications of the cluster analyses discussed in the previous chapter. Section 8.1 discusses the use of clustering for traffic monitoring. Sections 8.2, 8.3 and 8.4 respectively deal with the use of cluster analysis for traffic forecasting, traffic management and transport modelling. The chapter ends with conclusions.

8.1 Traffic monitoring

By analyzing historical traffic data insight can be obtained into existent traffic patterns. Cluster analysis is a way to structure historical data. The results of the cluster analyses can be used to monitor regular and irregular variations in daily flow profiles. Moreover, in case traffic data is collected for multiple years, also (long term) changes as a result of infrastructural or demographic developments can be monitored. This insight provides information on the time and location of bottlenecks in the network and can be used for traffic planning, land use planning and for taking adequate traffic management measures. In Almelo -or other cities for which cluster analysis is carried out- the results of the cluster analysis can be applied directly. Moreover, as discussed in Chapter 7, the recurrent traffic patterns can be explained for by variations in activity patterns or travel behaviour (modal choice). These variations also apply to other (Dutch) cities and thus can the results of this research also be used for traffic monitoring in other cities.

8.1.1 Direct use of results of cluster analysis

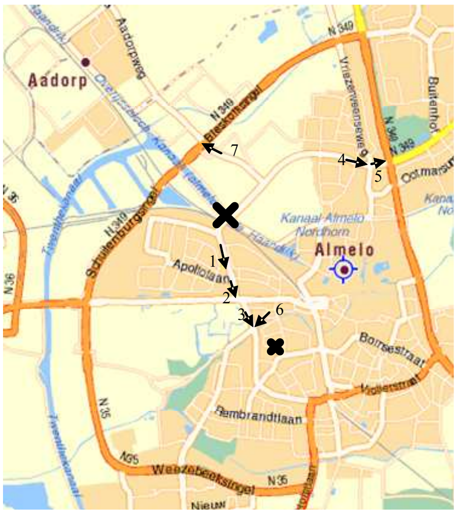
Local authorities estimate prevailing traffic loads by means of historical average daily traffic volumes or peak volumes. However, in Chapter 7 it was demonstrated that Fridays and days within (school) holiday periods show lower peak flows (especially a lower A.M. peak flow). As a consequence, Monday peak volume will be higher than the historical average working day peak volume and the historical average A.M. peak flow thus is not representative for a Monday morning peak. The classification of historical data enables a better estimation of the actual traffic volumes on different types of days (e.g. on different weekdays) on different locations. Also, insight is obtained into times of bottlenecks on different locations. Additionally, the classification of locations on the basis of their daily, weekly, seasonal and weather related variations provides insight into the locations of bottlenecks on different types of days. Finally, cluster analysis on a network level can provide more insight into the most important traffic patterns in the entire network or a part of the network the policy maker is interested in. As explained in Section 7.3, policy makers can define different sets of key locations dependent on the OD-relations or part of the network they want to analyse. The obtained insight into regular variations in traffic volumes can be used for traffic planning, land use planning and for taking adequate (traffic management) measures.

In Chapter 7 was concluded that traffic volumes -besides regular variations-also show irregular variations due to road works or events. A wide variety of techniques have been developed for the detection of outliers in the data that are caused by these atypical circumstances, varying from basic statistical methods to very advanced statistical, neural and machine learning techniques (for an overview see for example Hodge and Austin (2004)). The basic statistical methods (e.g. box plots) used in common practice can be used to detect dissimilar days, yet they are not able to detect different patterns that result from road works that last for multiple months. The same accounts for the detection of changes in traffic patterns due to infrastructural changes or land use developments. Since all these occasions are known by the local government, one could argue to link traffic data to data on road works etc, yet in most cities, data on road works and events are not collected in a central database (yet) (see Chapter 2). Besides, since there are many road works, events and infrastructural and land use development in cities, it will be complicated to built such a database that incorporates all processes and determines all effects on traffic volumes.

Outliers can also be detected by cluster analysis. Venkatanarayana et al. (2006) for example use cluster analysis for the detection of 'abnormal'[1] traffic patterns. For some locations in Almelo, one of the clusters appeared to show an 'abnormal' flow profile as well. Analysis has shown that road works or accidents were on the basis of these clusters. Moreover, in some cases, road works resulted in a separate cluster that showed a little higher, yet not an 'abnormal' traffic pattern. So cluster analysis can also be used for the detection of relatively small deviations in traffic volumes due to road works. In the long-term, cluster analysis probably also provides an easy way to detect changes in traffic patterns due to infrastructural or land use developments. It thus can be concluded that clustering provides an easy way to automatically detect irregular variations in traffic volumes. Moreover, since these days are classified to a separate cluster, effects of these situations on traffic volumes can be monitored and analysed as well (see Figure 8.1).

Summarized, the advantage of cluster analysis is that it enables a better estimation of the actual traffic volumes on a certain type of day and a certain location. Moreover, regular variations as well as the effect of (lengthy) road works, recurrent events and infrastructural and land use developments on traffic volumes can be monitored. Insight into traffic patterns on a network level enables adequate traffic management on network level. Note that all of the above concerns off-line traffic monitoring. In combination with real-time traffic data, the historical patterns can also be used for online traffic monitoring, for example for incident detection.

[1] An abnormal daily flow profile is defined as a flow profile that deviates substantially from daily flow profiles on most other working days.



source: www.viamichelin.com

From 28 February 2005 until 31 March 2005 the Elandsbrug (indicated by the big cross) was closed for traffic due to road works. This closure affected traffic volumes on different locations. At link 1, days from the road works period are classified to a separate cluster that shows much lower traffic volumes than the other clusters. Also at links 2, 3, 4 and 5 road works play a role in the classification, although days with road works are grouped together with holiday periods and Mondays and the differences in traffic volumes are smaller. At links 6 and 7, traffic volumes are higher as a result of road works. At location 6, days with road works are classified to a separate cluster, probably also due to road works at another alternative road (indicated by the small cross). The relative influence is smaller at location 7, at which one of the clusters contains relatively many, and not all, days from the road works period next to Thursdays.

Figure 8.1: Use of cluster analysis for analysing the effect of road works.

8.1.2 Application of results to other cities

As we concluded in Chapter 7, daily flow profiles vary by day of the week, holiday periods and – to a lesser extent – season. Since the underlying variations in activity patterns apply to the Netherlands in general, it is likely that also the variations in traffic volumes apply to other cities as well. So the results of the described analyses also provide insight into traffic patterns in other cities.

The amount of variation differs by location as it depends on the exact composition of the traffic. Moreover, the size of the effects in a city is also influenced by other factors, like the attractiveness on the city centre and the type of companies and industry. However, the results from the classification of locations on the basis of weekly patterns can be used as an estimation of the size of weekly variations for different types of roads. To make the results applicable to other cities, we calculated the percentage differences in relation to the average working day profile for different weekdays and different clusters:

$$\bar{\delta}^*_{kwt} = \frac{1}{Nl_k} \sum_{l=1}^{Nl_k} \frac{q_{lwt} - \bar{q}^{\text{work}*}_{lt}}{\bar{q}^{\text{work}*}_{lt}} \tag{8.1}$$

The resultant patterns are shown in Figure 8.2. Also the decrease of traffic in summer and the increase of traffic in case of rain on locations with relatively much short distance traffic most probably apply to other cities as well. Since the exact factors differed between the locations, it is difficult to quantify the effects for a new location, yet in general, car traffic is about 12% lower on summer days and about 9% higher on rainy days on locations with much (more than 50%) short distance traffic.

The resulting spatial traffic patterns can also be used for the estimation of average daily traffic volumes and peak volumes on the basis of short term traffic counts (also see for example Bellamy, 1978; Sharma and Werner, 1981; Hicks and Niemeier, 2001; Niemeier et al., 2002; Li et al., 2003). On the basis of the characteristics of a location and information on the type of travel that is served, one has to decide to what spatial group a link belongs. Subsequently, a short term traffic count can be adjusted using the average profile of the cluster to represent traffic volumes for an average working day or peak period. Hereby, traffic patterns on multiple temporal aggregation levels can be combined. Figure 8.3 shows an example of the use of spatial traffic patterns for the estimation of traffic volumes on a link on which short term traffic counts are executed.

8.2 Traffic forecasting

This section discusses the use of cluster analysis for traffic forecasting and imputation of missing or erroneous data. Subsection 8.2.1 describes the method

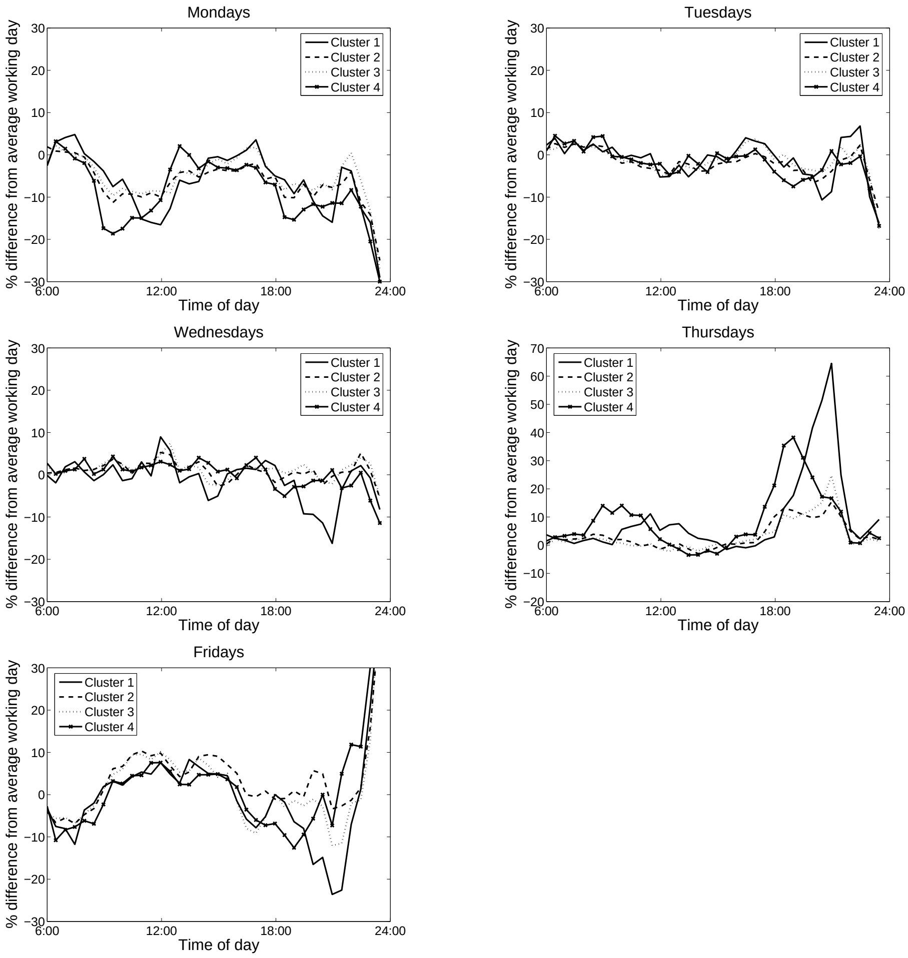


Figure 8.2: Percentage difference in relation to the average working day flow profile for different weekdays and different clusters.

Suppose we want to estimate the traffic volume between 8:00 and 9:00 on an average working day on a certain link l on the basis of the following short term traffic counts:

Monday	16:00 - 18:00	q = 600 vehicles
Tuesday	16:00 - 18:00	q = 700 vehicles

Further suppose that link l belongs to cluster 1 with regard to the daily flow profile and to cluster 2 regarding the weekly flow profile. When we assume that the average patterns of the clusters are applicable to link l, this results in the following factors:

$s_{\mathrm{l,work,16h-18h}}$ = 15.5%	$s_{\mathrm{l,work,8h-9h}}$ = 7.5%
$\delta_{\mathrm{l,Mon,16h-18h}}$ = -0.25	$\delta_{\mathrm{l,Tue,16h-18h}}$ = -0.03
$\hat{q}_{\mathrm{l,Mon,16h-18h}}$ = 600 veh	$\hat{q}_{\mathrm{l,Tue,16h-18h}}$ = 700 veh

The traffic volume from 16:00 to 18:00 on an average working day can be estimated using equations 4.18 and 4.26:

$$\hat{q}^{\mathrm{work}}_{\mathrm{l,16h-18h}} = \frac{Nt * s_{\mathrm{l,16h-18h}} * q_{\mathrm{l,w,16h-18h}}}{\delta_{\mathrm{l,w,16h-18h}} + Nt * s_{\mathrm{l,16h-18h}}}$$

where $w \in$ {Monday,Tuesday} and $Nt = 12$

To get a more accurate estimation, the volume is estimated using both traffic counts and the final estimation is the average of both values. This results in $\hat{q}^{\mathrm{work}}_{\mathrm{l,16h-18h}} = 702$ vehicles

Subsequently, the traffic volume between 8:00 and 9:00 on an average working day can be estimated using:

$$\hat{q}^{\mathrm{work}}_{\mathrm{l,8h-9h}} = \frac{s_{\mathrm{l,8h-9h}}}{s_{\mathrm{l,16h-18h}}} * \hat{q}^{\mathrm{work}}_{\mathrm{l,16h-18h}} = 340 \text{ vehicles}$$

Figure 8.3: Example of the use of the clustering of locations for estimating unknown traffic volumes.

that is subsequently applied in 8.2.2. Moreover, the method is assessed by comparing the results to the results of a classification according to weekdays and holiday periods. In 8.2.3 finally, a hybrid method is proposed that combines the results of the cluster analysis and a classification on the basis of the day of the week and holiday periods.

8.2.1 Method

The estimation of missing or future traffic volumes using temporal traffic patterns consists of two steps. First, it has to be determined to what cluster day d belongs. Subsequently, the missing or future traffic volume has to be estimated. These two steps are discussed here.

In case of a complete clustering, the selection of the most appropriate cluster is no problem. When for example, all Thursdays are classified to the same cluster, it is most likely that a new Thursday is also classified to this cluster. Unfortunately however, many of the clusters are not complete. The more incomplete the resulting clusters, the more difficult it is to select the right cluster. When for example, Thursdays are equally divided over 4 clusters and we do not know what factor is on the basis of this distribution, the probability is only 25% that we select the right cluster. In case traffic data is already collected for part of the day, pattern matching can be used to select the most appropriate cluster (Wild, 1994). Another possibility is to adjust the clustering in such a way that the classification becomes complete. In case that for example 80% of the Wednesdays are classified to the Monday and Tuesday cluster and 20% of the Wednesdays are classified to the Thursday cluster and the traffic volume on a future Wednesday has to be predicted, the future Wednesday could be classified to a new cluster containing all Mondays, Tuesdays and Wednesdays. Alternatively, new clusters could be created by a linear combination of the daily flow profiles of the clusters a certain type of day is classified to. In the example this would mean that a new profile is created of which the traffic volumes are 0.8*the average traffic volume of the Monday and Tuesday cluster + 0.2*the average traffic volume of the Thursday cluster. Figure 8.4 demonstrates these ways to adjust the clustering for one of the locations. Note that most methods for selecting the most appropriate cluster can be applied off-line.

After determination of the most appropriate cluster, the missing or future traffic volume has to be estimated. The most straightforward method is to use the average traffic volume over all days within the cluster for time period t. More advanced methods combine the average traffic volume of the cluster with available measurements on day d and/or upstream locations by applying a regression or ARIMA model (e.g. Danech-Pajouh and Aron, 1991; Van der Voort et al., 1996). Another possibility is to select the most similar traffic patterns within the cluster (pattern matching) and to estimate the traffic volume or travel time on day d as the average of the traffic volumes or travel times of these most similar patterns (e.g. Bajwa et al., 2004). Wild

Suppose that a cluster analysis at a certain link results in four clusters that show the following distribution over weekdays and holiday periods

Cluster	Mon	Tue	Wed	Thu	Fri	Holiday
1			2	1	23	
2	27	11	3	3		
3	4	24	31	31	13	
4	2	1	1		3	20

Different methods for selecting the most appropriate cluster result in the following:

Method 1	most likely cluster:	Mon, clu 2; Tue, clu 3 etc
Method 2	complete clustering:	4 clusters: Mon, Tue+Wed+Thu, Fri, Hol
Method 3	linear combination:	Mon: 27/33*clu1+4/33*clu2+2/33*clu3 etc.

Figure 8.4: Methods for selecting the most appropriate pattern.

(1997) combines both methods: he selects the most similar traffic pattern and combines the traffic volume of this pattern with the difference in the actual traffic volume between the pattern and the current day in case of a forecasting horizon of less than 1.6 hours.

A distinction in A.M. peak and P.M. peak traffic profiles may result in a more accurate prediction (Chung, 2003). In that case, a pattern thus is defined as a series of traffic counts that represent the A.M. peak period respectively the P.M. peak period on a certain day at a certain location. The described method can also be applied for traffic forecasting on the basis of peak period clusters, although the resulting groups might be different.

The described method can be applied both online and off-line. In general, in case of the imputation of missing or erroneous data, an off-line application satisfies (unless missing data has to be imputed in real-time), whereas in traffic forecasting an on-line application probably results in more accurate forecasts. An online application is needed when the most appropriate cluster is selected by means of pattern matching and/or when real-time traffic data are applied for the actual forecast.

8.2.2 Application and assessment

The method described in the previous subsection is demonstrated by applying it to a number of locations. Thereby, different methods for the selection of the most appropriate pattern are compared. Moreover, the method is assessed by comparing it to a classification on the basis of weekdays and holiday periods. The average standard deviation within the resulting groups is adopted as a measure for the quality of the classification. The smaller the standard deviation, the more representative the average daily flow profiles of the groups are for the days within the group.

Tables 8.1 and 8.2 show the average standard deviations after classification for six locations. From these tables a number of conclusions can be drawn. First

Table 8.1: Standard deviations resulting from cluster analysis for different locations. 0 is Standard deviation in case that the optimal cluster is selected in all cases. 1, 2 and 3 represent different methods for selecting the most appropriate pattern shown in Figure 8.4.

L	$\bar{q}$	Description of clustering			Standard deviation [veh/h]			
	[veh/h]	Nk	Factors	Complete	0	1	2	3
1	288	4	weekday, hol	quite	28.30	30.15	29.66	29.82
2	137	2	weekday	quite	21.94	22.18	22.09	22.18
3	140	3	week, hol, season	not	21.36	23.12	22.81	22.42
4	120	2	road works, week	very	21.95	23.02	22.96	22.09
5	155	2	road works, week	quite	23.08	23.48	23.36	23.17
6	396	4	week, hol, season	not	40.91	43.98	43.06	43.04

Table 8.2: Standard deviations from cluster analysis and classification by weekday and holidays.

Location	Average standard deviation [veh/h]		
	Cluster analysis		Weekday and holiday
	optimal	estimated	
1	28.30	29.82	28.67
2	21.94	22.18	20.29
3	21.36	22.42	21.23
4	21.95	22.09	24.98
5	23.08	23.17	24.41
6	40.91	43.04	42.50

of all, as expected, in case of an incomplete clustering, it is more difficult to estimate the most appropriate pattern. The difference in standard deviation between the optimal clustering and the adapted classification is larger for the locations with an incomplete clustering. Moreover, for most locations, the estimation of new daily flow profiles as a linear combination of the resultant clusters results in the lowest standard deviation. Therefore, this method is concluded to be the most appropriate for the selection of a cluster.

With regard to the difference in standard deviation between the clusters and the weekdays and holiday periods, it can be concluded that in general, both methods result in similar standard deviations. At locations 2 and 3, cluster analysis results in larger standard deviations. This is due to a small number of clusters[2]. A classification on the basis of weekday and holiday periods results in

[2]In case that the variation is relatively large between the most similar types of days, the optimal number of clusters is small (see Section 9.2).

