Proceedings of the 8th International
IEEE Conference on Intelligent Transportation Systems
Vienna, Austria, September 13-16, 2005

Analyzing highway flow patterns using cluster analysis

Wendy Weijermars and Eric van Berkum

Abstract—**Historical traffic patterns can be used for the prediction of traffic flows, as input for macroscopic traffic models, for the imputation of missing or erroneous data and as a basis for traffic management scenarios. This paper investigates the determination of historical traffic patterns by means of Ward's hierarchical clustering procedure. Days were clustered before and after a pre-classification into working days and non-working days, using two different definitions of a daily traffic profile. The results of the clustering after pre-classification are clearly better than before classification. Moreover, working days are easier to classify into distinctive, recurrent traffic patterns than non-working days. Finally, clustering on the basis of 15 minutes traffic flows resulted in a better classification of working days than the two-step clustering that used the total daily traffic flow, peak flows, peak times and ratios. The clustering on the basis of 15 minutes traffic flows resulted in a classification into five clusters that show distinct daily flow profiles and are representative for the days within the clusters. The day of the week and vacation periods are determinative for the cluster a working day is classified to.**

I. Introduction

Traffic monitoring is essential for the successful implementation of traffic management. Traffic can be monitored on-line (in real time) and off-line. The goal of off-line traffic monitoring is to obtain insight into the functioning of the traffic system by using historical traffic data. Hereby, both temporal and spatial variations in traffic flows and travel times can be analyzed.

Regarding temporal variations, a distinction can be made between within-day variations [1] and variations between days. This paper deals with the variation in traffic flows between days. Some research analyzes traffic flow variations between pre-defined types of days [2], [3]. Reference [2] concluded on the basis of an ANOVA analysis that daily traffic flow profiles vary statistically significant between core weekdays (Tuesdays, Wednesdays and Thursdays) on the one hand and Mondays, Fridays, Saturdays and Sundays on the other hand. Reference [3] defines four classes using a matching process that compares daily traffic profiles on the basis of an error measure: (1) Monday until Thursday, except holidays or days before holidays, (2) Fridays and days before holidays, (2) Saturday except holidays and (4) Sundays and holidays.

Another way to take variations between days into account is by defining different types of days on the basis of measured traffic flows or estimated travel times. In that case, (daily) traffic profiles are grouped by means of cluster analysis. Most research concerning clustering temporal traffic patterns is done in the field of short term traffic forecasting [4] – [6]. The emphasis is on the accuracy of the forecast. Only [4] discusses the daily characteristics that are on the basis of the patterns explicitly. He clustered travel times during the AM and PM period and concluded that the AM period could be grouped in weekday, Saturday and holiday (including Sunday) whereas for the PM period, each day should be treated separately.

It is interesting to obtain more insight into variations in daily travel demand patterns as well, besides travel time patterns. For the further development of traffic management it is important to know what typical daily travel demand profiles can be distinguished. Moreover, information about the shape of the flow profiles of the resultant clusters and the daily characteristics that are on the basis of these clusters makes the clusters also appropriate for forecasting travel demand on future days, traffic modeling, and as a basis for traffic management scenarios.

In the existing literature not much attention is paid to the influence of the design of the clustering procedure on the outcome. References [4], [5] and [6] all define daily traffic profiles on the basis of a time-series, whilst other definitions might lead to better results.

This paper discusses the clustering of travel demand profiles on a Dutch Highway location. Travel demand is expressed in terms of measured traffic flows. Two definitions of a daily traffic profile are compared and clustering is performed before and after a pre-classification into working days and non-working days. Section II deals with the available traffic data. In Section III the design of the clustering procedure is described. Section IV presents the results of the cluster analyses. Our conclusions are presented in the final section.

II. Traffic data

Traffic count data from a Dutch highway is used for the clustering. The selected detector is situated on Highway A50 and detects the traffic in the direction of Beekbergen, between the exit for Apeldoorn and the entry from Apeldoorn (Hectometre 206.2). From the dual loop detector, data on speed and flow is collected and aggregated into 15 minute intervals over both lanes. All data was collected in

Manuscript received June 20, 2005.