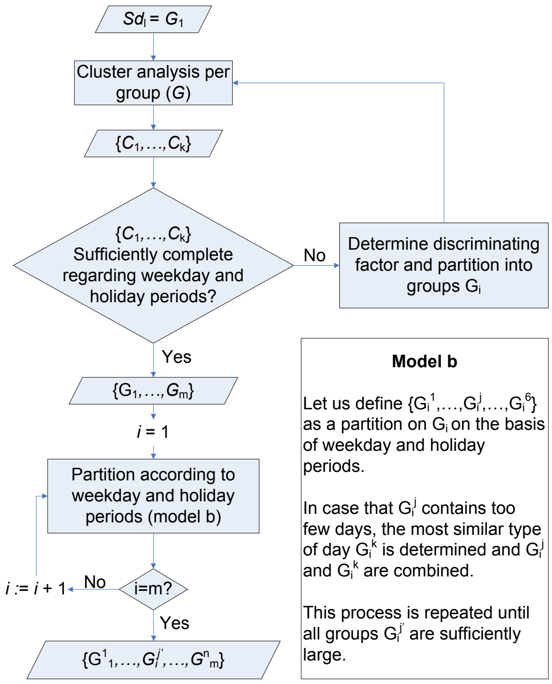


Figure 8.5: Model that combines cluster analysis and classification on the basis of weekday and holidays.

six groups, resulting in lower standard deviations. At locations 4 and 5, cluster analysis results in clearly smaller standard deviations, even though the number of clusters is small. At these locations, road works occurred. In a classification on the basis of weekday and holiday periods, these road works are not taken into account.

Summarized, the advantage of cluster analysis is that it can easily detect variations that are due to special circumstances like road works, whereas the disadvantage is that in some cases not enough groups are distinguished. A hybrid method that combines the strengths of the cluster analysis and the classification on the basis of weekday and holiday periods probably results in smaller variations within the groups. This method is described in Section 8.2.3.

8.2.3 Hybrid model

This section describes and evaluates a hybrid method that combines cluster analysis with a classification into weekdays and holiday periods in order to define distinctive, recurrent and representative traffic patterns. The method is depicted in Figure 8.5. When all clusters C_1 to C_k are sufficiently complete regarding weekday and holiday periods, the dataset is partitioned on the basis of weekdays and holiday periods. This partitioning is performed in model b. In case that the clusters are not sufficiently complete, it is investigated what factor is responsible for the classification and days are grouped according to this factor. New cluster analyses are executed for the resulting groups G_i. Again, if the resulting clusters are not sufficiently complete, it is determined what other factor is on the basis of the clustering and days are grouped according to this other factor. This procedure is repeated until all resulting clusters are sufficiently complete. Then the groups G_1 to G_m are further subdivided on the basis of weekdays and holiday periods (model b). Figure 8.6 demonstrates the model for location 4 from Tables 8.1 and 8.2. From this example it can be seen that this new method results in a smaller standard deviation. An additional advantage of the hybrid model is that it is easier to forecast to what cluster a certain new day belongs to, thus it is easier to select the right group.

8.3 Traffic management scenarios

The resultant clusters can also be used for traffic management. The most common traffic management measure in urban areas is signal control. Signals can be controlled locally (isolated intersection signal control) or in combination with other signals (interconnected intersection control) (Klein, 2001). The simplest local isolated intersection control is pre-timed control. In that case, signals are controlled according to a predefined signal plan, regardless of the current traffic situation. In the Netherlands, vehicle actuated control is more common. In that case, green times are determined on the basis of actual traffic data. Green times are bounded by minimum and maximum thresholds, Interconnected intersection control intends to tune signals in such a way that platoons of vehicles can proceed along an arterial without stopping (arterial systems) or that the level of service is optimized on a network level (network systems) (Klein, 2001). According to Klein (2001) interconnected signal control can function in two general ways. The first method selects the signal timing plan that best matches current conditions from among a library of pre-stored plans. The second group of methods generate signal timing plans online and updates the plan incrementally at each signal cycle (e.g. SCOOT (Hunt et al., 1981), SPOT/UTOPIA (http://www.peektraffic.pl/index.php?nodeid=231/). Other traffic management measures that are applied in urban areas in the Netherlands are Variable Message Signs (VMS) and tidal flow lanes. In the future, electronic tolling systems may be implemented. Variable Message Signs can be used to reroute traffic in case of congestion or an incident and tidal flow lanes provide a method to redistribute capacity over both directions of a link dependent of the traffic volumes. Finally, road charging fares can vary in time and place to influence travel behaviour in order to reduce congestion.

Cluster analysis can be used for an effective implementation of these traffic management measures. Smith et al. (2001) and Wang et al. (2005) for example describe the use of cluster analysis for the identification of Time-of-

At location 4 from Tables 8.1 and 8.2 the cluster analysis resulted in 2 clusters:

	Mon	Tue	Wed	Thu	Fri	holiday
Cluster 1	13	13	13	12	3	11
Cluster 2	13	13	14	15	25	2

Since all Mondays, Tuesdays etc are divided over both clusters, the resulting clusters are clearly not complete regarding weekdays. Therefore, it is determined what other factor(s) is (are) on the basis of the classification. This factor appears to be road works on an alternative route. Days with road works and most Fridays are classified to cluster 2.

Days are grouped into two groups: Road works (G_1) and non-road works (G_2). For these groups, new cluster analyses are executed. These result in the following classifications:

road works	Mon	Tue	Wed	Thu	Fri	holiday
Cluster 1	3	10	12	13	3	
Cluster 2	10	3				
Cluster 3					10	
Total	13	13	12	13	13	

non-road works	Mon	Tue	Wed	Thu	Fri	holiday
Cluster 1	10	12	12	13	3	1
Cluster 2	3	1	1			10
Cluster 3					12	2
Total	13	13	13	13	15	13

For both groups, the cluster analysis results in a quite complete classification. Furthermore, both for G_1 and for G_2 all weekday groups as well as the holiday group (for G_2) contain enough days. Therefore, G_1 is further subdivided into $G_1^{Mon}, ..., G_1^{Fri}$ and G_2 is further subdivided into $G_2^{Mon}, ..., G_2^{hol}$. Figure 8.7 shows an example for which the weekday groups do not contain enough days.

The average standard deviation after this classification is 18.68 veh/hour. This implies a decrease compared to the normal cluster analysis of more than 15% (the standard deviation in case of cluster analysis was 22.09).

Figure 8.6: Example of the use of the hybrid model for estimating unknown traffic volumes.

Suppose that a group (G_i) contains 11 Mondays, 8 Tuesdays, 11 Wednesdays, 12 Thursdays, 3 Fridays and 3 holidays. Further suppose that each group (G_i^j) should contain at least 10 days.

In that case, the Monday group is large enough. However, the Tuesday group is too small. The daily flow profile of an average Tuesday is compared to the average flow profiles of the other weekdays and holiday periods. As a measure for the distance between two profiles, the sum of squared deviations could be used. Suppose the sum of squared deviations is smallest for the Monday flow profile. Then, Tuesdays are grouped together with Mondays. The Monday-Tuesday group contains enough days. Also the Thursday group contains enough days. The Friday group is too small. Suppose that the average Friday flow profile appears to be most similar to the holiday profile. Then, Fridays are grouped with holidays. The Friday-holiday group is however still too small. Therefore, the average daily flow profile of this group is compared with the other flow profiles. Suppose the profile is most similar to the Wednesday profile. Then, the Friday-holiday group is combined with the Wednesday group.

The partitioning on the basis of weekday and holiday periods results in three groups ($G_i^{j'}$): (1) Mondays + Tuesdays, (2) Wednesdays + Fridays + holidays and (3) Thursdays.

Figure 8.7: Example of the use of model b of the hybrid model.

Day (TOD) break points for traffic signal plans. Cluster analysis is applied for the determination of instants at which traffic volumes on the approaching lanes change and a different signal timing plan should be operational. A pattern is defined as a series of traffic volumes (possibly combined with occupancies) for a certain time interval for multiple phases or approaching lanes, i.e:

$$\mathbf{P}_{IDt} = (q_{IDt,1}, ..., q_{IDt,p}, ..., q_{IDt,Np}) \tag{8.2}$$

where I is an index for intersection, p is an index for approaching lane/ turn and Np is the number of approaching lanes/ turns (also see Figure 8.8). A pattern thus defines a certain traffic state instead of a daily flow profile, and time intervals instead of days are grouped.

In addition, the results from the cluster analysis described in this thesis can be used for the development of signal plans for different types of days. TOD break points can be determined for the different types of days resulting from the cluster analysis or the hybrid model proposed in Section 8.2.3. Since the distribution of the traffic over the approaching lanes may differ by type of day, a new cluster analysis using a different pattern definition might lead to better results. To be able to take account of variations in distribution of the traffic over the approaching lanes, a pattern is defined as a series of daily flow profiles

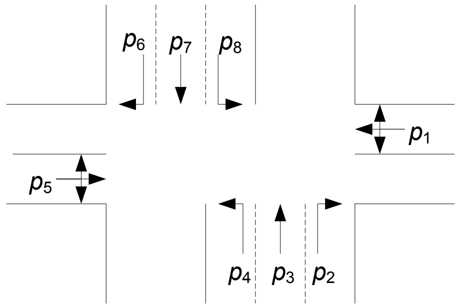


Figure 8.8: Approaching lanes at an intersection.

for different approaching lanes, i.e.:

$$\mathbf{P}_{Id} = (q_{Id,1,1}, ..., q_{Idpt}, ..., q_{IdNpNt}) \tag{8.3}$$

In case of isolated signal control, the resulting clusters can be used for the development of pre-timed signal plans or for the setting of maximum green times and other parameters that are used for actuated signal control. In case of interconnected signal control, the clusters can be used for the development of signal plans for the library of pre-stored plans.

Furthermore, the results of the cluster analysis described in this thesis can be applied on a more strategic level, for the selection of priority routes on different types of days or for the development of traffic management strategies that solve problems in one part of the network by taking measures in another part of the network (e.g. creating buffers at the border of the city to prevent congestion in the city centre). Besides traffic signal timing, rerouting strategies, tidal flow lanes, varying road prices and other measures (e.g. parking guidance, parking rates, public transport frequency and other travel demand management measures) could be part of these management strategies. Different scenarios could be developed for different types of day or for special circumstances like football matches of Heracles. Also for the classification of days on network level, probably a combination of clusters and day of the week and holiday groups provide the best results.

The management strategies for the different types of day can be developed off-line. Additionally, by using real time traffic data, the scenarios could be fine-tuned on-line according to the actual traffic volumes.

8.4 Transport modelling

The applications discussed thus far make direct use of the results of the cluster analysis. For all three applications, cluster analysis can also be applied indirectly, using a planning model. A planning model is a useful tool in transport planning, traffic management and traffic forecasting. The model provides information on traffic volumes on locations without monitoring and therefore can be used for a sound overview of the traffic situation on a network level. Some of the more advanced traffic information centres discussed in Chapter 2 also contain a traffic model to be able to provide traffic information for the entire network (e.g. MIZAR Automazione, 1998; Fellendorf et al., 2000; Kellerman and Schmid, 2000).

The most commonly used planning model is the classic four-stage transport model described by Ortuzar and Willumsen (1994) that consists of the following stages: (1) trip generation, (2) trip distribution, (3) modal split and (4) assignment. The model is calibrated for a certain situation that consists of a (1) traffic network, (2) type of day (e.g. working day, weekday) and (3) time period (e.g. 24-hour, A.M. peak, P.M. peak). The trip generation and distribution phase result in a OD-matrix. OD-matrices can be estimated for different transport modes. Here, only car traffic is considered. In that case, the matrix can be assigned to the network, resulting in link volumes. In common practice, information on trip generation (generations and attractions of each zone) is combined with deterrence functions to obtain a basis OD-matrix. This matrix is calibrated for a certain type of day and a certain time period using traffic counts and other available data (e.g. license plate surveys). In literature, multiple methods are proposed for this matrix calibration, for example the entropy maximizing approach or information minimization approach (Van Zuylen and Willumsen, 1980), the generalized least squares approach (Cascetta, 1984), a maximum likelihood model (Spiess, 1987) and the Bayesian statistical approach (Maher, 1983). The quality of the calibrated OD-matrix depends on the quality of the input data.

In case that the model is calibrated for (a certain time period of) an average working day and traffic data is collected continuously, cluster analysis can be used to filter out road works and other special circumstances that influence average working day traffic volumes. The resulting average traffic volumes provide a more accurate estimation of the actual situation (also see Section 8.1) and thus lead to a more accurately calibrated matrix for an average working day. In case that only short term traffic counts are available, the results from the clustering of locations can be applied to translate these short term traffic counts into working day averages (as explained in Figure 8.3).

Besides, cluster analysis can also be used for the definition of different day types for which the model is calibrated. In this case, average traffic volumes from different clusters (determined by a cluster analysis on network level) can be used for representing traffic counts on different types of days. Again, instead of clusters, also groups resulting from the hybrid model proposed in Section 8.2 can be used. Also in this case, the results of the clustering of locations can be used for the estimation of traffic volumes on certain types of days on the basis of short term traffic counts (see Figure 8.3). Furthermore, the accuracy of the calibrated model possibly can be improved by adding insight obtained from the cluster analyses as extra input. We found -for example- that on Friday, peak flows are lower and off-peak flows are higher than on an average working day. The differences for different types of locations are shown in Figure 8.2. These values apply for links, yet they also give an estimation of differences in the amount of trips from a certain type of origin or to a certain type of attraction. The attraction totals of all residential zones could for example be reduced by 6% for the P.M. peak period on Fridays.

The calibrated models for different types of days can be used for traffic monitoring, the estimation of travel times on different types of days and for the generation of traffic scenarios that can be used for the development of traffic management scenarios.

8.5 Conclusions

In this chapter potential applications of cluster analysis and the resultant clusters were discussed. First of all, the obtained insight into existent traffic patterns can be used for traffic monitoring. Cluster analysis can be used to monitor regular and irregular variations in traffic profiles and provides a better estimation of the actual traffic volume on a certain type of day. The advantage of cluster analysis over methods that group days on the basis of pre-defined types of days is that cluster analysis provides an easy way to detect atypical circumstances that influence traffic volumes. Besides, the effect of road works, recurrent events and infrastructural and spatial developments on traffic volumes can be monitored. The results of the cluster analysis carried out in this research can also be used for traffic monitoring in cities without archived traffic data, because part of the results can be explained for by variations in activity patterns and travel behaviour that also apply to other (Dutch) cities. The results of the spatial analysis provide more insight into temporal patterns for different types of locations. Besides, the spatial traffic patterns can also be used for the estimation of average traffic volumes on the basis of short term traffic counts.

Secondly, cluster analysis can be used for traffic forecasting and for the estimation of missing or erroneous data. First, the most appropriate cluster has to be selected and second, the missing or future traffic volume has to be estimated, possibly by combining the historical data on which the clusters are based with real time data. With regard to the first step, different methods are proposed for the selection of the most appropriate cluster. These methods are applied to six locations and the methods are assessed by comparing the average standard deviation within the clusters. Moreover, the standard deviation is compared to the standard deviation that result from a classification on the basis of weekday and holiday period. From this assessment it is concluded that, in normal circumstances, a classification on the basis of day of the week and holiday periods results in similar or even smaller standard deviations than a classification resulting from cluster analysis. This can be explained by a relatively small number of clusters that is due to relatively large variations between the most dissimilar groups of days. However, in case of special circumstances – like road works – cluster analysis results in smaller standard deviations, even when the number of clusters is small. Therefore, a hybrid model is proposed that combines the strengths of both classifications. Cluster analysis is applied to detect clusters that are due to atypical circumstances and when possible the resultant groups are further classified on the basis of weekday and holiday periods. The method is applied to one of the locations and results in a lower standard deviation.

The resultant clusters or groups resulting from the proposed hybrid model can also be used for traffic management scenarios. First of all, the groups can be used for the development of signal plans for different types of days or peak periods. Another pattern definition, featuring daily flow profiles for different turns of an intersection is more appropriate in this case. On a more strategic level, clusters or groups can be used for the development of traffic management strategies that optimize traffic on a network level. Besides traffic signal timing, rerouting strategies (using Variable Message Signs), tidal flow lanes, varying road prices and other (travel demand management) measures could be part of these management strategies.

Finally, the results of the cluster analysis can be used for transport modelling. In case that traffic data is collected continuously, the average traffic volumes on 'normal' days (days without road works or events) can be used for the calibration of an OD-matrix for an average working day. In case that only short term counts are available, the results from the clustering of locations can be applied to translate these short term traffic counts into working day averages. Furthermore, cluster analysis can be used for the definition of different day types for which the model is calibrated. Average traffic volumes of different clusters or groups resulting from the hybrid model can be used as a representation of traffic volumes on different types of days. Also in this case, the results of the clustering of locations can be used for the estimation of traffic volumes on certain types of days on the basis of short term counts. Finally, the accuracy of the calibrated model can possibly be improved by adding insight obtained from the cluster analysis as extra input. The model estimates traffic volumes on locations without monitoring and thus provides a clear picture of the traffic situation in the entire network that can be used for traffic monitoring, travel time forecasting and traffic management.

Chapter 9

Evaluation and Discussion

The reason for applying the method to the city of Almelo was twofold. Besides obtaining insight into urban traffic patterns, it is also important to verify whether the method functions adequately. This chapter evaluates the functioning of the proposed method and discusses whether the method produces plausible results and is useable. Furthermore, Section 9.2 discusses the choices with regard to the clustering method and optimal number of clusters, the consequences of the used data on the clustering results and the transferability of the results to other cities. The chapter ends with conclusions.

9.1 Functioning of the analysis framework

In general, in can be concluded that the analysis framework presented in Chapter 4 is functioning adequately. Clustering of the daily flow profiles at different links resulted in plausible classifications. The temporal patterns found in Chapter 7 can be explained by variations in activity patterns, route choice, and modal choice and agree with the literature discussed in Chapter 3. In principle, the method can easily be applied to other cities, although it is labour-intensive to execute cluster analysis for all links in the network.

Also most analyses of spatial traffic patterns led to satisfying results. The classifications of locations on the basis of their average daily flow profiles, their weekly variations, their seasonal variations and their rain factors resulted in plausible clusters that could be explained by differences in the distribution over travel motives and by the percentage of short distance traffic. Only the classification of the links on the basis of their temporal clusters did not lead to distinctive, recurrent and representative clusters. Not so much the types of days appear to differ between types of locations, yet the amount of variation between the types of days. Therefore, spatial variations in traffic patterns can best be analysed by clustering locations on the basis of their daily, weekly (and possibly seasonal) patterns.

Finally, also the clustering of daily flow profiles on a network level resulted in a plausible classification. In Chapter 4, two methods were proposed for the analysis of traffic patterns on a network level. One of the methods exploits the results of the clustering of links on the basis of their temporal clusters. Since this clustering did not lead to a satisfying classification, the method for the analysis of network patterns is not applied. The other method defines patterns on a network level by combining daily flow profiles of multiple locations that are selected as key location. The groups that result from the clustering of these patterns can be explained by variations in travel behaviour. It has to be noted that the results of the cluster analysis depend on the locations that are selected as key location and that locations with higher traffic loads have more influence on the classification than locations with lower traffic loads.

Besides obtaining insight into the main traffic patterns, the intention of applying cluster analysis was also to determine typical traffic patterns that serve as a basis for traffic forecasting, traffic management and transport modelling. In Chapter 8 it is concluded that the advantage of clustering over classification on the basis of pre-defined factors is, that it easily detects variations that are caused by irregular factors such as events and road works. The disadvantage is that in some cases, insufficient patterns are distinguished as a result of which the variation between the days within a cluster is still quite large. Therefore, a hybrid model is proposed that combines a classification on the basis of weekday and holiday periods with a classification on the basis of cluster analysis. Groups resulting from this hybrid model show smaller standard deviations and are thus useful as a basis for traffic forecasting, traffic management and transport modelling scenarios.

9.2 Discussion

9.2.1 Clustering algorithm

In this research a Ward's hierarchical clustering algorithm is applied. The main reason for choosing an hierarchical algorithm is that the optimal number of clusters does not have to be chosen in advance (see Section 4.1). The main disadvantage of an hierarchical method is that it possibly results in a non-optimal grouping, because results from previous steps of the clustering procedure are fixed. This section investigates the effect of the selection of the clustering algorithm on the resulting classifications.

To investigate the consequence of selecting an hierarchical instead of a partitional method, first it is determined what percentage of days is classified to a non-optimal cluster. A day is not classified to the most appropriate cluster in case that the distance to the average daily flow profile of the selected cluster is larger than the distance to the average daily flow profile of another cluster. For each location, it is determined what percentage of the days is not classified to the most appropriate cluster. On average, about 5% of the days appeared to be classified to a non-optimal cluster. However, in general, the difference in distance between the selected cluster and the most appropriate cluster appeared to be small.