Wendy Weijermars is with the Centre for Transport Studies, University of Twente, Postbus 217 7500 AE Enschede, The Netherlands (corresponding author to provide phone: +31 53 489 2449; fax: +31 53 489 4040; e-mail: w.a.m.weijermars@utwente.nl).

Eric van Berkum is with the Centre for Transport Studies, University of Twente, Postbus 217 7500 AE Enschede, The Netherlands (e-mail: e.c.vanberkum@utwente.nl).

2003.

Before the actual clustering was executed, days with congestion and days with missing data were excluded from the dataset. In case of congestion, traffic flows do not represent travel demand adequately. Therefore, days (1) on which the average speed was lower than 50 km/h during one or more 15 minute intervals or (2) on which the speed was less than 50 km/h during more than 25% of the time for one or more 15 minute intervals, were excluded from the clustering. As a result of congestion, five days from the original dataset were removed. On four of these days, congestion occurred during the A.M. peak period, on the fifth day, congestion occurred in the afternoon. Within these days are one Monday in December, two Tuesdays in October one Thursdays in October and one Thursday in November. Because of missing data, 199 more days from the original dataset were removed.

At the selected location, traffic data meets the requirements on 118 days. Within these days are 77 working days and 41 non-working days. Three of these non-working days are Public Holidays (New Year's day, Christmas Day and Boxing day) and five are special days (days between Boxing day and New Year's eve). The distribution of data over the seasons is uneven (see table 1), since most of the data is collected during the autumn and winter. The distribution over the days of the week is more even.

TABLE 1
DISTRIBUTION OF DAYS ON WHICH TRAFFIC DATA IS COLLECTED OVER SEASONS AND WEEKDAYS.

	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Tot
Jan	1	1	1	1	2	2	1	9
April	0	1	0	0	1	1	0	3
June	0	0	0	0	1	0	0	1
July	0	1	1	0	0	0	0	2
Sept	4	4	4	3	4	4	4	27
Oct	4	4	2	3	4	4	4	25
Nov	5	3	4	3	3	4	5	27
Dec	2	3	3	2	1	2	3	16
Total	16	17	15	12	16	17	17	110

Public Holidays and special days are not included in this table

III. DESIGN OF CLUSTERING PROCEDURE

Several clustering techniques have been developed, for an overview see for example [7], [8] or [9]. In general, a distinction can be made between hierarchical and partitional techniques. The advantage of the hierarchical techniques is that the number of clusters does not have to be chosen in advance but follows from the clustering. Therefore a hierarchical clustering procedure is chosen for this research. Since Ward's method has been frequently used and has proved to lead to the formation of homogeneous groups of objects [10], within the hierarchical procedures, Ward's method is selected. In Ward's clustering technique, every daily flow profile starts as a separate cluster and clusters are combined on the basis of the sum of squared deviations from the mean of the cluster.

The optimal number of clusters is selected by means of a dendogram (see fig. 1). A dendrogram visualizes the variation within the clusters for different steps of the clustering procedure. The optimal number of clusters is the number for which a further decrease of the number of clusters leads to a high increase in variation within the clusters (expressed by the rescaled distance cluster combine) and for which an increase of the number of clusters leads to only a small decrease in variation within the clusters.

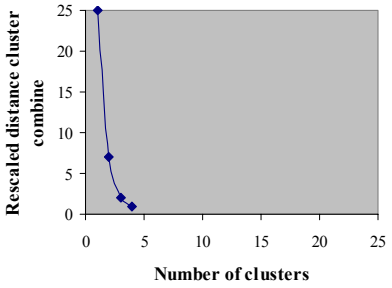

Fig. 1. Daily flow profiles of days within cluster 1. The optimal number of clusters is 3 in this case. A division into two clusters shows a much higher variation within the clusters, whereas with a division into 4 clusters only a small decrease in the variation within the clusters is obtained.

Cluster analysis requires traffic patterns to be defined mathematically, by a number of features. For the classification of daily flow profiles, it is important to take both differences in shape and in height into account. The simplest definition of a daily traffic profile that takes both these characteristics into account is a series of traffic counts as a function of the time of the day. Since we are interested in the general course of the daily flow profile, the daily profile is defined by 15-minute traffic flows.

Above definition of a daily flow profile does not provide information about the type of differences between traffic patterns. Therefore, a second definition is used that defines a traffic pattern by multiple features that each determines one characteristic of the daily traffic profile. Working day profiles are defined by the total daily traffic flow, the peak flows[1], the times of the peak periods and the ratios between peak flows and off-peak flow. Non-working days show a differently shaped pattern with only one peak and are therefore defined by the total daily traffic flow, the peak flow and the time of the peak period. Since flows and times are not measured on the same scale, they cannot be combined without any problems. Therefore, a two-step clustering procedure is applied. In the first step, the daily flow profiles are clustered on the basis of the different features separately. In that way, every day is classified to multiple clusters. In the second step, days are classified to final clusters on the basis of combinations of clusters resulting from the clustering procedures using separate features.

In some literature, a pre-classification is executed [6]. The

[1] The peak flows are defined as the average hourly traffic flow during the two successive busiest hours

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. To investigate the effect of pre-classification the results of clustering procedures before and after pre-classification into working days and non-working days are compared.

For all approaches it was investigated whether distinctive, recurrent daily traffic patterns can be detected. Besides, it was examined whether the average daily flow profiles of the patterns are representative for all days within the patterns by analyzing the variation within the clusters. Finally, it was determined what daily characteristics are on the basis of the traffic patterns found.

IV. Clustering Results

A. Before pre-classification

Clustering on the basis of 15 minutes traffic flows resulted in a classification into three clusters that are shown in fig. 2 and described in table 2. The clusters show clearly distinctive daily traffic profiles and working days are classified to a separate cluster from non-working days.

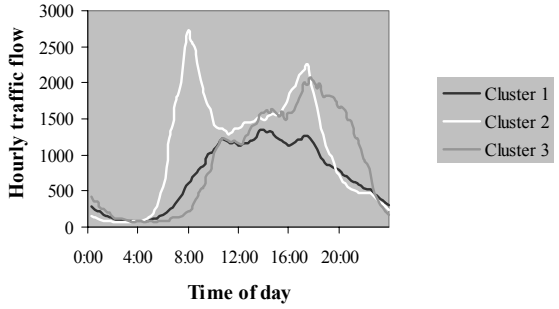

Fig. 2. Average daily flow profiles of the resultant clusters.

TABLE 2
Characteristics of Resultant Clusters

Cluster	*No of days*	*Types of days*
Cluster 1	39	Weekend days, Holidays, Special days, days within vacation period
Cluster 2	71	Working days
Cluster 3	8	Sundays

As mentioned in section III, the representativeness of the resultant clusters is examined by means of the variation within the clusters. Firstly the daily flow profiles of the days within the clusters were compared (see fig. 3 to 5). Next, the 10^{th} and 90^{th} percentiles of the clusters were determined. The width between the 10^{th} and 90^{th} percentiles provides information about the predictability of the traffic flow on a certain time of the day. The smaller the width, the better traffic flows can be predicted. Finally, the coefficients of variation were calculated for the peak flows to determine the predictability of the peak flows (see table 3).

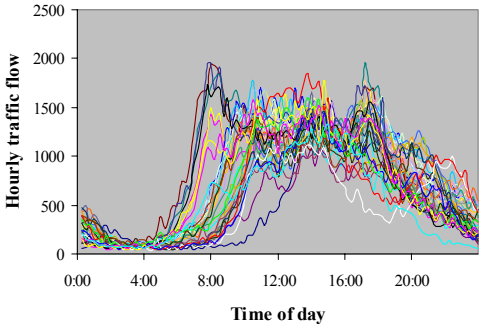

Fig. 3. Daily flow profiles of days within cluster 1.

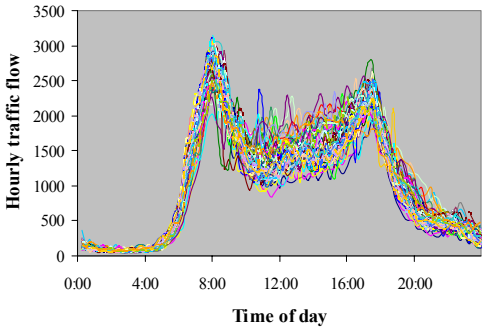

Fig. 4. Daily flow profiles of days within cluster 2.