Secondly, for five randomly selected locations, days are clustered by means of a K-means algorithm and the results of this commonly used partitional clustering algorithm are compared with the results of the Ward's hierarchical algorithm. Thereby, the number of clusters is adapted from the Ward's algorithm. Table 9.1 describes the locations and resulting classifications of the Ward's algorithm whereas Table 9.2 compares the results of the Wards algorithm with the K-means algorithm. First, the proportion of days that is classified to different clusters (Δ) is determined. Second, the factors on the basis of the clusters are compared for both methods. Third, the ratio F that represents the decrease in standard deviation after classification is compared.

Table 9.1: Locations for which the results of different clustering algorithms are compared.

Location	Description of results Wards algorithm		
	Nk	Factors	Complete
K24 from K38	4	holiday + weekday	quite
K30 from K36	3	events + combi of weekday, season and holiday	not
K31 to K49	2	road works	yes
K28 from K30	4	combination of holiday, weekday and season	not
K6 from K5	3	road works at two different locations	yes

Table 9.2: Comparison between Ward's hierarchical algorithm and K-means algorithm.

Location	Δ	Factors	Ratio F	
			Ward's	K-means
K24 from K38	22%	same	1.28	1.29
K30 from K36	30%	no events cluster	1.19	1.14
K31 to K49	0%			
K28 from K30	22%	same	1.20	1.21
K6 from K5	2%	same	1.33	1.34

Δ represents the proportion of days that are classified to different clusters. The ratio F is a measure for the decrease in standard deviation after classification (see Chapter 4).

From Table 9.2 it can be concluded that at two out of the five selected locations, the K-means algorithm resulted in similar clusters as the Ward's hierarchical algorithm. For the other three locations, the resulting clusters were dissimilar: 20% to 30% of the days were classified to a different cluster. However, the factors on the basis of the clusters are similar for both algorithms. Especially days from mixed and incomplete clusters are classified to different clusters. Also the ratio of the standard deviation before and after classification is similar for the K-means and the hierarchical algorithm. The K-means algorithm is a commonly used partitional algorithm and we expect the results of this algorithm to be representative for partitional algorithms in general. Since the main results of the K-means algorithm are similar to the results of the Ward's algorithm, we state that an hierarchical algorithm is appropriate for the research described in this paper.

To examine the effect of choosing a Ward's algorithm instead of another hierarchical clustering algorithm, the results of the Ward's algorithm are compared to results of the other commonly used hierarchical algorithms that were discussed in Chapter 4. The comparison is executed for the locations described in Table 9.1. Again the number of clusters is held constant. Table 9.3 shows the results. From the table can be concluded that the complete linkage

Table 9.3: Comparison between Ward's algorithm and other hierarchical algorithms.

Location	Method	Δ	Factors	Ratio F
K24	Ward's		holiday + weekday	1.28
	UPGMA	31%	same + outliers	1.25
	Within groups	16%	same	1.29
	Complete	16%	same	1.26
	Single	52%	holiday + outliers	1.10
K30	Ward's		events + combi	1.19
	UPGMA	55%	weekday + outliers	1.10
	Within groups	11%	same	1.16
	Complete	16%	same	1.17
	Single	45%	outliers	1.02
K31	Ward's		road works	1.41
	UPGMA	0%		
	Within groups	1%	same	1.40
	Complete	0%		
	Single	25%	outlier	1.02
K28	Ward's		combi	1.20
	UPGMA	54%	weekday + outliers	1.06
	Within groups	24%	same	1.17
	Complete	13%	same	1.19
	Single	60%	outliers	1.03
K6	Ward's		road works	1.33
	UPGMA	1%	same	1.33
	Within groups	4%	same	1.34
	Complete	6%	same	1.32
	Single	35%	road works + outliers	1.11

algorithm and the within groups algorithm result in similar classifications, whereas the single linkage algorithm and -to a lesser extent- the average linkage between groups (UPGMA) algorithm result in dissimilar classifications. In general, the single linkage algorithm results in a few small clusters with one or a few days and a large cluster with all remaining days. At three locations, the same applies for the between groups average linkage algorithm. The standard deviations after classification are clearly larger (expressed in a relatively small ratio F) for the single linkage algorithm and at some locations also for the between groups algorithm. Therefore it is concluded that these algorithms do not result in a satisfying classification. The other investigated methods - complete linkage and within groups clustering- result in similar classifications and are thus appropriate. In general, it can be concluded that methods that aim at minimizing the variation within the clusters and thus result in tight clusters of similar cases are most appropriate for this research.

9.2.2 Optimal number of clusters

In this research, the optimal number of clusters is determined by means of a dendrogram that is provided by SPSS. To determine the optimal number of clusters, the number of clusters is plotted against the rescaled distance cluster combine (see Figure 9.1 for an example). The optimal number of clusters is that number for which an extra step in the clustering procedure would lead to a more than proportional increase in the distance (the elbow criterium) (Nowotny et al., 2003). A consequence of this criterium is that the number of clusters

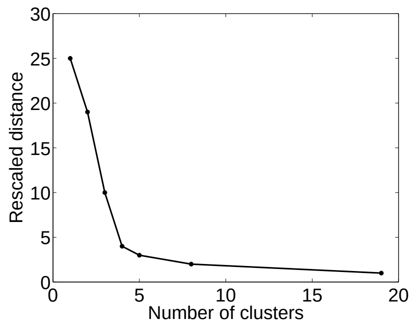


Figure 9.1: Example plot of the number of clusters against the distance coefficients. In this case the optimal number of clusters is 4.

depends on the magnitude of the differences in daily flow profiles that occur at a certain location. In case that for example road works took place that caused very low traffic volumes, the difference in daily flow profiles between the road works period and the period without road works is relatively large compared to other differences. In that case, the final step of the clustering procedure (from two clusters to one cluster) would result in a relatively high increase in distance and the optimal number of clusters would be two. Smaller differences between other types of days are not detected in that case.

The advantage of the dependence of the number of clusters on the size of the prevailing differences is that the cluster analysis determines the most distinctive traffic patterns at a certain location. This provides insight into the most important variations in traffic volumes. The disadvantage is that the variations within the clusters can still be quite large. This disadvantage mainly plays a role in case that the clusters are used as a basis for traffic forecasting, traffic management, or transport modelling scenarios.

In general it can be stated that the variance within the cluster decreases when the number of clusters increases, i.e. the higher the number of clusters, the more representative the average daily flow profile of a cluster for all days within the cluster. Thus, in case that clusters are used as a basis for traffic management, traffic forecasting or transport modelling scenarios, a distinction into more clusters is preferred, although the number of clusters should not be too high either to limit the number of scenarios. The number of groups can be increased by selecting a higher number of clusters. The main question in that case is what is the most appropriate number of clusters. Alternatively, in case of for example road works, new cluster analyses can be carried out for the road works group and the non-road works group. Finally, days could be grouped according to the hybrid model that is proposed in Chapter 8. The advantage of this hybrid model is that for each day it can be predicted to which group it should be classified, whereas in case of a cluster analysis this cannot always be predicted. Therefore, we recommend to apply the hybrid model proposed in Chapter 8 for the determination of typical traffic patterns that serve as a basis for traffic forecasting, traffic management of transport modelling scenarios.

9.2.3 Available traffic data

At the time of the data analysis, only one year of valid traffic data was available. Moreover, as a result of missing and erroneous data, not on all days valid traffic data is available for all links. Finally, in Chapter 5 it is shown that for most links, the available traffic data is unevenly distributed over the seasons, being less in summer and autumn. These limitations concerning the available traffic data limit the content of the analysis and possibly also distort the results.

With regard to the content of the analysis, first of all it is not possible to investigate long term changes in traffic patterns. Moreover, in case no or only few data is collected for one or more of the seasons, it is not possible to investigate the seasonal effects. The classification of locations on the basis of their seasonal patterns provides some insight into seasonal variations at different types of locations and can be used to estimate variations at locations for which no or only few data is available for one or more of the seasons. Finally, also within seasons traffic patterns might fluctuate. Variations between months (or weeks) within seasons were not investigated in this research due to the limited amount of available traffic data.

Missing or removed data might distort the results of the cluster analyses. This distortion probably is worse in case that the missing data is unevenly distributed over the seasons or days of the week. From Chapter 5 it was concluded that the available data is evenly distributed over the weekdays at all locations. However, the distribution over seasons is less even. The absence of summer or autumn days might mask the occurrence of a summer or autumn cluster. Besides, it might result in a different grouping of the other available data.

The sensitivity of the clustering for missing data and the effect of an uneven distribution of the available data is investigated by repeating the clustering for different sets of days at five locations for which relatively much valid traffic data was collected. The characteristics of these locations and their classifications are summarized in Table 9.4. The sensitivity of the clustering for missing data is examined by repeating the cluster analysis for 90% and for 50% of the original dataset that is randomly selected. The results of this analysis are summarized in Table 9.5. The effect of an uneven distribution of the data over the seasons is investigated by repeating the clustering for a new set of days that is evenly distributed over the seasons and the days of the week. The results of this analysis are summarized in Table 9.6.

Table 9.4: Locations for which the effects of different datasets are investigated.

Location	Nd	Nk	Factors
1: K32 from K33	159	3	road works + weekday
2: K30 from K3	122	3	weekday + holiday
3: K6 from K5	147	3	road works at two different locations
4: K35 westbound	137	4	weekday + holiday
5: K43 from K39	200	4	weekday + holiday

Table 9.5: Sensitivity of clustering results to missing data.

Loc	Resulting clusters				Δ		Ratio F		
	Nk		Factors						
	90%	50%	90%	50%	90%	50%	100%	90%	50%
1	3	3	same	same	1%	15%	1.29	1.27	1.30
2	$4^\star$	$4^\star$	+ season	+ season	1%	0%	1.22	1.29	1.29
3	3	$2^\dagger$	same	road works	2%	1%	1.33	1.29	1.22
4	4	$5^{\dagger\dagger}$	same	same	8%	24%	1.24	1.24	1.36
5	$3^\triangle$	4	same	same	16%	33%	1.22	1.18	1.28

*A cluster with Mon, Tue, Wed and hol is split into two clusters on the basis of a combination of weekday and season. †One of the road works clusters disappeared. ††One day is classified to a separate cluster. $^\triangle$A cluster with Tue, Wed and Thu is combined with a cluster with Thu and Fri and most Tue are classified to the Mon cluster.

From Table 9.5 it can be concluded that at individual locations, the results of the cluster analysis may change, especially when only 50% of the data is selected. In some cases, the new clustering results in a larger ratio, and thus in a larger decrease in standard deviation. In other cases, the ratio is smaller then for the original dataset. Moreover, in some cases, relatively

Table 9.6: Effect of an uneven distribution of the data over seasons on the results.

Location	Nd	Resulting clusters		Δ	Ratio F	
		Nk	Factors		Old	New
K32 from K33	64	3	same	22%	1.29	1.29
K30 from K3	52	4*	weekday, holiday, season	2%	1.22	1.27
K6 from K5	103	3	same	7%	1.33	1.27
K35 westbound	66	4	same	15%	1.24	1.28
K43 from K39	107	3†	same	12%	1.22	1.22

*A cluster with Mon, Tue, Wed and hol is split into two clusters on the basis of a combination of weekday and season. †A cluster with Tue, Wed and Thu is combined with a cluster with Thu and Fri and most Tue are classified to the Mon cluster.

large proportions of the days are classified to different clusters and sometimes clusters are combined or split up. However, in general, the factors that are determinative for the classification are similar. Moreover, the ratios F are in the same order of magnitude. Thus, the general conclusions are not distorted by missing data. From Table 9.6 it can be concluded that the general results of the cluster analysis are barely influenced by an uneven distribution over the data over the seasons, although also for this analysis the results of the individual classifications may differ. We have to note that we only examined locations for which traffic data is collected for more than 15 days in each season. As was mentioned before, in case that only few or none traffic data is collected during one of the seasons, a seasonal cluster might be undetected.

9.2.4 Influence of weather on traffic

Regarding the influence of weather on traffic, we expected a larger influence of rain on traffic volumes. Besides, a number of limitations with regard to the available data limit the content and possibly also the outcome of the analysis.

From the temporal analysis of traffic patterns it is concluded that at most locations, rain is not a decisive factor for the clustering. At only 10% of the locations, the distribution over wet and dry days differed between the clusters. It was expected that rain would have a larger influence, especially since Hagens (2005) found a modal shift from bike to car in case of rain, resulting in more car trips. The main reason for the fact that Hagens (2005) found a larger influence of rain appears to be that he focussed on short-distance traffic. In Chapter 7 we concluded that traffic volumes increase on wet days on locations with relatively much short distance traffic, whereas traffic volumes on locations with relatively little short distance traffic are hardly influenced by rain. Besides, regarding the results of the temporal clustering, again the optimal number of clusters plays a role. In case that the influence of weather on traffic volumes is small compared to other factors like weekday, wet and dry days are classified to the same cluster. Finally, also the limited number of pairs of wet and dry days may play a role.

The influence of rain is analysed by means of a matched pair analysis to correct for other factors. Moreover, since weather data from a station at approximately 20 km from Almelo is used for the analysis, only days on which rain occurred for more than 2 hours at an average intensity of > 0.5 mm/hour were classified as wet. As a consequence of these strict requirements, the number of pairs of wet and dry days is small for most locations. Therefore, no statistical tests could be applied to prove a significant influence of rain[1]. Moreover, the analyses are executed on a 24 hour aggregation level, so the time of rain can differ between wet days, being the night for one wet day and the morning for another wet day.

To investigate the effect of rain on the cluster a day is classified to more thoroughly, it is recommended to repeat the analyses when more data is available. Moreover, to obtain more insight into the influence of rain on traffic volumes it is recommended to use data on a lower temporal and spatial aggregation level (for example collected by amateur weather stations).

9.2.5 Congestion

As mentioned in Chapter 2, traffic volumes do not represent travel demand in case of congestion, because they are limited by capacity restraints. In case of congestion, a link cannot adequately process all traffic as a result of which traffic volume is constant at its capacity during the entire period congestion is present (for example a peak period). This results in a flattened daily flow profile. Directly after this period the delayed traffic is processed, resulting in relatively high traffic volumes for this period. In case of (recurrent) congestion, actual differences in travel demand patterns could be eliminated due to this deformation.

In Almelo, (recurrent) congestion hardly occurs. Therefore, the traffic volume patterns adequately represent present travel demand (and maybe sometimes variations in capacity in case of very low traffic volumes due to road works). Moreover, since the general travel demand patterns (that are caused by variations in activity patterns) apply to other cities as well, the found patterns can be exploited to derive travel demand patterns at locations at which congestion occurs. In this way, the amount of congestion can be estimated for these locations.

9.2.6 Application of results to other cities

As mentioned in the previous chapter (Section 8.1), part of the insight obtained in this research can be applied to other cities as well. However, it was also mentioned that the exact variations may differ between cities, since they probably depend on location specific factors like the attractiveness of the city center and the type of companies. In order to obtain more insight into the

[1] Chi square tests can only be applied in case that all expected frequencies are larger than 1 and maximum 20% of the expected frequencies are smaller than 5 (Huizingh, 2002).

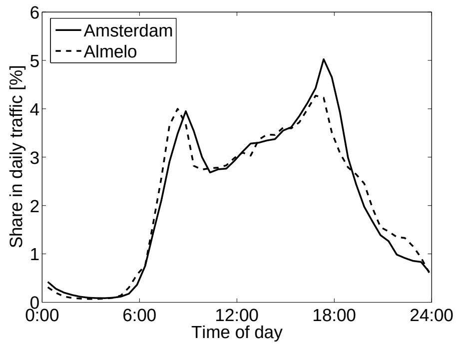


Figure 9.2: Comparison of the general shape of the daily flow profile between Almelo and Amsterdam Zuidoost. Adapted from Houtriet (2006).

influence of location specific factors on the general traffic patterns, the described analysis should be applied to different cities.

Houtriet (2006) studied variations in traffic volumes in Amsterdam Zuidoost and compared the results from this area to the results we found in Almelo. Figure 9.2 shows the average working day flow profiles and from this figure it can be concluded that the A.M. peak starts a little later and the P.M. peak is somewhat higher in Amsterdam. The difference in A.M. peak time is probably due to a difference in the type of commuter traffic. In Amsterdam, the employment in service industry is relatively high, whereas in Almelo, relatively many people work in industry. The cause of the difference in the height of the P.M. peak is less clear. Possibly the difference are partly due to rerouted traffic that avoids the congested highways around Amsterdam.

We also executed cluster analyses on the Amsterdam traffic data. The influence of events on the clustering results appears to be larger in Amsterdam. This is due to the fact that many major events (like national and international football matches and concerts) take place in Amsterdam Zuidoost. Moreover, during part of the analysis interval, major road works were carried out on a nearby highway, that had an impact on the traffic patterns in the area. As a result of these road works and events, cluster analyses at most locations result in a road works cluster, one or more event clusters and a cluster containing the remaining days.

The influence of road works and events were already expected to be location specific and to differ between cities. However, the regular (weekly) variations are expected to be general applicable. Figure 9.3 shows the general weekly profiles [2] of Amsterdam and Almelo and from this figure can be concluded that the weekly patterns are indeed comparable. In Almelo, traffic volumes are a little higher on Fridays, probably because the percentage of the employees taking a day off at Fridays is somewhat smaller than in Amsterdam. Moreover,

[2]The major events and road works are not included in this analysis.

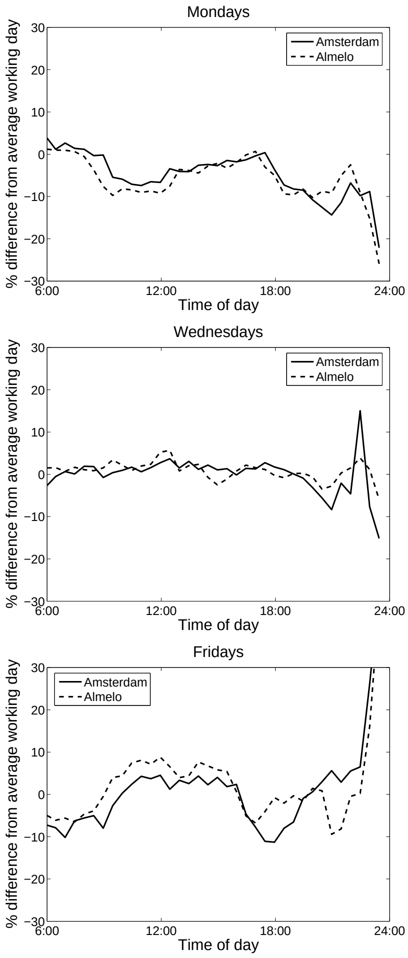


Figure 9.3: General relative weekly flow profiles for the cities of Amsterdam and Almelo. Adapted from Houtriet (2006).

Friday evening traffic volumes are relatively high in Amsterdam, probably due to a difference in the amount of leisure traffic

9.3 Conclusion

It is concluded that the method for the analysis of urban traffic patterns that is proposed in Chapter 4 functions adequately. Except for one, the cluster analyses resulted in plausible groups that can be explained by travel demand and supply factors. Only the grouping of locations on the basis of their resulting clusters did not result in clearly distinctive, recurrent and representative patterns. Not so much the type of variations appear to differ between different types of locations, but the amount of variations. The results of the analyses are not distorted by the selected method, available data or the occurrence of congestion, although summer and/or autumn patterns might be undetected at locations for which none or only few data is collected during these seasons. The clustering of daily flow profiles at all links in a network is labour-intensive, yet when cluster analyses are executed for some key locations along the main arterials, sufficient insight is obtained into the main traffic patterns in a city. Especially in case that cluster analyses have to be repeated regularly, it is also recommended to automate the selection of the most appropriate number of clusters.