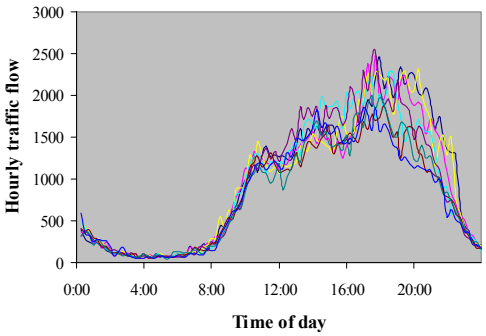

Fig. 5. Daily flow profiles of days within cluster 3.

From fig. 3 to fig. 5 it can be concluded that the variation within the clusters is quite large. Therefore, the average daily flow profiles of the clusters are not representative for all days within the clusters. Due to the large differences between working days and non-working days, the clustering algorithm is not able to detect smaller differences between days and a classification is made into too little clusters.

The two-step clustering resulted in a classification into ten clusters. Two of these clusters were relatively large, whilst the others contained five days or less. One of the large clusters contained only working days, the other only weekend days. The small clusters consisted of dissimilar working days or dissimilar weekend days. These small clusters were that small that they cannot be called recurrent traffic patterns. Moreover, the variation within some of the clusters was also for this clustering substantial. The

coefficients of variation were comparable to the coefficients resulting from the first clustering (see table 3). Therefore it can be concluded that also the variation within these clusters is too large. Also this clustering procedure thus does not lead to satisfying results.

TABLE 3
COEFFICIENTS OF VARIATION FOR A.M. PEAK AND P.M. PEAK FLOWS

Clustering	A.M. peak flow	P.M. peak flow
15 minutes flows	4.7 – 24.1	5.9 – 11.4
Two-step	6.7 – 25.5	4.6 – 11.2

Only clusters containing more than five days are taken into account

B. Working days

The clustering on the basis of 15 minutes traffic flow resulted in a classification into five clusters that are shown in fig. 6. The characteristics of the days within the clusters are summarized in table 4.

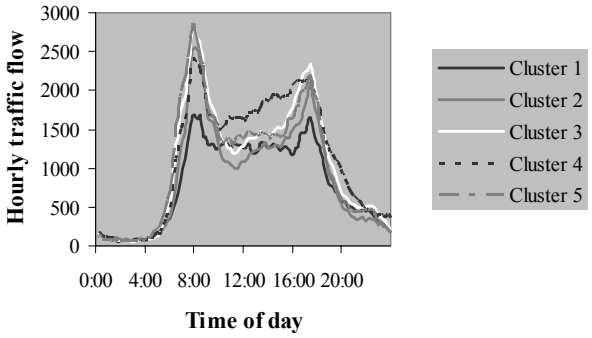

Fig. 6. Average daily flow profiles of resultant working day clusters.

TABLE 4
CHARACTERISTICS OF RESULTANT WORKING DAY CLUSTERS

Cluster	*No of days*	*Types of days*
Cluster 1	6	Vacation period
Cluster 2	5	6th of January - 10th of January
Cluster 3	38	Tuesdays, Wednesdays, Thursdays, 1 Monday
Cluster 4	15	Fridays
Cluster 5	13	Mondays

Cluster 1 does not show real A.M. and P.M peak periods. This cluster consists of days within vacation periods, on which less trips take place during the peak periods. On contrary, the other clusters show clear peak periods. Cluster 4 shows a higher off-peak flow compared to the other clusters and contains solely Fridays. This higher off-peak traffic on Fridays is probably caused by recreational traffic (see also [11]). Clusters 2, 3 and 5 are more similar, although the amount of traffic is somewhat lower during all day for cluster 2 and the P.M. peak flow is relatively high for cluster 3. Cluster 5 solely contains Mondays, whereas cluster 3 mainly contains core weekdays. Cluster 2 consists of all weekdays from the week of the 6th of January until the 10th of January. We could not find out why all working days of this week have been classified to a separate cluster.