The resultant clusters were aimed to also be used for traffic forecasting, traffic management and transport modelling scenarios. The disadvantage of cluster analysis in this respect is that the optimal number of clusters depends on the amount of variation that occurs at a certain location. Especially in case of relatively large variations between the most dissimilar types of days, the cluster analysis results in too few clusters, as a result of which the variation within the clusters is still quite large. The advantage of cluster analysis compared to the current method that groups days according to pre-defined factors as weekday is that variations due to road works and events are easily detected. In Chapter 8, a hybrid model is proposed that combines a classification according to weekday and holiday periods with a classification based on cluster analysis. Groups resulting from this hybrid model show smaller standard deviations and are thus useful for traffic forecasting, traffic management and transport modelling scenarios.

Finally, from a comparison between the traffic patterns in the cities of Almelo and Amsterdam it is concluded that the weekly variations in traffic volumes agree well between both cities. It looks as if the general traffic patterns can be applied to other cities as well. The general shape of the daily flow profile shows some differences and therefore is interesting to analyse further using data from more cities.

Chapter 10

Conclusions and Recommendations

10.1 Conclusions

In Chapter 1, three research objectives were presented. The first objective was to design a method for the analysis of temporal and spatial variations in urban traffic volumes using data from urban traffic information centres. The second objective was to apply the method to the city of Almelo in order to investigate whether the method functions adequately and the third objective was to obtain insight into urban traffic patterns in this city.

In this section, the main conclusions concerning these objectives are drawn. First, it is discussed whether and in what way traffic volume data from urban traffic information centres can be used for the analysis of urban traffic patterns. Second, the conclusions regarding the analysis of variations in urban traffic volumes are presented. The gaps in current research are summarized and the method designed in this thesis is briefly described and evaluated. Section 10.1.3 discusses the insight that is obtained into urban traffic patterns in Almelo and in general, and Section 10.1.4 discusses the use of the obtained insight and the resultant patterns.

10.1.1 Use of data from traffic information centres

As a consequence of the implementation of urban traffic information centres, more traffic data is becoming available for the urban network. This data is processed in order to be useable for research. In Chapter 2, we proposed a data validation procedure that detects invalid records of volume measurements applying minimum and maximum flow thresholds and the principle of conservation of vehicles.

The data validation procedure is evaluated using traffic data from the ViaContent system in Almelo and is found to function adequately, although the quality check based on the principle of conservation of vehicles has to be used with caution. From a visual inspection it is concluded that the percentage of records that pass data validation and include erroneous data is very small. Also during the application of the traffic data for the analysis of urban traffic patterns, almost no additional invalid data was detected. However, at a few locations, invalid data appeared to be on the basis of one of the clusters. These cases concern differences from average daily flow profiles instead of implausible profiles or profiles that deviate from profiles at other locations and could therefore not be detected by the data validation procedure. Therefore, it is concluded that cluster analysis is a useful addition to the data validation procedure.

For the case study described in this thesis, data from the ViaContent system in Almelo is applied. The data is collected by inductive loop detectors at signalized intersections. Field observations at six locations showed that the accuracy of the data is very high (the difference between manual counts and counts of ViaContent was less than 5% at all locations). Therefore, we state that the traffic data that passed the data validation procedure provides a good estimation of actual traffic volumes. For the analysis of urban traffic patterns, this traffic data is combined with calendar data and weather data. Moreover, information is available on road works, events and accidents. However, it is difficult to directly link these data to the traffic data, since they are often not stored in a central database and only affect traffic in part of the network.

10.1.2 Analysis of variations in urban traffic volumes

In Chapter 3 it is concluded that most existing literature on within and between day variations in urban traffic volumes deals with the general shape of a daily flow profile and the influence of the day of the week and the season on total daily traffic volumes. The insight into differences in daily flow profiles between different types of days is limited. Besides, possibly other factors also cause day-to-day variations in daily flow profiles. The insight into differences in temporal patterns between locations is very limited for the urban network. Finally, insight into typical traffic patterns on a network seems to be absent at all.

In Chapter 4 we propose an alternative approach that exploits a Ward's hierarchical clustering algorithm for the determination and analysis of urban traffic patterns. Days are grouped according to their daily flow profile and locations are grouped according to the temporal patterns they show. Finally, patterns are grouped on a network level. By means of basic statistical techniques it is investigated what factors are responsible for the resulting patterns.

From the application of the method to Almelo it is concluded that the proposed method functions adequately. Only the clustering of locations on the basis of their temporal patterns did not result in a clear grouping. This is due to the absence of clear types of locations: locations mainly appear to differ from each other with regard to the amount of variations instead of the type of variations. All other analyses resulted in recurrent, distinctive traffic patterns that can be explained by variations in travel demand and supply factors and agree well with the literature discussed in Chapter 3. In general, the variation between the days or locations within a cluster is clearly smaller than before classification, thus the cluster means are more representative for the days or locations within a cluster than the historical average. The results are not distorted by the selected method, available data or the occurrence of congestion, although summer and/or autumn patterns might be undetected at some locations as a consequence of the absence of traffic data for these seasons.

10.1.3 Insight into urban traffic patterns

The application to the city of Almelo also provides insight into traffic patterns in this city and in (Dutch) cities in general. The main conclusions with regard to the found patterns discussed in Chapter 7 are presented.

The clustering of non-working days generally resulted in a Saturday cluster and a Sunday cluster. Working days were clustered at 79 links. In most cases the clustering resulted in two to four clusters. Weekday, holiday periods and road works are found to be the main decisive factors. Also other location specific factors like events played a role. Rain played a role only at about 10% of the locations and also season was not a main determining factor for the resulting classification. In some cases a combination of season and weekday were on the basis of a cluster, e.g. a cluster contained winter days and Mondays and Tuesdays, or spring days and Thursdays. In most cases a cluster contained relatively many days of a certain type, yet not exclusively and not all. Clusters that were determined by a combination of factors were generally less complete than clusters that were determined by one factor.

Weekday and holiday clusters can be explained by variations in activity patterns (see for example Harms, 2003; Hilbers et al., 2004). Friday clusters show a relatively flat daily flow profile as a result of relatively little commuter traffic and much shopping and leisure traffic, whereas clusters with relatively many Mondays (and to a lesser extent Tuesdays) show a relatively peaked profile as a result of relatively much commuter and educational traffic and little shopping traffic. The Thursday cluster shows a peak in the evening due to late night shopping traffic to and from the city center. Road works result in very dissimilar profiles in some cases as a result of a decrease in capacity close to the monitoring location and in slightly dissimilar profiles in other cases as a result of re-routing of traffic due to road works elsewhere in the network.

Spatial traffic patterns are analysed at several aggregation levels. A classification on the basis of the average working day flow profile resulted in three common types of daily flow profiles. One of the clusters shows both an A.M. peak and a P.M. peak and contains links that are used by a mixture of travel motives. The other clusters show only an A.M. or a P.M. peak and are mainly used by commuter traffic to (only an A.M. peak) or from (only a P.M. peak) work. Also the results of the clustering of weekly patterns can be explained by differences in travel motives. Links that are used by traffic between residential areas on one hand and the city centre and zones with leisure activities on the other hand show relatively high traffic volumes on Saturdays compared to links that are mainly used by commuter traffic. Moreover, links that are used by traffic to and from the city centre show stronger weekly variations than links that are mainly used by commuter traffic. With regard to seasonal and weather influenced variations it is concluded that links with relatively much (more than 50%) short distance traffic show lower traffic volumes in summer and higher traffic volumes in case of rain, whereas links with relatively little short distance travel hardly show seasonal and weather related variations.

On a network level, the clustering of daily flow profiles at seven key locations, resulted in four clusters: (1) a cluster with many Mondays, Tuesdays, Wednesdays and winter days that shows a relatively peaked flow profile with a low off-peak flow at al locations, (2) a cluster that contains many Thursdays and spring days and shows relatively high traffic volumes throughout all day at all locations, (3) a cluster with a relatively flat daily flow profiles and low peak volumes at all locations that consists Fridays and days within holiday periods and (4) an events cluster that shows a peak in the evening at some of the locations caused by home-matches of the footballclub Heracles.

As mentioned before, weekday and holiday clusters can be explained by variations in activity patterns. These activity patterns are generally applicable. Besides, from a comparison between the weekly flow profiles of Almelo and Amsterdam it is concluded that both cities show similar weekly variations. Therefore, we conclude that it looks as if the regular traffic patterns can be applied to other (Dutch) cities as well.

10.1.4 Applications

The obtained insights that are discussed in the previous section can be applied to monitor regular and irregular variations in traffic volumes and provide a better estimation of the actual traffic volumes on a certain type of day at a certain (type of) location. The advantage of cluster analysis compared to classification into predefined types of days is that it provides an easy way to detect and analyse the effects of atypical circumstances such as road works and events. The obtained insights can be used for traffic planning, land use planning and for taking adequate (traffic management) measures.

Since part of the variations can be explained by general travel demand factors, the results are also applicable to other (Dutch) cities and thus provide insight into regular variations in traffic volumes in other cities as well. Additionally, the results of the classification of locations can be used for the estimation of average traffic volumes on the basis of short term traffic counts. In Chapter 9 traffic patterns found for Almelo are compared to traffic patterns found for Amsterdam Zuid-Oost. The general variations between weekdays appear to be similar for Amsterdam Zuid-Oost and Almelo. This indicates that the regular variations in traffic volumes are spatial transferable.

Typical traffic patterns can also be used for traffic forecasting and traffic management scenarios. For these applications, it is important that the cluster averages are representative for the days within the clusters. For six locations, the classifications that result from the cluster analysis are compared to classifications on the basis of weekday and holiday periods. This comparison shows that the advantage of cluster analysis is that it easily detects variations due to location specific, irregular factors such as events and road works. On the other hand, the disadvantage is that in some cases only few clusters are distinguished. As a consequence, the variation within the clusters may still be quite large. Moreover, since clusters generally contain mainly or relatively many days of a certain type instead of all days, it sometimes is difficult to determine in advance to what cluster a day should be classified to. We have proposed a hybrid model that combines the strengths of both methods. Cluster analysis is used as a filter; atypical, non-recurrent patterns are grouped in different clusters. Subsequently, the resulting groups are further classified according to weekday and holiday periods. With regard to traffic forecasting or the imputation of missing or erroneous data, the hybrid method is applied to select the most similar patterns. Subsequently, the missing or future traffic volume has to be estimated, possibly by applying ARIMA type of models or pattern matching. With regard to traffic management scenario's, for each typical type of day a different signal plan could be developed. Besides, on a more strategic level, traffic management strategies (including signal timing, rerouting strategies, tidal flow lanes, varying road-prices and/or other measures) could be developed that optimize traffic on a network level.

Finally, the results of the cluster analysis can be used for transport modelling. First, as cluster analysis is able to filter out road works and other special circumstances, the OD-matrix for an average working day can be calibrated more accurately. Besides, the matrix can also be calibrated for the typical types of days that result from the hybrid model. In case that traffic is not monitored continuously, the results from the clustering of locations can be applied to translate short term traffic counts into working day or weekday averages. The model estimates traffic volumes on locations without monitoring and thus provides a clear picture of the traffic situation in the entire network that can be used for traffic monitoring, travel time forecasting and traffic management.

10.2 Recommendations for practitioners

The method is easily applicable to other cities. However, it is labour-intensive to execute cluster analyses for all links in the network. Besides, since traffic patterns are subject to changes due to infrastructural and land use developments, new cluster analyses have to be executed from time to time to update the patterns. To improve the applicability, it is recommended to automate the process and/or to limit the amount of locations for which cluster analyses are carried out.

For the determination of typical traffic patterns that function as a basis for traffic forecasting, traffic management or transport modelling scenarios, it is recommended to apply the hybrid model. The determination of the patterns using the hybrid model should be automated entirely. Moreover, traffic patterns should be updated regularly by repeating the clustering procedures. These updates can be executed off-line, yet the historical database can be linked to an on-line database containing real time data for incident detection, online traffic forecasting and real-time traffic management.

With regard to traffic monitoring, part of the analysis cannot be automated, since the essence of the application is to obtain insight. However, automation of the selection of the most appropriate number of clusters and visualization using a GIS application simplifies the process and makes the results easier to interpret. In our opinion, the most valuable analyses regarding traffic monitoring are to detect and analyse irregular variations due to road works or events and to analyse daily and weekly flow profiles for different (types of) locations. Besides, to obtain more insight into the main patterns on a network level or for a specific part of the network, the method proposed in Section 4.4 can be applied. As mentioned before, policy makers can define different sets of key locations, depending on the patterns they are interested in. Besides, cluster analysis can also be carried out on other aggregation levels, e.g. on a lane level or intersection level in case this is more appropriate for the specific interest.

10.3 Further research

Three main directions of further research are discussed. First, when more data is becoming available, new analyses may be executed to obtain more insight into urban traffic patterns. Secondly, an in depth analysis of the underlying travel demand patterns enables a better estimation of future changes in traffic patterns. Finally, all four applications discussed in Chapter 8 require further research.

Further research on urban traffic patterns

The previous chapter discussed some limitations of the research that are due to a limited amount of available data. First, since relatively little data is collected in summer and autumn, seasonal variations could not be analysed for all locations. Although we expect seasonal variations to be relatively weak for urban areas, it is recommended to repeat the analyses when more data is available. Moreover, the seasons are defined arbitrary and traffic patterns might also vary between months or weeks within one season. Therefore, it is interesting to analyse month to month and week to week variations in traffic volumes in more detail.

Second, also weather effects could not be analysed fully. It is recommended to analyse the effects of weather on traffic in more detail using weather data from Almelo itself on a low temporal aggregation level.

Third, since only one year of traffic data was available at the time of research, long term variations in traffic patterns could not be analysed. As mentioned in Chapter 3, in the longer term, traffic patterns may change as a result of local land use and infrastructural changes and/or general demographic, social-economic and geographical developments. When more data is becoming available it would be interesting to analyse the effects of these local changes and general developments on urban traffic patterns.

Finally, only traffic patterns in Almelo are studied extensively. Chapter 9 shows that the general variations between weekdays are similar for Amsterdam Zuid-Oost, but that the average daily flow profile shows some differences. When traffic data is becoming available for more cities it would be interesting to compare the average daily and weekly flow profiles for different cities and possibly also for highways and to analyse the relationship between certain spatial factors and resulting flow profiles.

Analysis of underlying travel demand patterns

The observed variations in traffic volumes were explained by variations in travel behaviour. However, the underlying travel demand patterns are not specified, i.e. it is not clear what the characteristics (e.g. origin, destination, motive, personal characteristics) are of the road users. Specification of the travel demand patterns provides more insight into alternatives (regarding trip generation, destination, travel mode, departure time and route choice) of the road users. Besides, it is also important to have information on potential new road users (as a result of extra trips or changes in destination, travel mode, departure time or road choice). These insights enable better estimations of changes in traffic patterns as a result of traffic management measures or demographic, social-economic, infrastructural or land use developments.

As mentioned in the first chapter, travel diary data provides insight into travel demand patterns. However, the sample size of the available survey data is too small and the spatial aggregation level of the data is too high to analyse demand patterns on a local level. Moreover, it is expensive to execute extensive surveys to analyse local travel demand patterns. Methods for the synthesis of travel diary data may be useful for the estimation of local travel demand patterns using travel survey data. A method developed by Kager (2005) estimates individual travel diary data for non-observed respondents using a micro-simulation approach. The ALBATROSS model applies a set of decision rules that were derived from activity-diary data collected in two municipalities in The Netherlands (Arentze et al., 2002). Both the ALBATROSS model (Arentze et al., 2002) and the model developed by Kager (2005) have proven to be able to estimate the characteristics of trips made in a certain area. It would be interesting to apply (one of) these methods to Almelo in order to obtain more insight into the travel demand patterns that are responsible for the observed (variations in) traffic volumes.

Research related to potential applications

All four applications discussed in Chapter 8 require further research. With regard to traffic monitoring, the resulting traffic patterns could also be used for incident detection. To that end, research on pattern recognition and incident detection algorithms should be carried out. Regarding traffic forecasting, different forecasting algorithms (e.g. ARIMA type of models, pattern matching algorithms) – that are applied in the second step of the forecasting procedure – could be developed and assessed. With regard to traffic management, research should be done on the development of traffic management strategies to optimize traffic on a network level. The strategies could be evaluated by applying a simulation model. After selection of the most appropriate strategy for each situation, the benefit of distinguishing different typical traffic situations can be determined. Finally, in Chapter 8 was mentioned that the obtained insight can be used as an extra input in the matrix calibration process in traffic models. It has to be elaborated exactly how this insight can be applied.

Bibliography

Ancidei, V., Cera, E., Pisi, P. D., Francalanci, S., Landolfi, O. and Tomassini, M. (2000). Integrated ITS in the city of Rome, *7th World Congress on Intelligent Transport Systems*, Turin, Italy. CD-rom.

Andrey, J., Mills, B., Leahy, M. and Suggett, J. (2003). Weather as a chronic hazard for road transportation in canadian cities, *Natural Hazards* **28**: 319–343.

Arentze, T., Hofman, F., Van Mourik, H. and Timmermans, H. (2002). Spatial Transferability of the ALBATROSS Model System, *Transportation Research Record* **1805**: 1–7.

Aunet, B. (2000). Wisecounsin's Approach to Variation in Traffic Data, *North American Travel Monitoring Exhibition and Conference*, Wiscounsin, USA.

Axhausen, K. W., Zimmerman, A., Schnfelder, S., Findsfser, G. and Haupt, T. (2002). Observing the rhythms of daily life: A six-week travel diary, *Transportation* **29**: 95–124.

Bae, S. and Lee, B. G. (2000). A real-time traffic information service by dedicated fm broadcasting system in Seoul Korea, *7th World Congress on Intelligent Transport Systems*, Turin, Italy. CD-rom.

Bajwa, S. I., Chung, E. and Kuwahara, M. (2004). An adaptive travel time prediction model base on pattern matching, *11th World Congress on Intelligent Transport Systems*, Nagoya, Japan. CD-rom.

Batenburg, R. S. and Knulst, W. R. (1993). Sociaal-culturele beweegredenen; onderzoek naar de invloed van veranderende leefpatronen op de mobiliteitsgroei sinds de jaren zeventig, *Technical report*, Sociaal en Cultureel Planbureau, The Netherlands. in Dutch.

Bell, M. C., Bennett, L. D. and Evans, R. G. (1993). "The Instrumented City": a database for road transport information systems, *IEE Vehicle navigation & information systems conference*, Ottawa, Canada, pp. 11–15.

Bell, M. C., Evans, R., Boddy, R. and Hill, I. (1996). An introductory guide to the Instrumented City facility, *Traffic Engineering + Control* **37**(12): 698–703.

Bell, M. C. and Gillam, W. J. (1994). The 'Instrumented City': data provision for traffic management and research, *IEE Conference publication, Road traffic monitoring and control* **391**: 19–22.

Bell, M. G. H., Shield, C. M., Busch, F. and Kruse, G. (1997). A stochastic user equilibrium path flow estimator, *Transportation Research C* **5**(3/4): 197–210.

Bellamy, P. H. (1978). Seasonal Variations in Traffic Flow, *Technical Report Report SR437*, TRRL, Crowthorne, UK.

Bennett, C. R., Chamorro, A., Chen, C., de Solminihac, H. and Flintsch, G. W. (2005). Data collection technologies for road management, version 1, *Technical report*, East Asia Pacific Transport Unit, The World Bank.

Bexilius, S. and Kengen, H. P. A. M. (1997). Dagen met topdrukte in het verkeer, kenmerken en voorspelbaarheid, *Verkeerskundige werkdagen, publicatie 73*, CROW, Ede, The Netherlands, pp. 717–730. [In Dutch].

BGC (1986). Wetmatigheden in verkeersintensiteiten, *Technical Report IVV/226/09/Mr*, Bureau Goudappel Coffeng, in opdracht van het ministerie voor verkeer en waterstaat, Deventer, The Netherlands. [In Dutch].

BGC (1997). Marktprofiel van de filerijder, *Technical Report AVV105/Bq/1616*, Bureau Goudappel Coffeng, in opdracht van de Adviesdienst Verkeer en Vervoer, Deventer, The Netherlands. [In Dutch].

Budde, A. (2002). Düsseldorfs' traffic management system, *9th World Congress on Intelligent Transport Systems*, Chicago, USA. CD-rom.

Cascetta, E. (1984). Estimation of trip matrices from traffic counts and survey data: A generalized least squares estimator, *Transportation Research B* **18**: 289–299.

Changnon, S. A. (1996). Effects of summer precipitation on urban transportation, *Climatic Change* **32**: 481–494.

Chen, C., Kwon, J., Rice, J., Skabardonis, A. and Varaiya, P. (2003). Detecting Errors and Imputing Missing Data for Single Loop Surveillance Systems, *82th annual meeting of the Transportation Research Board*, Washington D.C., USA. CD-Rom.

Chen, L. and May, A. (1987). Traffic detector errors and diagnostics, *Transportation Research Record* **1132**: 82–93.

Chrobok, R., Kaumann, O., Wahle, J. and Schreckenberg, M. (2004). Different methods of traffic forecast based on real data, *European Journal of Operational Research* **155**: 558–568.

Chung, E. (2003). Classification of traffic pattern, *10th World Congress on Intelligent Transport Systems*, Madrid, Spain. CD-Rom.

Chung, E., Ohtani, O., Warita, H., Kuwahara, M. and Morita, H. (2005). Effect of rain on Travel Demand and Traffic Accidents, *8th IEEE Conference on Intelligent Transportation Systems*, Vienna, Austria. CD-Rom.

Chung, E. and Rosalion, N. (2001). Short term traffic flow prediction, *24th Australasian Transport Reseach Forum*, Hobart, Tasmania. http://www.atrf.info/forums/24th.htm.

Cleghorn, D., Hall, F. and Garbuio, D. (1991). Improved data screening techniques for freeway traffic management systems, *Transportation Research Record* **1320**: 17–31.

Codling, P. J. (1972). Weather and road accidents, *in* J. A. Taylor (ed.), *Climatic resources and economic activity: a symposium*, Newton Abbot, UK, pp. 205–222.

Coifman, B. and Dhoorjaty, S. (2002). Event Data Based Traffic Detector Validation Tests, *81th annual meeting of the Transportation Research Board*, Washington D.C., USA. CD-Rom.

Cone, R., L, J., Jeffery, D., Fisher, G. and Wainwright, A. (2002). Delivering ITS and travel information in Wales, *9th World Congress on Intelligent Transport Systems*, Chicago, USA. CD-Rom.

CROW (1996). ASVV, aanbevelingen voor verkeersvoorzieningen binnen de bebouwde kom, *Publicatie 110*, Stichting Centrum voor Regelgeving en Onderzoek in de Grond-, Water-, en Wegenbouw en de Verkeerstechniek, Ede, The Netherlands. [In Dutch].

Danech-Pajouh, M. and Aron, M. (1991). ATHENA: a method for short-term inter-urban motorway traffic forecasting, *Recherche Transports Sécurité, English issue* **6**: 11–16.

Deckers, L. (2001). RIO: inwinning op het onderliggend wegennet in Rotterdam, *Symposium DVM, Rotterdam*, Rotterdam, The Netherlands. in Dutch.

Di Taronto, C., Bigotti, R. and Biora, F. (2000). Network status estimation and traffic prediction in urban and sub-urban areas, *7th World Congress on Intelligent Transport Systems*, Turin, Italy. CD-Rom.

DLTR (2001). Transport Statistics Bulletin, Road Traffic Statistics: 2000, *Technical Report Statistics report SB(01)*, United Kingdom.

Du, J. and Aultman-Hall, L. (2007). Increasing the accuracy of trip rate information from passive multi-day GPS travel datasets: Automatic trip and identificaiton issues, *Transportation Research A* **41**(3): 220–232.

Edwards, J. B. (1999). Speed adjustment of motorway commuter traffic to inclement weather, *Transportation Research Part F* **2**(1): 1–14.

Erhunmwunsee, P. O. (1991). Estimating average annual daily traffic flow from short period counts, *ITE Journal* **61**(11): 23–30.

Evans, R. G. and Bell, M. C. (1996). The identification of recurrent urban traffic congestion, *IEEE Conference on Intelligent Transportation Systems*, pp. 183–187.

Faghri, A. and Hua, J. (1995). Roadway Seasonal Classification Using Neural Networks, *Journal of Computing in Civil Engineering* **9**(3): 209–215.

Fellendorf, M., Nökel, K. and Handke, N. (2000). VISUM-online -traffic management for the EXPO 2000 based on a traffic model, *7th World Congress on Intelligent Transport Systems*, Turin, Italy. CD-Rom.

Ferulano, G., Gortan, L., Sforza, A. and Tartaro, D. (2000). The ATENA project "Ambiente traffico telematica Napoli", *7th World Congress on Intelligent Transport Systems*, Turin, Italy. CD-Rom.

Festin, S. M. (1996). Summary of National and Regional Travel Trends 1970-1995, *Technical report*, US DOT, Federal Highway Administration, Washington D.C.

FHWA (2003). Traveler Information Systems in Europe, *Technical Report FHWA-PL-03-005*, US DOT, Federal Highway Administration, Washington D.C.

Flaherty, J. (1993). Cluster Analysis of Arizona Automatic Traffic Recorder Data, *The Science of the Total Environment* **1410**: 93–99.

Fox, K. and Clark, S. (n.d.). Evaluating the benefits of a responsive utc system using micro-simulation. http://www.its.leeds.ac.uk/projects/volumes/index.html (not published).

Goodwin, L. C. (2002). Weather impacts on arterial traffic flow. http://ops.fhwa.dot.gov/weather/best_practices/arterialimpactpaper.pdf (not published).

Gram, F. (1996). Time variations in traffic and traffic emissions, *The Science of the Total Environment* **189/190**: 115–118.

Hagens, A. (2005). *De auto laten staan: ook als het regent; De invloed van weer op de stedelijke verkeersvraag*, Msc thesis, University of Twente, Enschede, the Netherlands. [In Dutch].

Hanson, S. and Huff, J. O. (1988). Systematic variability in repetitious travel, *Transportation* **15**: 111–135.

Harms, L. (2003). Mobiel in de tijd, op weg naar een auto-afhankelijke maatschappij, 1975 - 2000, *Technical report*, Sociaal en Cultureel Planbureau, The Netherlands. [In Dutch].

Hasberg, P. and Serwill, D. (2000). Stadtinfoköln -a global mobility information system for the Cologne area, *7th World Congress on Intelligent Transport Systems*, Turin, Italy. CD-Rom.

Hassan, Y. A. and Barker, D. J. (1999). The impact of unseasonable or extreme weather on traffic activity within lothian region, scotland, *Journal of Transport Geography* **7**: 209–213.

Hellinga, B. R. (2002). Improving freeway speed estimates from single-loop detectors, *Journal of Transportation Engineering* **128**(1): 58–66.

Henriet, E. and Schmitz, P. (2000). Information strategies in the urban context, *7th World Congress on Intelligent Transport Systems*, Turin, Italy. CD-Rom.

Hicks, J. and Niemeier, D. A. (2001). Improving the resolution of gridded-hourly mobile emissions: incorporating spatial variability and handling missing data, *Transportation Research Part D* **6**: 153–177.

Hilbers, H., Van Eck, J. R. and Snellen, D. (2004). *Behalve de dagelijkse files, over betrouwbaarheid van reistijd*, NAI uitgevers, Ruimtelijk Planbureau, the Netherlands. [In Dutch].

Hodge, V. J. and Austin, J. (2004). A survey of oulier detection methodologies, *Artificial Intelligence Review* **22**: 85–126.

Hogema, J. H. (1996). Effects of rain on daily traffic volume and on driving behaviour, *Technical Report TM-96-B019*, TNO Human Factors Research Institute, Soesterberg, the Netherlands.

Houtriet, E. P. J. (2006). *Monitoring Amsterdam Zuidoost (MOZO); Het analyseren van VRI detectielusdata*, Msc thesis, University of Twente, Enschede, the Netherlands. [In Dutch].

Hoyer, R. and Herrmann, A. (2003). ITS in small innovative steps -succesful deployment in a medium-sized city, *10th World Congress on Intelligent Transport Systems*, Madrid, Spain. CD-Rom.

Hu, P., Wright, T. and Esteve, T. (1998). Traffic Count Estimates for Short-Term Traffic Monitoring Sites, Simulation Study, *Transportation Research Record* **1625**: 26–34.

Huisken, G. (2003). Reistijdschatten met GSM, *Symposium DVM*, Rotterdam, The Netherlands. in Dutch.

Huizingh, E. (2002). *Inleiding SPSS 11.0 voor Windows en Data Entry*, Academic Service, Schoonhoven, The Netherlands. [In Dutch].

Hunt, P. B., Robertson, D. I., Bretherton, R. D. and Winton, R. I. (1981). Scoot- a traffic responsive method of coordinating signals, *Technical Report LR 1014*, TRRL.

Iryo, T., Iwatani, A. and Asakura, Y. (2005). Classifiying of day-to-day variation of traffic flow with cluster analysis, *12th World Congress on ITS*, San Fransisco, USA. CD-Rom.

Ishak, S. (1990). Fuzzy-Clustering Approach to Quantify Uncertainties of Freeway Detector Observations, *Transportation Research Record* **1856**: 6–15.

Jacobsen, L., Nihan, N. and Bender, J. (1990). Detecting Erroneous Loop Detector Data in a Freeway Traffic Management System, *Transportation Research Record* **1287**: 151–166.

Jain, A. K., Murty, M. N. and Flynn, P. J. (1999). Data Clustering: A Review, *ACM Computing Surveys* **31**(3): 264–323.

Kager, R. (2005). *Design and implementation of a method for the synthesis of travel diary data*, PhD thesis, University of Twente.

Kamarianakis, Y. and Prastacos, P. (2003). Forecasting traffic flow conditions in an urban network: comparison of multivariate and univariate approaches, *Transportation Research Record* **1857**: 74–84.

Karl, C. A. and Trayford, R. S. (2000). Delivery of real-time and predictive travel time information: Experiences from a Melbourne trial, *7th World Congress on Intelligent Transport Systems*, Turin, Italy. CD-Rom.

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding Groups in Data: an introduction to cluster analysis*, Wiley.

Keay, K. and Simmonds, I. (2005). The association of rainfall and other weather variables with road traffic volume in Melbourne, Australia, *Accident Analysis and Prevention* **37**: 109–124.

Kellerman, A. and Schmid, A. (2000). Mobinet: Intermodal traffic management in Munich -control centre development, *7th World Congress on Intelligent Transport Systems*, Turin, Italy. CD-Rom.

Kikuchi, S. and Miljkovic, D. (1999). Method To Preprocess Observed Traffic Data for Consistency, Application of Fuzzy Optimization Concept, *Transportation Research Record* **1679**: 73–80.

Kirschfink, H., Riegelhoth, G. and Boero, M. (2000). Strategic management by the enterprice approach, *7th World Congress on Intelligent Transport Systems*, Turin, Italy. CD-Rom.

Kitamura, R. and Van der Hoorn, T. (1987). Regularity and irreversibility of weekly travel behaviour, *Transportation* **14**: 227–251.

Klein, L. A. (2001). *Sensor Technologies and Data Requirements for ITS*, Artech House ITS library.

Kroes, E., Hagemeier, F. and Linssen, J. (1999). A new, probe vehicle-based Floating Car Data system: Conceps, implementation and pilot study, *Traffic Engineering + Control* **40**: 200–204.

Kruse, G., Tannert, R. and Hasberg, P. (2000). Incident detection by MOTION for strategic control in the traffic management system "Statinfoköln", *7th World Congress on Intelligent Transport Systems*, Turin, Italy. CD-Rom.

Kwon, J., Chen, C. and Varaiya, P. (2004). Statistical Methods for Detecting Spatial Configuration Errors in Traffic Surveillance Sensors, *Transportation Research Record* **1870**: 124–132.

Leitsch, B. (2002). A Public-privat partnership for mobility -Traffic Management Center Berlin, *9th World Congress on Intelligent Transport Systems*, Chicago, USA. CD-Rom.

Li, M. T., Zhao, F. and Wu, Y. (2004). Application of Regression Analysis for Identifying Factors That Affect Seasonal Traffic Fluctuations in Southeast Florida, *Transportation Research Record* **1870**: 153–161.

Li, M. T., Zhao, F., Wu, Y. and Gan, A. (2003). Evaluation of agglomerative hierarchical clustering methods, *82nd Annual Meeting of the Transportation Research Board (TRB)*, Washington, USA. CD-Rom.

Lingras, P., Shama, S. C., Osborne, P. and Kalyar, I. (2000). Traffic Volume Time-Series Analysis According to the Type of Road Use, *Computer-Aided Civil and Infrastructure Engineering* **15**: 365–373.

Liu, Z. and Sharma, S. (2006). Statistical investigations of statutory holiday effect on traffic volumes, *Transportation Research Record* **1945**: 40–48.

Liu, Z., Sharma, S. and Wyatt, K. (2005). Traffic flow peaking characteristics during holiday periods, *Annual Conference of the Canadian Institute of Transportation Engineers*, Saskatoon, Canada. CD-Rom.

Lomax, T., Turner, S. and Margiotta, R. (2003). Monitoring Urban Roadways in 2001: Examining Reliability and Mobility with Archived data, *Technical Report FHWA-OP-03-141*, Texas Transportation Institute, USA.

Lomax, T., Turner, S. and Margiotta, R. (2004). Monitoring Urban Roadways in 2002: Using Archived Operations Data for Reliability and Mobility Measurement, *Technical Report FHWA-HOP-04-011*, Texas Transportation Institute, USA.

Lyly, S. (1968). *The variation patterns of traffic flow on the basis of traffic census methods*, PhD thesis, Finland's institute of Technology, Helsinki.

Maher, M. J. (1983). Inferences on trip matrices from observations on link volumes: A Bayesian statistical approach, *Transportation Research B* **17**: 435–447.

Mahmassani, H. S. (1997). Dynamics of cummuter behaviour: Recent research and continuing challenges, *in* P. Stopher and M. Lee-Gosselin (eds), *Understanding travel behaviour in an era of change Pergamon*, Pergamon.

Manchester Metropolitan University (2004). http://149.170.199.144/multivar/ca.htm.

May, A. D. (1990). *Traffic flow fundamentals*, Prentice-Hall.

Meyer, M. D. and Miller, E. J. (2001). *Urban transportation planning, second edition*, McGraw-Hill.

MinVenW (2004). *Nota Mobiliteit. Naar een betrouwbare en voorspelbare bereikbaarheid*, Ministerie van Verkeer en Waterstaat, VROM. [In Dutch].

MIZAR Automazione (1998). COSMOS (Congestion Management Strategies and Methods in Urban Sites) report on Torino's Demonstrator for Congestion and Incident Detection and Management, *Technical Report deliverable D04.3, Telematics Application Program Transport (TR1015).*

Nankervis, M. (1999). The effect of weather and climate on bicycle commuting, *Transportation Research A* **33**: 417–431.

Niemeier, D. A., Utts, J. M. and Fay, L. (2002). Cluster analysis for Optimal Sampling of Traffic Count Data: Air Quality Example, *Journal of Transportation Engineering* **128**(1): 97–102.

Nowotny, B., Asamer, J., Din, K. and Karim, R. (2003). Classification of traffic data time series by cluster analysis, artificial neural networks and ANOVA, *10th world congress on Intelligent Transport Systems and Services*, Madrid, Spain. CD-Rom.

Ortuzar, J. D. and Willumsen, L. G. (1994). *Modelling Transport*, Wiley.

Papageorgiou, P. (1991). *Concise Encyclopedia of Traffic and Transportation Systems*, Pergamon Press.

Pas, E. I. (1987). Intrapersonal variability and model goodness-off-fit, *Transportation Research A* **21**(6): 431–438.

Pribyl, O. and Pribyl, P. (2005). Mining data from induct loops, *8th international IEEE Conference on Intelligent Transportation Systems*, Vienna, Austria. CD-Rom.

Quiroga, C. A. and Bullock, D. (1998). Travel time studies with global positioning and geographic information systems: an integrated methodology, *Transportation Research C* **6**: 101–127.

Rakha, H. and Van Aerde, M. (1995). Statistical Analysis of Day-to-Day Variations in Real-Time Traffic Flow Data, *Transportation Research Record* **1510**: 26–34.

Reefhuis, M. (2005). *Incidentele gebeurtenissen op het wegennet*, Msc thesis, University of Twente, Enschede, the Netherlands. [In Dutch].

Richards, A., Piao, J. and McDonald, M. (2000). VMS in Southampton: A case study, *7th World Congress on Intelligent Transport Systems*, Turin, Italy. CD-Rom.

Roess, R. P., McShane, W. R. and Prassas, E. S. (1998). *Traffic Engineering, second edition*, Pretence Hall, USA.

Scharrer, R., Kippes, G., Glas, F. and Keller, H. (2003). An Innovative Road Side Driver Information System -NetzInfo-, *10th World Congress on Intelligent Transport Systems*, Madrid, Spain. CD-Rom.

Schatz, C., Franosch, M. and Ganser, M. (2003). PARTRAS -Pattern Recognition of Traffic States, *10th world congress on Intelligent Transport Systems and Services*, Madrid, Spain. CD-Rom.

Schlich, R. and Axhausen, K. W. (2003). Habitual travel behaviour: Evidence from a six-week travel diary, *Transportation* **30**: 13–36.

Schmidt, G. (1996). Hochrechnungsfaktoren für Kurzzeitzhlungen auf Innerortstrassen, *Strassenverkehrstechnik* **40**(11): 546–556. [In German].

Schoemakers, M. and Van Engelenburg, G. (2003). Een oplossing voor onbetrouwbare en onvolledige lusdata op het hoofdwegennet, *Colloquium Vervoersplanologisch Speurwerk (CVS), Innovatie: van inspiratie naar realisatie?* , Rotterdam, The Netherlands. In Dutch.

Sharma, S. C., Gulati, B. M. and Rizak, S. N. (1996). Statewide Traffic Volume Studies and Precision of AADT Estimates, *Journal of Transportation Engineering* **122**(6): 430–439.

Sharma, S. C., p j. Lingras, Hassan, M. U. and Murthy, N. A. S. (1986). Road Classification according to Driver Population, *Transportation Research Record* **1090**: 61–69.

Sharma, S. C. and Werner, A. (1981). Improved Method of Grouping Provincewide Permanent Traffic Counters, *Transportation Research Record* **815**: 13–18.

Smith, B. L., Scherer, W. T. and Hauser, T. A. (2001). Data-Mining Tools for the Support of Signal-Timing Plan Development, *Transportation Research Record* **1768**: 141–147.

Spiess, H. (1987). A maximum likelihood model for estimating Origin-Destination matrices, *Transportation Research B* **21**: 395–412.

Stathopoulos, A. and Karlaftis, M. (2001a). Spectral and Cross-Spectral Analysis of Urban Traffic Volumes, *4th IEEE Conference on Intelligent Transportation Systems*, Oakland, USA, pp. 820–825.

Stathopoulos, A. and Karlaftis, M. (2001b). Temporal and Spatial Variations of Real-Time Traffic Data in Urban Areas, *Transportation Research Record* **1768**: 135–140.

Stathopoulos, A. and Karlaftis, M. (2003). A multivariate state space approach for urban traffic flow modelling and prediction, *Transportation Research C* **11**: 121–135.

Tanner, J. C. (1952). Effect of Weather on Traffic Flows, *Nature* **4290**: 107.

Taylor, A. M. P., Young, W. and Bonsall, P. W. (1996). *Understanding Traffic Systems: Data, Analysis and Presentation*, Avebury Technical, UK.

Transpute (2000). Innovatie inwinning filegegevens, presentaties uit het materiaal, *Technical report*. [In Dutch].

Turner, S. (2001). Guidelines for Developing ITS Data Archiving Systems, *Technical Report Report 2127-3*, Texas Transportation Institute.

Turner, S. (2004). Defining and Measuring Traffic Data Quality, White Paper on Recommended Approaches, *Transportation Research Record* **1870**: 62–69.

Turner, S., Albert, L., Gajewski, B. and Eisele, W. (2000). Archived Intelligent Transportation System Data Quality, Preliminary Analyses of San Antonio TransGuide Data, *Transportation Research Record* **1719**: 77–84.

Turochy, R. E. and Smith, B. L. (2002). Measuring variability in traffic conditions using archived data, *81nd Annual Meeting of the Transportation Research Board (TRB)*, Washington, USA. CD-Rom.