Also for working days, the variation within the clusters is analyzed. Fig. 7 to 11 show the daily flow profiles of the days within the clusters.

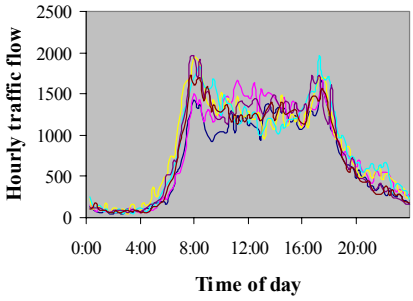

Fig. 7. Daily flow profiles of days within cluster 1.

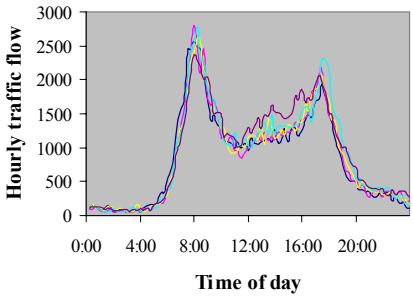

Fig. 8. Daily flow profiles of days within cluster 2.

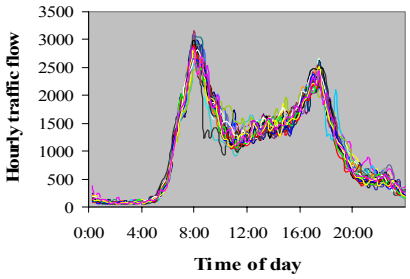

Fig. 9. Daily flow profiles of days within cluster 3.

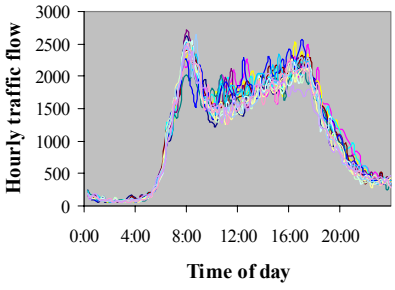

Fig. 10. Daily flow profiles of days within cluster 4.

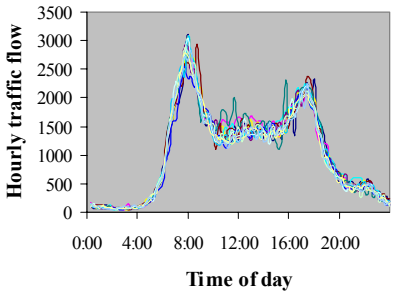

Fig. 11. Daily flow profiles of days within cluster 5.

The variation between the days is quite small within clusters 2, 3 and 5. For these clusters, the average daily flow profile is representative for the daily flow profiles of the days within the clusters. Traffic flows appear to show less clear daily patterns during the vacation period, as the variation is relatively large within cluster 1. Also the days within cluster 4 show substantial variations. This can also be seen in fig. 12 that shows the 10th and 90th percentiles of the daily flow profiles for the days within cluster 4.

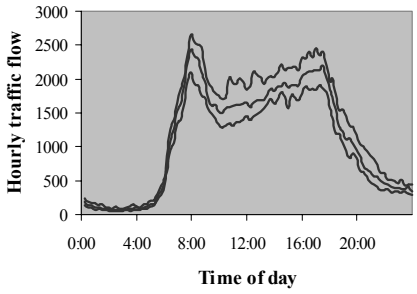

Fig. 12. 10th and 90th percentiles for days within cluster 4.

Table 5 shows the coefficients of variation for both the morning peak and the evening peak flow of the different clusters. For most clusters the coefficients of variation are clearly smaller than the coefficients of variation for all working days. Only for cluster 1 the coefficient of variation is higher for the morning peak flow.