US DOT (2001). Traffic Monitoring Guide, *Technical report*, Federal Highway Administration, Washington, USA.

Van der Voort, M., Dougherty, M. and Watson, S. (1996). Combining Kohonen Maps with ARIMA time series models to forecast traffic flow, *Transportation Research C* **4**(5): 307–318.

Van Wissen, L. J. G. and Meurs, H. J. (1989). The Dutch Mobility panel: Experiences and evaluation, *Transportation* **16**: 99–119.

Van Zuylen, H. J. and Willumsen, L. G. (1980). The most likely trip matrix estimated from traffic counts, *Transportation Research B* **14**: 281–293.

Vanajakshi, L. and Rilett, L. R. (2004). Loop Detector Data Diagnostics Based on Conservation-of Vehicles Principle, *Transportation Research Record* **1870**: 162–169.

Venkatanarayana, R., Snith, B. L., Zhang, G. and Tanikella, H. (2006). An exploration of advanced computing approaches to automate the Identification of Traffic Patterns in Large Data Sets, *85th annual meeting of the Transportation Research Board*, Washington, USA. CD-Rom.

Wall, Z. R. and Dailey, D. J. (2003). Algorithm for Detecting and Correcting Errors in Archived Traffic Data, *Transportation Research Record* **1855**: 183–190.

Wang, X., Cottrell, W. and Mu, S. (2005). Using K-means Clustering to Identify Time-of-Day Break Points for Traffic Signal Plans, *8th international IEEE Conference on Intelligent Transportation Systems*, Vienna, Austria. CD-Rom.

Wang, Y. and Nihan, N. L. (2000). Freeway traffic speed estimation using single loop outputs, *79th annual meeting of the Transportation Research Board*, Washington, USA. CD-Rom.

Webb, A. R. (2002). *Statistical pattern recognition*, Wiley.

Weijermars, W. A. M. and Van Berkum, E. C. (2005a). Analyzing highway flow patterns using cluster analysis, *8th international IEEE Conference on Intelligent Transportation Systems*, Vienna, Austria. CD-Rom.

Weijermars, W. A. M. and Van Berkum, E. C. (2005b). Road Traffic patterns, analysis of daily patterns on a Dutch Highway using cluster analysis, *Technical Report CEandM research report 2005R-001.VVR-001 (ISSN 1568-4652)*, University of Twente, Enschede, The Netherlands.

Weijermars, W. A. M. and Van Berkum, E. C. (2006a). Detection of invalid loop detector data in urban areas, *Transportation Research Record* **1945**: 82–88.

Weijermars, W. A. M. and Van Berkum, E. C. (2006b). Temporal traffic patterns in Almelo: Results of cluster analyses for all links along the main arterials, *Technical Report CEandM research report 2006W-001.VVR-002*, University of Twente, Enschede, The Netherlands.

White, J. (2001). From A to B: filling the gaps, *Traffic technology international* **8**: 30–32.

Wild, D. (1994). Pattern Based Forecasting, *Proceedings of the Second DRIVE-II Workshop on Short-Term Traffic Forecasting*, Delft, The Netherlands.

Wild, D. (1997). Short-term forecasting based on a transformation and classification of traffic volume time series, *International Journal of Forecasting* **13**: 63–72.

Williams, B. M. (2001). Multivariate Vehicular Traffic Flow Prediction: An Evaluation of ARIMAX modelling, *Transportation Research Record* **1776**: 194–200.

Wright, T., Hu, P. S., Young, J. and Lu, A. (1997). Variability in Traffic Monitoring Data, Final Summary Report, *Technical report*, Oak Ridge National Laboratory, prepared for the U.S. Department of Energy, USA.

Zhang, Y., Ou, X., Ren, J. and Yoa, D. (2002). An approach for urban traffic volume compression based on similarity attributes, *9th world congress on Intelligent Transport Systems and Services*, Chicago, USA. CD-Rom.

Zhao, F. and Chung, S. (2001). Contributing Factors of Annual Average Daily Traffic in a Florida County, Exploration with Geographic Information System and Regression Models, *Transportation Research Record* **1769**: 113–122.

Zhao, Y. (2000). Mobile Phone Location Determination and Its Impact on Intelligent Transportation Systems, *IEEE Transactions on Intelligent Transportation Systems* **1**(1): 55–64.

Zito, R., D'Este, G. and Taylor, M. A. P. (1995). Global positioning systems in the time domain: how useful a tool for intelligent vehicle-highway systems?, *Transportation Research C* **3**: 193–209.

Notation

Frequently used indices

m	monitoring detector	[-]
u	upstream detector	[-]
l	link	[-]
d	day	[-]
t	time interval of volume measurement	[min]
h	hour of the day	[-]
k	index for cluster	[-]
D	type of day (working day, Saturday, Sunday)	[-]
w	index for weekday (Mon, Tue, Wed, Thu, Fri)	[-]
f	index for month of the year	[-]
g	index for season	[-]

These indices are used in combination with variables in such a way that it is self-explanatory. The indexation is shortly explained on the basis of an example using traffic volume:

q_{ldt}	traffic volume for link l on day d during time interval t	[veh]
$\bar{q}_{lDkt}$	average traffic volume over all days in cluster k of day type D	[veh]
$\bar{q}_{lDt}$	average traffic volume over all days of day type D	[veh]

Frequently used variables

q	traffic volume	[veh]
σ	standard deviation	[-]
F	Ratio between standard deviations before and after clustering	[-]
s	proportion of daily traffic volume	[%]
π	ratio between q's for two types of days	[-]
δ	standardized difference between q on type of day and average day	[-]
$\mathbf{P}$	pattern, features used as a description of an entity that is clustered	
C	cluster, group of entities that show similar patterns	
Sd	set of days for which traffic data is collected	[-]
Sc	set of clusters	[-]
Nt	number of measurement intervals on each day	[-]
Nk	number of clusters	[-]
Nd	number of days	[-]
Nl	number of links	[-]
Nu	number of upstream detectors	[-]

Data validation

Q_i	i^{th} quality check	[-]
R_{md}	Record of traffic flow measurements on monitoring detector m and day d	[veh]
T_1	threshold for suspiciously high traffic volume	[-]
Su_m	set of upstream detectors that detect the same traffic as monitoring detector m	[-]
T_2	threshold regarding the number of vehicles queueing between two detectors	[-]
L_1, L_2	locations between which the principle of conservation of vehicles is applied	[-]
T_3	threshold regarding the maximum difference in traffic volume between L_1 and L_2	[%]
Sm	Set of monitoring detectors on which the principle of conservation of vehicles is applied	[-]

Analysis of traffic patterns

$\mathbf{P}_{ld}$	daily flow profile at link l on day d	[-]
C_{lk}	group k of days that show similar patterns on link l [-]	
ϕ	function that assigns a pattern $\mathbf{P}_{ld}$ to a cluster C_{lk}	[-]
Sc_l	set of clusters at link l [-]	
$\mathbf{P}_{lDk}$	average daily flow profile for cluster k at link l on daytype D	[-]
t_{lDk}^{AM},t_{lDk}^{PM}	peak times for cluster k	[-]
$\bar{q}_{lDk}^{AM}$,$\bar{q}_{lDk}^{PM}$	peak volumes for cluster k	[veh/h]
$\mathbf{P}_{lD}$	series of proportions s of daily traffic volume for day type D on link l	[-]
C_{Dk}	group k of links that show similar daily traffic profile	[-]
γ	function that assigns a pattern $\mathbf{P_{lD}}$ to a cluster C_{Dk}	[-]
$\mathbf{P}_l^{week1}$	set of ratios between Saturday and Sunday traffic and working day traffic for link l	[-]
$\mathbf{P}_l^{week2}$	series of 5*Nt differences δ that represent daily flow profiles on different weekdays compared to the average working day profile	[-]
$\mathbf{P}_l^{season}$	series of 12*Nt differences δ that represent daily flow profiles of different months compared to the average working day profile	[-]
$\mathbf{P}_l^{season'}$	series of 4*Nt differences δ that represent daily flow profiles of different seasons compared to the average working day profile	[-]
$\mathbf{P}_l^{weather}$	set of ratios between traffic volumes on wet and corresponding dry periods	[-]
$\mathbf{P}_l^c$	series of Nd_l clusters to which days are assigned to	[-]
Δ	Distance between two classifications: % of days that are classified to different clusters	[%]
$\mathbf{P}_d$	series of daily flow profiles at a set of selected links $Sl*$ on day d	[-]
$Sl*$	set of links that are selected as key locations	[-]

Appendix A

Quality control procedure

General formula for the i^th quality check:

$$Q_i\left(R_{md}\right) = \begin{cases} 0 & \text{if quality check is not passed} \\ 1 & \text{otherwise} \end{cases} \tag{A.1}$$

where

$$R_{md} = \left(q_{md,1}, ..., q_{mdt}, ..., q_{md,Nt}\right) \tag{A.2}$$

where q_{mdt} is a reported traffic volume for time interval t on day d and monitoring location m. Nt is the number of measurement intervals on a day.

Quality checks can be categorized into a check for missing data (Q_1), microscopic quality checks executed by the detector station (Q_2), basic macroscopic quality checks on the basis of maximum (Q_3) and minimum (Q_4) thresholds and macroscopic quality checks on the basis of the principle of conservation of vehicles between monitoring detectors and upstream detectors (Q_5) and between two sets of monitoring detectors (Q_6).

$$Q_1\left(R_{md}\right) = \begin{cases} 0 & \text{if } \exists_t q_{mdt} = \text{'missing'} \\ 1 & \text{otherwise} \end{cases} \tag{A.3}$$

$$Q_2\left(R_{md}\right) = \begin{cases} 0 & \text{if } \exists_t \text{state}_{mdt} > 0 \\ 1 & \text{otherwise} \end{cases} \tag{A.4}$$

where state is a variable that reports on the outcome of the micrsoscopic quality checks executed by the detector station itself

$$Q_3\left(R_{md}\right) = \begin{cases} 0 & \text{if } \exists_t \left(q_{mdt} > 3000 \vee \right. \\ & \left. \left(1000 \leq q_{mdt} \leq 3000 \wedge R_{md} \text{ looks abnormal}\right)\right) \\ 1 & \text{otherwise} \end{cases} \tag{A.5}$$

$$Q_{4a}\left(R_{md}\right) = \begin{cases} 0 & \text{if } \exists_t q_{mdt} < 0 \\ 1 & \text{otherwise} \end{cases} \tag{A.6}$$

$$Q_{4b}\left(R_{md}\right) = \begin{cases} 0 & \text{if } \sum_{h=8}^{19} q_{mdh} = 0 \\ 1 & \text{otherwise} \end{cases} \tag{A.7}$$

where q_{mdh} is the reported traffic volume during hour h on day d for m

$$Q_{4c}\left(R_{md}\right) = \begin{cases} 0 & \text{if } \exists_{h \in [8,19]} \left(q_{mdh} = 0 \wedge \left(\exists_{u \in Su_m} q_{udh} > 20 \vee \right.\right. \\ & \left.\left.\left(Su_m = \emptyset \wedge R_{md} \text{ looks abnormal}\right)\right)\right) \\ 1 & \text{otherwise} \end{cases} \tag{A.8}$$

where

$$q_{mdh} = \sum_{j=1}^{p} q_{md,(h-1)*p+1} \tag{A.9}$$

p is the number of measurement intervals in an hour, q_u are reported traffic volumes at location u that is part of Su_m, the set of upstream detectors that belong to monitoring detector m.

$$Q_{5a}\left(R_{md}\right) = \begin{cases} 0 & \text{if } Nu_{Su_m} \geq 2 \wedge \exists_{u \in Su_m} \left(T = 3 \vee T = 4\right) \wedge \\ & A_1\left(R_{md}\right) = 1 \wedge \forall_{u \in Su_m} \dfrac{\left|q_{md} - q_{ud}\right| - 20}{q_{md}} > 0.10 \\ 1 & \text{otherwise} \end{cases} \tag{A.10}$$

$$Q_{5b}\left(R_{md}\right) = \begin{cases} 0 & \text{if } Nu_{Su_m} \geq 2 \wedge \exists_{u \in Su_m} \left(T = 3 \vee T = 4\right) \wedge \exists_h \\ & \left(A_2\left(q_{mdh}\right) = 1 \wedge \forall_{u \in Su_m} \dfrac{\left|q_{mdh} - q_{udh}\right| - 20}{q_{mdh}} > 0.10\right) \\ 1 & \text{otherwise} \end{cases} \tag{A.11}$$

where Nu_{Su_m} is the number of upstream detectors that are part of Su_m and T is the type of detector

$$A_1\left(R_{md}\right) = \begin{cases} 0 & \text{if } \exists_{u \in Su_m} q_{md} - q_{ud} > 0 \wedge \exists_{u \in Su_m} q_{md} - q_{ud} < 0 \\ 1 & \text{otherwise} \end{cases} \tag{A.12}$$

$$A_2\left(q_{mhd}\right) = \begin{cases} 0 & \text{if } \exists_{u \in Su_m} q_{mhd} - q_{uhd} > 0 \wedge \exists_{u \in Su_m} q_{mhd} - q_{uhd} < 0 \\ 1 & \text{otherwise} \end{cases} \tag{A.13}$$

In Almelo, there are eight sets of monitoring detectors Sm^1 to Sm^8 for which the principle of conservation of vehicles can be applied. The sets are defined in such a way that the flow conservation law can be applied between Sm^1 and Sm^2, between Sm^3 and Sm^4, etc. To symplify the notation, the first set of detectors that is included in the check is denoted by Sm^a and the second set is denoted by Sm^b. This results in the following quality checks:

$$Q_{6a}\left(R_{md}\right) = \begin{cases} 0 & \text{if } m \in \{Sm^1, ..., Sm^8\} \wedge \\ & \dfrac{\left|q_{Sm^a d} - q_{Sm^b d}\right| - 20}{0.5 * \left(q_{Sm^a d} + q_{Sm^b d}\right)} > 0.05 \\ & \wedge \neg_{m \in Sm^a Sm^b} \exists_{i \in 1 \ldots 5} Q_i = 0 \\ 1 & \text{otherwise} \end{cases} \tag{A.14}$$

$$Q_{6b}\left(R_{md}\right) = \begin{cases} 0 & \text{if } m \in \{Sm^1, ..., Sm^8\} \wedge \\ & \exists_h \dfrac{\left|q_{Sm^a hd} - q_{Sm^b hd}\right| - 20}{0.5 * \left(q_{Sm^a hd} + q_{Sm^b hd}\right)} > 0.05 \\ & \wedge \neg_{m \in Sm^a S^b} \exists_{i \in 1 \ldots 5} Q_i = 0 \\ 1 & \text{otherwise} \end{cases} \tag{A.15}$$

A record is removed from further analysis when one or more of these quality checks are not passed, i.e.

$$\prod_i Q_i\left(R_{md}\right) = 0 \tag{A.16}$$

Appendix B

Daily flow profiles at main arterials

Henriette Roland Holstlaan

Figure B.1 shows the average working day flow profile for traffic in both directions. These flow profiles are from the location with the highest traffic load (K31). Most other locations along the route show similar daily traffic profiles, although the peak volumes vary between the locations (see Table B.1). The daily traffic profiles on Saturdays and Sundays show comparable patterns to the general pattern and therefore are not shown here.

From Figure B.1 it can be concluded that the distribution of the traffic over the day differs between the directions. Traffic in the direction of the highway shows a relatively high P.M peak, whereas traffic from the highway shows a relatively high A.M. peak. Moreover, locations between the Weezebeeksingel and the A35 show a more peaked flow profile than locations between the Weezebeeksingel and K7 (see Table B.1) This is probably due to a difference in the share of

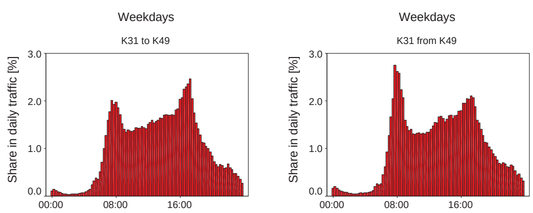


Figure B.1: Daily flow profile on an average working day to (left and from (right) the highway.

Table B.1: Peak volumes on H. Roland Holstlaan *=only data available at 30-minute level.

	To highway			From highway	
	A.M. peak	P.M. peak		A.M. peak	P.M. peak
K32 from K7	6.7%	8.9%	A35 to K49*	9.4%	9.4%
K33 from K32	6.8%	9.0%	K49 to K31*	9.2%	8.2%
K31	7.8%	9.4%	K31	10.4%	8.2%
K49 from K31*	8.0%	9.5%	K32 from K33	8.7%	8.5%
K49 to A35*	9.8%	9.5%	K7 to K32*	8.7%	8.3%

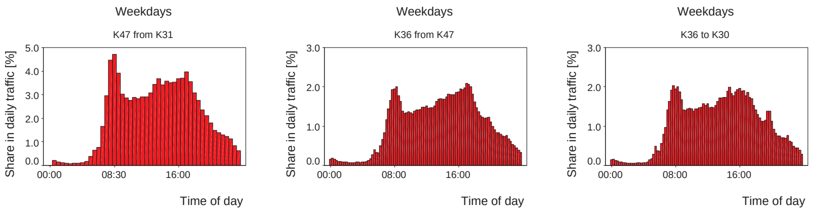


Figure B.2: Daily flow profile on Weezebeeksingel towards K30.

commuter traffic (being larger between the Weezebeeksingel and the A35).

Weezebeeksingel

For this route, it is not possible to show one typical daily flow profile for each direction. Figure B.2 shows the different working day flow profiles for traffic in the direction of de Wierdensestraat, whereas Figure B.3 shows the daily flow profiles for the traffic in opposite direction.

For the traffic towards the Wierdensestraat, the average daily flow profile of the traffic entering K47 does not show a real evening peak, whereas the flow profile of the traffic entering the next intersection (K36) shows an evening peak that is as high as the A.M. peak. After K36, the average daily flow profile changes again, showing higher traffic volumes all afternoon and a peak around 20:00.

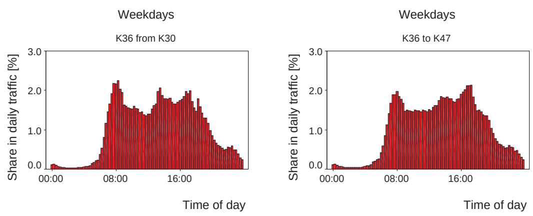


Figure B.3: Daily flow profile on Weezebeeksingel towards K31.

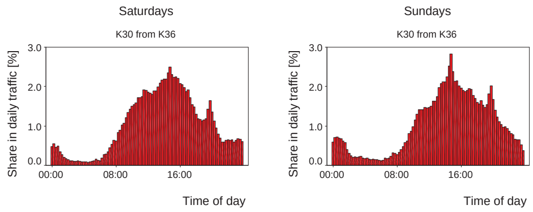


Figure B.4: Daily flow profiles on Saturdays and Sundays for traffic entering K30.

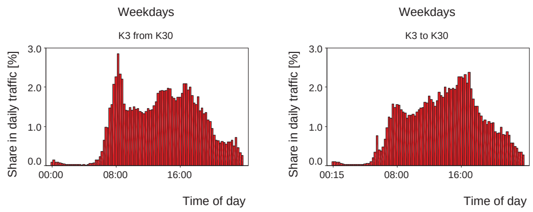


Figure B.5: Daily traffic profiles on Wierdensestraat on working days.

This peak is probably caused by traffic that leaves the hospital after visiting hours. The traffic that joins the Weezebeeksingel in the afternoon (at K47 and K36) originates from the residential area, the Hospital and the furniture mall.

The effects of the hospital and residential areas can also be seen for the traffic in opposite direction. Figure B.3 shows that relatively many vehicles leave the Weezebeeksingel at K36 during the A.M. peak and the afternoon period. Also on Saturdays and Sundays, the influence of the hospital on the traffic flows is visible. Both daily traffic profiles show peak periods that match with the visiting hours of the hospital (see figure B.4).

Wierdensestraat

All locations along the arterial show similar daily flow profiles. Figure B.5 shows the average working day flow profiles for both directions. Also for this route, the main direction of travel differs between the peak periods.

For the traffic in the direction of the city centre, the Saturday and Sunday flow profiles deviate from the general flow profiles (see figure B.6). Saturdays show a relatively peaked profile and Sundays show a relatively unsteady profile with peaks in the morning and afternoon.

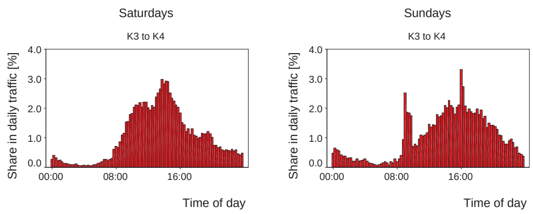


Figure B.6: Daily traffic profiles on Saturdays and Sundays for traffic towards the city centre.

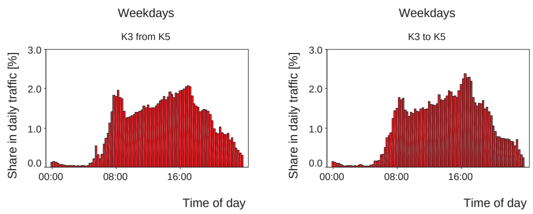


Figure B.7: Daily traffic profiles on the Schoolstraat on working days.

Egbert ten Catelaan - Schoolstraat

Figure B.7 shows the average daily flow profiles for working days in both directions for one of the locations along the route. Although the P.M. peak is higher than the A.M. peak for both directions, it can be seen that during the A.M. peak relatively a lot of traffic is traveling in the direction of K3, whilst during the P.M. peak traffic volumes are relatively high in opposite direction. These traffic patterns can be explained by commuter traffic departing from and returning to residential areas along the Schoolstraat. The relatively high off-peak and P.M. peak in the direction of K3 can be explained by home based trips with other purposes (e.g. shopping, school) and by through traffic towards the Aalderinkshoek (a residential area connected by the Aalderinkssingel). The Saturday and Sunday profiles are comparable to the general profiles on these days.

Aalderinkssingel

Figure B.8 shows the daily flow profiles for the traffic between K2 and K3. During the A.M. peak, relatively a lot of traffic is traveling from K3 towards K2. This traffic probably goes to the industrial areas along the Aalderinkssingel. The traffic in opposite direction shows a much smaller A.M. peak. This peak is probably caused by traffic from the residential area Aalderinkshoek as is the relatively high off-peak afternoon flow. During the P.M. peak, traffic

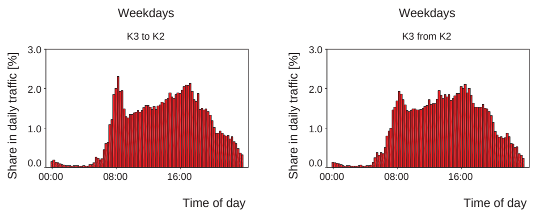


Figure B.8: Daily traffic profiles on the Aalderinkssingel on working days.

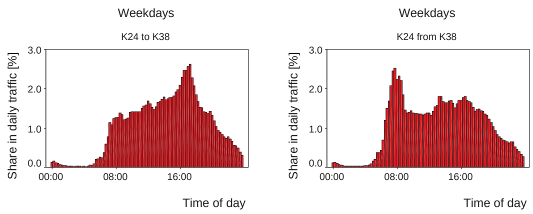


Figure B.9: Daily traffic profiles to and from K23.

loads are comparable for both directions. Traffic in the direction of the Aalderinkssingel probably is commuter traffic that returns to the residential areas, whereas traffic in the direction of the Wierdensestraat is commuter traffic that returns home from the industrial areas. Saturday and Sunday flow profiles are comparable to the general flow profiles.

Sluitersveldsingel

Also at the Sluitersveldsingel the amount of traffic in each direction varies largely between the peak periods (see figure B.9). These daily flow profiles can be explained by traffic that leaves the residential areas during the A.M. peak period (towards the Van Rechteren Limpurgsingel and Kolthofsingel that further direct the traffic towards the industrial areas or regional roads) and returns to the residential areas during the P.M. peak period. Daily traffic profiles on Saturdays and Sundays are comparable to the general profiles on these days, although the peak on Sunday is relatively high for the traffic in the direction of K35.

Schuilenburgsingel - Bleskolksingel

From Section 6.2 was already concluded that different parts of the route serve different traffic. This can also be seen from the daily flow profiles. Especially for the traffic in the direction of K43, the shape of the daily flow profile changes

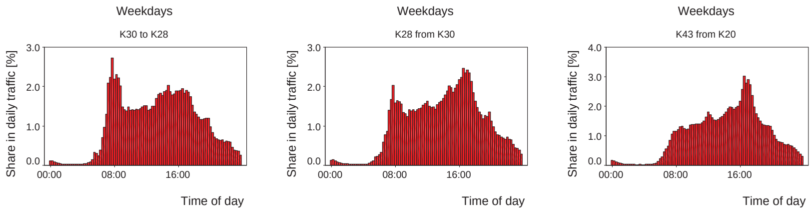


Figure B.10: Daily traffic profiles on Schuilenburgsingel - Bleskolksingel towards K43 on working days.

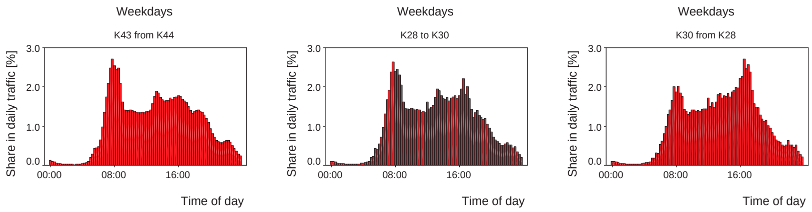


Figure B.11: Daily traffic profiles on Schuilenburgsingel - Bleskolksingel towards K30 on working days.

along the route (see Figure B.10). The traffic leaving K30 shows a high A.M. peak that can be explained by traffic towards industrial areas between K30 and K28. The traffic entering K28 shows a higher P.M. peak than A.M. peak, and further along the route the A.M. peak disappears and the P.M. peak is even higher. The high P.M. peak is probably caused by commuter traffic leaving the industrial areas. The traffic in the other direction shows an opposite pattern (see Figure B.11), the A.M. peak being higher for traffic between K43 and K28 and the P.M. peak being higher for traffic entering K30. Traffic entering K30 also shows a small A.M. peak, which is probably caused by traffic from small residential areas that commutes to areas outside Almelo. At most locations, Saturdays and Sundays show a somewhat more peaked flow profile than in general (see Figure B.12).

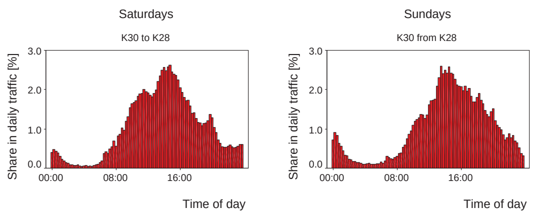


Figure B.12: Daily traffic profiles on Schuilenburgsingel - Bleskolksingel on Saturdays and Sundays.

Appendix C

Resulting clusters on network level

Table C.1: Cross-tabulation of weekdays with cluster.

	Mon	Tue	Wed	Thu	Fri	Tot (excl hol)
Cluster 1	5 days	4 days	5 days	1 day	0	15 days
Cluster 2	5 days	6 days	6 days	9 days	0	26 days
Cluster 3	0	0	0	0	4 days	4 days
Cluster 4	0	0	0	0	5 days	5 days
Total	10 days	10 days	11 days	10 days	9 days	50 days

Table C.2: Cross-tabulation of seasons and holiday periods with clusters.

	Winter	Spring	Tot (excl holidays)	holidays
Cluster 1	10 days	5 days	15 days	0
Cluster 2	3 days	23 days	26 days	0
Cluster 3	2 days	2 days	4 days	4 days
Cluster 4	1 day	4 days	5 days	0 days
Total	16 days	34 days	50 days	4 days

Table C.3: Indices of peak and off-peak volumes of clusters at different locations.

	Indices (average over all clusters = 100)			
	cluster 1	cluster 2	cluster 3	cluster 4
K49 from K31				
7:00 - 9:00	99	112	86	103
9:00 - 16:00	91	102	97	110
16:00 - 18:00	99	110	90	100
K30 to K36				
7:00 - 9:00	104	108	88	100
9:00 - 16:00	94	101	100	105
16:00 - 18:00	99	105	95	101
K30 to K3				
7:00 - 9:00	107	107	85	101
9:00 - 16:00	94	101	97	108
16:00 - 18:00	99	107	93	101
K5 from K6				
7:00 - 9:00	108	105	90	97
9:00 - 16:00	95	99	102	104
16:00 - 18:00	100	102	98	100
K23 from K24				
7:00 - 9:00	105	107	87	100
9:00 - 16:00	94	98	100	108
16:00 - 18:00	97	103	95	105
K30 from K28				
7:00 - 9:00	102	108	88	101
9:00 - 16:00	93	101	99	108
16:00 - 18:00	100	107	95	98
K20 to K43				
7:00 - 9:00	102	107	88	103
9:00 - 16:00	93	99	99	109
16:00 - 18:00	102	106	94	99

Summary

Mobility is still increasing, as are its corresponding negative side effects such as congestion and air pollution. To be able to take adequate measures to minimize these side effects, it is important to obtain insight into the functioning of the traffic system. In common practice, the traffic analysis process deals with average traffic volumes. However, also the variability of traffic volumes is of crucial importance, for example with regard to travel time reliability, the robustness of the road network and dynamic traffic management. Moreover, until recently, the main focus was on highway systems. Since mobility is also increasing in urban areas, it is also important to obtain insight into the functioning of the urban traffic system. This thesis aims at improving the insight into variations in urban traffic by analysing within and between day variations in traffic volumes.

The traffic volume data that is necessary for the analysis is provided by urban traffic information centres. The data is processed to make it suitable for research. We developed a quality control procedure that detects invalid daily records of volume measurements applying minimum and maximum volume thresholds and the principle of conservation of vehicles. The valid traffic data is subsequently linked to data on factors potentially causing variations in traffic volumes (i.e. calendar data, weather data, data on road works and events and accident data).

Most existing literature on within and between day variations in urban traffic volumes deals with the general shape of a daily flow profile and the influence of weekday, season, holiday periods and weather on total daily traffic volumes. However, also the shape of the daily flow profile may differ between different types of days. The insight into these differences is limited, especially for urban areas. Besides, other factors may also cause variations in traffic volumes. Regarding spatial variations, only little is known about differences in temporal patterns between different (types of) urban locations.

We propose an alternative approach for the analysis of temporal and spatial variations in urban traffic volumes, using cluster analysis. Temporal patterns are analysed by grouping days that show similar daily flow profiles. Thereby, a daily flow profile is defined as a series of traffic volume measurements. Cluster

analyses are executed for both working days and non-working days. By means of cross-tabulations and Chi-square tests it is investigated whether the factors (1) weekday, (2) season, (3) holiday periods and (4) rain are responsible for the classification. If the resulting groups cannot be explained by one or more of these factors, the clusters are analysed further to investigate whether other factors are responsible for the classification. Finally, by comparing the standard deviation before and after classification, it is investigated to what extent the clustering reduces the variation.

Spatial traffic patterns are analysed by grouping links (i.e. road segments that facilitate traffic in one direction) that show similar temporal variations. Links are grouped according to: (1) their average daily flow profile, (2) their traffic volumes on weekend days in relation to weekdays, (3) their weekly variations, (4) their seasonal variations, (5) the influence of rain on their traffic volumes and (6) the clusters that result from the clustering of daily flow profiles. By means of a Geographical Information System (GIS) it can be visualized which links are assigned to each cluster and it can be investigated what these links have in common. Also for the groups of locations, it is investigated to what extent the clustering reduces the standard deviation.

Finally, we propose two approaches for the analysis of traffic patterns on a network level. The first combines the results of the temporal and spatial analyses, whilst the second defines a pattern by series of daily flow profiles at several locations.

The proposed framework is applied to traffic data of Almelo, a city with approximately 70.000 inhabitants in the east of The Netherlands. The traffic data is provided by the ViaContent traffic information system. Traffic volumes are collected by inductive loop detectors at 23 signalized intersections throughout the city for the period September 2004 - September 2005. The quality control procedure is evaluated and appears to function adequately. Despite a significant amount of missing data, for most links valid traffic data is available for a considerable number of days (on average 118 working days). Unfortunately, the traffic data is unevenly distributed over the seasons; the amount of missing data is relatively high for summer and autumn months. The data appears to be very accurate, at least for the locations at which field observations were carried out.

From the application to Almelo it is concluded that the analysis framework functions adequately. The cluster analyses result in distinctive, recurrent and representative traffic patterns that can be explained by travel demand and supply factors. In most cases, after clustering the variation between the days within a cluster is clearly smaller than before clustering.

The clustering of non-working days in most cases results in a Saturday cluster and a Sunday cluster of which the Saturday cluster shows higher traffic volumes throughout all day. For working days, the results differ by location. In most cases, the clustering results in two to four clusters. Weekday, holiday periods and road works are the main decisive factors for the classification. Weekday and holiday clusters can be explained by activity patterns. Friday clusters for example show a relatively flat daily flow profile as a result of less commuter traffic and more shopping and leisure traffic compared to other working days. Road works in some cases result in very dissimilar daily flow profiles as the consequence of a decrease in capacity, whilst in other cases they result in slightly higher traffic volumes due to re-routing effects. Also other location specific factors like events and invalid data play a role in some classifications. Rain influences the clustering only at about 10% of the locations and also season is not a main decisive factor. In some cases, a combination of season and weekday is on the basis of a cluster. The resulting working day clusters are hardly ever totally homogeneous and complete. In most cases, a cluster contains relatively many days of a certain type, yet not exclusively and not all.

With regard to the clustering of links, the results of the cluster analyses according to daily flow profiles and weekly variations can be explained by differences in travel motives. Links that are used by a mixture of travel motives show both an A.M. and a P.M. peak whereas locations that are mainly used by commuter traffic show only an A.M. peak (towards work) or a P.M. peak (from work). With regard to weekly variations, links that contain a large proportion of shopping traffic show stronger variations than other links. The results of the clustering according to seasonal variations and the influence of weather can be explained by differences in the proportion of short distance traffic. Locations with relatively much short distance traffic show relatively low traffic volumes in summer and relatively high traffic volumes on rainy days as a result of a modal shift between bicycle and car. The clustering of locations on the basis of their temporal traffic patterns does not lead to clear groups of links. Classifications between pairs of links differ from each other in several ways, yet there are no clear groups of links that show classifications that are similar to each other and dissimilar from links in other groups. However, there are some groups of links that show specific patterns, like a Thursday cluster or a Heracles[1] cluster.

On a network level, a clustering on the basis of daily flow profiles at seven key locations results in four clusters, determined by home matches of Heracles and combinations of weekday and holiday periods and weekday and season.

The results of the analyses are not distorted by the selected method, available data or the occurrence of congestion. Moreover, from a comparison between the traffic patterns in Almelo and in Amsterdam it seems that the general traffic patterns are spatially transferable.

The obtained insight into existent traffic patterns can be used for traffic monitoring. Moreover, the average flow profiles of the resulting clusters can be used for traffic forecasting, traffic management and transport modelling scenarios. The advantage of cluster analysis over methods that group days on the basis of pre-defined factors is that cluster analysis provides an easy way to

[1]Heracles is a local football club that played in the First Division at the time of the data collection.

detect and analyse atypical circumstances, such as road works or events. The main disadvantage is that in some cases only few clusters are distinguished. We propose a hybrid model for the determination of typical traffic patterns that combines the strengths of both methods. Cluster analysis is used as a filter; atypical, non-recurrent patterns are grouped in different clusters. Subsequently, the resulting groups are further classified according to weekday and holiday periods.

In conclusion, in this thesis we propose and apply a general method for the determination and analysis of urban traffic patterns. The research provides insight into within and between day variations in urban traffic. Besides, we propose a hybrid model that can be used for the determination of typical traffic patterns that serve as a basis for traffic forecasting, traffic management and traffic modelling scenarios.

Samenvatting

Mobiliteit is een noodzakelijke voorwaarde voor economische groei en sociale ontwikkeling in Nederland (MinVenW, 2004), maar heeft ook een aantal negatieve neveneffecten zoals congestie en luchtvervuiling. Voor de ontwikkeling van maatregelen om deze te verminderen zonder de mobiliteit teveel te beperken is inzicht nodig in het functioneren van het verkeerssysteem. De huidige praktijk richt zich met name op gemiddelde intensiteiten op het hoofdwegennet. Echter, de variatie in verkeersintensiteiten is ook van cruciaal belang, bijvoorbeeld met betrekking tot de betrouwbaarheid van reistijden, de robuustheid van het wegennet en dynamisch verkeersmanagement. Daarnaast is het ook belangrijk maatregelen te ontwikkelen voor de stedelijke situatie. Daarom staat in dit proefschrift de analyse van dag-tot-dag variaties in stedelijke verkeersintensiteiten centraal.

Bestaand onderzoek naar variaties in verkeersintensiteiten richt zich met name op de invloed van (1) de dag van de week, (2) het seizoen, (3) de vakantie en (4) het weer op de totale dagintensiteit. Echter, ook de vorm van het intensiteitsprofiel zou kunnen variëren tussen verschillende typen dagen. Het inzicht in deze verschillen is beperkt, met name voor stedelijke gebieden. Daarnaast zouden ook andere factoren variaties in verkeersintensiteiten kunnen veroorzaken. Tot slot is weinig bekend over verschillen in dag-tot-dag variaties tussen verschillende typen stedelijke locaties.

In dit proefschrift is een alternatieve methode ontwikkeld voor de analyse van temporele en ruimtelijke variaties in stedelijke verkeersintensiteiten. Deze methode maakt gebruik van clusteranalyse. Door werkdagen en "niet-werkdagen" (weekenddagen, feestdagen en brugdagen) te groeperen op basis van hun intensiteitsprofiel (een serie van gemeten verkeersintensiteiten) wordt een aantal typische dagprofielen onderscheiden. Met behulp van kruistabellen en chi-kwadraattoetsen is vervolgens onderzocht of (1) weekdag, (2) seizoen, (3) vakantie en (4) regen bepalend zijn voor de clustering. Als de clustering niet verklaard kan worden door deze factoren, worden de clusters verder onderzocht om te bepalen welke andere factoren ten grondslag liggen aan de clustering. Door de standaarddeviatie voor en na clustering met elkaar te vergelijken wordt onderzocht in hoeverre de clustering de variatie vermindert. Ruimtelijke patronen worden geanalyseerd door links (weggedeelten) te groeperen op basis van hun temporele variaties, namelijk op basis van hun (1) gemiddelde dagprofiel, (2) intensiteit op zaterdag en zondag in relatie tot de gemiddelde werkdag intensiteit, (3) variaties in dagprofielen tussen werkdagen, (4) variaties tussen seizoenen, (5) de invloed van regen en (6) de resultaten van de temporele clustering. Met behulp van een Geografisch Informatie Systeem (GIS) wordt gevisualiseerd welke links in welk cluster worden ingedeeld. Ook voor deze analyses wordt bepaald in welke mate de clustering de standaarddeviatie verlaagt. Tenslotte worden twee methoden voorgesteld voor de analyse van verkeerspatronen op netwerkniveau. De eerste methode combineert de resultaten van de temporele en ruimtelijke analyses en de tweede methode definieert een patroon als een serie van intensiteitsprofielen op verschillende locaties.

De data die gebruikt wordt voor de analyses is afkomstig van stedelijke verkeersinformatiecentrales. Deze data moet gecontroleerd worden voor deze voor de analyse gebruikt kan worden. In dit proefschrift wordt een datacontrole-procedure gepresenteerd die gebruik maakt van een onder- en een bovenlimiet voor intensiteiten en het principe van het behoud van voertuigen. Daarna wordt de verkeersdata gekoppeld aan data met betrekking tot factoren die variaties in verkeersintensiteiten kunnen veroorzaken, zoals kalenderdata en data met betrekking tot wegwerkzaamheden).