TABLE 5
COEFFICIENTS OF VARIATION FOR A.M. PEAK AND P.M. PEAK FLOWS

Cluster	A.M. peak period	P.M. peak period
Cluster 1	12.2	6.7
Cluster 2	3.4	5.7
Cluster 3	3.4	2.5
Cluster 4	4.8	4.4
Cluster 5	3.3	2.4
All working days	11.7	10.0

The two-step clustering resulted in a classification into 47 clusters that were all very small. The clusters were that small that we cannot speak of recurrent traffic patterns. The classification into these small clusters is due to the fact that working days show highly comparable peak flows and peak times. Therefore, a classification into multiple clusters is made in the first step of the clustering, while in reality only one type of day exists.

C. *Non-working days*

The clustering of the 41 non-working days on the basis of 15 minutes traffic flows resulted in a classification into four clusters that are shown in fig. 13. The characteristics of the days within the clusters are summarized in table 6.

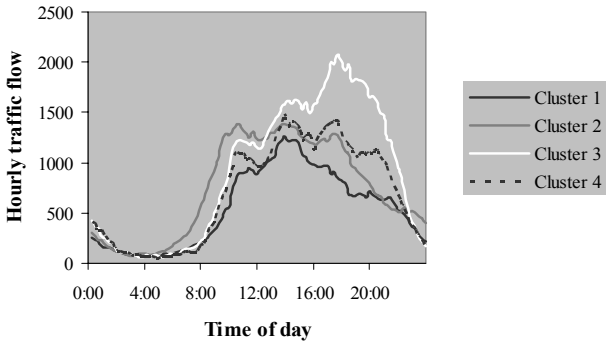

Fig. 13. Average daily flow profiles of resultant non-working day clusters.

TABLE 6
CHARACTERISTICS OF RESULTANT NON-WORKING DAY CLUSTERS

Cluster	*# days*	*Types of days*
Cluster 1	9	Sundays in January, Public Holidays, Special days, 1 Saturday
Cluster 2	19	Saturdays, special days
Cluster 3	8	Sundays in September and October
Cluster 4	5	Sundays in November

All Public Holidays are classified in cluster 1, whereas the days between Christmas and New Year's Eve are divided over clusters 1 and 2. In general, Saturdays are classified to a separate cluster from Sundays. On Saturdays, the peak period takes place earlier than on Sundays. The month of the year appears to be determinative for the cluster a Sunday is classified to. Sundays in September and October show a peak during the evening, probably caused by recreational traffic going home after spending a weekend in the north of the Netherlands.

The variation within the non-working day clusters is larger than the variation within the working day clusters. Especially cluster 1 shows large variation between the days within the cluster (see fig. 14).

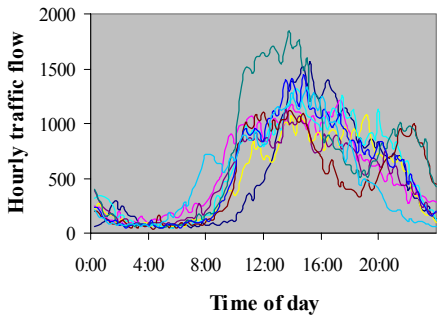

Fig. 14. Daily flow profiles of days within cluster 1.

The two-step clustering on the basis of the daily traffic

flow, peak flow and peak time resulted in a classification into eight small clusters. Only half of these clusters contained five or more days. Also within these clusters, the variation was quite large. Since the clusters were already very small, a classification into more clusters was no option. Therefore, it is concluded that the non-working days in this dataset do not exhibit that clear recurrent daily traffic patterns as working days. Probably this is caused by less fixed activity patterns on non-working days.

V. Conclusion

For a Dutch highway location, daily traffic profiles are grouped by means of a hierarchical Ward's clustering procedure using various approaches.

It is concluded that a pre-classification into working days and non-working days substantially improves the clustering result. Since the variation within the clusters resulting from the clustering without pre-classification is large, the average daily flow profiles of these clusters are not representative for all days within the clusters. This large variation is due to large differences between daily flow profiles of working days and non-working days. A pre-classification into working days and non-working days filters the largest differences out in advance and enables the clustering algorithm to detect smaller differences. These results correspond to the results of [6], that showed traffic forecasts were better in case of a pre-classification into day-groups and conditions.