De voorgestelde methoden zijn toegepast op verkeersgegevens uit Almelo. De verkeersdata is afkomstig van het verkeersinformatiesysteem ViaContent. Verkeersintensiteiten zijn gemeten middels inductielussen op 23 kruispunten met verkeerslichten. Voor het onderzoek is data van de periode september 2004 - september 2005 gebruikt. De datacontrole-procedure is geëvalueerd en blijkt goed te functioneren. Als gevolg van communicatiestoringen tussen het detectorstation en de centrale computer is niet op alle dagen voor alle kruispunten data verzameld. Ondanks deze storingen is voor de meeste links een aanzienlijke hoeveelheid data beschikbaar (gemiddeld 118 werkdagen per link). Helaas is de data niet gelijk verdeeld over de seizoenen; voor de zomer- en herfstmaanden is relatief weinig data beschikbaar. Om de nauwkeurigheid van de gegevens te onderzoeken zijn op een aantal locaties tellingen uitgevoerd. Voor deze locaties bleek de data uit ViaContent zeer nauwkeurig te zijn.

Uit de toepassing van de methode op de verkeersdata uit Almelo kan geconcludeerd worden dat de methode goed functioneert. De clusteranalyses leiden tot een indeling in duidelijk te onderscheiden, terugkerende en representatieve verkeerspatronen die verklaard kunnen worden door verkeersvraag en -aanbod factoren. In de meeste gevallen is de variatie na de clustering duidelijk lager dan voor de clustering.

De clustering van weekenddagen resulteert voor de meeste links in een zaterdag- en een zondagcluster. Het zaterdagcluster vertoont gedurende de hele dag hogere intensiteiten dan het zondagcluster. Voor werkdagen verschillen de resultaten per locatie. In de meeste gevallen resulteert de clustering in twee tot vier clusters. Weekdagen, vakantieperioden en wegwerkzaamheden zijn de belangrijkste verklarende variabelen. Verschillen tussen weekdagen onderling en tussen weekdagen en vakantieperioden kunnen verklaard worden door verschillen in activiteitenpatronen. Clusters met veel vrijdagen vertonen bijvoorbeeld een relatief vlak dagprofiel als gevolg van relatief weinig woon-werkverkeer en relatief veel winkel- en vrijetijdsverkeer vergeleken met andere werkdagen. Wegwerkzaamheden leiden in sommige gevallen tot zeer afwijkende dagprofielen als gevolg van een lagere capaciteit en in andere gevallen tot iets hogere intensiteiten als gevolg van een verandering in routekeuze. Ook andere locatiespecifieke factoren zoals evenementen en onjuiste data spelen een rol op een aantal locaties. Regen blijkt slechts op 10% van de locaties een rol te spelen in de clustering. Ook seizoen blijkt in het algemeen geen grote rol te spelen, al wordt een cluster in sommige gevallen bepaald door een combinatie van seizoen en werkdag. De resulterende clusters zijn bijna nooit helemaal homogeen en compleet: in de meeste gevallen bevat een cluster relatief veel dagen van een bepaald type, maar niet alle dagen en niet exclusief dagen van dat type.

Met betrekking tot de clustering van links, kunnen de resultaten van de clustering op basis van dag- en weekprofielen verklaard worden door verschillen in motiefverdeling. Links die worden gebruikt door een mix van reismotieven vertonen een duidelijke ochtend- en avondspits, terwijl links die vooral door woon-werkverkeer gebruikt worden alleen een duidelijke ochtendspits (verkeer richting werk) of een duidelijke avondspits (verkeer richting huis) laten zien. Links met relatief veel winkelverkeer laten meer variatie tussen weekdagen zien dan andere links. De resultaten van de clusteringen op basis van seizoenspatronen en de invloed van regen kunnen verklaard worden door verschillen in het percentage korteafstandsverkeer. Links met veel korteafstandsverkeer vertonen relatief lage intensiteiten in de zomer en relatief hoge intensiteiten op dagen met regen als gevolg van een 'modal-shift' tussen auto en fiets. De clustering van links op basis van hun temporele clusters leidt niet tot een indeling in duidelijk te onderscheiden, representatieve en terugkerende typische verkeerspatronen. Locaties verschillen op verscheidene manieren van elkaar wat betreft de clustering van dagprofielen, maar er zijn geen duidelijke groepen locaties die dezelfde clusters laten zien. Er is wel een aantal groepen links waarvoor een van de clusters overeenkomt (bijvoorbeeld een donderdagcluster of een cluster met dagen waarop de voetbalclub Heracles thuis speelde). Op netwerkniveau is een patroon gedefinieerd als een serie van dagprofielen op zeven locaties. De clustering resulteert in vier clusters, die verklaard kunnen worden door thuiswedstrijden van Heracles en combinaties van weekdag en vakantieperioden en weekdag en seizoen.

Uit nadere analyses blijkt dat de algemene resultaten niet beïnvloed zijn door de gekozen methode, beschikbare data en congestie. Daarnaast blijkt uit een vergelijking tussen verkeerspatronen in Almelo en in Amsterdam dat de reguliere patronen (clusters bepaald door weekdagen en vakantieperioden) overeenkomen. Deze algemene patronen zijn dus zeer waarschijnlijk ruimtelijk overdraagbaar.

Het verworven inzicht in bestaande verkeerspatronen kan gebruikt worden voor monitoring. Daarnaast kunnen de gemiddelde dagprofielen van de clusters gebruikt worden voor verkeersvoorspellingen, verkeersmanagement en verkeersmodellering. Het voordeel van clusteranalyse ten opzichte van methoden die dagen groeperen op basis van vooraf gedefinieerde factoren is dat bijzondere omstandigheden als wegwerkzaamheden en evenementen eenvoudig gedetecteerd en geanalyseerd kunnen worden. Het nadeel is dat in sommige gevallen te weinig clusters onderscheiden worden, waardoor de variatie binnen de clusters nog te groot is. We hebben een hybride model gepresenteerd dat de voordelen van clusteranalyse en een classificatie op basis van weekdag en vakantieperiode combineert. Clusteranalyse wordt daarbij als een filter gebruikt voor de detectie van bijzondere omstandigheden. Vervolgens worden de resulterende groepen verder onderverdeeld op basis van weekdag en vakantieperioden.

Concluderend hebben we een algemeen toepasbare methode ontwikkeld en toegepast voor het detecteren en analyseren van stedelijke verkeerspatronen. Het onderzoek levert inzicht in temporele en ruimtelijke variaties in stedelijk verkeer. Daarnaast hebben we een hybride model gepresenteerd dat gebruikt kan worden voor het definiëren van typische patronen en dat als basis kan dienen voor verkeersvoorspellingen, verkeersmanagement en verkeersmodellering.

About the author

Wendy Weijermars was born on 25 December 1977 in Rijnsburg, The Netherlands. From 1990 onwards she attended the KSE in Etten-Leur and received her VWO diploma in 1996. In the same year she started her academic education by studying Civil Engineering & Management at the University of Twente, The Netherlands. Wendy specialized in traffic and transport and followed some additional courses on traffic psychology at the Rijksuniversiteit Groningen, The Netherlands. For her internship she went to the University of Pretoria, South Africa where she carried out a research on traffic system performance indicators. She finished her Master in 2001 with a thesis on traffic safety on roundabouts. This thesis was rewarded with the first price for best Master's thesis of the Civil Engineering & Management programme for that year and the third price for best national Master's thesis on traffic and transport (Cuperusprijs 2003).

In February 2002 she started as a junior researcher at the Centre for Transport Studies of the University of Twente. Together with her promotor Eric van Berkum, she wrote a proposal for the PhD research on urban traffic patterns that is described in this thesis. Besides doing her PhD research she supervised Master students and taught courses on regional traffic management and on survey methods. From November 2006 until April 2007 she worked at the International Institute for Geo-Information Science and Earth Observation (ITC), while finishing her PhD thesis.

TRAIL Thesis Series

A series of The Netherlands TRAIL Research School for theses on transport, infrastructure and logistics.

Nat, C.G.J.M., van der, *A Knowledge-based Concept Exploration Model for Submarine Design*, T99/1, March 1999, TRAIL Thesis Series, Delft University Press, The Netherlands

Westrenen, F.C., van, *The Maritime Pilot at Work: Evaluation and Use of a Time-to-boundary Model of Mental Workload in Human-machine Systems*, T99/2, May 1999, TRAIL Thesis Series, Eburon, The Netherlands

Veenstra, A.W., *Quantitative Analysis of Shipping Markets*, T99/3, April 1999, TRAIL Thesis Series, Delft University Press, The Netherlands

Minderhoud, M.M., *Supported Driving: Impacts on Motorway Traffic Flow*, T99/4, July 1999, TRAIL Thesis Series, Delft University Press, The Netherlands

Hoogendoorn, S.P., *Multiclass Continuum Modelling of Multilane Traffic Flow*, T99/5, September 1999, TRAIL Thesis Series, Delft University Press, The Netherlands

Hoedemaeker, M., *Driving with Intelligent Vehicles: Driving Behaviour with Adaptive Cruise Control and the Acceptance by Individual Drivers*, T99/6, November 1999, TRAIL Thesis Series, Delft University Press, The Netherlands

Marchau, V.A.W.J., *Technology Assessment of Automated Vehicle Guidance - Prospects for Automated Driving Implementation*, T2000/1, January 2000, TRAIL Thesis Series, Delft University Press, The Netherlands

Subiono, *On Classes of Min-Max-Plus Systems and Their Applications*, T2000/2, June 2000, TRAIL Thesis Series, Delft University Press, The Netherlands

Meer, J.R., van, *Operational Control of Internal Transport*, T2000/5, September 2000, TRAIL Thesis Series, Delft University Press, The Netherlands

Bliemer, M.C.J., *Analytical Dynamic Traffic Assignment with Interacting User-Classes: Theoretical Advances and Applications using a Variational Inequality Approach*, T2001/1, January 2001, TRAIL Thesis Series, Delft University Press, The Netherlands

Muilerman, G.J., *Time-based Logistics: An Analysis of the Relevance, Causes and Impacts*, T2001/2, April 2001, TRAIL Thesis Series, Delft University Press, The Netherlands

Roodbergen, K.J., *Layout and Routing Methods for Warehouses*, T2001/3, May 2001, TRAIL Thesis Series, The Netherlands

Willems, J.K.C.A.S., *Bundeling van Infrastructuur, Theoretische en Praktische Waarde van een Ruimtelijk Inrichtingsconcept*, T2001/4, June 2001, TRAIL Thesis Series, Delft University Press, The Netherlands

Binsbergen, A.J., van, J.G.S.N. Visser, *Innovation Steps towards Efficient Goods Distribution Systems for Urban Areas*, T2001/5, May 2001, TRAIL Thesis Series, Delft University Press, The Netherlands

Rosmuller, N., *Safety Analysis of Transport Corridors*, T2001/6, June 2001, TRAIL Thesis Series, Delft University Press, The Netherlands

Schaafsma, A., *Dynamisch Railverkeersmanagement, Besturingsconcept voor Railverkeer op basis van het Lagenmodel Verkeer en Vervoer*, T2001/7, October 2001, TRAIL Thesis Series, Delft University Press, The Netherlands

Bockstael-Blok, W., *Chains and Networks in Multimodal Passenger Transport. Exploring a design approach*, T2001/8, December 2001, TRAIL Thesis Series, Delft University Press, The Netherlands

Wolters, M.J.J., *The Business of Modularity and the Modularity of Business*, T2002/1, February 2002, TRAIL Thesis Series, The Netherlands

Vis, F.A., *Planning and Control Concepts for Material Handling Systems*, T2002/2, May 2002, TRAIL Thesis Series, The Netherlands

Koppius, O.R., *Information Architecture and Electronic Market Performance*, T2002/3, May 2002, TRAIL Thesis Series, The Netherlands

Veeneman, W.W., *Mind the Gap; Bridging Theories and Practice for the Organisation of Metropolitan Public Transport*, T2002/4, June 2002, TRAIL Thesis Series, Delft University Press, The Netherlands

Van Nes, R., *Design of Multimodal Transport Networks: A Hierarchical Approach*, T2002/5, September 2002, TRAIL Thesis Series, Delft University Press, The Netherlands

Pol, P.M.J., *A Renaissance of Stations, Railways and Cities, Economic Effects, Development Strategies and Organisational Issues of European High-Speed-Train Stations*, T2002/6, October 2002, TRAIL Thesis Series, Delft University Press, The Netherlands

Runhaar, H., *Freight Transport: At Any Price? Effects of Transport Costs on Book and Newspaper Supply Chains in The Netherlands*, T2002/7, December 2002, TRAIL Thesis Series, Delft University Press, The Netherlands

Spek, S.C., van der, *Connectors. The Way beyond Transferring*, T2003/1, February 2003, TRAIL Thesis Series, Delft University Press, The Netherlands

Lindeijer, D.G., *Controlling Automated Traffic Agents*, T2003/2, February 2003, TRAIL Thesis Series, Eburon, The Netherlands

Riet, O.A.W.T., van de, *Policy Analysis in Multi-Actor Policy Settings. Navigating Between Negotiated Nonsense and Useless Knowledge*, T2003/3, March 2003, TRAIL Thesis Series, Eburon, The Netherlands

Reeven, P.A., van, *Competition in Scheduled Transport*, T2003/4, April 2003, TRAIL Thesis Series, Eburon, The Netherlands

Peeters, L.W.P., *Cyclic Railway Timetable Optimisation*, T2003/5, June 2003, TRAIL Thesis Series, The Netherlands

Soto Y Koelemeijer, G., *On the Behaviour of Classes of Min-Max-Plus Systems*, T2003/6, September 2003, TRAIL Thesis Series, The Netherlands

Lindveld, Ch.D.R., *Dynamic O-D Matrix Estimation: A Behavioural Approach*, T2003/7, September 2003, TRAIL Thesis Series, Eburon, The Netherlands

Weerdt, de M.M., *Plan Merging in Multi-Agent Systems*, T2003/8, December 2003, TRAIL Thesis Series, The Netherlands

Langen, de P.W, *The Performance of Seaport Clusters*, T2004/1, January 2004, TRAIL Thesis Series, The Netherlands

Hegyi, A., *Model Predictive Control for Integrating Traffic Control Measures*, T2004/2, February 2004, TRAIL Thesis Series, The Netherlands

Lint, van, J.W.C., *Reliable Travel Time Prediction for Freeways*, T2004/3, June 2004, TRAIL Thesis Series, The Netherlands

Tabibi, M., *Design and Control of Automated Truck Traffic at Motorway Ramps*, T2004/4, July 2004, TRAIL Thesis Series, The Netherlands

Verduijn, T. M., *Dynamism in Supply Networks: Actor Switching in a Turbulent Business Environment*, T2004/5, September 2004, TRAIL Thesis Series, The Netherlands

Daamen, W.,*Modelling Passenger Flows in Public Transport Facilities*,T2004/6, September 2004, TRAIL Thesis Series, The Netherlands

Zoeteman, A., *Railway Design and Maintenance from a Life-Cycle Cost Perspective: A Decision-Support Approach*, T2004/7, November 2004, TRAIL Thesis Series, The Netherlands

Bos, D.M., *Changing Seats: A Behavioural Analysis of P&R Use*, T2004/8, November 2004, TRAIL Thesis Series, The Netherlands

Versteegt, C., *Holonic Control For Large Scale Automated Logistic Systems*, T2004/9, December 2004, TRAIL Thesis Series, The Netherlands

Wees, K.A.P.C. van, *Intelligente Voertuigen, Veiligheidsregulering en Aansprakelijkheid. Een Onderzoek naar Juridische Aspecten van Advanced Driver Assistance Systems in het Wegverkeer*, T2004/10, December 2004, TRAIL Thesis Series, The Netherlands

Tampère, C.M.J., *Human-Kinetic Multiclass Traffic Flow Theory and Modelling: With Application to Advanced Driver Assistance Systems in Congestion*, T2004/11, December 2004, TRAIL Thesis Series, The Netherlands

Rooij, R.M., *The Mobile City. The Planning and Design of the Network City from a Mobility Point of View*, T2005/1, February 2005, TRAIL Thesis Series, The Netherlands

Le-Anh, T., *Intelligent Control of Vehicle-Based Internal Transport Systems*, T2005/2, April 2005, TRAIL Thesis Series, The Netherlands

Zuidgeest, M.H.P., *Sustainable Urban Transport Development: A Dynamic Optimisation Approach*, T2005/3, April 2005, TRAIL Thesis Series, The Netherlands

Hoogendoorn-Lanser, S., *Modelling Travel Behaviour in Multimodal Networks*, T2005/4, May 2005, TRAIL Thesis Series, The Netherlands

Dekker, S., *Port Investment - Towards an integrated planning of port capacity*, T2005/5, June 2005, TRAIL Thesis Series, The Netherlands

Koolstra, K., *Transport Infrastructure Slot Allocation*, T2005/6, June 2005, TRAIL Thesis Series, The Netherlands

Vromans, M., *Reliability of Railway Systems*, T2005/7, July 2005, TRAIL Thesis Series, The Netherlands

Oosten, W., *Ruimte voor een democratische rechtsstaat. Geschakelde sturing bij ruimtelijke investeringen*, T2005/8, September 2005, TRAIL Thesis Series, Sociotext, The Netherlands

Le-Duc, T., *Design and control of efficient order picking*, T2005/9, September 2005, TRAIL Thesis Series, The Netherlands

Goverde, R., *Punctuality of Railway Operations and Timetable Stability Analysis*, T2005/10, October 2005, TRAIL Thesis Series, The Netherlands

Kager, R.M., *Design and implementation of a method for the synthesis of travel diary data*, T2005/11, October 2005, TRAIL Thesis Series, The Netherlands

Boer, C., *Distributed Simulation in Industry*, T2005/12, October 2005, TRAIL Thesis Series, The Netherlands

Pielage, B.A., *Conceptual Design of Automated Freight Transport Systems*, T2005/14, November 2005, TRAIL Thesis Series, The Netherlands

Groothedde, B., *Collaborative Logistics and Transportation Networks, a modeling approach to network design*, T2005/15, November 2005, TRAIL Thesis Series, The Netherlands

Valk, J.M., *Coordination among Autonomous Planners*, T2005/16, December 2005, TRAIL Thesis Series, The Netherlands

Krogt, R.P.J. van der, *Plan Repair in Single-Agent and Multi-Agent Systems*, T2005/17, December 2005, TRAIL Thesis Series, The Netherlands

Bontekoning, Y.M., *Hub exchange operations in intermodal hub-and-spoke networks. A performance comparison of four types of rail-rail exchange facilities*,T2006/1, February 2006, TRAIL Thesis Series, The Netherlands

Lentink, R., *Algorithmic Decision Support for Shunt Planning*, T2006/2, February 2006, TRAIL Thesis Series, The Netherlands

Ngoduy, D., *Macroscopic Discontinuity Modeling for Multiclass Multilane Traffic Flow Operations*, T2006/3, April 2006, TRAIL Thesis Series, The Netherlands

Vanderschuren, M.J.W.A., *Intelligent Transport Systems for South Africa. Impact assessment through microscopic simulation in the South African context*, T2006/4, August 2006, TRAIL Thesis Series, The Netherlands

Ongkittikul, S., *Innovation and Regulatory Reform in Public Transport*, T2006/5, September 2006, TRAIL Thesis Series, The Netherlands

Yuan, J., *Stochastic Modelling of Train Delays and Delay Propagation in Stations*, T2006/6, October 2006, TRAIL Thesis Series, The Netherlands

Viti, F., *The Dynamics and the Uncertainty of Delays at Signals*, T2006/7, November 2006, TRAIL Thesis Series, The Netherlands

Huisken, G., *Inter-Urban Short-Term Traffic Congestion Prediction*, T2006/8, December 2006, TRAIL Thesis Series, The Netherlands

Feijter, R. de, *Controlling High Speed Automated Transport Network Operations*, T2006/9, December 2006, TRAIL Thesis Series, The Netherlands

Makoriwa, C., *Performance of Traffic Networks. A mosaic of measures*, T2006/10, December 2006, TRAIL Thesis Series, The Netherlands

Miska, M., *Microscopic Online Simulation for Real time Traffic Management*, T2007/1, January 2007, TRAIL Thesis Series, The Netherlands

Chorus, C., *Traveler Response to Information*, T2007/2, February 2007, TRAIL Thesis Series, The Netherlands

Weijermars, W.A.M., *Analysis of Urban Traffic Patterns Using Clustering*, T2007/3, April 2007, TRAIL Thesis Series, The Netherlands