Secondly, working days are easier to classify than non-working days. The variation within the non-working day clusters is quite large and the average daily traffic profiles thus do not represent the daily traffic profiles of all days within the clusters. Moreover, the clusters are already relatively small, so a classification into more clusters is no option. Probably non-working days show less fixed activity patterns compared to working days. More data from concessive years will probably lead to a better clustering result. Especially since the cluster a non-working day is classified to, seems to be dependent on the season.

Moreover, for working days, the clustering on the basis of 15 minutes traffic flows resulted in a better classification than the two-step clustering. The latter resulted in a classification into too many small clusters that do not represent recurrent traffic patterns. The classification into these small clusters is probably due to the fact that working days show highly similar daily traffic profiles. Clustering on the basis of some of the separate features resulted in a classification into multiple clusters, while in reality only one group exists.

Finally, on the basis of the resultant classification of working days it is concluded that four types of working days can be distinguished: (1) Mondays, (2) core week days, (3) Fridays and (4) days within vacation periods. These results correspond to the results of [2], who also found differences between Mondays, core week days and Fridays and partly to the results of [3], who classified Fridays in a separate class from the other weekdays.

The resultant working day patterns can be used as input for macroscopic traffic models and as a basis for traffic management scenarios. Moreover, when predicting traffic flows on the basis of historical data, a pre-classification into vacation day, Mondays, core week days and Fridays can be made. Finally, these patterns can be used to detect and replace erroneous data and to impute missing data.

The cluster analysis is only executed for one location. Different Highways serve different types of traffic, so one has to be careful translating these results to other locations.

Especially for traffic management and traffic forecasting purposes it can also be useful to classify A.M. and P.M peak periods. For these classifications, other aggregation levels might be more appropriate but the clustering procedures described in this paper can also be used for these classifications. Moreover, it is recommended to extend the method for the analysis of congestion patterns as well, since traffic management is most useful in these situations. In case of congestion, traffic flows do not provide sufficient information. A definition of a traffic situation on the basis of flows, speeds and densities is probable more appropriate.

Acknowledgment

The authors thank the AVV Transport Research Centre of the Dutch Ministry of Transport, Public Works and Water Management for providing the traffic data.

References

[1] A. Stathopoulos and M.G. Karlaftis, "Spectral and Cross-Spectral Analysis of Urban Traffic Flows" *IEEE Intelligent Transportation Systems Conference Proceedings*, pp820-825, Oakland USA, August 25-29 2001

[2] H. Rakha and M. Van Aerde " Statistical Analysis of Day-to-Day Variations in Real-Time Traffic Flow Data", *Transportation Research Record* **1510** pp26-34, Washington, 1995.

[3] R. Chrobok, O. Kaumann, J. Wahle and M. Schreckenberg "Different methods of traffic forecast based on real data", *European Journal of Operational Research* **155**, pp558-568, 2004.

[4] E. Chung, "Classification of traffic pattern", presented at the 10th World Congress on Intelligent Transport Systems, Madrid, November 16-20, 2003, Paper 3233.

[5] M. Danech-Pajouh, and M. Aron, "ATHENA: a method for short-term inter-urban motorway traffic forecasting", Recherche Transports Sécurité, English issue (6), pp11-16, 1991.

[6] D. Wild, "Short-term forecasting based on a transformation and classification of traffic volume time series", International Journal of Forecasting **13**, pp 63-72, 1997.

[7] A.K. Jain, M.N. Murty and P.J. Flynn, "Data Clustering: A Review", *ACM Computing Surveys*, **31(3)**, pp264-323, 1999.

[8] A.R. Webb, *Statistical pattern recognition*. Chichester: Wiley, 2002.

[9] R.O. Duda, P.E. Hart and D.G. Stork, *Pattern Classification,* John Wiley & Sons, inc., 2001.

[10] B. Nowotny, J. Asamer, D. Kashif and R. Karim, "Classification of traffic data time series by cluster analysis, artificial neural networks and ANOVA" presented at 10th World Congress on Intelligent Transport Systems, Madrid, 2003, Paper 4114.

[11] L. Harms, *Mobiel in de tijd, op weg naar een auto-afhankelijke maatschappij, 1975-2000*. Sociaal en Cultureel Planbureau, Den Haag, 2003 (In Dutch).